Su et al. BMC Systems Biology 2012, 6(Suppl 3):S11
http://www.biomedcentral.com/1752-0509/6/S3/S11

BMC Systems Biology

RESEARCH Open Access

Statistical aspects of omics data analysis using the random compound covariate

Pei-Fang Su[1], Xi Chen[1,2], Heidi Chen[1,2], Yu Shyr[1,2*]

From The International Conference on Intelligent Biology and Medicine (ICIBM)
Nashville, TN, USA. 22-24 April 2012

Abstract

Background: Dealing with high dimensional markers, such as gene expression data obtained using microarray chip technology or genomics studies, is a key challenge because the numbers of features greatly exceeds the number of biological samples. After selecting biologically relevant genes, how to summarize the expression of selected genes and then further build predicted model is an important issue in medical applications. One intuitive method of addressing this challenge assigns different weights to different features, subsequently combining this information into a single score, named the compound covariate. Investigators commonly employ this score to assess whether an association exists between the compound covariate and clinical outcomes adjusted for baseline covariates. However, we found that some clinical papers concerned with such analysis report bias p-values based on flawed compound covariate in their training data set.

Results: We correct this flaw in the analysis and we also propose treating the compound score as a random covariate, to achieve more appropriate results and significantly improve study power for survival outcomes. With this proposed method, we thoroughly assess the performance of two commonly used estimated gene weights through simulation studies. When the sample size is 100, and censoring rates are 50%, 30%, and 10%, power is increased by 10.6%, 3.5%, and 0.4%, respectively, by treating the compound score as a random covariate rather than a fixed covariate. Finally, we assess our proposed method using two publicly available microarray data sets.

Conclusion: In this article, we correct this flaw in the analysis and the propose method, treating the compound score as a random covariate, can achieve more appropriate results and improve study power for survival outcomes.

Introduction

High-dimensional omics data

Personalized medicine is expected to enable a more predictive discipline, in which therapies are targeted toward the molecular constitution of individual patients and their disease; thus, molecular biomarkers are widely expected to revolutionize the current practice of medicine. For example, the progress of genomics has made it possible to evaluate molecular signatures to predict cancer metastasis [1,2]. Various technological breakthroughs have led to a plethora of high-dimensional omics data to support personalized medicine, and these data have a common characteristic: the numbers of features greatly exceeds the number of biological samples. Because biological phenomena are the result of sets of features (e.g., concerted expression of multiple genes), the analysis of a group of related features (e.g., genes) may be more effective and may provide more directly interpretable results than the analysis of individual genes.

As high-dimensional omics research has advanced, the compound covariate (or compound score) has generally been held as a simpler and more straightforward approach. After selecting biologically relevent genes in training cohort, such a score is often a useful device in medical applications to define the information contained in a single set of data and to summarize the association of a set of variables with disease. Tukey [3] first advocated the use of compound covariates in the clinical trial setting. To develop a compound score, the individual

* Correspondence: yu.shyr@Vanderbilt.Edu
[1]Center for Quantitative Sciences, Vanderbilt University, Nashville, TN, USA
Full list of author information is available at the end of the article

covariates are summed; the association between such a compound covariate and outcome then is evaluated via regression analysis. Tomasson [4] used a compound score for binary outcomes, via fitting a logistic regression. Later, Hedenfalk [5] successfully applied the compound covariate method to class prediction analysis for breast cancer data. Because the use of the compound covariate is intuitive and seems useful, many other leading researchers also have applied this method for analyzing omics data sets [6-9].

Problem statements

A compound covariate is a linear combination of the basic covariates being studied, with each covariate having its own coefficient or weight. For survival outcomes, a commonly used scheme is to 1) compute the univariate Cox regression [10] for each gene of interest, 2) assign a weight to each gene (typically, the estimated regression coefficients or Wald statistics from the univariate Cox regressions), and 3) combine the weighted genes in a linear model that incorporates gene expression levels in each sample. This method of modeling weighted genes is believed to reflect the importance of each individual gene to the outcome; the higher the weight assigned, the more significant a particular gene is.

However, the linear combination of the group of genes, with each gene having its own estimated weight, should not be treated as an observed covariate or fixed covariate. Because selected weights are estimated through computing the univariate Cox regression of each individual gene, the compound "covariate" should be treated as a random covariate that includes estimated error. Besides, for the purpose of assessing whether an association exists between the compound covariate and survival outcomes, Cox regression is typically used to evaluate the significance level of the parameter of the compound score. However, bias concerns arise when the same data set, training cohort, is used for a double purpose: to construct the compound covariate and then to test it. This framework results in an over-fitting problem. As shown in Figure 1, we simulate 50 observations with 3, 5, 10, and 30 non-causal genes used to create a compound covariate. The Cox regression model is then used to test whether an association exists between the compound scores and survival outcomes in the same dataset. The distribution of p-values should be uniform in the interval [0, 1]. In our simulation, however, p-values are concentrated around zero, especially as the number of genes increases. This demonstrates that type I error rates are inflated and consequently not controlled. Obviously, double using the training cohort casue overfitting problem and bias p-values arised. We found that some medical papers report inappropriate p-values for testing training cohort data [11,6]. Although the proposed bias p-values in their training cohort do not affact their final research results inferred from another independent testing data set, these observations motive us to study a more appropriate testing procedure for the compound covariate if the investagtors whish to report a testing result in the training cohort.

In this paper, we first contend the compound covariate should be treated as a random observation. Our idea is based on that proposed by Prentice [12], who analyzed covariates with measurement error and used a partial likelihood function technique to infer whether the parameter for the covariate was significant. In addition, if a training data set is used for a double purpose (i.e., to construct the compound covariate and then to test it), the resulting over-fitting means the p-value is not reliable when testing the regression parameter. Therefore, we use a 2-fold method (e.g., [13,14]), splitting all observations in the training cohort into two parts, one part for assigning gene weights, and another part for testing the regression parameter through a partial likelihood score test. The remainder of this paper is organized as follows: We outline creation of the compound score using a random covariate approach. Then, we investigate the accuracy of the asymptotic distribution of the proposed tests. We thoroughly assess the performance of two commonly used estimated weights, "estimated coefficient" and "Wald statistic", for the Cox proportional hazards model. Finally, we illustrate the implementation of the proposed methods through two real data sets, and offer concluding remarks.

The proposed method

The compound covariate

In this section, we formally define some notations for compound covariate and introduce a procedure to identify whether a set of genes is truly associated with survival times in the training cohort. Let T_j denote the survival time and C_j denote the censoring time independent of T_j for the jth patient, with $j = 1, 2, ..., n$. When some observations are right-censored, one can observe only random variables $X_j = \min(T_j, C_j)$ and $\delta_j = I(T_j \le C_j)$, where $I(A)$ is an indicator function of event A, assuming the value 1 if event A occurs and the value 0 otherwise. Let $x_j = (x_{j1}, x_{j2}, ..., x_{jp})$ be the standardized selected gene's intensity in the jth patient, where p is the size of the gene set. For creating the compound covariate, one first fits the univariate Cox regression model for each individual gene, that is,

$$h(t|x_k) = h_{0k}(t) \exp(x_k \beta_k), k = 1, 2, \ldots, p \quad (1)$$

where $h_{0k}(t)$ is a baseline hazard function of each gene k, and β_k is a corresponding parameter to be estimated. Let $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ be the estimators of $\beta_1, \beta_2, ..., \beta_p$, and $\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_p$ be the Wald statistics obtained from (1). A compound covariate commonly used by clinicians is defined as

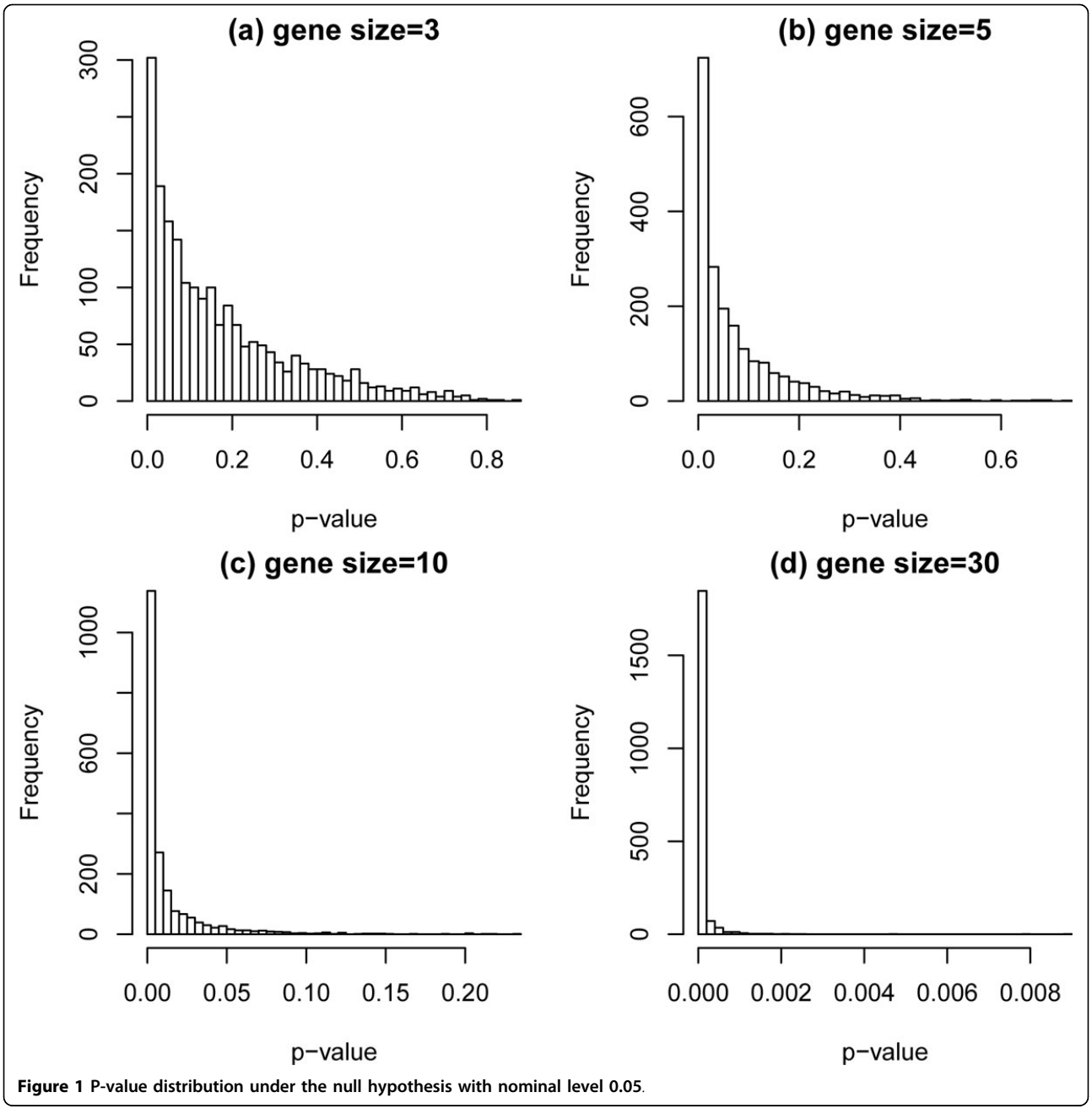

Figure 1 P-value distribution under the null hypothesis with nominal level 0.05.

$$z_j = \sum_{k=1}^{p} x_{jk}\hat{\beta}_k,$$

or another possible weighting policy depends on Wald statistics,

$$z_j = \sum_{k=1}^{p} x_{jk}\hat{w}_k \text{sign}\left(\hat{\beta}_k\right),$$

for each patient j, j = 1, 2, ..., n. From the perspective of biology, the weighting policy is believed to reflect the importance of each individual gene to survival outcome, the higher the weight, the more important the gene is. In other words, the score can be regarded as a condensed index, representing the collective effects of gene expression.

To identify whether this gene expression pattern is truly associated with survival in training cohort, investigators prefer to use Cox regression analysis. That is, after fitting model (1), they construct

$$h(t \mid z) = h_0(t) \exp(\gamma_0 z)$$

where $h_0(t)$ is a baseline hazard function and γ_0 is a corresponding parameter for the compound covariate z; they then use the same data set to test the null hypothesis H_0 : $\gamma_0 = 0$. Under the null hypothesis, however, the method results in uncontrolled type I error, because the training data set has been used twice, both for building the model and for testing the regression parameter. If independent data are available, carry $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ from training data set and test on another independent data set is possible, allowing unbiased model validation to prevent over-fitting. However, if investagtors whish to report a testing result in the training cohort, an alternative, though less optimal, study design is using k-fold method or split the training cohort data randomly (2-fold), with 50% of the data being assigned to develop the score, compound covariate, and 50% to evaluate its performance. The limitation of this approach is that it requires a relatively large sample size. With this method, Kaplan-Meier survival curves [15] for the two sets should be examined to ensure no significant difference by the random selection of those two sets from training cohort data.

Cox regression with a random compound covariate

The measuring mechanism makes the compound covariate an estimation and not a fixed observable. Naturally, such a covariate should be treated as a random covariate, and the variance of each score needs to be taken into account. To fit a Cox regression model with a random covariate, we use the idea advocated by Prentice, which presents the Cox model as a multiplicative hazards model, with a relative risk at time t,

$$E\left[\exp\left(\gamma_0 z\right)\right]. \tag{2}$$

This is a weighted average of a possible relative risk given the covariate z. The Cox regression model then can be written as

$$h\left(t|z\right) = h_0\left(t\right) E\left[\exp\left(\gamma_0 z\right)\right].$$

Because omics data sets involve a large number of features, we assume the distributions of the scores follow normal distribution. That is, for each patient j, assuming z_j is drawn from a normal distribution with mean μ_j and variance σ_j^2, (2) can be derived as

$$\exp\left(\gamma_0\mu_j + \frac{1}{2}\sigma_j^2\gamma_0^2\right).$$

Thus, a quadratic term, $\sigma_j^2\gamma_0^2/2$, is added to the relative risk function, accounting for the variance in the random covariate. In addition, with both random covariates z (in this case, the compound covariate) and observed covariates $\mathbf{w}$ with dimension d (in this case, the clinical variable), it is easy to incorporate the fixed covariate effects into the Cox model, as:

$$h\left(t|z, \mathbf{w}\right) = h_0\left(t\right) E\left[\exp\left(\gamma_0 z + \boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{w}\right)\right],$$

where γ_0 is the parameter for the compound covariate, $\boldsymbol{\gamma}_1$ is the corresponding parameters for fixed observations and $\boldsymbol{\gamma}_1^{\mathrm{T}}$ is the transpose of $\boldsymbol{\gamma}_1$.

A partial likelihood function and score test

Suppose now that $t_1 < \ldots < t_l$ are the ordered distinct survival times in the sample, and let $R(t_i)$ and $F(t_i)$ denote the risk set prior to t_i and the set of subjects failing at t_i, respectively. The partial likelihood function is:

$$L\left(\boldsymbol{\gamma}\right) = \prod_{i=1}^{l} \frac{\prod_{j\in F(t_i)} E\left[\exp\left(\gamma_0 z + \boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{w}\right)\right]}{\sum_{j\in R(t_i)} E\left[\exp\left(\gamma_0 z + \boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{w}\right)\right]^{m_i}},$$

where m_i is the number of failures at time $t_i (i = 1, 2, \ldots l)$. Let $a = E\left[\exp\left\{\gamma_0 z + \boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{w}\right\}\right]$, $b = \partial a/\partial\boldsymbol{\gamma}$ and $c = \partial b/\partial\boldsymbol{\gamma}$ where $\boldsymbol{\gamma} = \left[\gamma_0, \boldsymbol{\gamma}_1^{\mathrm{T}}\right]^{\mathrm{T}}$ (The explicit forms of a, b and c are shown in Additional file 1). The score statistic then can be derived as

$$v = \frac{\partial \log L\left(\boldsymbol{\gamma}\right)}{\partial\boldsymbol{\gamma}} = \sum_{i=1}^{l}\left(\sum_{j\in F(t_i)}\frac{b_{ij}}{a_{ij}} - m_i\frac{\sum_{j\in R(t_i)} b_{ij}}{\sum_{j\in R(t_i)} a_{ij}}\right) \tag{3}$$

with the observed information matrix $\boldsymbol{V}$

$$-\frac{\partial^2 \log L\left(\boldsymbol{\gamma}\right)}{\partial\boldsymbol{\gamma}^2} = \sum_{i=1}^{l} m_i\frac{\sum_{j\in R(t_i)} c_{ij}}{\sum_{j\in R(t_i)} a_{ij}} - \sum_{i=1}^{l} m_i\left(\frac{\sum_{j\in R(t_i)} b_{ij}}{\sum_{j\in R(t_i)} a_{ij}}\right)^2 - \sum_{i=1}^{l}\sum_{j\in F(t_i)}\left(\frac{c_{ij}}{a_{ij}} - \frac{b_{ij}^2}{a_{ij}^2}\right). \tag{4}$$

Consequently, under the null hypothesis $H_0 : \boldsymbol{\gamma} = 0$, the partial likelihood score test $v^{\mathrm{T}}\boldsymbol{V}^{-1}\, v$ will have an asymptotic χ^2_{d+1} distribution when $\boldsymbol{V}$ is nonsingular. In addition, (3) can be used in a standard manner for $\boldsymbol{\gamma}$estimation. In practice, we can calculate a p-value by replacing μ_j with z_j and σ_j^2 with $\mathrm{Var}(z_j)$ where

$$\mathrm{Var}\left(z_j\right) = \sum_{k=1}^{p}\mathrm{Var}\left(x_{jk}\hat{\beta}_k\right) = \sum_{k=1}^{p} x_{jk}^2 s_{\beta_k}^2,$$

is derived based on approximation and $s_{\beta k}^2$ is the variance estimation of $\hat{\beta}_k$ by using estimated coefficient as weight or

$$\mathrm{Var}\left(z_j\right) = \sum_{k=1}^{p}\mathrm{Var}\left(x_{jk}\hat{w}_k\mathrm{sign}\left(\hat{\beta}_k\right)\right) = 3\sum_{k=1}^{p} x_{jk}^2$$

by using Wald statistics as weight. We show the derivation in more detail in Additional file 2.

Multiple gene sets

Further, we extended the compound covariate to multiple gene sets. If there are q given independent gene sets for the jth patient, $\mathbf{x}_j^{(1)}, \mathbf{x}_j^{(2)}, \ldots, \mathbf{x}_j^{(q)}$, where $\mathbf{x}_j^{(s)} = \left(x_{1j}^{(s)}, x_{2j}^{(s)}, \ldots, x_{p_s j}^{(s)}\right)$, $s = 1, 2, \ldots, q$, the compound covariates can be written as vector

$$\mathbf{z}_j = \left(z_j^{(1)}, z_j^{(2)}, \ldots, z_j^{(s)}\right) = \left(\sum_{k=1}^{p_1} x_{jk}^{(1)}\hat{\beta}_k, \sum_{k=1}^{p_2} x_{jk}^{(2)}\hat{\beta}_k, \ldots, \sum_{k=1}^{p_s} x_{jk}^{(s)}\hat{\beta}_k\right), \text{ or}$$

$$\mathbf{z}_j = \left(\sum_{k=1}^{p_1} x_{jk}^{(1)}\hat{w}_k \text{sign}\left(\hat{\beta}_k\right), \sum_{k=1}^{p_2} x_{jk}^{(2)}\hat{w}_k \text{sign}\left(\hat{\beta}_k\right), \ldots, \sum_{k=1}^{p_s} x_{jk}^{(s)}\hat{w}_k \text{sign}\left(\hat{\beta}_k\right)\right)$$

where p_s is the number of genes of the sth gene set. Then, the partial likelihood function can be written as

$$L(\boldsymbol{\gamma}) = \prod_{i=1}^{l} \frac{\prod_{l \in F(t_i)} E\left[\exp\left(\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{z} + \boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{w}\right)\right]}{\sum_{l \in R(t_i)} E\left[\exp\left(\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{z} + \boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{w}\right)\right]^{m_i}}.$$

Let $a = E\left[\exp\left(\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{z} + \boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{w}\right)\right]$, $b = \partial a/\partial\boldsymbol{\gamma}$ and $c = \partial b/\partial\boldsymbol{\gamma}$, where $\boldsymbol{\gamma} = \left[\boldsymbol{\gamma}_0^{\mathrm{T}}, \boldsymbol{\gamma}_1^{\mathrm{T}}\right]^{\mathrm{T}}$. The score statistic and the observed information matrix can be further derived as (3) and (4) as well. Consequently, under the null hypothesis $H_0 : \boldsymbol{\gamma} = 0$, the partial likelihood score test $v^{\mathrm{T}}V^{-1}v$ has an asymptotic χ^2_{d+q} distribution when V is nonsingular. If we reject the null hypothesis, we can conclude that the covariate vector is associated with survival time.

Simulation results

To assess the performance of the proposed testing procedure for compound covariate, we conducted simulation studies under various scenarios to study type I error rate and power. For the scenario of split training data set as two parts and the consideration of compound scores as random covariates, we denoted the compound score using $\hat{\beta}$ as a weight function as SRC_B, and the compound score using the Wald statistic as a weight function as SRC_W. The corresponding notations, SC_B and SC_W, refer to split data but without treating the compound covariate as a random covariate (i.e., typical Cox regression [10]). The notations DC_B and DC_W refer to compound scores with double use of the training data set, both for building the model and then testing for the same data set. Assume there are p selected genes in a gene set. Gene expression data were generated from a multivariate normal distribution with mean vector $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_p]^{\mathrm{T}}$, and variance-covariance matrix equal to the identity matrix for n cases. Generated survival times were associated with gene expression via a proportional hazard model, $\exp(x^{\mathrm{T}}\boldsymbol{\beta})$. All tests with nominal significance level 0.05 were applied and empirical rejection probability was obtained based on 2000 simulation runs.

For comparing empirical type I error rates, the value of $\boldsymbol{\beta}$ was set to 0. The total sample size was set to 50, 75, or 100. After gene selection process, we assume the total number of disease relative genes was set to 10, 30, 50, or 70. Censoring times (denoted as cen.) were generated from an exponential distribution, and the overall censoring fraction in either setup was fixed at 10% or 40%. Table 1 shows the empirical type I error rates. As shown, the proposed procedure for SRC_B and SRC_W preserve reasonable type I error rates. The methods DC_B and DC_W, however, which make double use of the data set, do not control type I error. Note that we do not show type I error for SC_B and SC_W methods, because these methods are typical Cox regression.

Because type I error rates are preserved for both the SRC_B/SRC_W and SC_B/SC_W methods, we compared the power under each method in Table 2. In this simulation, censoring times were also generated from an exponential distribution, and the overall censoring fraction in either

Table 1 Empirical type I error rates

Method	*n*	cen.	The total number of genes			
			10	30	50	70
SRC_B	50	10%	0.052	0.057	0.051	0.048
		40%	0.041	0.047	0.045	0.046
	75	10%	0.052	0.048	0.045	0.046
		40%	0.044	0.046	0.050	0.046
	100	10%	0.056	0.049	0.052	0.052
		40%	0.045	0.044	0.048	0.050
SRC_W	50	10%	0.058	0.046	0.052	0.050
		40%	0.034	0.046	0.036	0.043
	75	10%	0.046	0.042	0.051	0.051
		40%	0.044	0.038	0.044	0.040
	100	10%	0.051	0.046	0.048	0.060
		40%	0.044	0.041	0.046	0.048
DC_B	50	10%	0.937	1.000	1.000	1.000
		40%	0.910	1.000	1.000	1.000
	75	10%	0.944	1.000	1.000	1.000
		40%	0.946	1.000	1.000	1.000
	100	10%	0.957	1.000	1.000	1.000
		40%	0.952	1.000	1.000	1.000
DC_W	50	10%	0.926	1.000	1.000	1.000
		40%	0.916	1.000	1.000	1.000
	75	10%	0.920	1.000	1.000	1.000
		40%	0.936	1.000	1.000	1.000
	100	10%	0.929	1.000	1.000	1.000
		40%	0.933	1.000	1.000	1.000

Empirical type I error rates for comparing SRC_B, SRC_W DC_B and DC_W methods. The value of β was set to 0.

Table 2 Power comparison under two different scenario

n	cen.	Scenarios 1				Scenarios 2			
		SRC_B	SC_B	SRC_W	SC_W	SRC_B	SC_B	SRC_W	SC_W
Strong effect: $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_{30}]^T = [1, 1, ..., 1]^T$									
50	10%	0.757	0.742	0.675	0.650	0.600	0.599	0.723	0.692
	30%	0.624	0.580	0.536	0.490	0.480	0.422	0.546	0.505
	50%	0.448	0.350	0.381	0.312	0.350	0.250	0.359	0.294
75	10%	0.960	0.956	0.907	0.902	0.876	0.870	0.944	0.942
	30%	0.883	0.864	0.828	0.766	0.783	0.771	0.875	0.822
	50%	0.758	0.626	0.690	0.526	0.607	0.494	0.694	0.580
100	10%	0.998	0.997	0.985	0.982	0.974	0.973	0.996	0.995
	30%	0.982	0.974	0.948	0.917	0.940	0.905	0.966	0.955
	50%	0.928	0.846	0.868	0.730	0.806	0.695	0.883	0.802
Low effect: $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_{30}]^T = [0.5, 0.5, ..., 0.5]^T$									
50	10%	0.666	0.625	0.594	0.576	0.266	0.242	0.326	0.305
	30%	0.498	0.487	0.440	0.430	0.206	0.165	0.224	0.214
	50%	0.362	0.285	0.328	0.244	0.144	0.122	0.162	0.124
75	10%	0.930	0.924	0.859	0.850	0.492	0.466	0.574	0.570
	30%	0.824	0.756	0.756	0.688	0.370	0.325	0.432	0.421
	50%	0.642	0.553	0.571	0.469	0.263	0.206	0.312	0.224
100	10%	0.992	0.990	0.964	0.950	0.662	0.654	0.796	0.792
	30%	0.959	0.944	0.918	0.850	0.558	0.505	0.652	0.594
	50%	0.852	0.760	0.802	0.654	0.412	0.319	0.472	0.370

We compared the power under each method SRC_B/SRC_W and SC_B/SC_W. The first scenario considers 30 disease related genes. The second scenario considers 30 disease related genes, with the other 27 genes considered no effect.

setup was fixed at 10%, 30%, or 50%. We then simulated 30 total genes in one gene set under two different scenarios. The first scenario considers 30 disease related genes. We designed two different levels of effect, strong effect and low effect, as

$$\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_{30}]^{\mathrm{T}} = [1, 1, \ldots, 1]^{\mathrm{T}},$$

$$\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_{30}]^{\mathrm{T}} = [0.5, 0.5, \ldots, 0.5]^{\mathrm{T}},$$

respectively. The second scenario considers 3 disease related genes, with the other 27 genes considered "noise" (i.e., no effect). Strong effect and low effect in this case were set as

$$\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_{30}]^{\mathrm{T}} = [1, 1, 1, 0 \ldots, 0]^{\mathrm{T}},$$

$$\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_{30}]^{\mathrm{T}} = [0.5, 0.5, 0.5, \ldots, 0]^{\mathrm{T}}.$$

Results are shown in Table 2.

For scenario 1, all 30 genes have effects. As expected, the power of the tests increases with increase in total sample size and gene effect, but decreases as the censoring proportion grows. Under the first scenario, the power of the SRC_B method is always better than that of SC_B, and SRC_W is always better than SC_W. This result indicates that treating the compound score as a random covariate yields higher power than treating the score as a fixed covariate. When the sample size is 100, the power average increases 10.6, 3.5, and 0.4 percentage points for 50%, 30%, and 10% censoring, respectively. This is a reasonable result, because the random covariate approach involves fitting a quadratic Cox regression model, $\exp\left(\gamma_0\mu_j + \sigma_j^2\gamma_0^2/2\right)$, instead of $\exp(\gamma_0, \mu_j)$. The quadratic form takes into account the variance of each score, use of the compound score without acknowledgement of covariate error yields lower power.

To further illustrate the effect of treating the compound score as a random covariate, in Figure 2, we show the power curves of the SRC_B, SC_B, SRC_W, and SC_W methods for gene effect varying from -1 to 1. The total sample size was set to 100, and the censoring fraction was fixed at 50%. As shown in Figure 2, difference in power between SRC_B and SC_B and between SRC_W and SC_W increases as gene effect size increases, when gene effect size increases, the quadratic term can more accurately account for variance in the effect. Thus, treating the compound score as a random covariate in the Cox regression model provides greater power.

For the first scenario, tests based on SRC_B have higher power than those based on SRC_W. In the second scenario, however, SRC_W yields higher power than SRC_B. This pattern change seems to relate to the increase in noise in the gene set. For Figure 3, we fixed the total number of genes at 30, sample size, 50; censoring fraction, 10% in one gene set, and show the power curves as gene effect and number of noise genes increase. There are eight lines in Figure 3. All solid lines indicate power curves for SRC_B while all long-dash lines indicate power curves for SRC_W. Situation *a* has 30 disease related genes, *b* has 10 disease related genes and 20 noise genes; *c* has 5 disease related genes and 25 noise genes, and *d* has 3 disease related genes and 27 noise genes. As gene effect increases, all powers increase. As the number of noise genes increases (from *a* to *d*), however, the powers of SRC_B and SRC_W decrease. With no noise genes (case *a*), the power of SRC_B is always greater than that of SRC_W. As the number of noise genes increases, however, the power of SRC_W gradually improves over that of SRC_B. This results from the fact that the compound covariate SRC_W takes into account the variance of $\hat{\beta}$. Similar results were obtained with larger sample size (75, 100) and censoring fraction (30%, 50%). Consequently, if prior biological knowledge indicates many noise genes in a given gene set/pathway, we recommend use of the compound covariate SRC_W over SRC_B.

Figure 4 shows the power curve for SRC_B with different sample sizes and different numbers of disease related genes. In this simulation design, the censoring fraction was fixed at 30%. The effect of all disease related genes

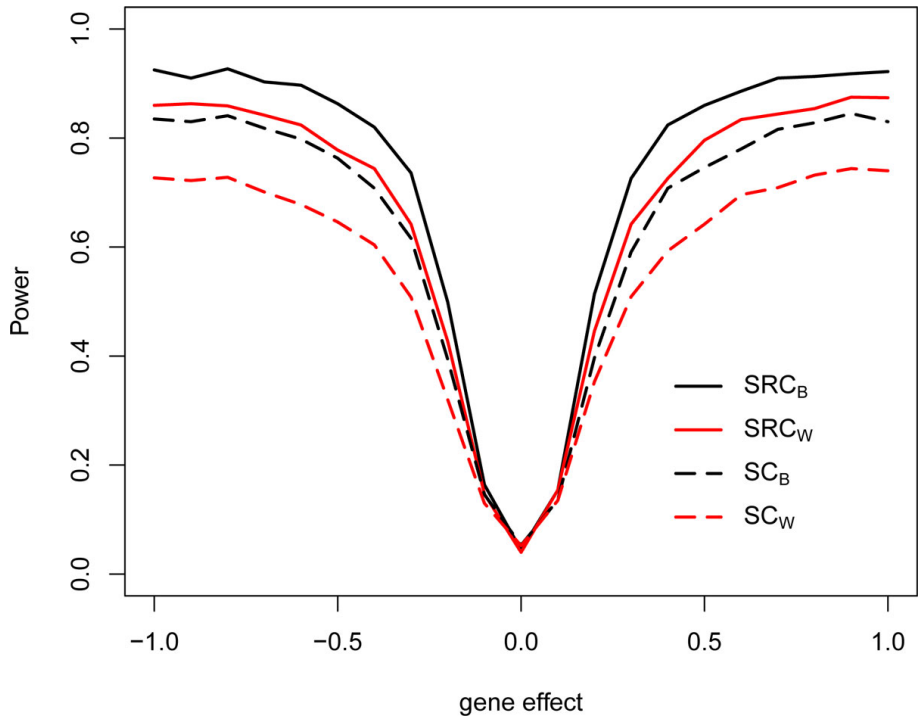

Figure 2 Power curves with varying gene effect.

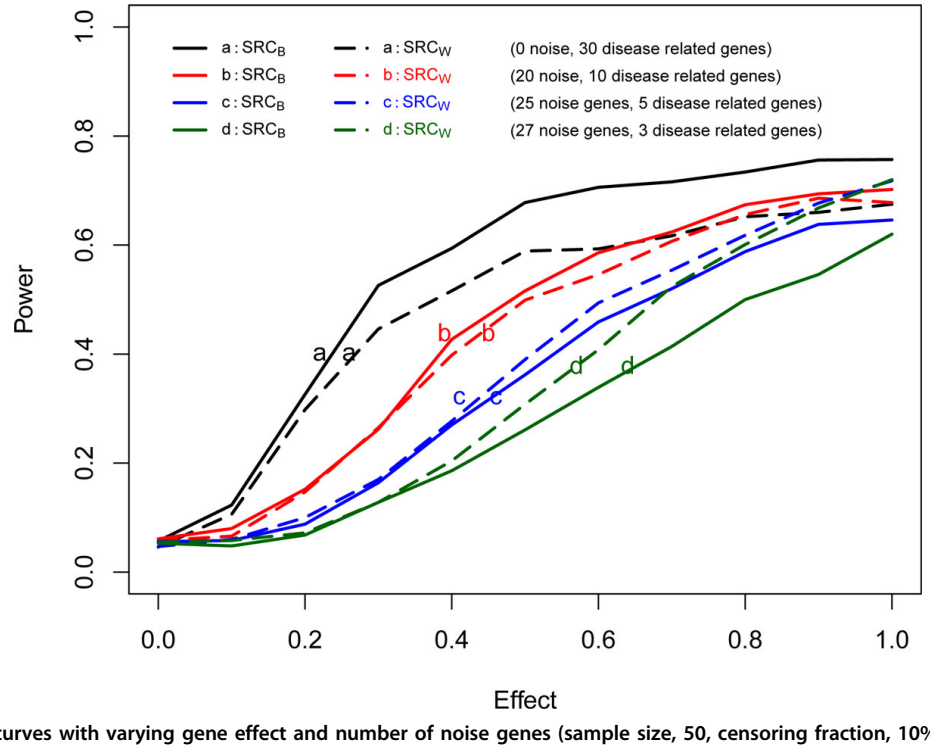

Figure 3 Power curves with varying gene effect and number of noise genes (sample size, 50, censoring fraction, 10%).

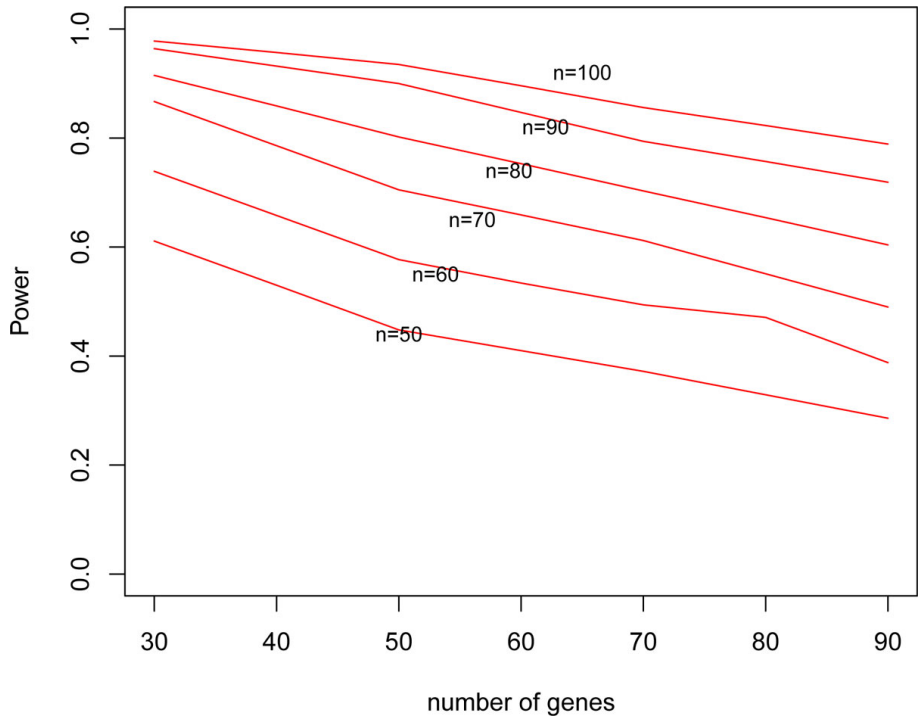

Figure 4 Power curves with different numbers of genes and sample sizes.

was set to 0.5. As expected, the power increases as sample size grows. Power decreases, however, as the number of disease genes increase. Under this specified setting, if we need 80% power and have sample size 100, we can include about 90 disease related genes in this analysis. If we have sample size 60, however, we can only include fewer than 30 disease related genes. This result indicates the need for greater sample size to preserve power when a a gene set includes a large number of disease genes.

Examples

In this section, we demonstrate our methodology using two examples, an Amsterdam 70-gene breast cancer gene signature [1] and a data set involving two pathways for non-small-cell lung cancer. All tests with nominal level 0.05 were applied to the training cohort. The R code for obtaining p-values for the proposed testing procedure is available from the authors upon request.

Breast cancer data set

The well-known Amsterdam 70-gene breast cancer gene signature was published by Van't Veer [1]. To evaluate the previously established 70-gene prognosis file, Van de Vijver [16] further classified an additional 295 consecutive patients with stage I or II breast cancer to validate the breast cancer gene signature. Because the 295 patients are independent of the original data, we re-analyzed this data set using our methodology. In this data set, patients were followed for a median of 7.2 years, with 79 observed deaths. The survival curve is shown in Figure 5 (a) and the testing results, including estimated coefficients (Coef.), relative risk (RR), and p-values are given in Table 3.

Although all coefficients and relative risks are very close, the p-values are very different. When using DC_B and DC_W, the p-values are 8.6×10^{-13} and 1.1×10^{-13}, respectively. When treating the compound covariate as fixed, the p-values of SC_B and SC_W are 1.1×10^{-7} and 1.3×10^{-7}. When using our procedure, the p-values of SRC_B and SRC_W are 1.9×10^{-8} and 1.8×10^{-8}. Although the results remain significant regardless of method, we achieve appropriate p-values for the training cohort, showing that the 70-gene prognosis signature can be used to evaluate early events in breast cancer patients. We get consistent conclusion with the other researches [17,16].

Non-small cell lung cancer data set

We also tested our method by applying it to a publicly available non-small-cell microarray data set downloaded from National Center for Biotechnology Information Gene Expression Omnibus (GSE14814). There are 90 gene expression profiling conducted on mRNA isolated from frozen tumor samples. In this example, two well-known cancer-related pathways were used to test association with survival outcomes for demonstration purposes. The first signaling pathway is the p53 pathway, which is induced by a number of stress signals, including DNA

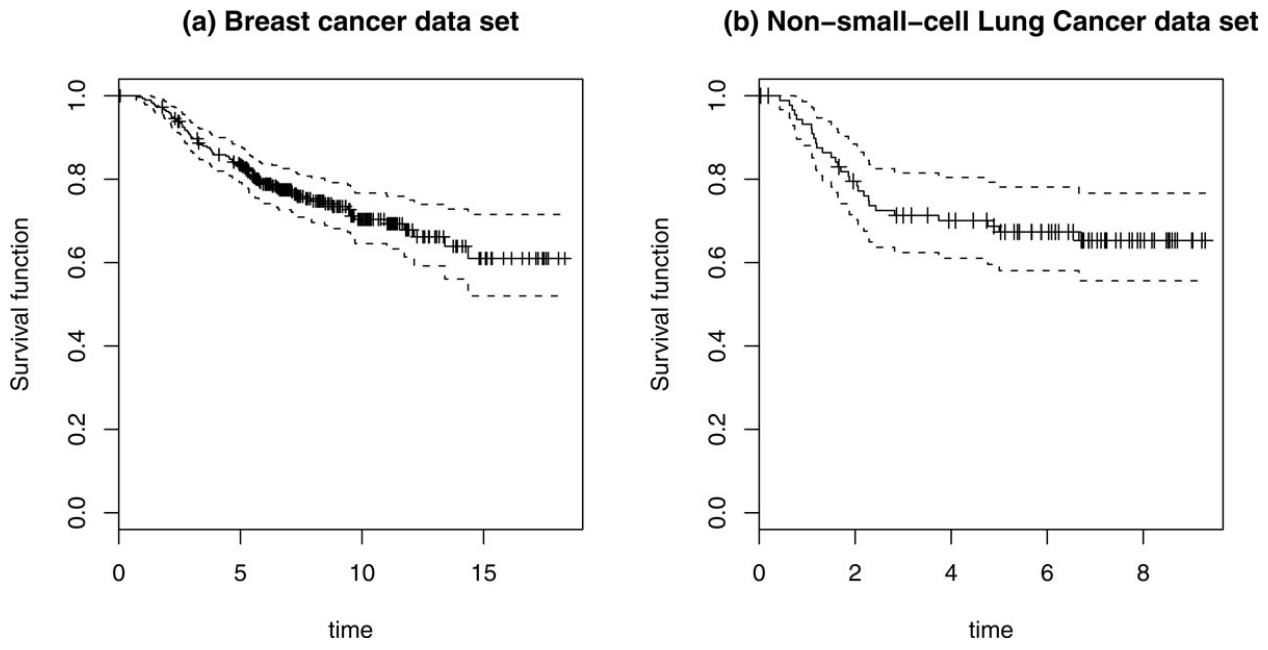

Figure 5 Kaplan-Meier curves for two data sets.

damage, oxidative stress, and activated ontogenesis. The other pathway is the NOD-like receptor signaling pathway, which has been associated with an increased risk for the development of different types of cancer [18]. There are 61 genes in p53 pathway and 54 genes in NOD-like receptor signaling pathway, respectively. The median follow-up time of these patients was 5.4 years, and the number of observed deaths was 29. Figure 5 (b) shows the survival curve, and the test results are given in Table 4.

To summarize all the information, two compound covariates were used. As shown, conventional Cox regression yields overall p-values that are strongly statistically significant (2.29×10^{-6} for DC_B and 1.85×10^{-5} for DC_W). When treating the compound score as a fixed covariate and using a split data set, however, the p-values of SC_B and SC_W become 0.432 for both. When treating the compound score as a random covariate, the p-values of SRC_B and SRC_W become 0.236 and 0.358, respectively. Such divergent p-values suggest that an inappropriate method may well lead to misleading results.

Table 3 Breast cancer data set analysis

Method	Coef	RR	p-value
SRC_B	0.052	1.12	1.9×10^{-8}
SRC_W	0.022	1.12	1.8×10^{-8}
SC_B	0.093	1.10	1.1×10^{-7}
SC_W	0.040	1.04	1.3×10^{-7}
DC_B	0.078	1.08	8.6×10^{-13}
DC_W	0.015	1.02	1.1×10^{-13}

To evaluate the established 70 breast cancer gene signature published by Van't Veer with ther proposed method.

Concluding remarks

In this paper, we focused on survival outcomes and proposed a feasible and correct method for testing the compound covariate to evaluate its association with survival outcomes for training cohort data. We have described the use of a random covariate, SRC_B/SRC_W, to achieve correct testing results for training cohort data and moderately improve power as compared to the use

Table 4 Non-small-cell lung cancer data set analysis

Method	Pathway	Coef	RR	p-value	Overall p-value
SRC_B	NOD	0.033	1.0013	0.59	0.236
	P53	0.037	1.0044	0.67	
SRC_W	NOD	0.016	1.0063	0.37	0.358
	P53	0.001	1.0002	0.99	
SC_B	NOD	0.077	1.08	0.36	0.432
	P53	0.015	1.01	0.90	
SC_W	NOD	0.034	1.03	0.24	0.432
	P53	-0.01	0.99	0.74	
DC_B	NOD	0.072	1.07	0.37	2.29×10^{-6}
	P53	0.314	1.37	0.003	
DC_W	NOD	0.019	1.02	0.21	1.85×10^{-5}
	P53	0.055	1.06	0.006	

To evaluate the 90 gene expression profiling from National Center for Biotechnology Information Gene Expression Omnibus (GSE14814). The first signaling pathway is the p53 pathway. The other pathway is the NOD-like receptor signaling pathway.

of SC_B/SC_W. Simulation study shows that our proposed method performs consistently better than SC_B/SC_W, because the quadratic term utilized in the SRC_B/SRC_W method takes into account error in the compound covariate. We further found that an increase in sample size improves power when there is a high proportion of censored data. If the gene set of interest includes noise genes, we suggest that the compound covariate SRC_W is a better choice than SRC_B; whether noise genes or non-functionally related genes are hidden in gene set is a judgment call for a geneticist. In addition, we contend that a flaw of biomedical papers concerned with such topics report an bias p-value based on flawed compound covariate analysis for the same training data set. In this paper, we use a well-known 2-fold concept, with one part of the data to built compound covariate and the remainder part for testing if there is association between survival outcomes and the score to ensure correct p-values in the training data set. Note that we need to check the proportional hazards assumption.

Our method can simultaneously test for more than one gene set in a training cohort data. More generally, this procedure can be applied not only for survival outcomes, but also for binary or continuous outcomes. The weighted flexible compound covariate method WFCCM [19], an extension of the compound covariate, also allows for use of the method of statistical analysis presented here. In addition, this method can easily be extended to consider the interaction between random covariates and clinical observed covariates, as

$$h(t|z, \mathbf{w}) = h_0(t)\, E\left[\exp\left(\gamma_0 z + \boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{w} + \boldsymbol{\gamma}_2^{\mathrm{T}}(z\mathbf{w})\right)\right]$$

using the same analysis procedure. The chosen weight $\hat{\beta}$ or $\hat{w}$ can be adjusted by the other clinical observed covariates in the proposed framework. Our method, however, cannot be used to test the interaction between two random covariates, because of the complexity of specifying the distribution for the interaction between two random covariates; this is an area worthy of further investigation. Another one potential practical concern of the proposed method is that sample size must not be too small, higher fractions of censored data create the need for further increased sample size. Similarly, to achieve high power when studying a large number of genes, greater sample size is needed. When studying a large number of genes, ignoring the covariance that exists between genes does not influence the type I error rate, however, taking the covariance into account may increase power. Further research is required to address these limitations. Note that the permutation test (e.g., [20]) might be another method to calculate an appropriate p-value for the training dataset, however, with the permutation test, weights are not easily adjusted by the other covariates. Even for a small gene set, this approach may appear too expensive in computer time.

Additional material

Additional file 1: The explicit forms of *a*, *b* and *c*. Additional file 1 is a PDF file which shows the explicit forms of *a*, *b* and *c*. Then, the score statistic can be derived.

Additional file 2: The derivation of variance for compound covariates. Additional file 2 is a PDF file which shows the derivation of variance for compound covariates.

Acknowledgements
The authors wish to thank editorial assistants, Lynne Berry and Yvonne Poindexter, for their editorial work on this manuscript. This work was supported by National Cancer Institute (grant numbers P30 CA068485, P50 CA095103, P50 CA090949, P50 CA098131).
This article has been published as part of *BMC Systems Biology* Volume 6 Supplement 3, 2012: Proceedings of The International Conference on Intelligent Biology and Medicine (ICIBM) - Systems Biology. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcsystbiol/supplements/6/S3.

Author details
[1]Center for Quantitative Sciences, Vanderbilt University, Nashville, TN, USA. [2]Department of Biostatistics, Vanderbilt University, Nashville, TN, USA.

Authors' contributions
PFS developed the mathematical derivations, designed and performed the simulation, and drafted the manuscript. Xi and HC performed the experimence about microarray analysis and gave valuable advice related to the compound score issues. YS supervised the research and managed this project. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Published: 17 December 2012

References
1. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend S: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415**:530-536.
2. Wang Y, Klijn J, Zhang Y, Sieuwerts A, Look M, Yang F, Talantov D, Timmermans M, Meijer-van Gelder M, Yu J, Jatkoe T, Berns E, Atkins D, Foekens J: **Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.** *Lancet* 2005, **365**:671-679.
3. Tukey JW: **Tighening the clinical trial.** *Control Clin Trials* 1993, **14**:266-285.
4. Tomasson H: **Risk scores from logistic regression: unbiased estimates of relative and attributable risk.** *Stat Med* 1995, **14**:1331-1339.
5. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A, Trent J, Raffeld M, Yakhini Z, Ben-Dor A, Dougherty E, Kononen J, Bubendorf L, Fehrle W, Pittaluga S, Gruvberger S, Loman N, Johannsson O, Olsson H, Sauter G: **Gene-expression profiles in hereditary breast cancer.** *N Engl J Med* 2001, **344**:539-548.
6. Lossos IS, Czerwinski DK, Alizadeh AA: **Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes.** *N Engl J Med* 2004, **350**:1828-1837.
7. Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, Cheng CL, Wang CH, Terng HJ, Kao SF, Chan WK, Li HN, Liu CC, Singh S, Chen WJ, Chen JJ, Yang PC: **A five-gene signature and clinical outcome in non-small-cell lung cancer.** *N Engl J Med* 2007, **356**:11-20.
8. Hsu YC, Yuan S, Chen HY, Yu SL, Liu CH, Hsu PY, Wu G, Lin CH, Chang GC, Li KC, Yang PC: **A four-gene signature from NCI-60 cell line for survival**

pre-diction in non-small cell lung cancer. *Clin Cancer Res* 2009, **15**:7309-7315.
9. Beer DG, Kardia SLR, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JMG, Iannettoni MD, Orringer MB, Hanash S: **Gene-expression profiles predict survival of patients with lung adenocarcinoma.** *Nat Med* 2002, **8**:816-824.
10. Cox DR: **Regression models and life-tables.** *J R Statist Soc B* 1972, **34**:187-220.
11. Salmon S, Chen H, Chen S, Herbst R, *et al*: **Classification by mass spectrometry can accurately and reliably predict outcome in patients with non-small cell lung cancer treated with erlotinib-containing regimen.** *J Thorac Oncol* 2009, **4**:689-696.
12. Prentice RL: **Covariate measurement errors and parameter estimation in a failure time regression model.** *Biometrika* 1982, **69**:331-342.
13. Zhao H, Tibshirani R, Brooks J: **Gene expression profiling predicts survival in conventional renal cell carcinoma.** *PLoS Med* 2005, **3**:115-124.
14. Efron B, Tibshirani R: **On testing the significance of sets of genes.** *Ann Appl Stat* 2007, **1**:107-129.
15. Kaplan EL, Meier P: **Nonparametric estimator from incomplete observations.** *J Amer Stat Assoc* 1958, **53**:457-481.
16. van de Vijver MJ, He YD, van 't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R: **A gene-expression signature as a predictor of survival in breast cancer.** *N Engl J Med* 2002, **347**:1999-2009.
17. Buyse M, Loi S, van't Veer L, Viale G, Delorenzi M, Glas AM, Saghatchian d'Assignies M, Bergh J, Lidereau R, Ellis P, Harris A, Bogaerts J, Therasse P, Floore A, Amakrane M, Piette F, Rutgers E, Sotiriou C, Cardoso F, Piccart MJ: **Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer.** *J Nat Cancer Inst* 2006, **98**:1183-1192.
18. Rosenstiel P, Till A, Schreiber S: **NOD-like receptors and human diseases.** *Microb Infect* 2007, **9**:648-657.
19. Shyr Y, Kim K: **Weighted flexible compound covariate method for classifying microarray data.** *A practical approach to microarray data analysis* Norwell: Kluwer Academic Publishers, New York; 2003, 186-200.
20. Rdmacher MD, McShane LM, Simon R: **A paradigm for class prediction using gene expression profiles.** *J Comput Biol* 2002, **9**:505-511.

doi:10.1186/1752-0509-6-S3-S11
Cite this article as: Su *et al.*: Statistical aspects of omics data analysis using the random compound covariate. *BMC Systems Biology* 2012 **6** (Suppl 3):S11.