BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 12 2014, pages 1690–1697
doi:10.1093/bioinformatics/btu065

Gene expression

Advance Access publication February 14, 2014

Discrete mixture modeling to address genetic heterogeneity in time-to-event regression

Kevin H. Eng[1,*] and Bret M. Hanlon[2]

[1]Department of Biostatistics and Bioinformatics, Roswell Park Cancer Institute, Elm and Carlton Streets, Buffalo, NY 14263, USA and [2]Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53705, USA

Associate Editor: Gunnar Ratsch

ABSTRACT

Motivation: Time-to-event regression models are a critical tool for associating survival time outcomes with molecular data. Despite mounting evidence that genetic subgroups of the same clinical disease exist, little attention has been given to exploring how this heterogeneity affects time-to-event model building and how to accommodate it. Methods able to diagnose and model heterogeneity should be valuable additions to the biomarker discovery toolset.

Results: We propose a mixture of survival functions that classifies subjects with similar relationships to a time-to-event response. This model incorporates multivariate regression and model selection and can be fit with an expectation maximization algorithm, we call Cox-assisted clustering. We illustrate a likely manifestation of genetic heterogeneity and demonstrate how it may affect survival models with little warning. An application to gene expression in ovarian cancer DNA repair pathways illustrates how the model may be used to learn new genetic subsets for risk stratification. We explore the implications of this model for censored observations and the effect on genomic predictors and diagnostic analysis.

Availability and implementation: *R* implementation of CAC using standard packages is available at https://gist.github.com/programeng/8620b85146b14b6edf8f Data used in the analysis are publicly available.

Contact: kevin.eng@roswellpark.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 13, 2013; revised on September 6, 2013; accepted on January 28, 2014

1 INTRODUCTION

In cancer genomic studies, unobserved heterogeneity obfuscates the effort to build accurate descriptive models of risk stratification. In ovarian cancer, for example, Vaughan *et al.* (2011) argue that the set of patients with the same clinical disease have distinct molecular diseases. With respect to inference, this implies that the same regression models may not be valid for every patient and further that it is unclear which patients should be considered together (Köbel *et al.*, 2008). Therefore, a major statistical task is to organize patients into previously unknown classes while simultaneously fitting their time-to-event models.

The examples throughout the article are taken from our ongoing analysis of data from The Cancer Genome Atlas (TCGA), which has the goal of cataloging all of the genomic alterations in cancer. For each patient, there is a tremendous amount and variety of data: 12 000 genes in expression arrays, 1 million single-nucleotide polymorphism genotypes, exome and whole-genome sequences, methylation of thousands of CpG islands and the expression of microRNA. From this plurality of data, we anticipate that exploratory methods will serve to extract and characterize genetic subgroups relevant to survival time clinical outcomes.

For summarizing the impact of a genetic signature, one often stratifies patients to demonstrate separation between class-defined survival curve estimates. Unfortunately, as Na *et al.* (2009) review, current methods either awkwardly dichotomize a continuous score at a *post hoc* threshold or rely on hierarchical clustering to define subgroups with no necessary relation to survival. In that sense, it is desirable to have a method that identifies genetic subgroups supervised by their survival times.

The standard methods for dealing with non-homogenous time-to-event data do not apply when our goal is to discover unknown subgroups. Continuous frailty models (Aalen, 1988) treat all individuals separately and therefore do not produce subgroups. O'Quigley and Stare (2002) emphasize the use of random effects and stratified regression models when subgroups are known. Classification and regression tree methods have been adapted for survival responses (Lostritto *et al.*, 2012; Segal, 1988). However, these methods partition the predictor space to form a single piecewise functional estimate, and our interest is in the subgroups that exist in similar regions.

Some treatment of heterogeneity relevant to survival time where a subgroup of patients does not expire appears in the cure rate model literature (Farewell, 1982). Our situation is distinct in three ways: a subgroup may have variable time-to-event outcomes (cured patients have infinite survival times), the variables of interest to each mixture component may be distinct and the set of patients in each subgroup is not known.

In this article, we propose a discrete mixture regression model that synergizes with potential heterogeneity in time-to-event data. Concretely, we assume that observations belong to unlabeled classes with class-specific proportional hazards (PH) regression models relating their genetic covariates to survival time outcomes (Section 3). This conditional semi-parametric model leads to a surprising variety of model effects, which we illustrate

*To whom correspondence should be addressed.

in Section 2. In Section 4, we describe an algorithm and the considerations for fitting the model. Simulations highlight the use of censored data (Section 5) and a data analysis demonstrates a single-pathway hypothesis-driven model (Section 6). Our discussion (Section 7) again emphasizes the exploratory role that this analysis may address.

2 ILLUSTRATIONS

2.1 Genetic heterogeneity

There is evidence that distinct molecular subgroups lead to the same clinical presentation of ovarian cancer (Cooke *et al*., 2010; Köbel *et al*., 2008; Konstantinopoulos *et al*., 2008; Vaughan *et al*., 2011). This form of genetic heterogeneity may arise because the commonalities leading to cancer may aggregate within pathways and not on the level of genes (Jones *et al*., 2008). In the following case study, we highlight the ability of the mixture to produce unusual associations between covariates and survival and how it may augment our understanding of subgroup discovery.

Suppose that $X \sim \mathcal{N}(0,1)$ represents a single typical normalized gene expression measurement and that patients do have survival times arising from two distinct hazard models, $h_1(t|x) = h_0(t)\exp(-2x)$ and $h_2(t|x) = h_0(t)\exp(+2x)$. These hazards represent an extreme version of heterogeneity in expression; in one class, the gene has a protective effect and in the other it is equally deleterious.

Assuming the baseline hazard is exponential ($h_0(t) = 1$), we generate 1000 complete survival times, Y, under each of these hazards and plot them on the log scale with their randomly generated expression in Figure 1A. Without knowledge of the true classes, fitting a standard Cox regression to these data finds no significant relationship between Y and X ($\hat{\beta} = -0.0314,\ P = 0.1$). This effect is strong enough that the relationship is easily identified if the true classes are known ($\hat{\beta}_1 = 1.93,\ P < 0.001$ and $\hat{\beta}_2 = -2.12,\ P < 0.001$).

A standard diagnostic technique to estimate non-linear relationships between Y and X is to use a smoothing estimate on the added variable plot (Fig. 1B). In this case, it fails to identify any important effects. Estimating a time-varying effect is another diagnostic for assessing PH (Grambsch and Therneau, 1994). Again it does not discern any non-proportional effect ($P = 0.116$) or time-varying effect (Fig. 1C). So, by the standard analyses, this important gene would not be identified for further study.

With respect to gene expression analyses, this is a case where differential expression (DE) models that look for mean difference will fail: there is no true underlying survival difference between classes that may be attributed to X (Fig. 1D). This means that models that try to estimate a rule $1_{\{X>c\}}$ that can classify patients are not applicable. So, the mixture reflects a different way to

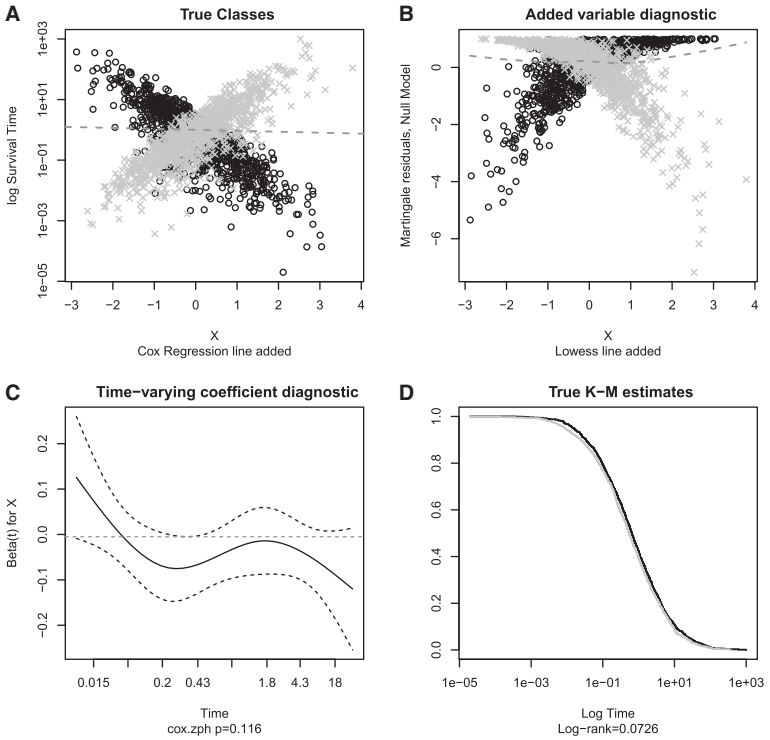

Fig. 1. Heterogeneity example. (**A**) Log simulated survival times by covariate; underlying heterogeneity is represented by dark and light classes. The marginal Cox model estimated relationship with covariate X is indicated by the dashed line. (**B**) An added variable plot, the estimated functional relationship between survival time and X (dashed) fails to detect unknown subgroups. (**C**) A check for non-PH finds no significant deviation. (**D**) Even though the generating models are different, the oracle-based survival estimates show no significant difference

model risk in gene expression. The question remains about whether this kind of example exists in real data. We present a gene that does just this in the following illustration.

2.2 Cytoreduction and epiregulin

We draw known class labels from a measured surgical covariate, which is presently used as a biomarker. Cytoreduction, the amount of tumor remaining after surgery, is a measure of success of surgical debulking, which is a component of primary therapy in ovarian cancer. Patients who are suboptimally cytoreduced have a clinically significant amount of residual tumor that will seed future recurrent disease and progression (Bhoola and Hoskins, 2006).

Using the TCGA study to be introduced in Section 6, we separate the patients into optimal and suboptimal categories and provide a kernel smoothing estimate of their hazards (Fig. 2). The estimated hazards are clearly non-proportional [$P = 0.007$, Grambsch and Therneau (1994)], reflecting the early protective effect of optimal cytoreduction and its transient nature.

Fitting separate models in each subgroup, we searched for genes whose relationship with survival inverts over classes finding epiregulin (interaction $P = 0.004$), which has been recently highlighted as a progression marker (Amsterdam *et al*., 2011). In optimally cytoreduced patients, increased expression is detrimental to survival ($\hat{\beta} = 0.156, P = 0.014$); in suboptimal patients, increased expression is protective ($\hat{\beta} = -0.452, P = 0.012$). In Figure 2, we have plotted the estimated survival for high and low epiregulin expression for optimal and suboptimal patients. Noting that epiregulin has been shown to inhibit epithelial tumor cells and stimulate cancer-associated fibroblasts (Toyoda *et al*., 1995), the surgical outcomes may indicate a more epithelial or more fibrous tumor; a fair biological explanation is that epiregulin expression leads to the inhibition of tumor burden (a better prognostic outcome), or the stimulation of fibroblasts that leads to cancer progression.

Thus, these effects do exist and imply surprisingly deep biological connections. Following a genomic survey, this type of effect is an ideal target for functional studies. Given that we want to identify genetic subgroups with different prognoses, we should favor a model that admits unknown and possibly dramatically different survival experiences. The mixture model should let us estimate labels and should be able to resolve non-PH.

3 METHODS

Let (Y_i, δ_i, x_i), $i = 1, \ldots, n$ be an independent right-censored sample with regression covariates $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$, where $\delta_i = 1$ indicates that the complete time has been observed. We will denote the collections of survival times, censoring indicators and covariate vectors as $Y = (Y_1, \ldots, Y_n)$, $\Delta = (\delta_1, \ldots, \delta_n)$ and $\boldsymbol{x} = (x_1, \ldots, x_n)$, respectively.

To account for heterogeneity, we propose that each patient arises from one of K latent classes with probability $\pi_k, k = 1, \ldots, K$, $\sum_k \pi_k = 1$. We assume Cox's PH model (Cox, 1972) within each class k, so that the covariate vector x enters the model log-linearly via a class-specific hazard: $\log h_k(t|x) = \log h_{0k}(t) + x'\beta_k$. For a general introduction to the Cox regression, see Hosmer *et al*. (2011). In particular, recall that a right-censored observation following a PH model has the density

$$f_k(y, \delta|x) = [h_{0k}(y)\exp(x'\beta_k)]^{\delta}\exp[-H_{0k}(y)\exp(x'\beta_k)], \tag{1}$$

where $h_{0k}(t)$ and $H_{0k}(t)$ are the baseline hazard and baseline cumulative hazard for the kth class. The mixture density may be written as

$$f(Y, \Delta|\boldsymbol{x}) = \prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k f_k(Y_i, \delta_i|x_i). \tag{2}$$

If we also observe the latent class $U = (U_1, U_2, \ldots, U_n)$, where $U_i \sim \text{Multinomial}(\pi)$, $U_i \in \{1, 2, \ldots, K\}$ and $u_{ik} = 1_{\{U_i = k\}}$, we may write the density of the complete data as

$$f(Y, \Delta|\boldsymbol{x}, U) = \prod_{i=1}^{n}\prod_{k=1}^{K}[\pi_k f_k(Y_i, \delta_i|x_i)]^{u_{ik}}. \tag{3}$$

To estimate the regression coefficients and baseline hazard parameters, we propose maximizing this likelihood via the expectation maximization (EM) procedure (Dempster *et al*., 1977) described in Section 4.

The discrete mixture leads to the model's interpretation as organizing observations into clusters that are not known *a priori*. This type of clustering should not be confused with *clustered* survival data, which typically refers to the case where class labels identifying multiple observations from the same source (e.g. treatment centers or year of diagnosis) are known. Instead, observations are gathered according to their best-fitting regression model.

Additionally, our mixture relaxes the PH assumption; we only need to assume that hazards are proportional within their given clusters. The practical interpretation of this property is highlighted in the example given in Section 2.2. Continuous frailty (Aalen, 1988) and random effects

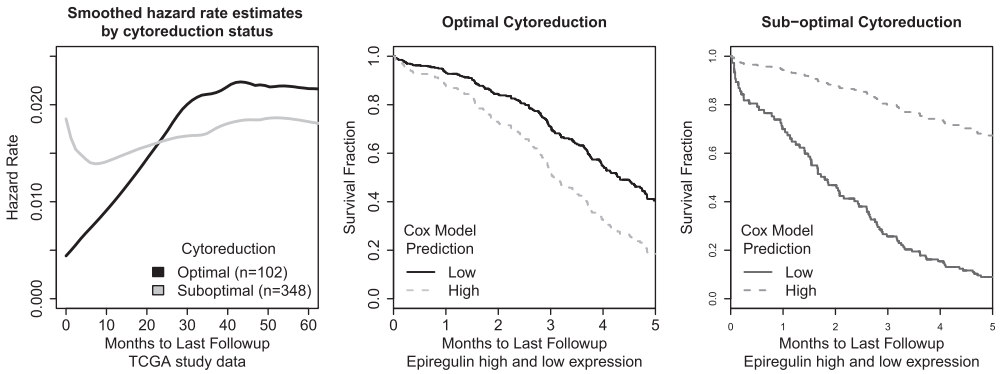

Fig. 2. Smoothed hazard rate estimates stratified on optimal and suboptimal cytoreduction classes show a non-proportional relationship. We identify gene epiregulin whose relationship to survival inverts across these underlying classes

models (O'Quigley and Stare, 2002) are other common methods for accommodating heterogeneity, but they still rely on the PH framework.

In addition to presenting a novel approach for handling heterogeneity and subgroup discovery for time-to-event data, our approach offers a new contribution to the finite mixture model literature (McLachlan and Peel, 2000; Frühwirth-Schnatter, 2006). The problem of finite mixtures has been explored in mixed effects models (Qin and Self, 2006), generalized linear models (Wedel and DeSarbo, 1995) and discrete-time survival models (Muthén and Masyn, 2005); our approach extends the idea to PH regression.

4 ALGORITHM: COX-ASSISTED CLUSTERING

We refer to the following algorithm for maximizing the mixture likelihood as Cox-assisted clustering (CAC). For convenience, we write the parameters to be estimated as $\pi = (\pi_1, \pi_2, \ldots, \pi_K)$, the mixing proportions, $\mathbf{h} = \{h_{01}(t), \ldots, h_{0K}(t)\}$, the set of baseline hazard functions and $\beta = (\beta_1, \ldots, \beta_K)$, the coefficient vectors. We further abbreviate the hazards at their evaluation points: $h_{0ki} = h_{0k}(Y_i)$ and $H_{0ki} = H_{0k}(Y_i)$.

The complete data likelihood with mixing parameters π and class-specific parameters $\mathbf{h}$ and β may be separated into a mixing distribution part and a component distribution part:

$$\log L(\pi, \mathbf{h}, \beta;\, Y, \Delta, U|\boldsymbol{x}) = \log L_1(\pi;\, U) + \log L_2(\mathbf{h}, \beta;\, Y, \Delta|U, \boldsymbol{x}).$$

The first is simply

$$\log L_1(\pi;\, U) = \sum_{k=1}^{K} \left(\sum_{i=1}^{n} u_{ik} \right) \log \pi_k, \tag{4}$$

and the likelihood associated with the component distributions is

$$\log L_2(\mathbf{h}, \beta;\, Y, \Delta|U, \boldsymbol{x}) = \sum_{k=1}^{K} \sum_{i=1}^{n} \delta_i u_{ik} \log h_{0ki} + \delta_i u_{ik} x_i' \beta_k - u_{ik} H_{0ki} \exp(x_i' \beta_k). \tag{5}$$

To compute the maximum likelihood estimate, we follow an EM approach that estimates and optimizes the observed data log likelihood by plugging $\hat{u}_{ik} = E(u_{ik}|Y_i, \delta_i, \boldsymbol{x})$ into the complete data likelihood. Supposing that the current values of the parameters at the mth iteration are $\pi_k^{(m)}$, $h_{0ki}^{(m)}$, $H_{0ki}^{(m)}$ and $\beta_k^{(m)}$, the algorithm proceeds as follows.

In the E-step, conditional mean is

$$\hat{u}_{ik} = \frac{\pi_k^{(m)} \left[h_{0ki}^{(m)} \exp\left(x_i' \beta_k^{(m)} \right) \right]^{\delta_i} \exp\left[-H_{0ki}^{(m)} \exp(x_i' \beta_k^{(m)}) \right]}{\sum_{k'} \pi_{k'}^{(m)} \left[h_{0k'i}^{(m)} \exp\left(x_i' \beta_{k'}^{(m)} \right) \right]^{\delta_i} \exp\left[-H_{0k'i}^{(m)} \exp(x_i' \beta_{k'}^{(m)}) \right]} \tag{6}$$

after the application of Bayes rule. We note that, unlike the standard Cox regression setting, computing the baseline hazard is necessary to compute the conditional means. It can be shown that if we assume a common baseline hazard across clusters, the E-step update depends only on the current estimates of π and β.

In the M-step, the update for mixing proportions π_k is straightforward:

$$\pi_k^{(m+1)} = \frac{\sum_{i=1}^{n} \hat{u}_{ik}}{n}. \tag{7}$$

To update $\mathbf{h}$, we make a profile likelihood argument that leads to a partial likelihood (Johansen, 1983). Suppose we hold β_k constant. Maximizing over h_{0k}, we obtain profile estimates of the hazards as a function of the $\beta_k^{(m+1)}$ that are similar to Breslow (1974):

$$h_{0k}^{(m+1)}(Y_i) = \frac{\hat{u}_{ik}}{\sum_{j:\, Y_j \geq Y_i} \hat{u}_{jk} \exp(x_j' \beta_k^{(m+1)})} \tag{8}$$

$$H_{0k}^{(m+1)}(Y_i) = \sum_{l:\, Y_l \leq Y_i} \frac{\hat{u}_{lk}}{\sum_{j:\, Y_j \geq Y_l} \hat{u}_{jk} \exp(x_j' \beta_k^{(m+1)})}. \tag{9}$$

The profiled M-step objective is a partial likelihood weighted by the $\hat{u}_{ik}$:

$$\log L_2\left(\mathbf{h}(\beta), \beta;\, Y, \delta, \hat{U}|\boldsymbol{x} \right) = \sum_{k=1}^{K} \sum_{i=1}^{n} \delta_i \left\{ \hat{u}_{ik} x_{i'} \beta_k - \log \sum_{j:\, Y_j \geq Y_i} \exp\left[\hat{u}_{jk} x_{j'} \beta_k \right] \right\}. \tag{10}$$

Each component indexed by k may be maximized separately to obtain the $\beta_k^{(m+1)}$ update using standard statistical software. The M-step is operationally equivalent to fitting K-weighted Cox models. Finally, one iterates between the E and M step until the increment in log-likelihood is small.

4.1 Starting conditions and number of classes

As an iterative procedure, the EM algorithm requires initial values $\beta^{(0)}, \pi^{(0)}$. For analyses that begin with strong biological hypotheses, the corresponding parameters may be set directly. An alternative is to choose starting parameters by assigning observations to specific classes and estimating the initial $\beta^{(0)}$ and $\pi^{(0)}$. This is equivalent to setting an initial value for every $\hat{u}_{ik}$ and running the algorithm forward. This assignment may be random; one may set a randomly selected $\hat{u}_{ik}$ to 0.8, say, and divide the remaining weight among the other classes. In practice, we use multiple random starts and pick the best by the fitted log-likelihood.

As in other clustering problems (Fraley and Raftery, 1998), we select the number of classes using the Bayesian information criterion (BIC). Let $L(K) = L(\mathbf{h}_K, \beta_K, \pi_K;\, Y, \delta, U|\boldsymbol{x})$, where we have added the K subscript to emphasize the dependence. The BIC criterion is expressed as

$$BIC(K) = -2 \log L(K) + pK \log(n), \tag{11}$$

where p is the dimension of X and n is the number of observed patients. The value of K minimizing $BIC(K)$ is a penalized compromise between fit and complexity. Also, while Volinsky and Raftery (2000) propose weighting by the number of observed events, $\log(\sum_i \delta_i)$, because $\log(n)$ is always larger, the standard BIC is a more conservative criterion.

As a measure of model sensitivity to additional clusters, we consider an adaption of the DFBETA statistic (Hamilton, 1992). Given a model fit for K classes, each patient i can be assigned a

parameter vector $\beta(\tilde{u}_i^{(K)})$ given their assigned class, $\tilde{u}_i^{(K)}$. For each component j, we simply consider the average change over K:

$$DFBETA(j,K) = n^{-1}\sum_{i=1}^{n}\frac{\left|\beta_j\left(\tilde{u}_i^{(K)}\right)-\beta_j\left(\tilde{u}_i^{(K-1)}\right)\right|}{\left|\beta_j\left(\tilde{u}_i^{(K-1)}\right)\right|} \quad (12)$$

This statistic will be large when the coefficients change dramatically between cluster numbers. Conversely, if the (K + 1)th cluster simply subdivides an existing cluster, the statistic will be small.

5 SIMULATION STUDIES

Although there are several properties of the model and algorithm to highlight, we focus on its treatment of censored data and a demonstration of its estimation ability. Let true class indicator $U_i \in \{1,2\}$ be evenly split among $2n$ observations with a single covariate $(X_1, \ldots, X_{2n}) \sim \mathcal{N}(\mu 1_{2n}, I_{2n})$ independent normal with mean $\mu \geq 0$ and variance 1. As is common in gene expression studies, we will work with scaled and centered X, so μ reflects the sensitivity of the analysis to this standardization.

The relationship between survival and X is controlled by $\beta \geq 0$, where the first class has $\beta_1 = \beta$ and the second class has $\beta_2 = -\beta$. The survival time for the ith patient is then $T_i = \frac{\epsilon_i}{\exp(X_i\beta_{(U_i)})}$ where $\epsilon_i \sim$ Exponential(1). The censoring time is generated from $C_i \sim$ Uniform$(0,\lambda)$, where λ depends on the choice of μ and β and a target censoring rate. Finally, the observed survival time is $Y_i = \min(T_i, C_i)$.

We set $n = 200$ patients in each class and set $\beta = 3$ so that $\beta_1 = +3$ and $\beta_2 = -3$. We target 40% censoring by setting $\lambda = \exp(0.99)$ for $\mu = 0$ and $\lambda = \exp(12.83)$ for $\mu = 5$. For this simulation, we run our algorithm at the true number of clusters $K = 2$.

We study the same scenarios over 1000 simulations. In Table 1, we report the estimated β_k (choosing $\hat{\beta}_1 \leq \hat{\beta}_2$ for identifiability) alongside the oracle estimator that knows the true classes. Intuitively, if the data are perfectly classified, the oracle estimate will have properties consistent with the well-studied Cox model estimate. Thus, the accuracy, the proportion of patients assigned to their true class, is an ideal measure of loss of performance due to uncertainty. Considering the standard Cox model in this heterogeneity setting, the median parameter estimate for the $\mu = 0$ case is 1.8e-05 (range −0.26–0.28) implying that heterogeneity has masked all detectable association with the covariate of interest.

Table 1. Simulation study results demonstrate accurate model fitting with CAC

Parameter	Scenario	
	$\mu = 0$	$\mu = 5$
β_{Cox} (SD)	0.00 (0.09)	0.00 (0.08)
$\beta_{CAC,1}$ (SD)	3.45 (0.53)	3.24 (0.58)
$\beta_{CAC,2}$ (SD)	−3.46 (0.52)	−2.14 (0.55)
$\beta_{oracle,1}$ (SD)	3.05 (0.34)	3.03 (0.28)
$\beta_{oracle,2}$ (SD)	−3.03 (0.33)	−3.12 (0.57)
Accuracy (range)	0.87 (0.78–0.94)	0.91 (0.63–1.00)
Censoring (range)	0.39 (0.31–0.47)	0.39 (0.36–0.44)

Note: SDs are standard deviations over 1000 simulations. Accuracy is the proportion of observations assigned to their correct class. β_{Cox} refers to K = 1 component Cox model estimates and were of order 1.0e-05.

The results imply that the clustering algorithm works well despite heavy censoring and mean mis-specification. We note a bias toward larger absolute parameter estimates ($\beta_{CAC,1} > \beta_{oracle,1}$) that we believe comes from the algorithm greedily reinforcing what it has already learned. A censoring bias in the $\mu = 5$ scenario appears as $\beta_{CAC,2}$ is smaller than it should be; this group is more likely to be censored so it has a lower effective sample size. Variation in per cluster censoring rates is a novel data consideration, and we recommend tracking the number of events in each cluster (see for example Section 6). When the cluster sizes are generated by $\sum_i U_i \sim$ Binomial $(2n, 1/2)$, the estimates are similar and the standard errors increase reflecting the variation in U_i (Supplementary Material).

To consider the ability of CAC to identify heterogeneity, we tested the above $\mu = 0$ scenario with the DFBETA statistic noting that DFBETA(K = 2) is greater than DFBETA(K = 3) in all cases; the median DFBETA(K = 2) was 60 (IQR: 92), implying the model was much more sensitive to two clusters than one, whereas the median DFBETA(K = 3) was 0.91 (IQR: 0.21), implying that two clusters are sufficient. Conversely, if we generate data where all patients come from the same class, DFBETA(K = 2) is larger than DFBETA(K = 3) in 61.2% of cases with medians 0.83 and 0.81, respectively. This implies that, along with appropriate context and judgment, the DFBETA statistic is a useful tool for diagnosing heterogeneity.

6 DNA REPAIR EXPRESSION SUBGROUPS IN OVARIAN CANCER

Because of its frequency among gynecological cancers, its high lethality and poor options for treatment (Vaughan *et al*., 2011), serous ovarian cancer was a pilot target for molecular characterization in TCGA (The Cancer Genome Atlas Research Network, 2011). The study collected banked surgical samples from $n = 503$ patients with highly annotated clinical follow-up whose cancers had been surgically debulked and who had been treated with platinum-based chemotherapy (Bhoola and Hoskins, 2006).

Platinum resistance is an important concept in the treatment of ovarian cancer because these cancers respond poorly to any type of chemotherapy (Bookman, 2005). Although resistance is not an ideal predictive marker because it is defined through treatment, the development of an independently queried molecular model is precisely the promise of a large repository study like TCGA.

One unaddressed complication is the expectation of genetic heterogeneity: patients with similar survival outcomes may have dissimilar molecular profiles. If this heterogeneity appears to take the form of subgroups and mixtures (as in the illustrations), we anticipate that our model and algorithm will be able to address it.

Therefore, we demonstrate the use of our model to explore possibly heterogenous data by modeling a potential mechanism of platinum resistance in TCGA patients. Because recent reviews of resistance highlight the homologous repair pathway for repairing DNA damage (Martin *et al*., 2008; Cooke and Brenton, 2011), we focus on modeling the function of this set of genes. The homologous repair pathway is defined by Kyoto Encyclopedia of Genes and Genomes annotation (hsa:03440) (Kanehisa *et al*., 2010) and corresponds to 27 unique gene symbols.

We fit our model for $K = 2, \ldots, 10$ using 100 random starts for each K and selecting the best fit by log likelihood. By BIC, we select $K = 5$ clusters (Supplementary Material). Survival times are truncated at 60 months of observation to reduce the influence of 76 patients who are observed beyond the time of interest. In total, 186 of 503 (37%) patients are censored before 60 months.

Table 2 describes the quality of the cluster fits by the relative weight of each cluster ($\sum_i \hat{u}_{ik}$), the number of patients assigned (n) and the mean posterior probability for patients in their assigned clusters. The number of events in each cluster and the restricted mean (up to 60 months) are listed.

We observe that, although they have the largest number of patients assigned, clusters 4 and 5 have the smallest mean posterior probabilities, implying that their members are less similar internally. The clustering appears to be driven by the poor prognosis patients in clusters 1, 2 and 3. Noting the presence of crossing survival functions, the five class log-rank test is significant ($P = 1.26\text{e-}09$).

With respect to low posterior probabilities, we observe that the algorithm makes intuitively reasonable use of the censored observations. We plot the maximum posterior probability for each patient by their survival time in Figure 3A. Because censored patients do not have definitive events, the algorithm is less certain about which cluster to assign them. Patients with the least follow-up time have maximum posterior probability close to 0.2 (i.e. 1 of 5), and as they are observed, longer the certainty of their maximum assigned cluster rises. To wit, the hardest to classify patients are the least observed.

Within each cluster, we check model diagnostics and look for influential points. Noting that there was moderate evidence ($P = 0.059$) for non-PH (Grambsch and Therneau, 1994) when considering all the patients as a single group, after fitting our model, the within-cluster tests are all strongly insignificant. All influence statistics for all genes in each cluster are smaller than 1 standard deviation, implying no leverage points.

Because the mixture allows different clusters to have different baseline hazard functions, in Figure 3B, we used a kernel smoothing algorithm to visualize their estimates (Müller and Wang, 1994). We emphasize the non-proportionality of the hazards for clusters 3 and 5: patients in cluster 3 have a sudden acceleration in their hazard after 30 months, which may be consistent with the loss of effect in platinum chemotherapy.

In Figure 3C, we have plotted the estimated coefficients for clusters 3 and 5. Keeping in mind that the linear predictor in the Cox model scales hazard relative to the cluster-specific baseline hazard, we highlight three genes. RPA4's coefficients $(\beta_3, \beta_5) = (+7.13, -5.90)$ imply that it has a strong deleterious effect in cluster 3 (exacerbating the jump in hazard), while it has a strong protective effect in cluster 5. Contrast this change with RPA3 $(-0.96, +6.24)$, which only increases risk in cluster 5, and TOP3A $(-7.54, -0.54)$, which is protective in cluster 3 only. At this point in the analysis, these genes are good candidates for follow-up studies: we have identified their effect specific to a subgroup of patients.

Table 2. Fitted cluster diagnostics for $K = 5$, homologous repair model

Cluster	4	5	2	1	3
Survival time	34.16	26.54	15.21	12.69	5.90
Standard error	3.04	3.40	3.11	3.07	1.83
N	213.00	112.00	67.00	58.00	53.00
Events	50.00	46.00	50.00	45.00	50.00
Alive at 60 months	35.00	19.00	11.00	8.00	3.00
Weight	103.22	69.07	57.55	51.47	51.05
Mean $\hat{u}$	0.48	0.62	0.86	0.89	0.96

Note: Clusters are ordered by mean months of survival following surgery estimated by restricted mean.

Further, the clustering model may still recover a sense of DE for survival data. Because we have learned risk classes, we may consider DE across clusters. We focus on DE across clusters 3 and 5, SHFM1 (Bonferroni adjusted t-test, $P = 0.011$), RPA3 ($P = 0.011$) and RAD51L3 ($P = 0.020$), all have significant shifts in expression. Notably, RPA4 and TOP3A do not show significant DE, implying that they fit into the class of variables that only differ in regression model effects.

Representative of prognostic signature development, Kang *et al*. (2012) conducted a study of selected DNA repair pathways and produced a risk signature in this dataset. To compare with their model (using 151 genes across eight pathways) and a typical Cox model approach (using the 27 gene homologous repair pathway), we stratified patients based on survival to 3.7 years [as in Kang *et al*. (2012)] and produced receiver operating characteristic plots for all of the signatures (Fig. 4A). While the standard Cox model underperforms at area under the curve (AUC) 0.61 (comparable with a clinically derived model reported by Kang and colleagues), the $K = 5$ CAC model has AUC 0.73 comparable with Kang's model (AUC 0.70).

The key advantage of the CAC approach is the ability to describe risk sets. By itself, the Cox model describes continuous risk and must be dichotomized *post hoc* for survival curve plots of high- and low-risk subsets. In Figure 4B, we have illustrated the typical continuous risk score plot that describes sensitivity of high- and low-risk sets to the cutpoint. The CAC model naturally separates patients into risk classes (clusters 3 and 5 are highlighted, Fig. 4C) and these can be further described by processing their continuous risk scores.

Based on our analysis, we conjecture that we are able to identify a subgroup of patients (cluster 3) who experience a significant increase in hazard around month 30. We are able to identify genes whose expression leads to increased risk specific to a subgroup or whose relationship inverts across clusters. There is a tremendous amount of untapped information remaining in the fitted model. For example, every pairwise comparison between clusters should be informative as well as their holistic

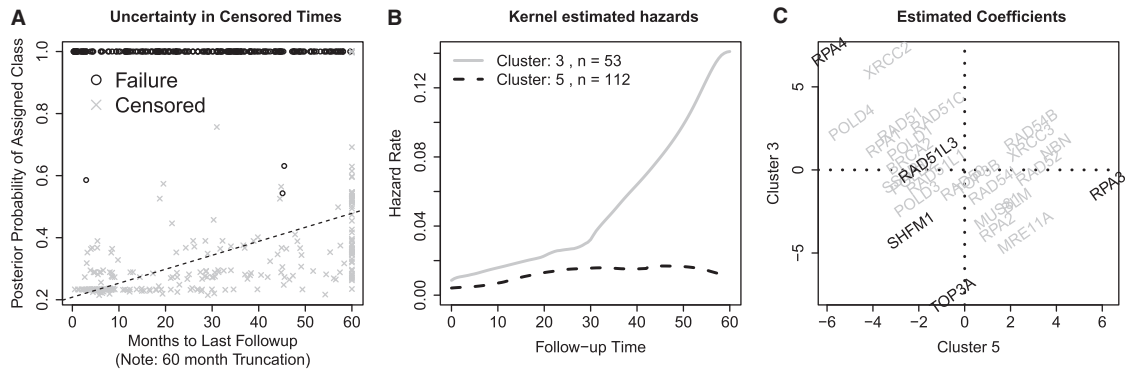

Fig. 3. (**A**) Uncertainty in censored observations with regression line for the censored points added shows that certainty increases with follow-up. (**B**) Hazards and (**C**) estimated coefficients for clusters 3 and 5 show non-PH and heterogenous effects. Genes highlighted in the text are shown in bold

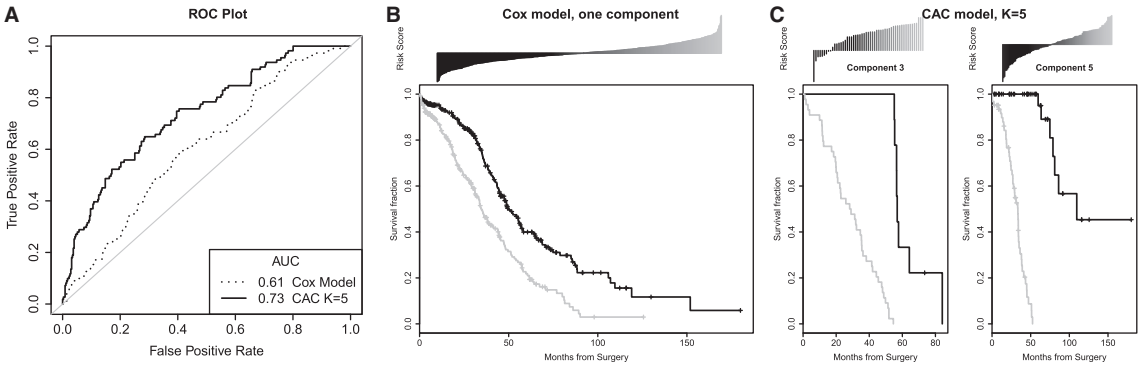

Fig. 4. (**A**) Receiver operating characteristic plot with AUC estimates. (**B**) The one component Cox model risk score has to be dichotomized for high-/low-risk Kaplan–Meier estimates. (**C**) For comparison, two components of the K = 5 CAC model show more prognostic ability

interpretation, foreshadowing the utility of this methodology for exploratory data analysis.

7 DISCUSSION

In this article, we have presented a model for heterogeneity in time-to-event data. Although its actual formulation is straightforward, the treatment of unknown classification, a consideration of the implications for censoring, the effect on genomic predictors and diagnostic analysis have not been previously considered. Finally, we have presented a novel and informative analysis in Section 6, which begins to identify the set of survival-associated and subgroup-dependent alterations in expression.

Admirably, this model relaxes the whole model PH assumption to conditional PH given cluster membership. In an exploratory situation, the utility of this flexibility cannot be overstated. Both our simulated and applied analysis highlight that our understanding of censoring has been augmented and the use of information in the model is intuitively simple.

In a data analytic view, informaticists are familiar with unsupervised clustering analysis and class-label supervised clustering analysis. Our algorithm may be seen as a way to use a survival time (possibly censored) to supervise the clustering of gene expression data. This clustering property is distinct from ensemble-type methods [e.g. Jordan and Jacobs (1994)], where covariate information may be used to reweight components. As we saw with its treatment of posterior class probabilities, the mixture believes each observation comes from a single component, while averaging over weak learners or hidden layers may be relevant to a more admixed sample.

With respect to developing ovarian cancer biomarkers, our data analysis has shown an example where class identification leads to risk stratification. We might further identify high- and low-risk classes within the assigned clusters as is standard practice, but this is no longer a required post-processing part of the expression analysis. The CAC algorithm has also given us its posterior weights allowing a concrete measure of uncertainty for downstream analyses.

In the TCGA project, as in many other cancer genomic studies, there are issues of both high-dimensional data and variable selection. As presented, the CAC regression framework does not incorporate these, but it can be extended with additional study.

ACKNOWLEDGEMENTS

The results published here are in whole or part based on data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov/.

Funding: This work was supported by Roswell Park Cancer Institute and National Cancer Institute [CA016056, CA157219] and the National Library of Medicine [LM007359].

Conflict of Interest: none declared.

REFERENCES

Aalen,O. (1988) Heterogeneity in survival analysis. *Stat. Med.*, **7**, 1121–1137.

Amsterdam,A. *et al.* (2011) Epiregulin as a marker for the initial steps of ovarian cancer development. *Int. J. Oncol.*, **39**, 1165–1172.

Bhoola,S. and Hoskins,W. (2006) Diagnosis and management of epithelial ovarian cancer. *Obstet. Gynecol.*, **107**, 1399–1410.

Bookman,M. (2005) Standard treatment in advanced ovarian cancer in 2005: the state of the art. *Int. J. Gynecol. Cancer*, **15**, 212–220.

Breslow,N. (1974) Covariance analysis of censored survival data. *Biometrics*, **30**, 89–99.

Cooke,S. and Brenton,J. (2011) Evolution of platinum resistance in high-grade serous ovarian cancer. *Lancet Oncol.*, **12**, 1169–1174.

Cooke,S. *et al.* (2010) Genomic analysis of genetic heterogeneity and evolution in high-grade serous ovarian carcinoma. *Oncogene*, **29**, 4905–4913.

Cox,D. (1972) Regression models and life-tables. *J. R. Stat. Soc. Ser. B (Methodol.)*, **34**, 187–220.

Dempster,A. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)*, **39**, 1–38.

Farewell,V. (1982) The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, **38**, 1041–1046.

Fraley,C. and Raftery,A. (1998) How many clusters? which clustering method? answers via model-based cluster analysis. *Comput. J.*, **41**, 578–588.

Frühwirth-Schnatter,S. (2006) *Finite Mixture and Markov Switching Models*. Springer Science, NY, New York.

Grambsch,P. and Therneau,T. (1994) Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515–526.

Hamilton,L.C. (1992) *Regression with Graphics: a Second Course in Applied Statistics*. Duxbury Press, Belmont, CA.

Hosmer,D.W. *et al.* (2011) *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Wiley-Interscience, Hoboken, NJ.

Johansen,S. (1983) An extension of Cox's regression model. *Int. Stat. Rev.*, **51**, 165–174.

Jones,S. *et al.* (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, **321**, 1801.

Jordan,M.I. and Jacobs,R.A. (1994) Hierarchical mixtures of experts and the em algorithm. *Neural Comput.*, **6**, 181–214.

Kanehisa,M. *et al.* (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.

Kang,J. *et al.* (2012) A DNA repair pathway–focused score for prediction of outcomes in ovarian cancer treated with platinum-based chemotherapy. *J. Natl Cancer Inst.*, **104**, 670–681.

Köbel,M. *et al.* (2008) Ovarian carcinoma subtypes are different diseases: implications for biomarker studies. *PLoS Med.*, **5**, e232.

Konstantinopoulos,P. *et al.* (2008) Gene-expression profiling in epithelial ovarian cancer. *Nat. Clin. Pract. Oncol.*, **5**, 577–587.

Lostritto,K. *et al.* (2012) A partitioning deletion/substitution/addition algorithm for creating survival risk groups. *Biometrics*, **68**, 1146–1156.

Martin,L. *et al.* (2008) Platinum resistance: the role of DNA repair pathways. *Clin. Cancer Res.*, **14**, 1291–1295.

McLachlan,G.J. and Peel,D. (2000) *Finite Mixture Models*. Wiley-Interscience, NY, New York.

Müller,H. and Wang,J. (1994) Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics*, **50**, 61.

Muthén,B. and Masyn,K. (2005) Discrete-time survival mixture analysis. *J. Educ. Behav. Stat.*, **30**, 27–58.

Na,Y. *et al.* (2009) Ovarian cancer: markers of response. *Int. J. Gynecol. Cancer*, **19**, S21.

O'Quigley,J. and Stare,J. (2002) Proportional hazards models with frailties and random effects. *Stat. Med.*, **21**, 3219–3233.

Qin,L. and Self,S. (2006) The clustering of regression models method with applications in gene expression data. *Biometrics*, **62**, 526–533.

Segal,M.R. (1988) Regression trees for censored data. *Biometrics*, **44**, 35–47.

The Cancer Genome Atlas Research Network. (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609–615.

Toyoda,H. *et al.* (1995) Epiregulin a novel epidermal growth factor with mitogenic activity for rat primary hepatocytes. *J. Biol. Chem.*, **270**, 7495–7500.

Vaughan,S. *et al.* (2011) Rethinking ovarian cancer: recommendations for improving outcomes. *Nat. Rev. Cancer*, **11**, 719–725.

Volinsky,C.T. and Raftery,A.E. (2000) Bayesian information criterion for censored survival models. *Biometrics*, **56**, 256–262.

Wedel,M. and DeSarbo,W. (1995) A mixture likelihood approach for generalized linear models. *J. Classif.*, **12**, 21–55.