RESEARCH ARTICLE Open Access

Identifying common prognostic factors in genomic cancer studies: A novel index for censored outcomes

Sigrid Rouam*[1,2], Thierry Moreau[3] and Philippe Broët[1,2]

Abstract

Background: With the growing number of public repositories for high-throughput genomic data, it is of great interest to combine the results produced by independent research groups. Such a combination allows the identification of common genomic factors across multiple cancer types and provides new insights into the disease process. In the framework of the proportional hazards model, classical procedures, which consist of ranking genes according to the estimated hazard ratio or the p-value obtained from a test statistic of no association between survival and gene expression level, are not suitable for gene selection across multiple genomic datasets with different sample sizes. We propose a novel index for identifying genes with a common effect across heterogeneous genomic studies designed to remain stable whatever the sample size and which has a straightforward interpretation in terms of the percentage of separability between patients according to their survival times and gene expression measurements.

Results: The simulations results show that the proposed index is not substantially affected by the sample size of the study and the censoring. They also show that its separability performance is higher than indices of predictive accuracy relying on the likelihood function. A simulated example illustrates the good operating characteristics of our index. In addition, we demonstrate that it is linked to the score statistic and possesses a biologically relevant interpretation.

The practical use of the index is illustrated for identifying genes with common effects across eight independent genomic cancer studies of different sample sizes. The meta-selection allows the identification of four genes (*ESPL1*, *KIF4A*, *HJURP*, *LRIG1*) that are biologically relevant to the carcinogenesis process and have a prognostic impact on survival outcome across various solid tumors.

Conclusion: The proposed index is a promising tool for identifying factors having a prognostic impact across a collection of heterogeneous genomic datasets of various sizes.

Background

In clinical cancer research, recent advances in genome-wide technologies have enabled researchers to identify large-scale genomic changes having a potential prognostic impact on time-to-event outcomes. The growing number of public repositories for high-throughput genomic data facilitates the retrieval and combination of various datasets produced by independent research groups (for a few: *GEO* [1], *Oncomine* [2], *ArrayExpress* [3]). These databases potentially represent valuable resources for identifying genomic factors that have a common prognostic impact on clinical outcomes (e.g. time to local or distant recurrence) across multiple cancer types. However, the joint analysis of these heterogeneous datasets is difficult due to the fact that they are usually of varying sample size, investigate different survival outcomes or are related to different tumors entities. In this context, defining a procedure for identifying common genomic risk factors across multiple heterogeneous datasets is a promising but very challenging task. In recent years, several authors [4-7] have proposed meta-profiling methods for class comparison, designed to identify common transcriptional features of the tumoral process (normal versus tumor state).

In the framework of the widely used Cox model [8] for analyzing possibly censored time-to-event or survival data,

* Correspondence: sigrid.rouam@inserm.fr
[1] Computational and Mathematical Biology, Genome Institute of Singapore, Singapore 138672, Singapore
Full list of author information is available at the end of the article

different procedures for feature selection across multiple gene expression datasets can be defined. Basically, each gene expression measurement is included in a simple Cox model, giving rise to an estimation of the corresponding hazard ratio and to a statistic for testing the null hypothesis of no association between survival outcome and gene expression changes. Simple procedures, frequently used in practice, consist of ranking the genes in each dataset from the highest (or lowest) value to the lowest (or highest) value according to either the estimated hazard ratio or quantities derived from the test statistic (e.g. p-value), and finally to select those that appear at the intersection of the lists using a defined thresholding procedure [9]. However, these approaches suffer serious drawbacks that are mostly related to the chosen selection criteria. Choosing the estimated hazard ratio clearly ignores the variability of the data, while the choice of quantities derived from test statistics leads to emphasize large datasets, since it is well known that every test statistic increases with the sample size.

In meta-selection of heterogeneous genomic datasets, taking into account both the magnitude of the prognostic impact of factors and the variability of the data without being highly dependent on the sample size is likely to be more biologically relevant. Addressing this issue led us to propose a novel index designed for genomic survival analysis that provides information about the capability of a genomic factor to separate patients according to their time-to-event outcome. Our work shares conceptual links with the framework of predictive ability measures that aim to determine which covariates have the greatest explanatory interest. For censored data, two main frameworks have been proposed for quantifying the predictive ability of a variable to separate patients: (i) concordance, which quantifies the degree of agreement among the ranking of observed failure times according to the explained variables and is used to assess the discriminatory performance of a model [10,11]; (ii) proportion of explained variation, which quantifies the relative gain in prediction ability between a covariate-based model and a null model (without explained variables) by analogy with the well-known linear model. In this latter case, two approaches have been considered. The first one focuses on comparing empirical survival functions with and without covariates [12-15]. The second one considers statistical quantities which are directly or indirectly related to the likelihood function [16-19]. In this paper, we propose a novel index that is linked to the approach discussed above. It is related to the score statistic and well-suited for meta-selection of genomic datasets. Our index is interpreted as the ability of a gene to separate patients observed to experience the event of interest from those who do not experience the event among the risk set at every observed failure time. As shown in this study, increasing values of the index correspond to a higher effect due to the gene variable. In contrast to a test statistic, our index is not highly sensitive to sample size variation which makes it well-suited for meta-selection from datasets with various sample sizes.

We report and discuss the statistical properties of the index obtained from simulation experiments, and compare it to Allison's index [16] and its modified version [18], Nagelkerke [17] and Xu and O'Quigley's [19] indices. In addition, the properties of these indices are illustrated on a fictitious example, where data are simulated so as to mimic a real study combining datasets of different sample sizes. We then illustrate the capability of the index for combining the results of eight cancer studies of different sample sizes and with different outcomes.

Results

Statistical properties of the proposed index and comparison with classical indices

Simulation Scheme

A simulation study was performed to evaluate the behavior of the proposed index, denoted $\mathbf{D}_0^*$ and compare it to Allison's index [16], a modified version of Allison's index [18], Nagelkerke [17] and Xu and O'Quigley's [19] indices denoted $\rho_N^2, \rho_k^2, R_N^2$ and ρ_{XOQ}^2 respectively (see the Methods Section for the description of the five indices) under proportional and non-proportional hazards regression models, using different values of the regression parameter, different covariate distributions and different sample sizes. Scenarios with various independent censoring distributions were also considered.

The simulation protocol was as follows. For each subject i, $i = 1,\cup, n$, we considered one covariate Z with either a discrete (Bernoulli $V(0.5)$) or a continuous (uniform $\mathcal{U}$ [0, $\sqrt{(3)}$]) distribution. These two distributions of Z were standardized to have the same variance. Survival times T were generated with the survival function $S(t; z) = \exp(-te^{\beta Z})$ (proportional hazards model) or $S(t; z) = (1 + t \cdot e^{\beta Z})^{-1}$ (proportional odds model). For these two survival distributions, the hazard ratios were $HR = e^{\beta}$ for the proportional hazard model and $HR = [1 + (e^{\beta} - 1)S_0(t)]^{-1}$ for the proportional odds model, $S_0(t)$ refering to the baseline survivor function. In our simulation scheme, e^{β} was set to 1 (null effect), to small values; 1.25, 1.5, 1.75, medium values; 2, 3, and high values; 4, 5. The sample sizes n of the data were taken equal to 50, 100, 500 and 1, 000.

The censoring mechanism was assumed to be independent from T given Z and the distribution of the censoring variable C_i, $i = 1,\cup, n$ was either uniform $C_i \sim \mathcal{U}\{0, r\}$ or exponential $C_i \sim \mathcal{E}\{\gamma\}$. The calculation of the parameters r and γ as functions of the expected overall percentage of censoring p_c is described in Additional file 1. The percent-

age of censoring was taken equal to 0%, 25% and 50%. For each configuration 1,000 repetitions were generated.

Simulation Results

The table in Additional file 2 displays the results of the simulations for $\mathbf{D}_0^*$ for four different sample sizes and two different covariate distributions, considering a Cox proportional hazards model. As seen from Additional file 2, when β = 0, i.e. in the absence of covariates, our index approaches 0 for n = 50 to 1, 000; the separability is close to 0. The index increases towards 1 with |β|, the separability increases with the effect of the covariate. When $\beta \neq 0$, the value of $\mathbf{D}_0^*$ for the different sample sizes is fairly stable, in particular for moderate or high effects ($e^{\beta} \geq 1.5$). The mean values of our index for n = 50 to 500 are close to the mean values obtained for n = 1, 000 which is assumed to approach its asymptotic limit. The standard errors of $\mathbf{D}_0^*$ (indicated in brackets in Additional file 2) are small even when censored, and, as expected, decrease when n increases. Our index is slightly sensitive to the censoring rate, especially for high values of hazard ratio. Similar comments can be made when dealing with an exponential censoring mechanism (results not shown).

Figures 1, 2, 3 and 4 display, for a Cox model, the differences δ between the mean of $\mathbf{D}_0^*$ and the mean of $\rho_N^2, \rho_k^2, R_N^2$ and ρ_{XOQ}^2 respectively, for n = 100, for different percentage of censoring p_c, different covariate distributions and with a uniform censoring mechanism. The means of the differences δ are always positive. They are close to zero for small hazard ratios and increase with higher hazard ratios. The differences between $\mathbf{D}_0^*$ and R_N^2 increase with the percentage of censoring, which is not surprising since the Nagelkerke's index is known to be sensitive to censoring [15]. The two indices ρ_k^2 and ρ_{XOQ}^2 have a similar behavior relatively to $\mathbf{D}_0^*$. This is expected since O'Quigley *et al* [18] propose to use ρ_k^2 as a simple working approximation of their index. The same results are obtained for n = 50, 500 and 1, 000 and for an exponential censoring mechanism (results not shown). For $e^{\beta} \geq 2$, the 95% confidence interval for the differences of the three graphs does not comprise 0, thus in each case the difference δ is significant. The table in Additional file 3 and Figures 5, 6, 7 and 8 display the results of the simulations under a proportional odds model. The mean values of the different indices are lower than in the case of a proportional hazards model. All indices are more sensitive to censoring. Our index shows higher mean values than the other indices, especially in case of a Bernoulli distribution.

Evaluation of the index in meta-selection

Simulation Scheme

In this subsection, using a basic example we evaluated the practical interest of our index $\mathbf{D}_0^*$ when combining the information contained in two studies with different sample sizes. The method used to generate the two datasets was inspired by Bair and Tibshirani [20] but modified in order to resemble the structure of real genomic data. The two datasets mimicked the analysis of the prognosis impact of transcriptional changes for a set of 1, 000 genes. The two datasets were of unequal size and composed of n = 150 and 50 individuals, respectively. To each individual i, i = 1,∪, n; n = 150 or 50, we associated a survival time T_i, a censoring time C_i and vector of 1,000 quantitative values $Z_i = \{ Z_i^{(g)} ; g = 1,\cup, 1, 000\}$ (e.g. expression measurement).

To perform a fair evaluation of our index, we simulated survival data with either an exponential distribution given by $S(t) = \exp(-te^{\xi})$ (proportional hazard model) or a log-logistic survival distribution given by $S(t) = (1 + t \cdot e^{\xi})^{-1}$ (non-proportional hazard model). For individuals i such as $1 \leq i \leq n/2$ (n = 150 or 50), the parameter ξ was equal to 0. For individuals i such as $n/2 + 1 \leq i \leq n$ (n = 150 or 50), e^{ξ} was equal to 3 and 5. We defined individuals i = 1 to $n/2$ as belonging to the group of patients with low risk of occurrence of the event of interest and individuals from $i = n/2 +$ 1 to n to the group with high risk of occurrence of the event.

For each dataset, censoring times C_i (i = 1,∪, n; n = 150 or 50) were considered independent from survival times and with a uniform distribution on {0, r}, r chosen in order to have an expected percentage of censoring of 30%.

The observed time to follow-up T_i^* (i = 1,∪, n; n = 150 or 50) was equal to the minimum between the two previously defined times T_i and C_i.

For the two datasets, for each individual i, 1,000 gene expression values $X_i^{(g)}$ (i = 1,∪, n; n = 150 or 50; g = 1,∪, 1, 000) were generated, according to the simulation scheme shown on the figure in Additional file 4. Gene expression values from g = 1 to 50 for individuals i = 1 to $n/2$ (n = 150 or 50) followed a log-normal distribution Log-$\mathcal{N}$ (μ = 4, σ = 1.5) with $E(X) = e^{\mu+0.5\sigma^2}$ and $V(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$. For the rest of the individuals (i = $n/2$ + 1,∪, n; n = 150 or 50), gene expression values followed a distribution Log-$\mathcal{N}$ (0, 1.5). Gene expression values from g = 51 to 100 for individuals i = 1 to $n/2$ (n = 150 or 50) followed a log-normal distribution with parameters μ = 3 and σ = 1.5 Log-$\mathcal{N}$ (3, 1.5). For the rest of the individuals (i = $n/2$ + 1,∪, n; n = 150 or 50), gene expression values followed a distribution Log-$\mathcal{N}$ (0, 1.5). For gene

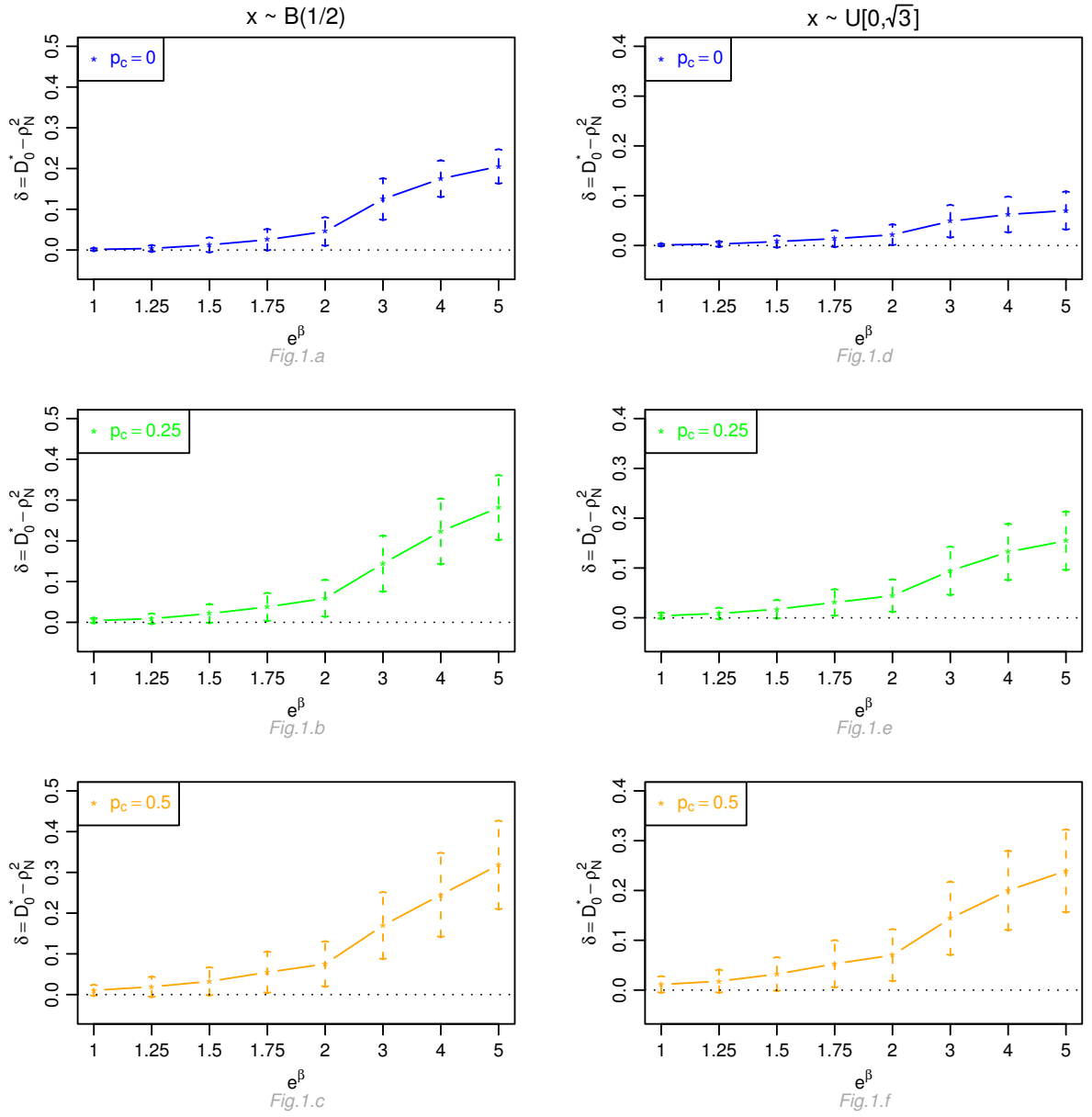

Figure 1 Graphic of the differences *δ* between the mean values of $\mathbf{D}_0^*$ and the mean values of ρ_N^2 as a function of the hazard ratio, for a Cox proportional hazards model. Mean of $\mathbf{D}_0^* - \rho_N^2$ as a function of the relative risk e^{β}, for different percentages of censoring p_c, for a covariate with Bernoulli $\mathbf{V}(1/2)$ or uniform $\mathcal{U}$ [0, $\sqrt{(3)}$] distribution, $n = 100$ and a uniform censoring mechanism.

expression values from $g = 100$ to 150 and for 40% individuals randomly selected among the n (n = 150 or 50), the $X_i^{(g)}$ (i 8 N, g = 100,∪, 150) followed a log-normal distribution Log-$\mathcal{N}$ (1, 1.5), whereas for the remaining individuals, they followed a log-normal distribution Log-$\mathcal{N}$ (0, 1.5). For gene expression values from $g = 151$ to 250 and for 50% individuals randomly selected among the n, the $X_i^{(g)}$ (i 8 N; g = 151,∪, 250) followed a log-normal distribution Log-$\mathcal{N}$ (0.5, 1.5), whereas for the remaining individuals, they followed a log-normal distribution Log-$\mathcal{N}$ (0, 1.5). For gene expression values from $g = 251$ to 350 and for 70% individuals randomly selected among the n, the

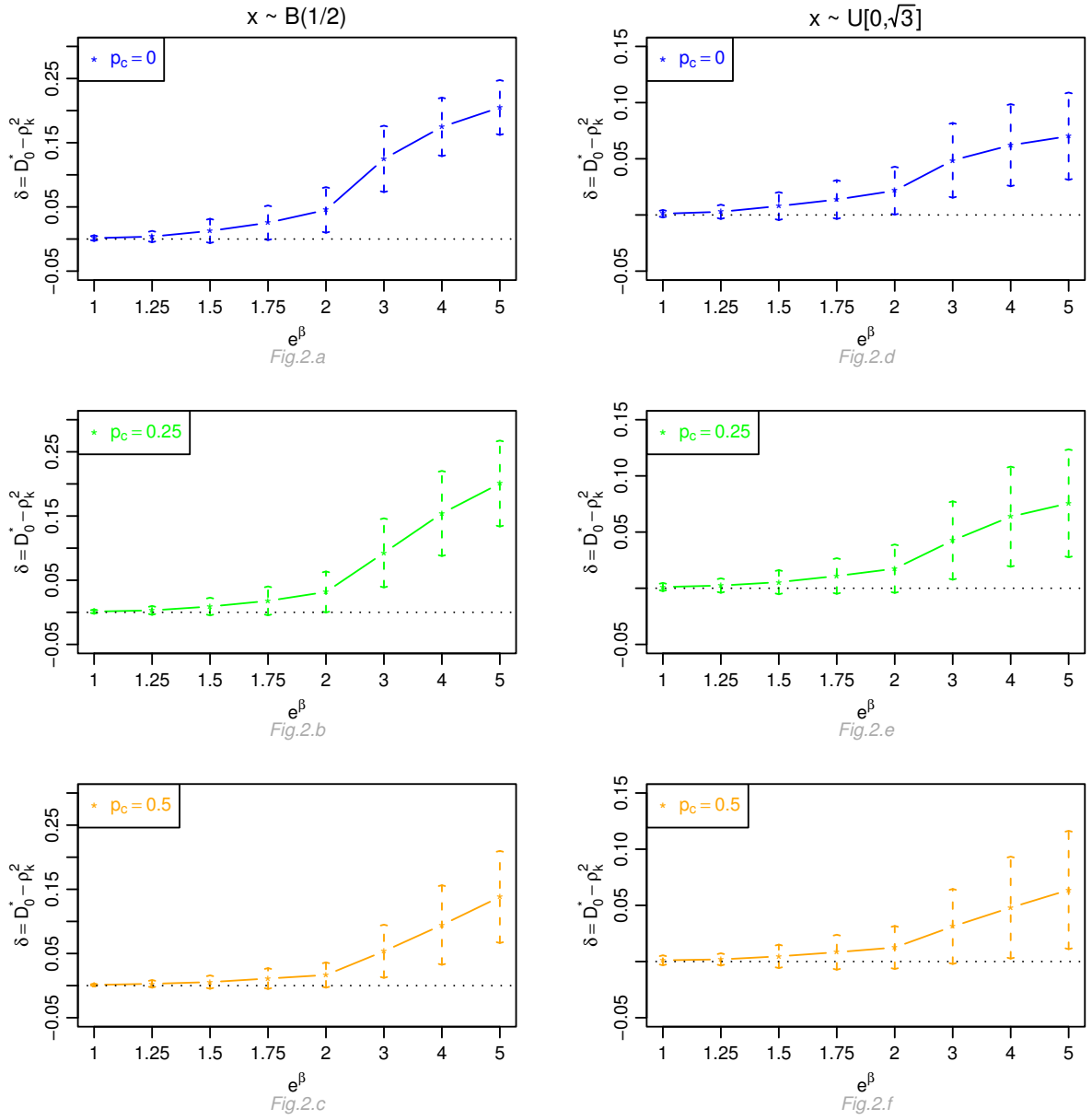

Figure 2 Graphic of the differences *δ* between the mean values of $\mathbf{D}_0^*$ and the mean values of ρ_k^2 as a function of the hazard ratio, for a Cox proportional hazards model. Mean of $\mathbf{D}_0^* - \rho_k^2$ as a function of the relative risk e^{β}, for different percentages of censoring p_c, for a covariate with Bernoulli $\mathbf{V}(1/2)$ or uniform $\mathcal{U}$ [0, $\sqrt{(3)}$] distribution, $n = 100$ and a uniform censoring mechanism.

$X_i^{\{g\}}$ (i 8 N; g = 251,∪, 350) followed a log-normal distribution Log-$\mathcal{N}$ (0.1, 1.5), whereas for the remaining individuals, they followed a log-normal distribution Log-$\mathcal{N}$ (0, 1.5). Finally, for gene expression values from g = 351 to 1, 000, the $X_i^{\{g\}}$ (i = 1,∪, n; g = 351,∪, 1, 000) followed a log-normal distribution Log-$\mathcal{N}$ (0, 1.5) for all individuals.

As genes involved in the same or related pathway are likely to be coexpressed, we introduced correlations between genes. To evaluate the behavior of our index in the context of dependent data, we generated datasets with so-called "clumpy" dependence (gene measurements are

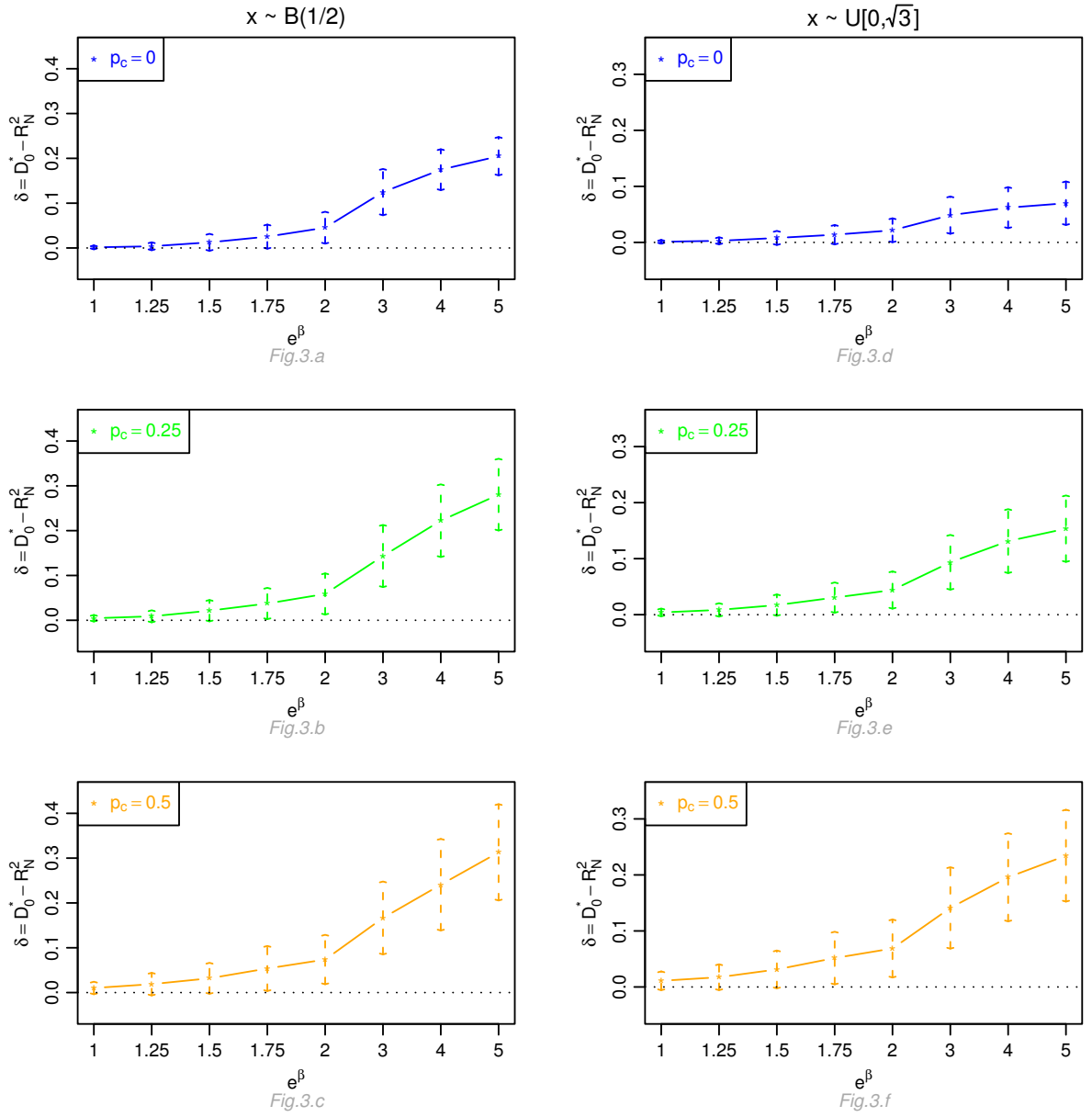

Figure 3 Graphic of the differences *δ* between the mean values of $\mathbf{D}_0^*$ and the mean values of R_N^2 as a function of the hazard ratio, for a Cox proportional hazards model. Mean of $\mathbf{D}_0^* - R_N^2$ as a function of the relative risk e^{β}, for different percentages of censoring p_c, for a covariate with Bernoulli $\mathbf{V}(1/2)$ or uniform $\mathcal{U}$ [0, $\sqrt{(3)}$] distribution, $n = 100$ and a uniform censoring mechanism.

dependent in small groups, but each group is independent from the others). We applied the following protocol [21,22]. For each group of ten genes indexed by l, $l = 1,\cup, 100$, a random vector $A = a_{il}$, $i = 1,\cup, n$, was generated from a standard log-normal distribution Log-$\mathcal{N}(0, 1)$. The data matrix Z was then built so that $Z_{il}^{(g)} = \sqrt{\rho} \cdot A_{il} + \sqrt{1-\rho} \cdot X_{il}^{(g)}$ with ρ equal to 0.25, 0.5 or 0.75. Finally and in order to show the behavior of our index in situations close to real genomic data analysis, we standardized the dataset using classical quantile normalization [23].

In this simulation scheme, the first hundred genes were differentially expressed between the low and high risk

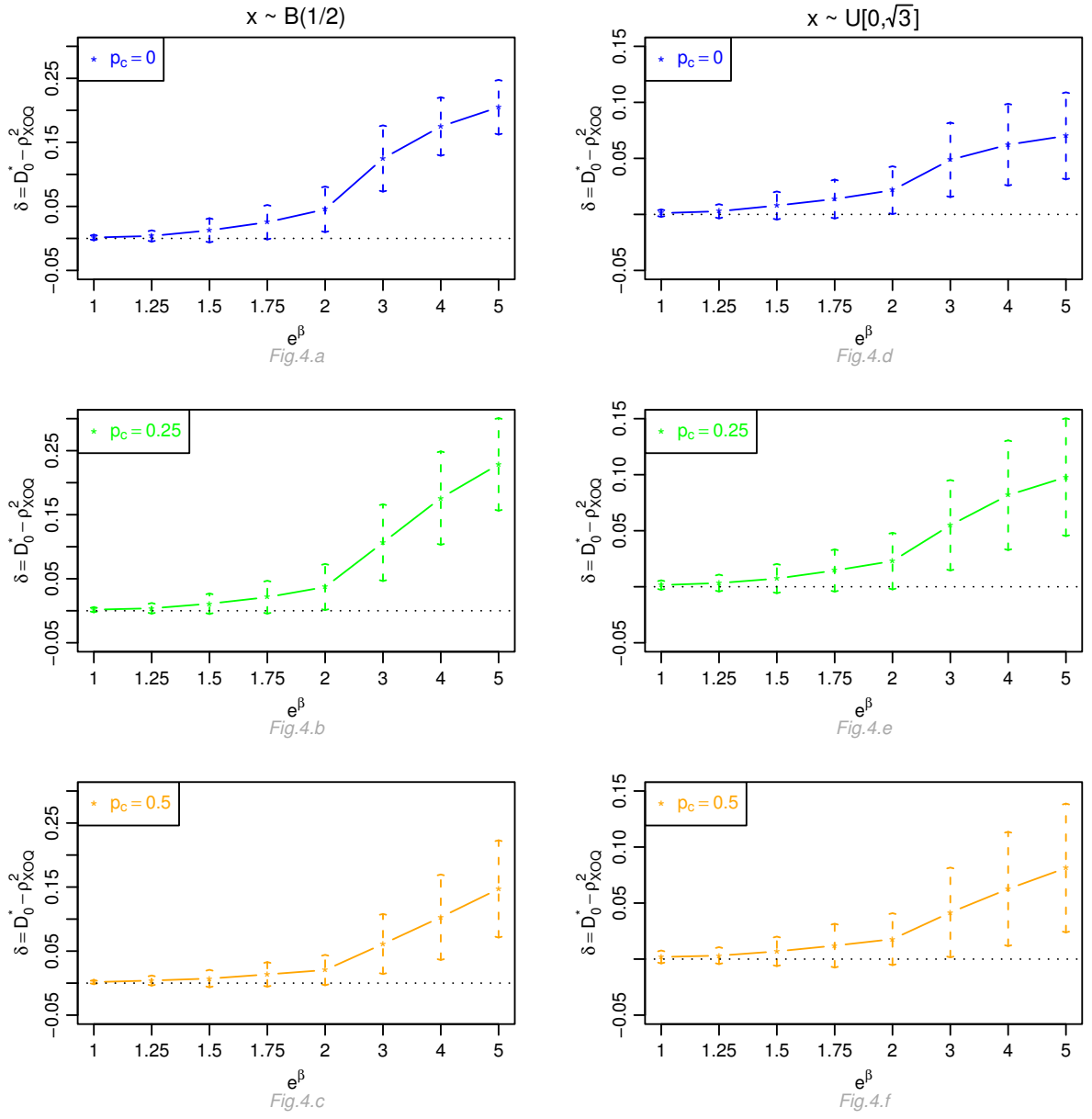

Figure 4 Graphic of the differences *δ* between the mean values of $\mathbf{D}_0^*$ and the mean values of ρ_{XOQ}^2 as a function of the hazard ratio, for a Cox proportional hazards model. Mean of $\mathbf{D}_0^* - \rho_{XOQ}^2$ as a function of the relative risk e^{β}, for different percentages of censoring p_c, for a covariate with Bernoulli $\mathbf{V}(1/2)$ or uniform $\mathcal{U}$ [0, $\sqrt{(3)}$] distribution, $n = 100$ and a uniform censoring mechanism.

group of patients. The other 250 genes were not linked to the low and high risk status, but were distributed differentially according to a binary factor (with various means) unlinked to the low/high risk status. The remaining genes were not linked to the low and high risk status.

For a given threshold, we calculated the number of genes common to the two simulated datasets with the five indices, for the different survival distributions, the different hazards ratio values and the different correlations between genes. We estimated the true positive fraction (TPF, number of true positives found divided by the number of truly prognostic genes) and the true negative fraction (TNF, number of true negatives divided by the number of truly non-prognostic genes) obtained with the five indices, $\mathbf{D}_0^*$,

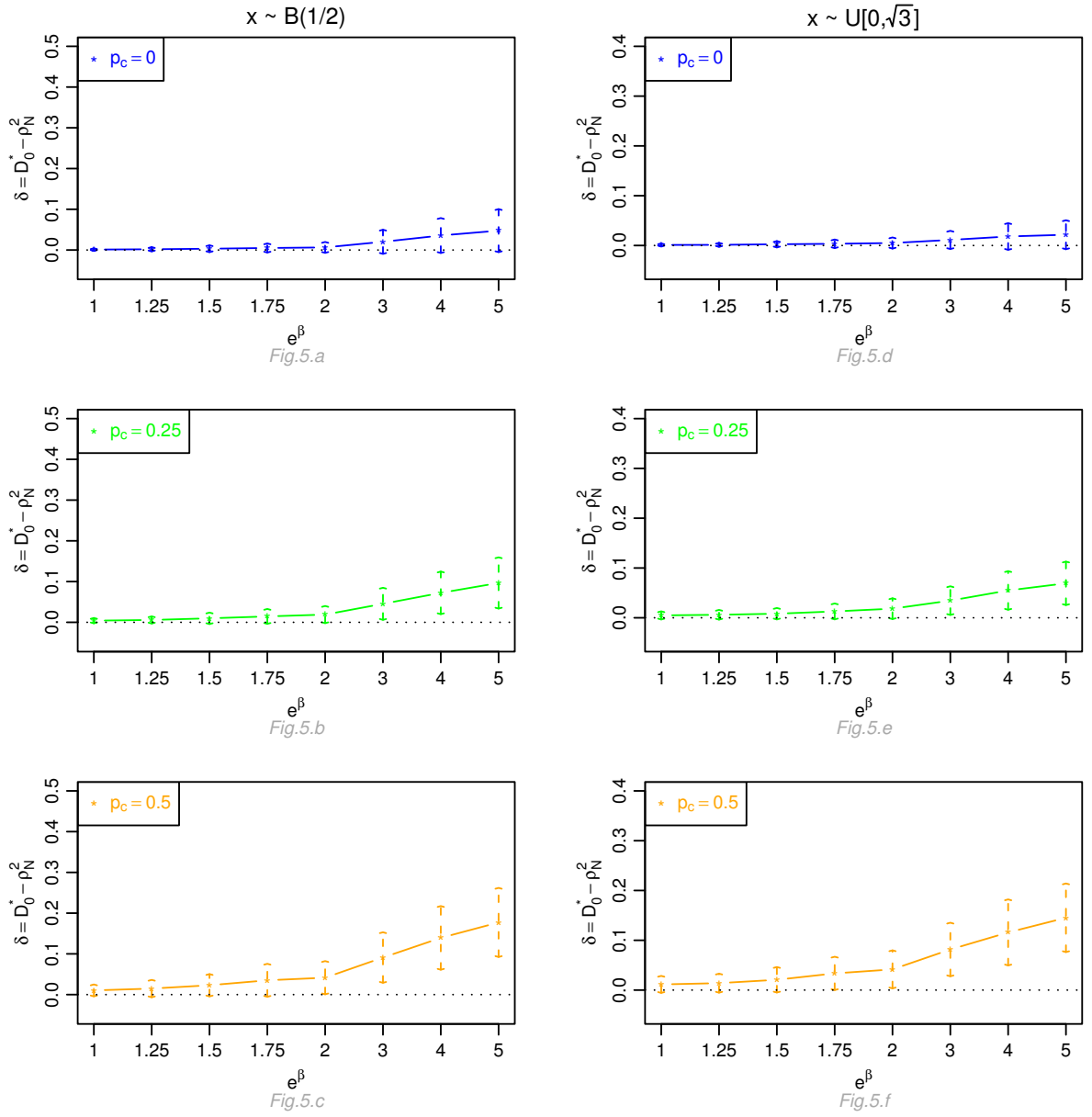

Figure 5 Graphic of the differences *δ* between the mean values of $\mathbf{D}_0^*$ and the mean values of ρ_N^2 as a function of the odds ratio, for a proportional odds model. Mean of $\mathbf{D}_0^* - \rho_N^2$ as a function of the odds ratio e^{β}, for different percentages of censoring p_c, for a covariate with Bernoulli $\mathbf{V}(1/2)$ or uniform $\mathcal{U}$ [0, $\sqrt{(3)}$] distribution, $n = 100$ and a uniform censoring mechanism.

$\rho_N^2, \rho_k^2, R_N^2$ and ρ_{XOQ}^2 as a function of the threshold target value. These criteria were estimated by the mean over one hundred iterations of: (i) the proportion of correct selection (i.e. when the selected genes g belonged to $\{0,\cup, 100\}$) among the modified genes; (ii) the proportion of correct 'non-selection' (i.e. when the selected genes g belonged to $\{101,\cup, 1, 000\}$) among the non-modified genes, respectively.

Considering this procedure, the most successful criterion was the one that achieve the best operating characteristics.

Simulation Results

Figure 9 displays the true positive fraction versus the false negative fraction (number of false positives found divided by the number of truly non-prognostic genes) for four con-

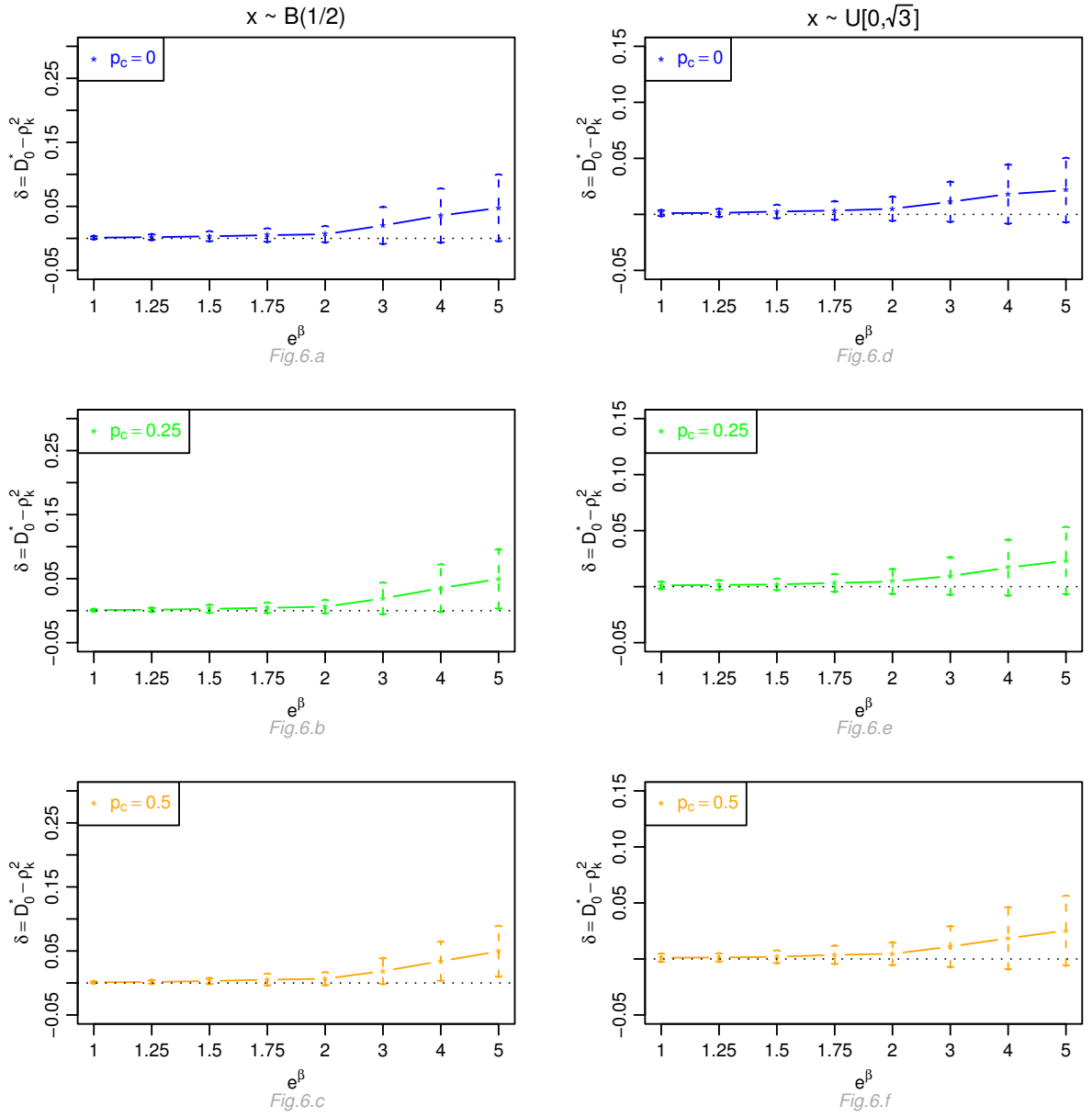

Figure 6 Graphic of the differences *δ* between the mean values of $\mathbf{D}_0^*$ and the mean values of ρ_k^2 as a function of the odds ratio, for a proportional odds model. Mean of $\mathbf{D}_0^* - \rho_k^2$ as a function of the odds ratio e^{β}, for different percentages of censoring p_c, for a covariate with Bernoulli $\mathbf{V}(1/2)$ or uniform $\mathcal{U}$ [0, $\sqrt{(3)}$] distribution, $n = 100$ and a uniform censoring mechanism.

figurations: $\rho = 0.5$, $e^{\xi} = 3$ and 5 and for a proportional and non-proportional model. For the five indices, higher operating characteristics are obtained under a proportional hazards model (Figure 9.a and 9.b) as compared to a proportional odds model (Fig 9.c and 9.d). Moreover, for a given distribution and a given threshold, our index gives the best results with higher true positive and true negative fractions. Results for the four other indices are very close to each other. Results with other levels of correlation ($\rho = 0.25$ and 0.75) are very close to those obtained with $\rho = 0.5$ (curves not shown here).

Application of the index on real data

Datasets

In this section, we exemplify the use of the proposed index by identifying transcriptomic prognostic factors common to

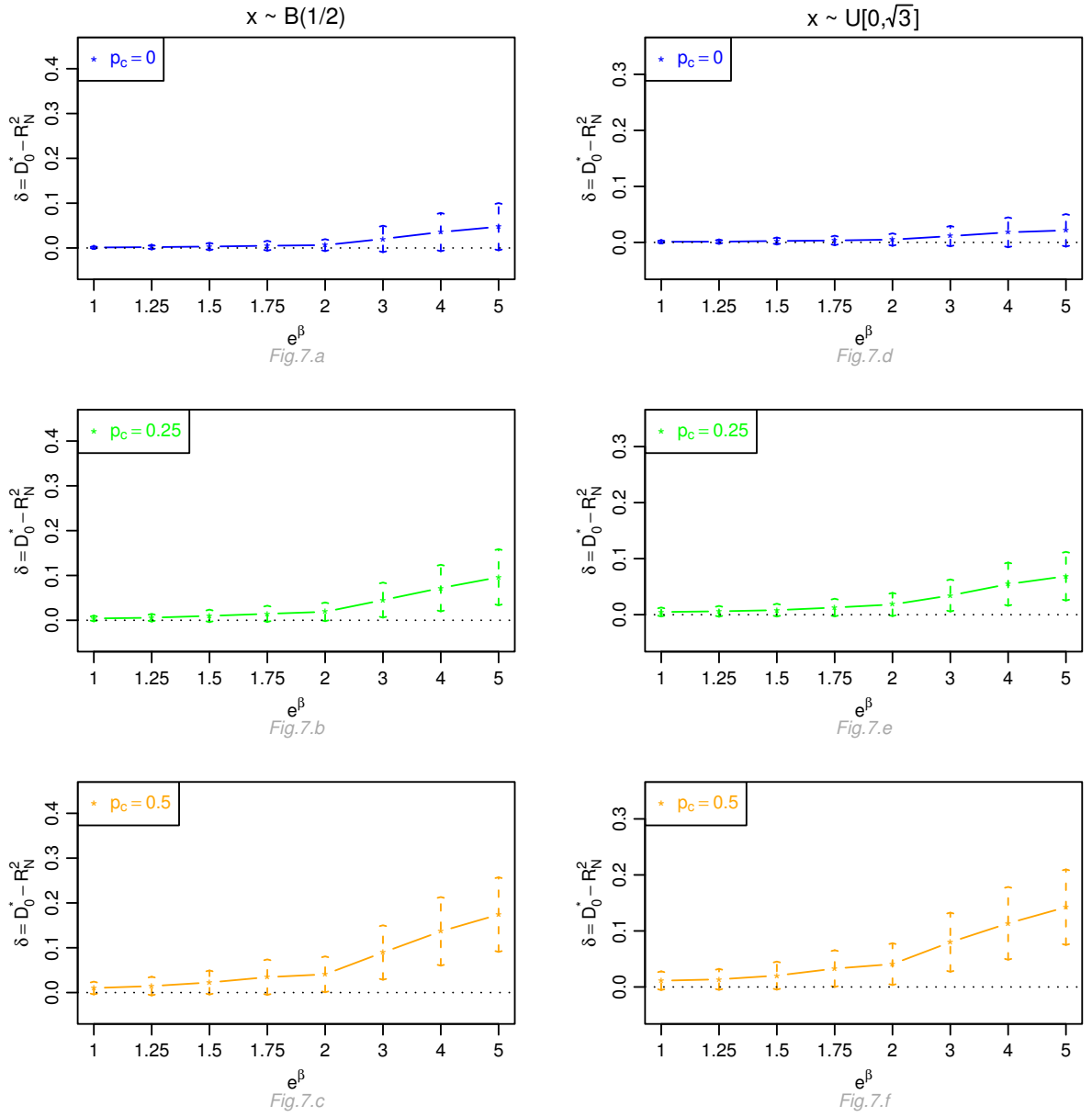

Figure 7 Graphic of the differences *δ* between the mean values of $\mathbf{D}_0^*$ and the mean values of R_N^2 as a function of the odds ratio, for a proportional odds model. Mean of $\mathbf{D}_0^* - R_N^2$ as a function of the odds ratio e^{β}, for different percentages of censoring p_c, for a covariate with Bernoulli $\mathbf{V}(1/2)$ or uniform $\mathcal{U}$ [0, $\sqrt{(3)}$] distribution, $n = 100$ and a uniform censoring mechanism.

eight studies corresponding to five different solid tumors such as breast, lung, bladder cancer, glioma and melanoma. We compared our index to the four indices and two classical test-based criteria (q-values derived from the log-likelihood ratio and robust score statistics). The data consisted of eight independent genomic studies [24-31], with different survival outcomes and different sample sizes which samples were hybridized on a same platform (Affymetrix HU133 Plus 2.0 or HU133A ; Affymetrix, Santa Clara, CA, USA). The datasets are publicly available on the GEO site under the labels GSE2034, GSE1456, GSE11121, GSE4573, GSE5287, GSE4271, GSE4412 and GSE19234, respectively, and they are briefly described below.

GSE2034 cohort, breast cancer [24] This series includes 286 lymph-node negative patients, among which 106 have developed a metastasis which is the event of interest in this

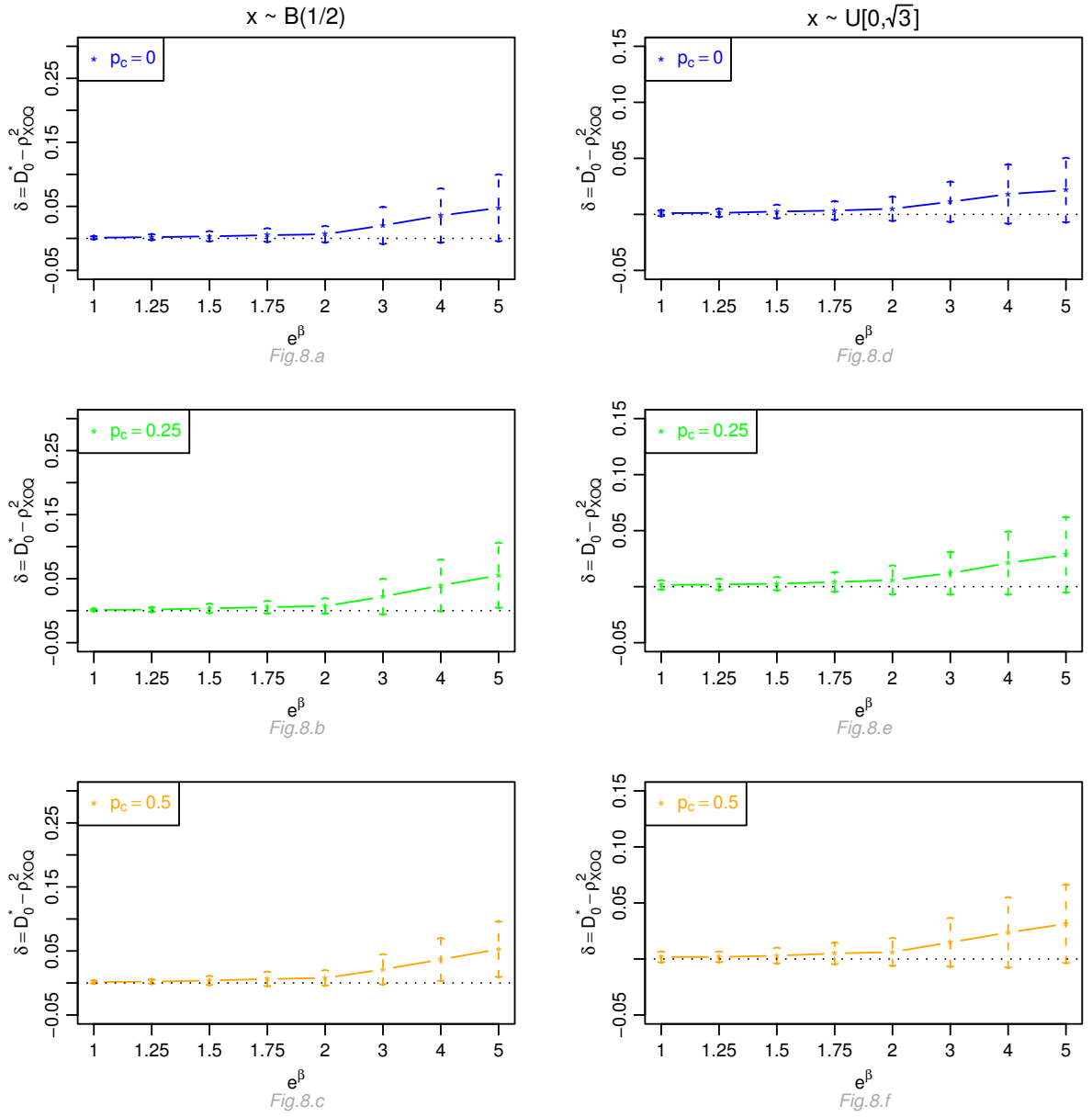

Figure 8 Graphic of the differences *δ* between the mean values of $\mathbf{D}_0^*$ and the mean values of ρ^2_{XOQ} as a function of the odds ratio, for a proportional odds model. Mean of $\mathbf{D}_0^* - \rho^2_{XOQ}$ as a function of the odds ratio e^{β}, for different percentages of censoring p_c, for a covariate with Bernoulli $\mathbf{B}(1/2)$ or uniform $\mathcal{U}$ [0, $\sqrt{(3)}$] distribution, $n = 100$ and a uniform censoring mechanism.

study. Metastasis-free survival was defined as the time interval from treatment until the apparition of distant relapse or last follow-up. The median metastasis-free survival time was 80 months. The two years metastasis-free survival was 83.9% [79.8%; 88.3%], and the five years metastasis-free survival was 66.7% [61.4%; 72.4%].

GSE1456 cohort, breast cancer [25] This series comprises 159 primary breast cancer patients (referred as Stockholm cohort). Metastasis-free survival measured the time from initial therapy until the first metastasis or last follow-up. The median metastasis-free survival time was 80 months. The two years metastasis-free survival was 87.9%

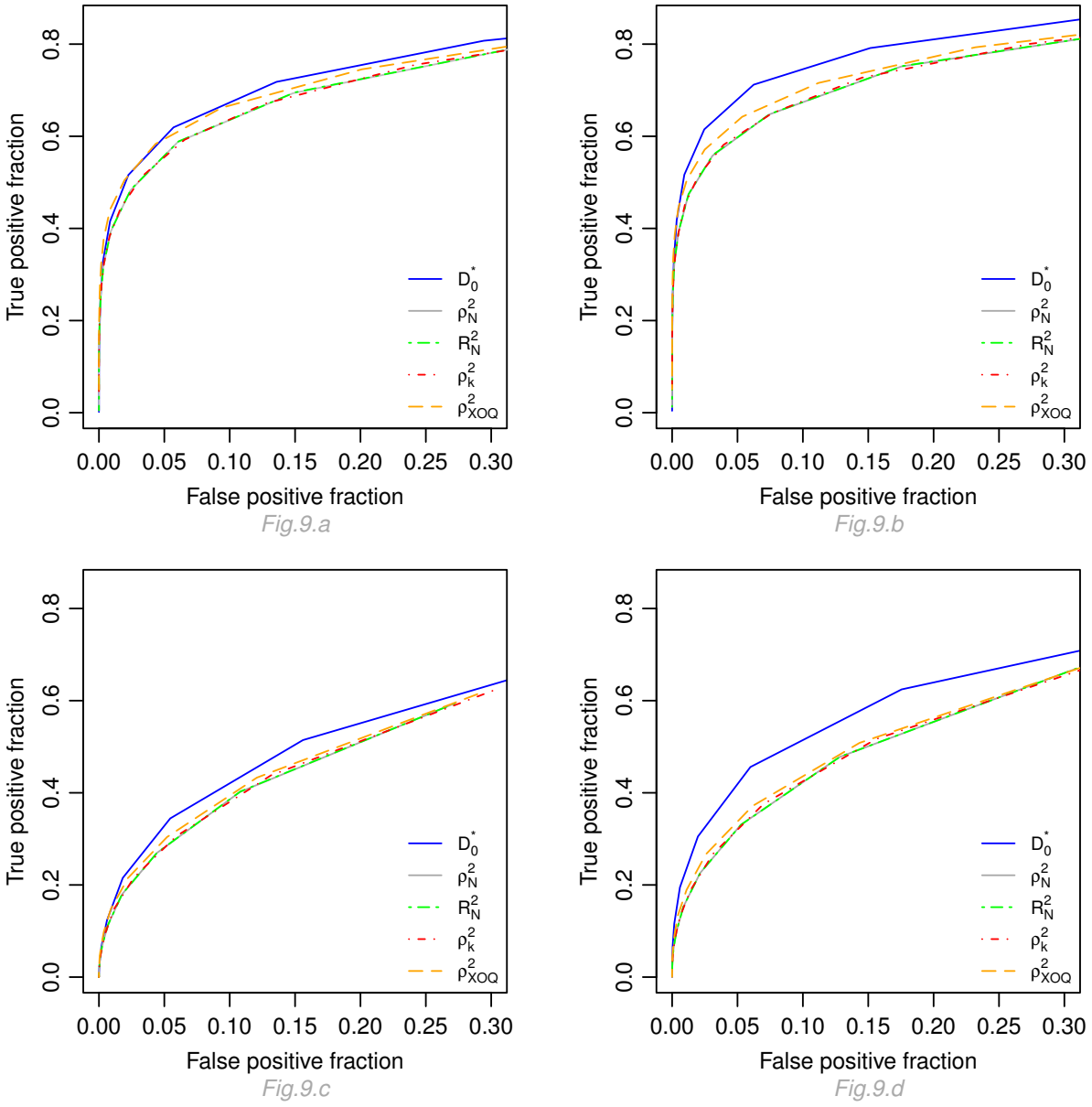

Figure 9 Operating characteristics of $\mathbf{D}_0^*$, $\rho_N^2, \rho_k^2, R_N^2$ and ρ_{XOQ}^2. Graphic of the true positive fraction versus the true negative fraction calculated for the five indices with different thresholds. Fig 9.a and 9.b display the results for a proportional hazard model, for $e^{\xi} = 3$ and 5 respectively ; Fig 9.c and 9.d display the results for a proportional odds model, for $e^{\xi} = 3$ and 5, respectively.

[83.0%, 93.2%], and the five years metastasis-free survival was 77.6% [71.3%, 84.4%].

GSE11121 cohort, breast cancer [26] This series is composed of 200 lymph node-negative breast cancer patients who were not treated by systemic therapy after surgery. Metastasis-free survival was defined as the interval from the date of therapy to the date of diagnosis of metastasis or last follow-up. The median metastasis-free survival time was 149 months. The two years metastasis-free survival was 92.9% [89.3%; 96.5%], and the five years metastasis-free survival was 85.4% [80.6%; 90.6%].

GSE4573 cohort, lung cancer [27] This series comprises 129 patients with different stages of squamous cell carcinomas, who underwent surgery resection of the lung. Overall survival was defined as the time from surgery until death or last follow-up. The median overall survival time was 63 months. The two years overall survival was 70.5% [63.1%; 78.9%], and the five years overall survival was 56.8% [48.3%; 66.7%].

GSE5287 cohort, bladder cancer [28] This series is composed of 30 patients who received chemotherapy. Overall survival was defined as the time from first chemotherapy to death or last follow-up. The median overall survival time was 47 months. The two years overall survival was 96.7% [90.5%; 100%], and the five years overall survival was 46.7% [31.8%; 68.4%].

GSE4271 cohort, glioma [29] This study comprises 77 patients with high-grade gliomas who underwent surgery (resection) of the brain. The overall survival was measured from initial surgical resection to death or last follow-up. The median overall-survival was 21 months. The two years overall-survival was 45.5% [35.6%, 58.1%], and the five years overall-survival was 22.6% [14.7%, 34.9%].

GSE4412 cohort, glioma [30] This series includes 85 patients who suffered of glioma of grade III or IV of any histologic type. The overall survival corresponded to the time from inclusion for surgical treatment to death or last follow-up. The median overall-survival was 13 months. The two years overall-survival was 33.2% [24.3%, 45.3%], and the five years overall-survival was 22.1% [12.9%, 37.7%].

GSE19234 cohort, melanoma [31] The authors considered 44 metastatic melanoma tissue samples. Overall survival was referred as the time from excision of the metastatic lesion to death or last follow-up. The median overall-survival was 46 months. The two years overall-survival was 76.7% [65.0%, 90.4%], and the five years overall-survival was 56.5% [43.1%, 74%].

For these studies, the hybridizations were performed on the Affymetrix GeneChip HU133A, except for the melanoma cohort where they were performed on HU133 Plus 2.0 (HU133A+HU133B). For each patient, we considered the information obtained from 22,283 transcripts (HU133A).

For selecting a threshold target value, we considered the intersection procedure introduced by Blangiardo and Richardson [32]. The main steps of this procedure were as follows. We first ranked the genes according to a measure of interest on probability scale (e.g. the p-value or the q-value). For each experiment and for a given threshold, we counted the number of differentially expressed genes in common between the different experiments. This number was then compared to the expected number of genes in common, calculated under the hypothesis of independence between the experiments. The ratio between these two numbers was calculated for all possible thresholds. Finally, the threshold considered in the intersection selection procedure was such as the ratio was superior to 2 with a clinically relevant survival difference. Here, we used this procedure, with the following criteria: (1) our index D_0^*; (2) Allison's index ρ_N^2; (3) the modified version of Allison's index ρ_k^2; (4) Nagelkerke's index R_N^2; (5) Xu and O'Quigley's index ρ_{XOQ}^2; (6) the q-value associated to the FDR (False Discovery Rate) calculated on the robust score statistic and estimated according to a non-parametric method [21]; (7) the q-value associated to the FDR calculated on the log-likelihood ratio statistic, estimated with the same method.

Selection of the Variables

The proposed index was calculated for the 22,283 gene expression measures for the eight datasets. The intersection procedure [32] led to a threshold equal of 0.07 for $\mathbf{D}_0^*$.

For $\mathbf{D}_0^* \geq 0.07$ (which corresponds from our simulations to a hazard ratio value around 1.5), we selected 5 transcripts related to four genes (Table 1).

We identified *HJURP* and *LRIG1* genes that are directly involved in tumorigenesis. *HJURP* encodes an indispensable factor for chromosomal stability in immortalized cancer cells. It is up-regulated in lung cancer [33]. *LRIG1* encodes a protein that acts as a growth suppressor in breast cancer [34]. Its expression decreases in human breast cancer and the majority of ErbB2+ breast tumors show underexpression of *LRIG1*. In our series, the increase of *HJURP* and decrease of *LRIG1* gene expressions are associated with a worse prognosis.

Our selection process also brought two genes involved in cell cycle regulation. Gene *KIF4A* encodes a protein critical for mitotic regulation including chromosome condensation, spindle organization and cytokinesis. It possesses a functional and physical link with the gene product of *BRCA2* (breast cancer 2, early onset) [35]. Gene *ESPL1* plays a central role in chromosome segregation at the onset of anaphase. Its over-expression induces aneuploidy and tumorigenesis [36]. The article of Zhang *et al* [36] showed that the *ESPL1* transcript is over-expressed in human breast tumors. It is worth noting that *ESPL1* and *KIF4A*, have been previously discussed in a meta-analysis conducted by Carter *et al* [37]. For these two genes, over-expression, leading to a cell proliferation, is associated with a worse prognosis.

Finally, for each gene from our selection, the hazard ratios were in the same direction in each of the eight studies.

With a same threshold of 0.07, Allison's index ρ_k^2 selected 3 transcripts corresponding to genes *KIF4A* and *ESPL1* and Xu and O'Quigley's index ρ_{XOQ}^2 selected 2 transcripts corresponding to gene *ESPL1*. The transcripts identified with these two indices are all included in our selected subset. For ρ_N^2 and R_N^2 with a threshold of 0.07, no transcript was selected. No transcript was selected rely-

Table 1: Top survival related genes across the eight studies, for $D_0^* \geq 0.07$.

AffyID	Gene symbol	UniGene Name	Cytoband	value of HR
211596-s-at	*LRIG1*	leucine-rich repeats and immunoglobulin-like domains 1	3p14	< 1
218355-at	*KIF4A*	kinesin family member 4A	Xq13.1	> 1
218726-at	*HJURP*	Holliday junction recognition protein	2q37.1	> 1
204817-at	*ESPL1*	extra spindle pole bodies homolog 1 (S. cerevisiae)	12q13.13	> 1
38158-at	*ESPL1*	extra spindle pole bodies homolog 1 (S. cerevisiae)	12q13.13	> 1

AffyID, Affymetrix identification code for each probe set ; HR, hazard ratio
If *HR* > 1, the over-expression of the gene is associated with a worse prognosis. If *HR* < 1, the over-expression of the gene is associated with a better prognosis.

ing on the q-value calculated with the robust score or the log-likelihood ratio statistics with a threshold of 0.40.

Discussion

Combining heterogeneous genomic datasets to select relevant genomic factors having a common prognostic impact across various tumor entities raises some concerns regarding the choice of the statistic to be considered. In particular, the use of hypothesis testing criteria across different datasets, such as p-values or related criteria, does not seem convenient due to its sensitivity to sample size. In this paper, we propose a novel index that is well suited for a combined analysis of heterogeneous genomic datasets and which allows a selection of features with a similar prognostic impact on outcome across studies.

The index possesses the four following properties: (1) it has a straightforward and meaningful interpretation in terms of percentage of separability between patients observed to experience the event of interest and those observed not to experience the event, according to their gene expression levels. (2) It increases with the ability to separate patients according to the gene variable from 0 to 1. (3) The index is not highly dependent on the sample size. (4) It is linked to the robust score statistic derived from the partial log-likelihood which has a known asymptotic distribution, and multiple testing criteria (e.g. FDR) can easily be calculated.

Our index shares a common framework with Allison's index, its modified version, Nagelkerke and Xu and O'Quigley's indices. Indeed, these latter indices are closely related to likelihood ratio statistics whereas ours relies on the score statistic. Moreover, our index is directly interpreted in terms of separability, whereas the other indices lack intuitive interpretation.

Simulation studies show that the separability performance of our index are better than for Allison's index, its modified version, Nagelkerke and Xu and O'Quigley's indices. In our simulated example, we illustrate the good operating characteristics of our index as compared to the classical ones. However, more extensive simulations work would be necessary to evaluate its performance in various real-world scenarios.

In this work, a meta-selection performed from different solid tumors allows the identification of a small set of genes (*ESPL1*, *KIF4A*, *HJURP*, *LRIG1*) that are biologically relevant to the carcinogenesis process and show a similar ability to separate patients according to time-to-event outcomes. It would be worth conducting further studies to validate or invalidate the prognostic impact of these genes. It is important to note that for the analysis of these data we have considered a very stringent method, which relies on finding the intersection set across the different studies. If necessary, less restrictive methods can be adopted. We have to highlight that our index was primarily designed for a proportional hazard model, but, as seen from our simulations, it performs well in other contexts such as proportional odds models. This last model corresponds to frequently encountered situations where the patient population becomes more and more homogeneous as time goes on and the prognostic effect decreases with time and disappears eventually. Future studies are needed to investigate other non-proportional hazard situations.

Finally, the proposed index may be appealing for time-to-event data in other medical fields such as auto-immune and infectious diseases in which identifying prognostic factors among different entities is a promising challenge.

Conclusion

In conclusion, we propose a novel index for identifying factors having a prognostic impact across collection of heterogeneous datasets that relies on the concept of separability and is not substantially affected by the sample size of the study. As the number of public available datasets obtained from independent studies keeps growing, our index is a promising tool which can help researchers to select a list of features of interest for further biological investigations.

Methods

Notations

Let $Z_i^{(g)}$ denote the value of a covariate Z for the i^{th} subject ($i = 1,\cup, n$) associated to the g^{th} gene ($g = 1,\cup, G$). For each patient i, let the random variables T_i and C_i be the survival and censoring times, which are assumed to satisfy the classical condition of independent censoring [38]. In practice, we observe $T_i^* = \min(T_i, C_i)$. Here we consider the possibility of the presence of ties among the uncensored failure times and we assume that there are N distinct times (of failure or censoring) and k distinct failure times ($k \leq N \leq n$). For $j = 1,\cup, N$, let $D(t_j)$ be the set of individuals failing at time t_j, $R(t_j)$ the risk set at t_j and $E(t_j)$ the set of individuals failing or censored at t_j. We denote d_j, n_j and e_j the cardinals of these three sets, respectively. We also define $R^*(t_j)$ as the risk set without the subjects failing at t_j and $R^*(t_{l(-j)})$ (for $t_l < t_j$) as the risk set at time t_l without the subjects failing or censored at t_j. Let $\eta_j = \mathbf{1}_{d_j} \geq 1$ be the indicator of at least one death at t_j (where 1 is the indicator function).

The hazard function at time t for gene g can be written in a semi-parametric proportional hazards form as [8]

$$\lambda^{(g)}(t;z) = \lambda_0^{(g)}(t)\exp\left[\beta^{(g)}Z^{(g)}(t)\right]$$

where $\lambda_0^{(g)}(t)$ is an unknown baseline hazard function, and $\beta^{(g)}$ is the regression parameter to be estimated. In the presence of ties, the partial log-likelihood of the Cox model [39] can be approximated according to the Peto and Breslow method [40,41]

$$\log\{\mathcal{L}^{(g)}(\beta^{(g)})\} = \sum_{j=1}^{N}\eta_j\left[\beta^{(g)}\sum_{l\in D(t_j)}Z_l^{(g)}(t_j) - d_j\log\left\{\sum_{l\in R(t_j)}\exp\{\beta^{(g)}Z_l^{(g)}(t_j)\}\right\}\right]$$

The first derivative of the partial log-likelihood, or score, is:

$$U^{(g)}(\beta^{(g)}) = \frac{\partial\log\{\mathcal{L}^{(g)}(\beta^{(g)})\}}{\partial\beta^{(g)}} = \sum_{j=1}^{N}U_j^{(g)}(\beta^{(g)})$$

$$= \sum_{j=1}^{N}\eta_j\left[\sum_{l\in D(t_j)}Z_l^{(g)}(t_j) - d_j\sum_{l\in R(t_j)}\frac{Z_l^{(g)}(t_j)\exp\{\beta^{(g)}Z_l^{(g)}(t_j)\}}{\sum_{i\in R(t_j)}\exp\{\exp\{\beta^{(g)}Z_i^{(g)}(t_j)\}}\right]$$

In the following, the exponent $^{(g)}$ is omitted in order to facilitate the reading. Consequently, β will refer to $\beta^{(g)}$, Z_i, to $Z_i^{(g)}$, ? to $?^{(g)}$ and U_j to $U_j^{(g)}$.

Proposed index

The proposed index is based on the interpretative property of the score deduced from the partial log-likelihood under the Cox model as recalled above. At each time $t = t_j$, $j = 1,\cup, N$, we consider the quantities U_j calculated under the null hypothesis (for $\beta = 0$) from the approximated Breslow partial log-likelihood

$$U_j = \eta_j\left[\sum_{l\in D(t_j)}Z_l(t_j) - d_j\sum_{l\in R(t_j)}\frac{Z_l(t_j)}{n_j}\right]$$

$$= \frac{\eta_j\cdot d_j(n_j - d_j)}{n_j}\left[\sum_{l\in D(t_j)}\frac{Z_l(t_j)}{d_j} - \sum_{l\in R^*(t_j)}\frac{Z_l(t_j)}{n_j - d_j}\right]$$

From this latter expression, it appears that, for a given covariate Z, and at each event time t_j, the U_j can be expressed as differences between the means of the covariates of the group $D(t_j)$ of patients observed to experience the event of interest, and the group $R^*(t_j)$ of those observed to not experience the event. The U_j provide a measure of

separability between the two groups of patients $D(t_j)$ and $R^*(t_j)$ at time t_j. Differences close to zero indicate a weak or null separability; large differences indicate that the two groups are well separated.

Hence, a global statistic over time can be computed as the sum of these differences: $\Delta_0 = \sum_{j=1}^{N} U_j$. The statistic Δ_0 is large if the two groups are well separated over time or for a few time points with large values but with the same directional effect (proportional hazard assumption).

For distributional reasons which will appear later, instead of the U_j, $j = 1,\cup, N$, we use closely related quantities W_j derived from the paper by Lin and Wei [42]. In the presence of ties, we propose the following formula for W_j

$$W_j = U_j - \hat{E}U(t_j)$$
$$= \eta_j \left[\sum_{l \in D(t_j)} Z_l(t_j) - \frac{d_j}{n_j} \sum_{l \in R(t_j)} Z_l(t_j) \right]$$
$$- \sum_{l=1}^{j} \frac{\eta_l d_l}{n_l} \left[\sum_{r \in E(t_j)} Z_r(t_l) - \frac{e_j}{n_l} \sum_{r \in R(t_l)} Z_r(t_l) \right]$$

The term $\hat{E}U_j$ is a weighted average of the score calculated at times t_l prior to time t_j ($t_l < t_j$). The sum of the so-called "robust" W_j, $j = 1,\cup, N$ is identical to the sum of the U_j, but, as shown by Lin and Wei, the W_j are independent and identically distributed, while the U_j are not. Simple calculations show that the W_j can be rearranged as in the following expression:

$$W_j = c_j \left[\sum_{l \in D(t_j)} \frac{Z_l(t_j)}{d_j} - \sum_{r \in R^*(t_j)} \frac{Z_r(t_j)}{n_j - d_j} \right]$$
$$- \sum_{l=1}^{j} \omega_{jl} \left[\sum_{r \in E(t_j)} \frac{Z_r(t_l)}{e_j} - \sum_{r \in R^*(t_{l(-j)})} \frac{Z_r(t_l)}{n_l - e_j} \right]$$

with

$$c_j = \frac{\eta_j (n_j - d_j) d_j}{n_j} \text{ and } \omega_{jl} = \frac{\eta_l d_l}{n_l} \cdot \frac{(n_l - e_j) e_j}{n_l}$$

The usual global robust score is computed as the sum of the differences W_j, $j = 1,\cup, N$ (which is also equal to the sum of the U_j). So, Δ_0 can be re-expressed as the sum of the W_j:

$$\Delta_0 = \sum_{j=1}^{N} W_j = \sum_{j=1}^{N} U_j$$

In Additional file 5, we show that $\mathbf{D}_0 = \Delta_0^2 / k = \left(\sum_{j=1}^{N} W_j \right)^2 / k$ ranges from 0 (null separability under the proportional hazard model) to $\mathbf{D}_{max} = \sum_{j=1}^{N} W_j^2$ (maximal separability). The value $\mathbf{D}_{max}$ is a theoretical maximum of $\mathbf{D}_0$, which corresponds to the case where β tends to infinity.

Finally,

$$\mathbf{D}_0^* = \frac{\mathbf{D}_0}{\mathbf{D}_{max}} = \frac{1}{k} \frac{\left(\Sigma_{j=1}^{N} W_j \right)^2}{\Sigma_{j=1}^{N} W_j^2}$$

gives a meaningful index that can be interpreted as the percentage of separability over time between the event/non-event groups. It is equal to 0 in the absence of separability and increases toward 1 as the separability rises. To a factor k, the index can also be interpreted as the robust score statistic ($\mathbf{S}_0 = k \cdot \mathbf{D}_0^*$) [43], whose distribution under the null hypothesis is an asymptotic chi-square distribution with 1 degree of freedom. Multiple error criteria can thus be computed using a parametric or non-parametric approach.

Existing indices

Several indices of predictive accuracy have been proposed in the literature. Here, only indices with direct or indirect links to the likelihood ratio function and with a known distribution after transformation under the null hypothesis are considered.

The indices are the following: (i) Allison's index ρ_N^2 [16], based on a transformation of the partial log-likelihood ratio test; (ii) a modified version of Allison's index ρ_k^2 proposed by O'Quigley *et al* [18]; (iii) Nagelkerke's index R_N^2 [17], which is a modification of Allison's index dividing it by its maximum value, and (iv) Xu and O'Quigley's index ρ_{XOQ}^2 [19] based on a transformation of the Kullback-Leibler distance between the null and the alternative models.

The expressions of these four indices for one given gene g; $g = 1,\cup, G$ are reminded here:

(i) Allison's index:

$$\rho_N^2 = 1 - \exp\left(-\frac{2}{N} \times \left[\log \mathcal{L}(\beta) - \log \mathcal{L}(0) \right] \right)$$

(ii) Modified version of Allison's index:

$$\rho_k^2 = 1 - \exp\left(-\frac{2}{k} \times \left[\log \mathcal{L}(\beta) - \log \mathcal{L}(0) \right] \right)$$

In this version of the index, the log-likelihood ratio is divided by the number of failures k. As discussed by O'Quigley *et al* [18], the original version is more sensitive to censorship than the modified one. In particular, O'Quigley *et al* show that ρ_N^2 approaches 0 as the percentage of censored observation approaches 100%.

(iii) Nagelkerke's index:

$$R_N^2 = \frac{\rho_N^2}{R_{max}^2}$$

with

$$R_{max}^2 = 1 - \exp\left(\frac{2}{N} \times \log \mathcal{L}(0) \right)$$

This index was initially proposed to fully exploit the range [0, 1], which is not the case with the original version of Allison's index.

(iv) Xu and O'Quigley's index:

$$\rho_{XOQ}^2 = 1 - \exp\left\{ -\Gamma(\beta) / \sum_{j=1}^{N} V(t_i) \right\}$$

with

$$\hat{\Gamma}(\hat{\beta}) = 2 \sum_{j=1}^{N} V(t_i) \sum_{i=1}^{n} \pi_i(t_j; \hat{\beta}) \log\left(\frac{\pi_i(t_j; \hat{\beta})}{\pi_i(t_j; 0)} \right)$$

where $V(t_i) = S(t_j^+) - S(t_j)$ and $\hat{S}$ is the Kaplan-Meier estimator of the distribution function of T.

The term $\hat{\Gamma}(\hat{\beta})$ is derived from twice the Kullback-Leibler distance between the null model ($\beta = 0$) and the model taking the covariates into account ($\beta \neq 0$).

The conditional probability $\pi_i(t_j; \beta)$ that the individual indexed by i is selected for failure at the time t_j is given by

$$\pi_i(t_j; \beta) = \frac{\exp\left(\beta Z_i(t_j) \right)}{\sum_{l \in R(t_j)} \exp\left(\beta Z_l(t_j) \right)}$$

Additional material

Additional file 1 Calculation of the parameters of the different censoring mechanisms. We explain the procedure adopted for the calculation of the parameters of uniform and exponential censoring mechanisms as functions of the distribution of the covariates and the percentage of censoring.

Additional file 2 Mean values of $\mathbf{D}_0^*$ in the framework of a Cox proportional hazards model, for different relative risks e^{β}, different percentages of censoring p_c and different sample sizes n, calculated for a covariate with Bernoulli B(1/2) or a uniform $\mathcal{U}$ [0, $\sqrt{(3)}$] distribution, for a uniform censoring mechanism (1,000 repetitions). The standard errors are indicated in brackets. Table with the mean values of $\mathbf{D}_0^*$ in the framework of a Cox proportional hazards model for different configurations.

Additional file 3 Mean values of $\mathbf{D}_0^*$ in the framework of a proportional odds model, for different odds ratios e^{β}, different percentages of censoring p_c and different sample sizes n, calculated for a covariate with Bernoulli B(1/2) or a uniform $\mathcal{U}$ [0, $\sqrt{(3)}$] distribution, for a uniform censoring mechanism (1,000 repetitions). The standard errors are indicated in brackets. Table with the mean values of $\mathbf{D}_0^*$ in the framework of a proportional odds model for different configurations.

Additional file 4 Representation of the simulated example. Simple representation of the simulation plan.

Additional file 5 Proof establishing that $\mathbf{D}_0 \left(\sum_{j=1}^{N} U_j \right)^2 / k = \left(\sum_{j=1}^{N} W_j \right)^2 / k$ ranges from 0 to $\mathbf{D}_{max} \sum_{j=1}^{N} W_j^2$. We show that $0 \leq \left(\sum_{j=1}^{N} W_j \right)^2 / k \leq \sum_{j=1}^{N} W_j^2$.

Authors' contributions
SR, TM and PB developed the original index. PB coordinated the project and is SR's PhD thesis advisor. All authors read and approved the final manuscript.

Acknowledgements
We thank Dr. Krishna Karuturi (Genome Institute of Singapore) for comments on the work. We also thank the following institutions for general funding: the Genome Institute of Singapore (Singapore) and the French Ministry of Higher Education and Research (France).

Author Details
[1]Computational and Mathematical Biology, Genome Institute of Singapore, Singapore 138672, Singapore, [2]Univ Paris-Sud, JE2492, Villejuif, F-94807 France and [3]Inserm, U780, Villejuif, F-94807 France; Univ Paris-Sud, Villejuif, F-94807 France

Received: 21 July 2009 Accepted: 24 March 2010
Published: 24 March 2010

References

1. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R: **NCBI GEO: mining millions of expression profiles-database and tools.** *Nucleic Acids Research* 2005, **33:**D562-D566.
2. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, Barrette TR, Anstet MJ, Kincead-Beal C, Kulkarni P, Varambally S, Ghosh D, Chinnaiyan AM: **Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles.** *Neoplasia* 2007, **9(2):**166-180.
3. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ, Adamusiak T, Brandizi M, Burdett T, Coulson R, Krestyaninova M, Kurnosov P, Maguire E, Neogi SG, Rocca-Serra P, Sansone SA, Sklyar N, Zhao M, Sarkans U, Brazma A: **ArrayExpress update-from an archive of functional genomics experiments to the atlas of gene expression.** *Nucleic Acids Research* 2009, **37:**D868-D872.
4. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.** *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101(25):**9309-9314.
5. Basil CF, Zhao Y, Zavaglia K, Jin P, Panelli MC, Voiculescu S, Mandruzzato S, Lee HM, Seliger B, Freedman RS, Taylor PR, Hu N, Zanovello P, Marincola FM, Wang E: **Common cancer biomarkers.** *Cancer Research* 2006, **66(6):**2953-2961.
6. Xu L, Geman D, Winslow RL: **Large-scale integration of cancer microarray data identifies a robust common cancer signature.** *BMC Bioinformatics* 2007, **8:**275.
7. Lu Y, Yi Y, Liu P, Wen W, James M, Wang D, You M: **Common human cancer genes discovered by integrated gene-expression analysis.** *PLoS One* 2007, **2(11):**e1149.
8. Kalbfleisch JD, Prentice RL: *The statistical analysis of failure time data. Wiley series in Probability and Mathematical Statistics* New York: Wiley; 2002.
9. Allison DB, Cui X, Page GP, Sabripour M: **Microarray data analysis: from disarray to consolidation and consensus.** *Nature Review Genetics* 2006, **7:**55-65.
10. Harrell F, Califf R, Pryor D, Lee K, Rosati R: **Evaluating the yield of medical tests.** *Journal of the American Medical Association* 1982, **247(18):**2543-2546.
11. Antolini L, Boracchi P, Biganzoli E: **A time-dependent discrimination index for survival data.** *Statistics in Medicine* 2005, **24(24):**3927-3944.
12. Korn E, Simon R: **Measures of explained variation for survival data.** *Statistics in Medicine* 1990, **9(5):**487-503.
13. Schemper M: **The explained variation in proportional hazards regression.** *Biometrika* 1990, **77:**216-218.
14. Schemper M, Henderson R: **Predictive accuracy and explained variation in Cox regression.** *Biometrics* 2000, **56:**249-255.
15. Schemper M, Stare J: **Explained variation in survival analysis.** *Statistics in Medicine* 1996, **15(19):**1999-2012.
16. Allison PD: *Survival Analysis Using SAS: A Practical Guide* SAS Publishing; 1995.
17. Nagelkerke N: **A note on a general definition of the coefficient of determination.** *Biometrika* 1991, **78(3):**691-692.
18. O'Quigley J, Xu R, Stare J: **Explained randomness in proportional hazards models.** *Statistics in Medicine* 2005, **24(3):**479-489.
19. Xu R, O'Quigley J: **A R2 type measure of dependence for proportional hazards models.** *Journal of Nonparametric Statistics* 1999, **12:**83-107.
20. Bair E, Tibshirani R: **Semi-supervised methods to predict patient survival from gene expression data.** *PLoS Biology* 2004, **2(5):**E108.
21. Dalmasso C, Broët P, Moreau T: **A simple procedure for estimating the false discovery rate.** *Bioinformatics* 2005, **21(5):**660-668.
22. Qiu X, Klebanov L, Yakovlev A: **Correlation between gene expression levels and limitations of the empirical bayes methodology for finding differentially expressed genes.** *Statistical Applications in Genetics and Molecular Biology* 2005, **4:**. Article34
23. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19(2):**185-193.
24. Wang Y, Klijn JGM, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EMJJ, Atkins D, Foekens JA: **Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.** *Lancet* 2005, **365:**671-679.
25. Pawitan Y, Bjöhle J, Amler L, Borg AL, Egyhazi S, Hall P, Han X, Holmberg L, Huang F, Klaar S, Liu ET, Miller L, Nordgren H, Ploner A, Sandelin K, Shaw PM, Smeds J, Skoog L, Wedrén S, Bergh J: **Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts.** *Breast Cancer Research* 2005, **7(6):**R953-R964.
26. Schmidt M, Böhm D, von Törne C, Steiner E, Puhl A, Pilch H, Lehr HA, Hengstler JG, Kölbl H, Gehrmann M: **The humoral immune system has a key prognostic impact in node-negative breast cancer.** *Cancer Research* 2008, **68(13):**5405-5413.
27. Raponi M, Zhang Y, Yu J, Chen G, Lee G, Taylor JMG, Macdonald J, Thomas D, Moskaluk C, Wang Y, Beer DG: **Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung.** *Cancer Research* 2006, **66(15):**7466-7472.
28. Als AB, Dyrskjot L, Maase H von der, Koed K, Mansilla F, Toldbod HE, Jensen JL, Ulhoi BP, Sengelov L, Jensen KME, Orntoft TF: **Emmprin and survivin predict response and survival following cisplatin-containing chemotherapy in patients with advanced bladder cancer.** *Clinical Cancer Research* 2007, **13(15):**4407-4414.
29. Phillips HS, Kharbanda S, Chen R, Forrest WF, Soriano RH, Wu TD, Misra A, Nigro JM, Colman H, Soroceanu L, Williams PM, Modrusan Z, Feuerstein BG, Aldape K: **Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis.** *Cancer Cell* 2006, **9(3):**157-173.
30. Freije WA, Castro-Vargas FE, Fang Z, Horvath S, Cloughesy T, Liau LM, Mischel PS, Nelson SF: **Gene expression profiling of gliomas strongly predicts survival.** *Cancer Research* 2004, **64(18):**6503-6510.
31. Bogunovic D, O'Neill DW, Belitskaya-Levy I, Vacic V, Yu YL, Adams S, Darvishian F, Berman R, Shapiro R, Pavlick AC, Lonardi S, Zavadil J, Osman I, Bhardwaj N: **Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in predicting patient survival.** *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106(48):**20429-20434.
32. Blangiardo M, Richardson S: **Statistical tools for synthesizing lists of differentially expressed features in related experiments.** *Genome Biology* 2007, **8(4):**R54.
33. Kato T, Sato N, Hayama S, Yamabuki T, Ito T, Miyamoto M, Kondo S, Nakamura Y, Daigo Y: **Activation of Holliday junction recognizing protein involved in the chromosomal stability and immortality of cancer cells.** *Cancer Research* 2007, **67(18):**8544-8553.
34. Miller JK, Shattuck DL, Ingalla EQ, Yen L, Borowsky AD, Young LJT, Cardiff RD, Carraway KL, Sweeney C: **Suppression of the negative regulator LRIG1 contributes to ErbB2 overexpression in breast cancer.** *Cancer Research* 2008, **68(20):**8286-8294.
35. Wu G, Zhou L, Khidr L, Guo XE, Kim W, Lee YM, Krasieva T, Chen PL: **A novel role of the chromokinesin Kif4A in DNA damage response.** *Cell Cycle* 2008, **7(13):**2013-2020.
36. Zhang N, Ge G, Meyer R, Sethi S, Basu D, Pradhan S, Zhao YJ, Li XN, Cai WW, El-Naggar AK, Baladandayuthapani V, Kittrell FS, Rao PH, Medina D, Pati D: **Overexpression of Separase induces aneuploidy and mammary tumorigenesis.** *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105(35):**13033-13038.
37. Carter SL, Eklund AC, Kohane IS, Harris LN, Szallasi Z: **A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers.** *Nature Genetics* 2006, **38(9):**1043-1048.
38. Fleming TR, Harrington DP: *Counting Processes and Survival Analysis* Wiley; 1991.
39. Cox DR: **Regression models and life-tables.** *Journal of the Royal Statistical Society Series B* 1972, **34:**187-220.
40. Breslow N, Crowley J: **A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship.** *Annals of Statistics* 1974, **2(3):**437-453.
41. Peto R: **Contribution to the discussion of the paper by DR Cox.** *Journal of the Royal Statistical Society Series B* 1972, **34:**205-207.
42. Lin DY, Wei LJ: **The robust inference for the Cox proportional hazards model.** *Journal of the American Statistical Association* 1989, **84:**1074-1078.
43. Lachin JM: *Biostatistical Methods: The assessment of relative risks. Wiley series in Probability and Mathematical Statistics* New York: Wiley; 2000.

doi: 10.1186/1471-2105-11-150
Cite this article as: Rouam *et al.*, Identifying common prognostic factors in genomic cancer studies: A novel index for censored outcomes *BMC Bioinformatics* 2010, **11**:150