Article:
Alqahtani, K, Taylor, CC orcid.org/0000-0003-0181-1094, Wood, HM orcid.org/0000-0003-3009-5904 et al. (1 more author) (2022) Sparse modelling of cancer patients' survival based on genomic copy number alterations. Journal of Biomedical Informatics, 128. 104025. ISSN 1532-0464

https://doi.org/10.1016/j.jbi.2022.104025

Sparse modelling of cancer patients' survival based on genomic copy number alterations

Khaled Alqahtani[a,b], Charles C. Taylor[b], Henry M. Wood[c], Arief Gusnanto[b,*]

[a] *Department of Mathematics, College of Science and Humanitarian Studies, Prince Sattam Bin Abdulaziz University, Al Kharj, Saudi Arabia*
[b] *Department of Statistics, University of Leeds, Leeds LS2 9JT, United Kingdom*
[c] *Leeds Institute of Medical Research at St. James's, University of Leeds, Leeds LS9 7TF*

Abstract

Copy number alterations (CNA) are structural variation in the genome, in which some regions exhibit more or less than the normal two chromosomal copies. This genomic CNA profile provides critical information in tumour progression and is therefore informative for patients' survival. It is currently a statistical challenge to model patients' survival using their genomic CNA profiles while at the same time identify regions in the genome that are associated with patients' survival. Some methods have been proposed, including Cox proportional hazard (PH) model with ridge, lasso, or elastic net penalties. However, these methods do not take the general dependencies between genomic regions into account and produce results that are difficult to interpret. In this paper, we extend the elastic net penalty by introducing additional penalty that takes into account general dependencies between genomic regions. This new model produces smooth parameter estimates while simultaneously performs variable selection via sparse solution. The results indicate that the proposed method shows a better prediction performance than other models in our simulation study, while enabling us to investigate regions in the genome that are associated with the patients' survival with sensible interpretation. We illustrate the method using a real dataset from a lung cancer cohort and simulated data.

Keywords: Cox proportional hazard, Regression, Sparse solution, Copy

*Corresponding author

number alterations, Lung cancer

1. Introduction

1.1. Motivation

Copy number alteration (CNA) is a type of structural variation in the genome [1]. It refers to the duplication or deletion of DNA segments larger than 1 kb [2]. It is therefore common to observe CNA estimates in long segments [3] or smooth segments [4] that describe the transition of copy number across genomic regions in an individual. As a result, neighbouring genomic regions are correlated; the CNA estimate in one particular genomic region is not likely to differ much from its adjacent regions. In the medical context, CNA has been extensively observed in tumorigenesis and speculated to drive tumor progression in multiple cancers [5]. Therefore, the CNA patterns carry valuable information on patients' survival and it is currently a statistical challenge to model the patients' survival based on CNA genomic profiles. From medical view point, prediction of cancer patients' survival is important for the care and management of cancer patients' well being.

From a statistical modelling view point, the task of modelling poses some inter-related challenges. The first one is not an uncommon feature of data produced by current technologies: the number of variables (genomic regions) far exceeds the number of patients. Facing this challenge, it is immediately clear that the original Cox proportional hazard (PH) model [6] is not suitable due to over-parameterisation. The second challenge is how to perform a variable selection: we expect that only some genomic regions are associated with the patients' survival while the other genomic regions are not. The third and final challenge is the fact that the variables (genomic regions) exhibit moderate to high 'block' correlation as discussed in Section 2.2. These three challenges are inter-related and we believe that it is sub-optimal to deal with those challenges separately. For example, ignoring the correlation between genomic regions in the variable selection creates a problem in the interpretation of the results. If a

single genomic region is identified to be associated with the patients' survival, then we expect the adjacent regions (to some extent) shall also be associated. This study proposes a novel method on how to deal with these challenges in a single modelling framework.

1.2. Survival analysis on high-dimensional genomic data

In dealing with the high-dimensionality of the data, some authors have considered Cox PH model with penalty function [7]. In the model, a ridge penalty is introduced to log partial likelihood and provide a shrinkage to model parameter estimates. The model does not produce a so-called 'sparse solution'. The term 'sparse' refers to the situation where some model parameters are zero estimated and the other parameters are estimated away from zero. In effect, a variable selection is carried out. The lasso model was proposed to specifically deal with this problem [8]. In [8], Cox PH model is equipped with a lasso penalty to perform a variable selection.

One concern of the lasso penalty is that it tends to select one variable from a group of correlated variables [9]. In our context of CNA genomic profile, having a lasso penalty in the Cox PH model would give results that are difficult to interpret (see Section 3.1). Since the data exhibit blocks of correlated regions, the model tends to select one or two variables per small correlated regions across the genome. To mitigate this situation, Zou and Hastie [9] consider a mixture of ridge and lasso penalties called the 'elastic net'. This results in a penalty function that is expected to combine the best aspects: the lasso penalty would result in variable selection while the ridge penalty would tend to group variables [9]. The Cox PH model with elastic net penalty has been used in some applications [10]. A previous study by Waldron *et al.* [11] compared the normal (ridge), elastic net, and lasso penalty in the Cox PH model using several types of genomic datasets and recommended the first two penalties. When we consider the elastic net penalty in our study, the results are better than those of lasso penalty, but their interpretation is still not satisfactory (see Section 3.1). In the context of the investigation of genomic regions, we find that the model is not

adequate since it deals with the first two challenges but not with the third one (dependencies between genomic regions).

To achieve our objective, we extend the elastic-net penalty in Cox PH model by introducing an additional penalty that imposes a smoothness in the parameter estimates. This additional penalty is specifically included to deal with the dependencies between genomic regions where second differences of model parameters of neighbouring regions are moderated to deviate around zero. With this formulation, the results show that the parameter estimates are sparse: many of the genomic regions have zero estimates, while the other regions have non-zero estimates. The non-zero estimates are grouped within some bigger regions and have smooth transition within them. To estimate the model parameters, we present a full gradient algorithm to maximise penalized partial likelihood (PPL) by generalising the idea of Goeman [12]. Coupling this with a Newton-Raphson algorithm near an optimal solution, this method is faster and can be done without the need for a high-performance computing facility. We illustrate the proposed method on a real dataset from a cohort of 89 lung cancer patients with approximately 14,000 genomic regions.

The rest of the paper is organised as follows. Section 2 describes the data involved. The proposed method of sparse-smoothed Cox PH model and its computational algorithm will be described in Section 3. Section 4 discusses our simulation study and Section 5 presents the results of real-data analysis. Section 6 contains the discussion and the conclusion is presented in Section 7.

2. Dataset

2.1. Patients data

Eighty-nine patients with early-stage lung squamous cell carcinoma (SCC) who had surgery at the Department of Thoracic Surgery, St. James Hospital, Leeds, UK, between 1994 and 2003 were included in the study [13]. Various clinical characteristics, such as age, gender, and stage of disease (Stage T and

Stage N) of the patients were recorded. The summary of the patient cohort in our study are presented in Table 1 and Figure 1.

Table 1: *Description of the lung cancer dataset*

Characteristics	Value
No. of patients	89
No. of censored obs.	23 (25.8%)
Time after surgery	Range: 34 days-12.5 years
	Median: 2.35 years
	95% CI: $(1.83, 4.31)$
Age at surgery	39 to 84 years
Gender	Male: 63; Female: 26
Stage T	T1: 23 patients, 5 censored
	T2: 59 patients, 16 censored
	T3: 7 patients, 2 censored
Stage N	N0: 47 patients, 14 censored
	N1: 35 patients, 8 censored
	N2: 7 patients, 1 censored

Table 1 shows that, among the 89 patients, approximately a quarter of survival times are censored. The median survival time is at 2.35 years, as also indicated in Figure 1. Figure 1 (top right panel) also shows the estimated cumulative baseline hazard function, which roughly indicates that Cox PH model is reasonable for the data. The Kaplan-Meier estimates of survivor function based on the T-staging (tumour size) and N-staging (location of affected lymph node) are presented in the bottom panels of Figure 1. The figures indicate that there are differences in the estimated survivor functions between different levels in each of Stage-T and Stage-N. This indicates that Stage-T and Stage-N are important factors for patients' survival.

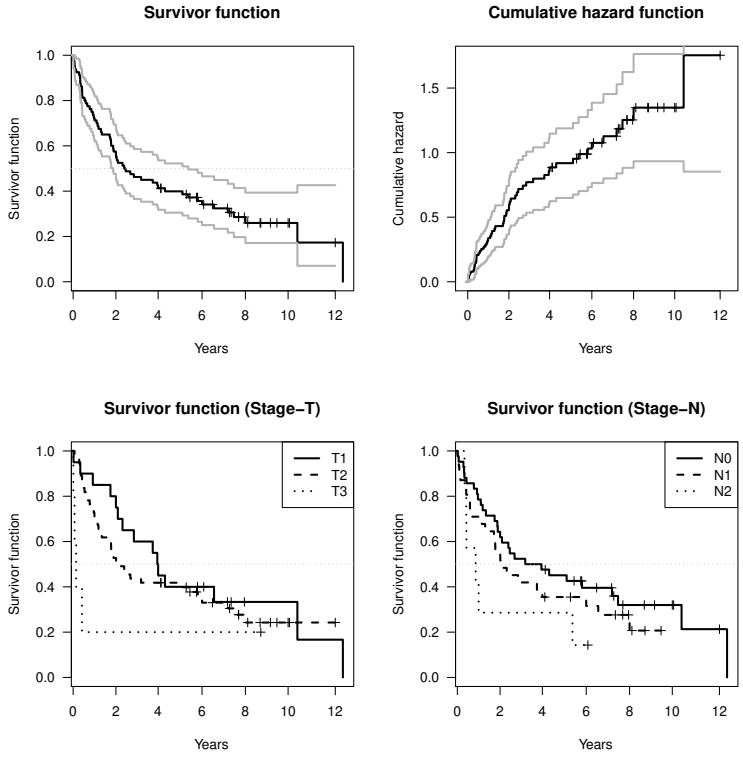

Figure 1: Top panels: *Kaplan-Meier (KM) estimates of survivor function and cumulative hazard function with* 95% *confidence interval (grey lines).* Bottom panels: *KM estimates of the survivor function based on Stage-T and Stage-N groupings. Log rank tests on the bottom panel figures indicate significant differences between groups (p-value less than* 0.05*). Horizontal dotted grey line in the estimated survivor functions marks the* 50% *probability.*

For our purpose of modelling, the information on patients' age, tumours' N-stage and T-stage will serve as predictors (or input) in the model. However, their treatment in the model will be different to that of the genomic CNA data as described in Section 3.1.

2.2. Genomic sequence and CNA estimate

The patients' DNA samples were sequenced in the next-generation sequencer Illumina GAII. The sequencer produced short sequences (usually called 'reads') that were aligned to assembly hg19 of the human reference genome using the

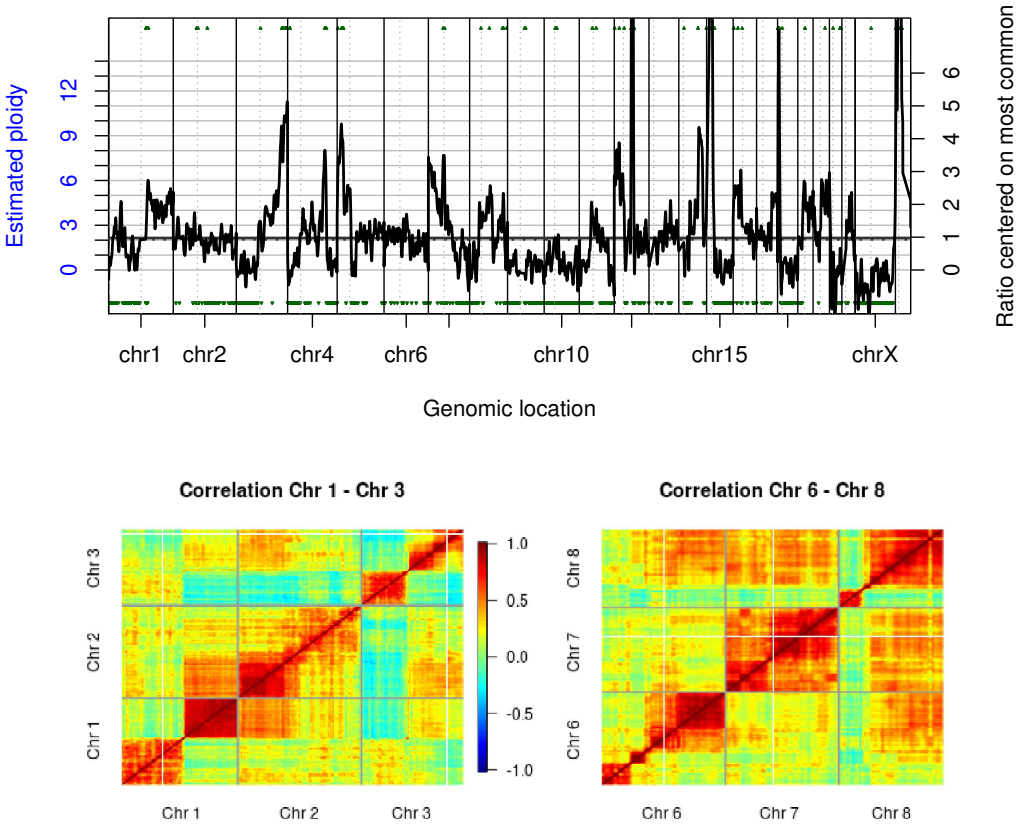

Figure 2: *Top panel: Estimated copy number alteration (CNA) as smooth segmentation line across the genome for patient LS170 (solid black line). One data point in the line corresponds to the CNA from a genomic region or 'window' of size 200 kb. The horizontal grey lines correspond to the expected copy number (left axis) or expected ratio (right axis). The horizontal line corresponds to the ratio one. The vertical solid lines separate the chromosomes and vertical dotted lines indicate the centromere regions. Bottom panel: Correlation of CNA's between genomic regions in chromosomes 1-3 and 6-8 from 89 patients in our dataset.*

Burrows-Wheeler Alignment suite version 0.5.9-r16 [14]. Further details are elaborated in Belvedere *et al.* [13].

The genome wide CNA profile from each patient is calculated by depth-of-coverage from their sequences. This involves counting the number of reads per fixed-size non-overlapping genomic-region ('window') in each sample. We estimated the optimal window size to be 200 kb [15], and this gives 15,490 windows to cover the human genome. However, due to missing data in certain parts of the genome, e.g. centromeres, we only include information from 13,968 windows in our analysis. The sequence data across patients are not directly comparable because tumour samples are inevitably contaminated with normal cells by different degrees. To deal with this problem, we performed a normalisation using the *CNAnorm* package [16] to obtain the CNA estimates, which are shown in Figure 2 (top panel) for one individual.

The figure shows the CNA profile estimated as smooth segmented lines [4] for patient LS170, where one data point corresponds to a genomic region ('window') of size 200 kb. When we combine the CNA estimates (e.g. solid line in Figure 2) across the 89 patients, we create a dataset where the genomic regions serve as predictor ('input') variables from modelling perspective. This indicates that the number of predictor variables from CNA estimation is in the order of 14 thousands (approximately) to cover the whole genome. Figure 2 (bottom panel) shows the correlation between genomic regions in chromosomes 1-3 and 6-8 from 89 patients in the dataset. The figure indicates blocks of correlation between genomic regions at different sizes in the dataset. This is an important characteristic of CNA data, which occurs because the alterations frequently manifest in segments as indicated in Figure 2. As such, if a genomic region occurs as a gain, then the adjacent regions are expected to occur as gains as well. We shall discuss how to deal with this characteristic in the modelling as discussed in Section 3.3.

CNA have been shown to be clinically meaningful for lung cancer patients' survival [17] and used as diagnostic tool for lung cancer [18]. Furthermore, CNA have also been shown to distinguish different cancer types [19, 20] or cancer

subtypes [21, 22, 23]. Our approach here will jointly utilise CNA and clinical information to make prediction of cancer patients' survival, as described in the next section.

3. Methods

In this section, we shall cover the relevant methodologies involved. First, we describe the Cox proportional hazard (PH) model that we employ for modelling in Section 3.1. In this section, we introduce the mathematical notation used throughout this manuscript. Secondly, we describe in Section 3.2 the current approaches in estimating model parameters, particularly in the context of penalised Cox PH model to model patients' survival on high-dimensional data. In this section, we shall describe why the current approaches are not satisfactory to achieve our objective, in particular the way we interpret the outcome. In Section 3.3, we describe the novel model extension that we consider to address the drawbacks of current approaches, in light of the characteristics of genomic CNA data. Subsequent sections will cover the estimation of model parameters.

3.1. Cox proportional hazard model

To proceed with the Cox proportional hazard modelling, let t be the time since surgery until a patient dies due to the cancer. Let T denote the corresponding random variable, which is defined on the positive real line with an underlying probability density function $f(t)$. Denote $S(t)$, $h(t)$, and $H(t)$ as the survivor, hazard, and cumulative hazard function, respectively [24, 25]. In the Cox PH model, we do not model t directly. Instead, we model the hazard, which is defined as the *instantaneous* rate of occurrence of death (due to the disease) [6].

Denote δ_i as the event indicator for the i-th patient, $i = 1, 2, \ldots, n$, where $\delta_i = 1$ if the i-th survival time, t_i, is uncensored and $\delta_i = 0$ if t_i is censored. We define X to be a matrix of p fixed predictors of size $n \times p$, and X_i to be the i-th row of X (a p-vector). We also define β to be a p-vector of model parameters

associated with X. In our application, the columns of X correspond to the patients' clinical characteristics such as age, gender, and tumour stages, while the rows correspond to the different patients.

Let Z be a matrix of CNA of size $n \times q$ (n patients and q genomic regions), and b be its corresponding vector of parameters. Similarly, we also denote Z_i to be the i-th row of Z (a q-vector of CNA from i-th patient). Therefore, in our application, the matrix Z contains the CNA estimates from 89 patients (rows) at approximately 14 thousands genomic regions (columns).

In the Cox PH model formulation, we model the hazard function of the i-th patient, $h_i(t)$, as

$$h_i(t|X, Z) = h_0(t) \exp\{X_i\beta + Z_i b\}, \tag{1}$$

where $h_0(t)$ is the baseline hazard function. The baseline hazard function may vary over time and is not a function of predictors. The predictors (or 'input' variables) are the clinical phenotypes contained in X and the genomic CNA data contained in Z.

3.2. Current approaches

The main challenge now from model formulation (1) is how to estimate the parameters β and b. To do this, we consider the log partial likelihood [6, 26, 27]

$$\ell_{\mathrm{p}}^{\mathrm{cox}}(\beta, b) = \sum_{i=1}^{n} \left[\delta_i (X_i\beta + Z_i b) - \delta_i \log\Bigg(\sum_{j \in R(t_i)} \exp(X_j\beta + Z_j b) \Bigg) \right] \tag{2}$$

where $R(t_i)$ is the risk set or a set of patients who are at-risk at time t_i. The parameters are not directly estimable due to over parameterisation in b, which correspond to the genomic regions. To deal with the challenges, we are interested to moderate and perform variable selection in the estimation of b while keeping standard estimation on β. Previously, Tibshirani [8] introduced the lasso penalty to the log partial likelihood

$$\ell_{\mathrm{p}}^{\mathrm{lasso}}(\beta, b, \lambda) = \ell_{\mathrm{p}}^{\mathrm{cox}}(\beta, b) - \lambda \sum_{k=1}^{q} |b_k| \tag{3}$$

for a positive λ.

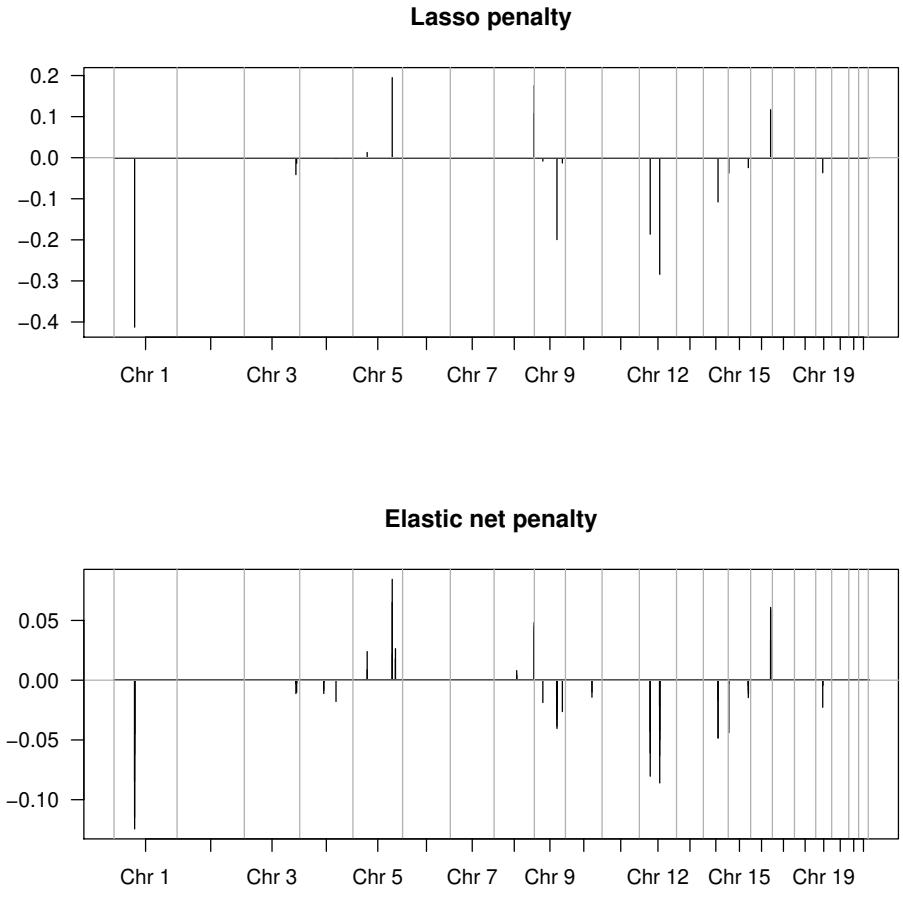

Figure 3: *Estimates of parameters for CNA profile $\widehat{(b)}$ under the lasso penalty (top panel) and elastic net penalty (bottom panel).*

However, when we estimate b using the penalised likelihood (3) as shown in Figure 3 (top panel), the estimates $\widehat{b}$ are difficult to interpret. All of the non-zero estimates are from isolated single windows, and not spanning over several windows. The first few non-zero estimates correspond to window no. 1914, 3197, 3210, and 4200, in chromosomes 2, 3, 3, and 5, respectively (window 4200 is the first window in chromosome 5 and the estimate for this window is not visible in Figure 3 because it coincides with the grey line to mark the start of the chromosome). It can be shown that almost all of the non-zero estimates are within blocks of correlation. It is therefore difficult to understand why those particular single windows are identified to have non-zero estimates and not their adjacent windows, for example. One might think to consider a whole block, represented by its corresponding non-zero estimates, to be significant. However, this raises another problem on the block size to consider. Furthermore, some of the non-zero estimates are relatively close such as windows 3197 and 3210. We have difficulties to understand whether we should consider them as a single region of interest (i.e. a single region between window 3198 and window 3209) or two separate regions of interest.

The above results are well known to be the characteristics of lasso penalty. To have some sort of 'grouping effect', Zou and Hastie [9] propose an elastic net penalty, which consists of ridge and lasso penalties. With a slight reparameterisation, the log partial likelihood with elastic net penalty can be written as

$$\ell_{\mathrm{p}}^{\mathrm{enet}}(\beta, b, \theta) = \ell_{\mathrm{p}}^{\mathrm{cox}}(\beta, b) - \left[\alpha\left(\frac{1}{2} b^T D(\theta)^{-1} b\right) + \frac{1-\alpha}{\sqrt{\theta}}\left(\sum_{k=1}^{q} |b_k|\right)\right] \quad (4)$$

where $D(\theta) \equiv \theta I_q$, I_q is the identity matrix of size q, α is a weight (set to 0.5 for our illustration), and θ is a positive 'tuning' parameter.

When we consider this model to our real data, the situation improves, in that the non-zero estimates form some sort of grouping (see Figure 3, bottom panel). The group size varies between two to 16. However, we find that the above issues in the interpretation remain, although in a lesser degree. Isolated single windows with non-zero estimates are still produced. The presence of short

gaps between regions with non-zero estimates also makes it difficult to identify genomic regions of interest as we do not know whether we should consider them to be a single big region or separate regions of interest.

The above case indicates that the elastic net is still not adequate for our purpose. To model the hazard while performing variable selection with sensible interpretation for genomic data, an additional penalty is needed to moderate or penalise the estimation on neighbouring windows as described in the next section.

3.3. Sparse-smoothed Cox PH model (SSCox)

We now discuss our proposal to extend the above sparse Cox PH model by introducing an additional penalty function in the log partial likelihood function to deal with the above interpretative issues. We still consider the model in Eq. (1), where the hazard is modelled as a function of the clinical phenotypes and the genomic CNA data. The overall model building is described in details in the Supplementary Material. In this section, we just describe the resulting penalty function and the relevant log partial likelihood function for simplicity.

The main idea is that each of the above penalty functions can be considered to have come from distributional assumptions on the parameter b in a random effects setting [28]. The lasso penalty, as discussed in the previous section, is the result of assuming a Laplace distribution on the parameter b [8] and the elastic-net penalty is the result assuming a mixture of Laplace and normal distributions on b [9]. The extension that we propose is to put an additional distributional assumption on the *differences* of parameters associated with adjacent windows in the genome. Specifically, we assume that differences of random effects for adjacent windows to follow a Cauchy distribution. This way the dependencies between adjacent windows are taken into account in the estimation of b. Further details on the model building are described in the Supplementary Material.

As we previously have seen from Figure 3, the non-zero estimates of b produced by the elastic net can have gaps or isolated single windows. To avoid this, we need to consider the case where the estimate of one window is not too

different with its adjacent windows. In this study, we consider second differences of consecutive windows

$$b_{j+1} - 2b_j + b_{j-1} \tag{5}$$

for $j = 2, \ldots, q-1$ to be not too far from zero. Previous authors [7, 28] formulated this to create the penalty

$$\sum_{j=2}^{q-1} \left(b_{j+1} - 2b_j + b_{j-1}\right)^2, \tag{6}$$

or, after reparameterisation with a tuning parameter θ,

$$\frac{1}{2} b^T \Sigma(\theta)^{-1} b \tag{7}$$

where $\Sigma(\theta)^{-1} \equiv \theta^{-1} R^{-1}$ and

$$R^{-1} = \begin{pmatrix} 1 & -2 & 1 & \cdots & \cdots & \cdots & 0 \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ 0 & \cdots & \cdots & 0 & 1 & -2 & 1 \end{pmatrix}. \tag{8}$$

The above penalty (7) is less suitable for our context with genomic CNA data that have sudden changes [4]. The penalty results in a slow changes of estimates b_j's between windows. Motivated by [4], we consider the penalty

$$\frac{q+1}{2} \log\left\{1 + b^T \Sigma(\theta)^{-1} b\right\}. \tag{9}$$

instead of that in Eq. (7). This is the result of assuming the second differences of the random effects for adjacent windows to follow a Cauchy distribution as described in the Supplementary Material. This additional penalty has an interesting characteristic. Up to a multiplicative constant, both Eqs. (7) and (9) apply similar amount of penalty for small $\widehat{b}$. However, as $\widehat{b}$ increases, the latter put less penalty. Therefore, using the penalty in (9), a sudden change

(either a 'jump' or 'drop') in the estimates of b is permitted while keeping them smooth.

Denoting ψ to represent other parameters and θ, we then combine this additional penalty to arrive at the log partial likelihood for β and b

$$\ell_p(\beta, b, \psi) = \sum_{i=1}^{n} \left[\delta_i (X_i\beta + Z_i b) - \delta_i \log \left(\sum_{j \in R(t_i)} \exp(X_j\beta + Z_j b) \right) \right] - \left[w_1 \left(\frac{1}{2} b^T D(\theta)^{-1} b \right) + w_2 \left(\frac{q+1}{2} \log \left\{ 1 + b^T \Sigma(\theta)^{-1} b \right\} \right) + \frac{w_3}{\sqrt{\theta}} \left(\sum_{k=1}^{q} |b_k| \right) \right] \tag{10}$$

where

1. the first line in the right hand side of the equation corresponds to $\ell_p^{\text{cox}}(\beta, b)$ in Eq. (2), and
2. $0 \leq w_r \leq 1$ for $r = 1, 2, 3$, are the weights for each penalty with $\sum_{r=1}^{3} w_r = 1$.

The estimation of the model parameters β, b, and $\psi = (\theta, w_1, w_2)$ can now be done by estimating β and b at fixed ψ as described in the next section. The estimation of the tuning parameter ψ is done via five-fold cross-validation partial likelihood as described in Section 3.6. Note that the Cox PH model with ridge penalty [7] is a special case of the above model with $w_1 = 1$, the lasso model [8] is a special case with $w_3 = 1$, and the elastic-net model [29] is a special case with $w_2 = 0$.

3.4. Estimation of β and b for fixed ψ

In this section, we describe the estimation of β and b at fixed (given) ψ. The estimation is done by alternating the estimation of β and b until convergence, using starting values $\beta^{(0)}$ and $b^{(0)}$.

3.4.1. Estimation of β

We can derive the estimation of β by taking the first partial difference of the log partial likelihood $\ell_p(\beta, b, \psi)$ of Eq. (10) with respect to β at fixed ψ. The

fixed effects β are then estimated as the solution of

$$\sum_{i=1}^{n} \delta_i \left[X_i - \frac{\sum_{j \in R(t_i)} X_j \exp(X_j\beta + Z_j b)}{\sum_{j \in R(t_i)} \exp(X_j\beta + Z_j b)} \right] = 0. \tag{11}$$

3.4.2. Estimation of b

To estimate the random effects b, the standard Newton-Raphson algorithm is not applicable since the partial log likelihood in Equation (10) is not twice differentiable everywhere. We therefore estimate b via gradient ascent algorithm by generalising the idea of Goeman [12] who estimated model parameters from a likelihood function with a Laplace penalty only. In our application, we involve a 'trick' as described below.

Consider the partial log likelihood function in Equation (10) as a target function to be optimised with respect to b. The function can be written as a sum of two terms

$$\ell_p(\beta, b) = \ell_{\text{pnc}}(\beta, b) - \frac{w_3}{\sqrt{\theta}} \left(\sum_{k=1}^{q} |b_k| \right), \tag{12}$$

where

1. the term $\ell_{\text{pnc}}(\beta, b)$ corresponds to all except the last term on the right hand side of Equation (10) or the log partial likelihood of the standard Cox PH model with the normal and Cauchy penalty parts only, and
2. the second term corresponds to the Laplace penalty part in Equation (10). It is less well behaved: concave and continuous, but only differentiable at $b_k \neq 0, k = 1, \ldots, q$.

It is important to note that, in the proposed (extended) Cox PH model, $\ell_{\text{pnc}}(\beta, b)$ is a regular function, i.e. it is concave and at least twice differentiable everywhere. With this in mind, the overall estimation algorithm will employ both the Newton-Raphson algorithm and the gradient ascent algorithm. The Newton-Raphson algorithm is employed to maximise $\ell_{\text{pnc}}(\beta, b)$, while the gradient ascent algorithm will 'encapsulate' the overall estimation to estimate $\ell_p(\beta, b)$ with the lasso penalty. The details of the estimation method are presented in the Supplementary Material.

3.5. Estimation of the variance of the random effects

The inverse of the negative Hessian matrix H can be used as an approximate covariance matrix of $\widehat{b}$. However, we consider the formulation by Gray [30] where the estimate of covariance matrix of b is taken as $H^{-1}I_{PL}H^{-1}$, where I_{PL} is the standard Cox PH model information matrix. From a simulation with bootstrap (not shown), we find that the latter formulation is more accurate.

3.6. Estimating ψ via cross validation

To estimate $\psi \equiv \{\theta, w_1, w_2\}$, we consider five-fold cross-validation (CV) by identifying ψ that maximises CV log partial likelihood as described below [31]. Let E_d and V_d respectively denote the estimation and validation subsets of observations for the d-th CV fold, $d = 1, \ldots, 5$. We calculate the cross validation partial log likelihood as

$$\ell_p^{\text{cv}}(\psi) = \sum_{d=1}^{5} \sum_{i \in V_d} \left[\delta_i (X_i \widehat{\beta}_\psi^{E_d} + Z_i \widehat{b}_\psi^{E_d}) - \delta_i \log \left(\sum_{j \in R(t_i)} \exp(X_j \widehat{\beta}_\psi^{E_d} + Z_j \widehat{b}_\psi^{E_d}) \right) \right], \tag{13}$$

where, in each fold, the parameter estimates $\widehat{\beta}_\psi^{E_d}$ and $\widehat{b}_\psi^{E_d}$ are obtained from fitting the model in the d-th fold estimation set, and $\ell_p^{\text{cv}}(\psi)$ are calculated from X_i's and Z_i's in the d-th fold validation set using $\widehat{\beta}_\psi^{E_d}$ and $\widehat{b}_\psi^{E_d}$ obtained from its corresponding estimation set.

In our application, we consider values of ψ which always include the ridge ($w_1 = 1$), lasso ($w_3 = 1$), and elastic net ($w_2 = 0$) models because they are special cases of the proposed model. This means that, in our analysis and simulation study (Section 4 below), we easily compare the proposed model to those models.

3.7. Estimation of survivor function and model diagnostic

One of the main interests in the modelling using the Cox PH model is the estimation of survivor function for an individual with clinical characteristics x^* and CNA profile z^*. We first estimate the baseline hazard function $h_0(t)$ after

we obtain the model parameter estimates $\widehat{\beta}$ and $\widehat{b}$. We consider an extension of the Breslow's estimator [32] to include the random predictors and effects as

$$\hat{h}_0(t_i) = \frac{1}{\sum_{j \in R(t_i)} \exp(X_j \widehat{\beta} + Z_j \widehat{b})}.$$

The cumulative hazard function $H_0(t)$ can then be estimated as

$$\widehat{H}_0(t) = \sum_{t_i \leq t} \frac{1}{\sum_{j \in R(t_i)} \exp(X_j \widehat{\beta} + Z_j \widehat{b})},$$

and the baseline survivor function as $\widehat{S}_0(t) = \exp\{-\widehat{H}_0(t)\}$. The predicted survivor function for a new individual with a known clinical characteristics x^* (a p-vector) and CNA profile z^* (a q-vector) is then given by $\widehat{S}(t; x^*, z^*) = \widehat{S}_0(t)^{\exp\{x^* \widehat{\beta} + z^* \widehat{b}\}}$.

For model diagnostic, we consider the Cox-Snell residuals [33]. Further details are presented in the Supplementary Material.

4. Simulation study

To understand the proposed model's performance, we carry out a simulation study. As mentioned in Section 3.6, we automatically compare the proposed model with Cox PH models on ridge, lasso, and elastic-net penalty. In estimating $\psi \equiv \{w_1, w_2, \theta\}$, we include the values of w_1, w_2, and $w_3 = (1 - w_1 - w_2)$ that correspond to those models and, in all cases, the optimal values for w_1, w_2 and w_3 always correspond to the general proposed model. Therefore, we compare the proposed model to the sparse partial least squares (SPLS) model as described in [34] to model survival data. Lee *et al.* [34] consider SPLS with L_1 lasso penalty (SPLS-L1) and hierarchical likelihood penalty (SPLS-HL). In our simulation study, all comparisons are based on 100 simulation datasets. In each dataset, we do not simulate nor include the matrix of clinical characteristics X, so that we work only with the matrix of random predictors Z of size $n \times q$ ($n = 100, q = 200$).

4.1. Simulated data

We follow the simulation setting proposed in [35], [36], and [34]. First, the matrix of random predictors Z is generated from a multivariate normal distribution with zero mean vector and covariance matrix Σ of size 200×200. We assume Σ to be a block diagonal matrix, with the diagonal blocks to be Σ_c of size 20×20 for $c = 1, \ldots, 10$. The (sub-) matrix Σ_c has diagonal elements $\sigma_c^2 = 1$ and off-diagonal elements $\rho\sigma_c^2$ for all $c = 1, \ldots, 10$. In this simulation, we set $\rho = 0.9$ because it is a good reflection of the real lung cancer data and Lee *et al.* [34] argue that their method works better with higher correlation. The latter suggests that we compare the proposed method to their method at its expected optimal working condition.

For $k = 1, \ldots, 20$, we set $b_k = \exp(-\alpha(j-1))$. For $k = 21, \ldots, 40$, we set $b_k = -b_{k-20}$ to get the same pattern on the opposite sign. For $k = 41, \ldots, 200$, we set $b_k = 0$. Here we have used an exponential decay with parameter α, which we take as $\alpha = 0.0141$. We generate survival time T_i, $i = 1, \ldots, n$, from a Weibull distribution with baseline hazard rate $h_0(t) = 5t^4$ and the censoring time C_i assumed to follow Uniform$(0, 3)$ distribution, which gives censoring rate of approximately 35%.

4.2. Simulation analysis

For each data set, we employ three methods: our proposed sparse-smoothed Cox PH model, sparse PLS with lasso (L_1) penalty (SPLS-L1) and hierarchical likelihood penalty (SPLS-HL) as described in [34]. We evaluate the three methods in the aspects of variable selection and prediction in a single cross validation as follows.

1. In terms of the variable selection, we calculate the average sensitivity and specificity [37]. Sensitivity is defined as the proportion of those parameters that are truly non-zero and estimated as non-zero. Specificity, on the other hand, is defined as the proportion of parameters that are truly zero and estimated as zero.

2. In terms of prediction power, we compute a cross-validation [35]

$$-2\text{PL} = -2\ell_p^{\text{cv}} \tag{14}$$

where ℓ_p^{cv} is the cross-validation log partial likelihood in Equation (13) with the relevant tuning parameters optimised (using five-fold cross validation). For the purpose of the comparison between our proposed method with SPLS-L1 and SPLS-HL methods, we calculate the difference between -2PL of each method and -2PL of the true model (where $\widehat{b}$ is taken at the true value).

3. We also consider different measures of prediction performance as suggested by [34]. In the cross-validation, we actually know the true failure time in the validation set (T_i^V). On the other hand, we can predict median survival time in the validation set ${}_m\widehat{T}_i^V$ as

$${}_m\widehat{T}_i^V = \widehat{H}_0^{-1}\big[-\log(1/2)\exp\big(-Z_i^V\widehat{b}^E\big)\big],$$

where $\widehat{H}_0(t)$ is the estimated cumulative baseline hazard function for the validation set using $\widehat{b}^E$ from the estimation set.

Using T_i^V and ${}_m\widehat{T}_i^V$, we compute the sum of squared prediction error (SSPE)

$$\text{SSPE} = \sum_{i\in V}\big(T_i^V - {}_m\widehat{T}_i^V\big)^2$$

and the sum of absolute prediction error (SAPE)

$$\text{SAPE} = \sum_{i\in V}|T_i^V - {}_m\widehat{T}_i^V|$$

as indicators of prediction performance.

The prediction performance is considered good when -2PL, SSPE, and SAPE are relatively small.

4.3. Simulation results

The results of the simulation study are summarised in Figure 4 and Table 2. Figure 4 indicates that the sparse-smoothed Cox PH model (SSCox) has similar

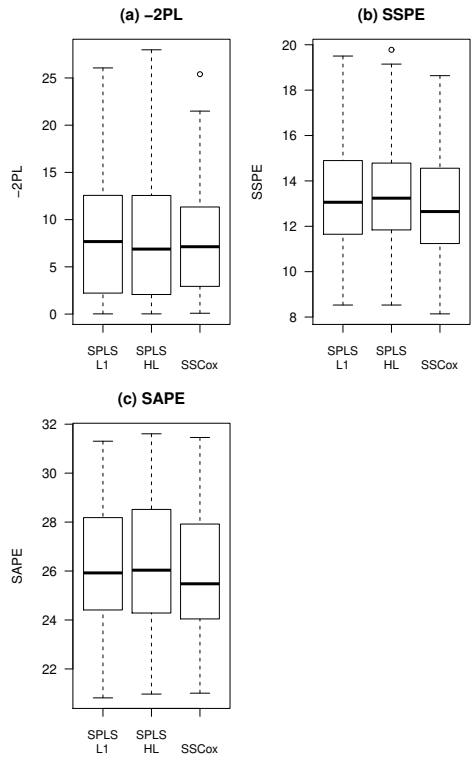

Figure 4: *Results from the simulation study:* (a) *boxplots of the absolute difference of* $-2PL$ *in Equation (14) between the true model and each of SPLS-L1, SPLS-HL, and the proposed sparse-smoothed Cox PH model,* (b) *boxplots of sum of squared prediction error (SSPE), and* (c) *boxplots of sum of absolute prediction error (SAPE). Lower values on all of boxplots indicates better prediction performance. In the simulation, we automatically compare our results with the Cox PH models with normal (ridge), Laplace (lasso), and elastic net penalties. Using paired t-test, we can conclude that the SSCox model has similar* $-2PL$ *as, but lower SSPE and SAPE than, the SPLS-L1 and SPLS-HL models. More than 99% of simulated times are between zero and one, hence higher median in SAPE than SSPE.*

Table 2: *Sensitivity and specificity for the variable selection in the simulation study, between sparse PLS with L1 and HL penalty (SPLS-L1 and SPLS-HL, repsectively) and the proposed sparse-smoothed Cox PH model (SSCox). Sensitivity is defined as the proportion of those parameters that are truly non-zero and estimated as non-zero. Specificity, on the other hand, is defined as the proportion of parameters that are truly zero and estimated as zero.*

Method	Sensitivity	Specificity
SPLS-L1	0.998	0.273
SPLS-HL	0.985	0.575
SSCox	0.978	0.592

$-2PL$ as, but lower SSPE and SAPE than, the sparse PLS models (SPLS-L1 and SPLS-HL). For SSPE, the means of 100 simulation are $13.28, 13.42,$ and 12.73 for SPLS-L1, SPLS-HL and SSCox respectively, while for the SAPE the means are $26.09, 26.21,$ and 25.66, respectively.

In term of sensitivity and specificity, the ordinary PLS and Cox PH model with ridge penalty always have zero specificity because they do not produce a sparse solution. However, this is not what we would expect from all of the methods in our comparison. We can infer from Table 2 that all methods have relatively high sensitivity and the proposed SSCox method has the highest specificity. The proposed model SSCox has a slightly lower sensitivity compared to the sparse PLS methods. This is really an artefact because the smoothness imposed in the SSCox method would force the estimates around the change of sign of the true parameter (such as b_{20} to b_{21}) to have smooth transition. This smooth transition sometimes requires the estimate to be zero just before changing sign for the next window.

It is important to note that in our simulation, we automatically consider the Cox PH model with normal (ridge) penalty, Laplace (lasso) penalty, and the elastic-net penalty. In all of the simulated data, the proposed sparse-smoothed Cox PH model is selected, i.e. has the highest cross-validation log partial likelihood across all values of θ evaluated.

5. Results for real data

5.1. Estimating ψ

An important parameter to be estimated in the proposed Cox PH model (SSCox) is the 'tuning' parameter $\psi = (\theta, w_1, w_2)$. These parameters are important in the interpretation. Firstly, when θ goes to zero (in limit terms), the estimates of the random effects will be zero and no information in CNA are taken into account in the model. Secondly, w_3 controls the sparseness imposed on the random effects estimates, while w_2 controls their smoothness. To estimate ψ, one could solve over a 3-dimensional grid of θ, w_1, and w_2. However, we find this to be computationally impractical and does a poor job of model selection, as pointed out by Simon *et al.*[38]. Instead, we fix the mixing parameters (w's) and, for each combination of w's, compute the cross-validation log partial likelihood for a path of θ values. We begin the path with θ sufficiently small to set $b = 0$ (and call this $\theta_{\min}$) and increase θ until we are near the unregularised solution.

Figure 5 shows the cross-validation log partial likelihood $\ell_p^{\text{cv}}(\theta; w_1 = 0.3, w_2 = 0.2)$, with their one standard-error bar obtained from the cross validation. Among several combinations of w's, the highest $\ell_p^{\text{cv}}(\theta)$ is obtained at $w_1 = 0.3, w_2 = 0.2$, and $w_3 = 0.5$ in the lung cancer data. To estimate θ, we select the maximum value θ that is within one standard-error from $\ell_p^{\text{cv}}(\theta_{\min}; w)$ [38]. Using this criterion, the figure suggests to select $\log(\theta) = -6.075$, which corresponds to $\theta = 0.0023$.

5.2. Model fit: Fixed effects estimates

Using the optimal θ, the estimates of the fixed effects and their inference can be seen in Table 3, where Stage T1 and Stage N0 are part of the baseline. The table indicates that Age, Stage-T, and Stage-N are statistically significant at the 5% significance level. The estimates indicate that the hazard ratio increases by about six percent ($e^{0.055} \approx 1.06$) as age-at-operation increases by one year. The estimate of Stage-T3 indicates that large tumour size increases the hazard

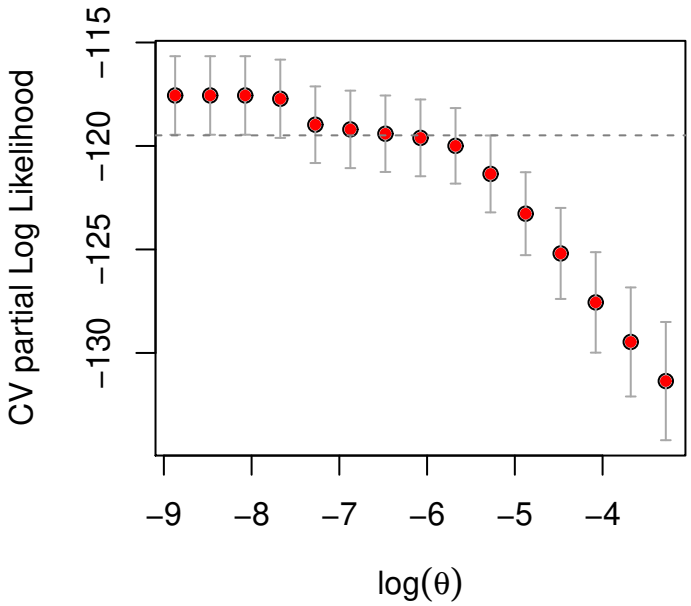

Figure 5: *Cross-validation log partial likelihood* $\ell_p^{cv}(\theta; w)$ *for different* θ *using* $w_1 = 0.3$ *(normal penalty part),* $w_2 = 0.2$ *(Cauchy penalty part), and* $w_3 = 0.5$ *(Laplace penalty part) with their one standard-error. The horizontal dashed line indicates one standard error from* $\ell_p^{cv}(\theta_{min}; w)$.

Table 3: *Summary of fixed effects estimates. Stage-T1 and Stage-N0 are part of the baseline.*

Predictor	Estimate	Exp	Std.Error	z-value	p-value
Age	0.0551	1.06	0.0164	3.37	0.0008
StageT2	0.1817	1.20	0.3215	0.57	0.5679
StageT3	1.7622	5.83	0.6392	2.76	0.0058
StageN1	0.3616	1.43	0.3019	1.20	0.2301
StageN2	1.3653	3.92	0.4824	2.83	0.0047

ratio by almost six times from the baseline. Similarly, the estimate of Stage-N2 indicates that a wider spread of cancer cells to far-away lymph nodes increases the hazard almost four times relative to the baseline.

5.3. Model fit: Random effects estimates

The random effect estimates $\widehat{b}$ of the proposed sparse-smoothed Cox PH model are presented in Figure 6. The figure shows the estimates are sparse: almost all of the estimates of b are zero and only some regions in the genome that have non-zero estimates. The non-zero estimates are grouped within bigger regions, and exhibit smooth transition across neighbouring windows within them. A more detailed view of the random effects estimates in each chromosome is presented in the Supplementary Material.

The pattern of the estimates enables us to investigate the genome in relation to the patients' survival. Genomic regions with positive estimates are associated with genes that are involved in the progression of lung cancer (poor prognosis) while negative estimates with those that are protective. In Figure 6, more regions have positive estimates than negative estimates, which indicate that most identified genes are those that increase the risk. For example, Flacco *et al.*[39], Antoniou *et al.*[40], and Pelosi *et al.*[41] show that *TERC* copy number gain in chromosome 3 is associated with early-stage non-small lung cancer. Moreover, there are some more studies that show the involvement of chromosome 7 in non-small lung cancer [42, 43, 44].

An interesting finding is that some regions in chromosomes 8 and 12 have

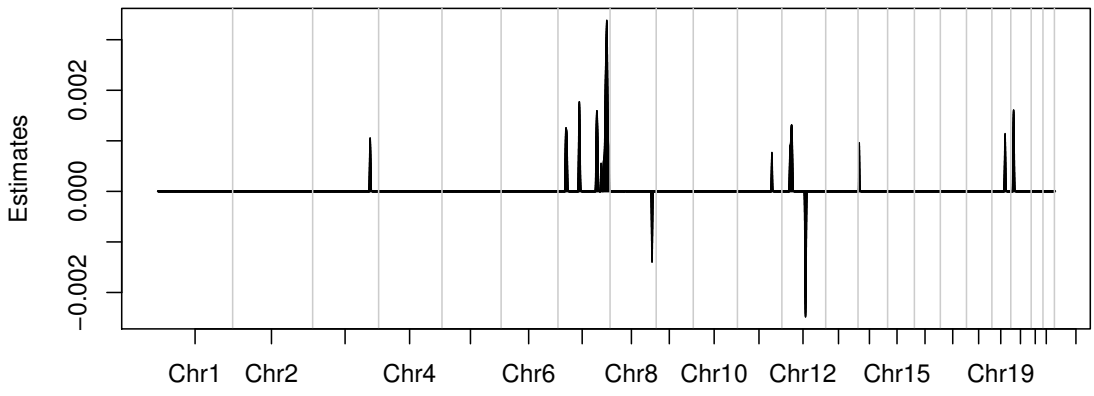

Figure 6: *Random effects estimates* $\widehat{b}$ *in the full model, using CNA profiles. Genomic windows with missing values (for example in the centromere regions) were excluded from analysis. Vertical grey lines separate the chromosomes. A more detailed view of the random effects estimates in each chromosome is presented in the Supplementary Material.*

negative random effects estimates. We expect that there are some tumour repressing genes in those regions. We found three genes in chromosome 8 that fall in this category, although not in lung cancer specifically. A study by Schemionek *et al.*[45] reported that *MTSS* gene reduces tumour growth in leukemia, while a study by Yue *et al.*[46] shows that the gene *ZHX2* is involved in the reduction of proliferation of liver cancer. The gene *NDUFB9* was shown in a study by Li *et al.*[47] to be suppressor of breast cancer cell proliferation, migration and invasion. Several other studies also show some genes in chromosome 12 with negative random effects that are involved in tumour suppression [48, 49].

6. Discussion

We have investigated the extension of Cox PH model to accommodate and investigate genomic CNA profiles. Two key parameters in the model are θ, which controls the amount of information in CNA profiles used in the model fitting, and w's, which control the degree of sparseness and smoothness of the parameter estimates. In the estimation of those parameters, we always include the ridge ($w_1 = 1$), lasso ($w_3 = 1$), and elastic net ($w_2 = 0$) models because they are special cases of the proposed model. We find, both in the real data and simulated data, that the proposed model with $w_1 > 0$, $w_2 > 0$, and $w_3 > 0$ are always preferred than those special cases. This indicates that our distributional assumption on the random effects are more suitable for genomic profiles, than the simpler models.

The proposed formulation of the extended Cox PH model enables us to interpret the parameter estimates sensibly. Some genes can be identified in genomic regions with non-zero estimates and their relevance can be confirmed from past studies. This does not mean that there is no other genes in different regions that are related to patients' prognosis. Our results simply indicate that, with limited number of patients to estimate thousands of parameters, the model enables this limited information to be 'channeled' to identify associated genomic regions that are best supported by the data. Within the identified regions, there

are some other genes that have not been determined whether they are related to prognosis or not, because there have been no previous study that neither confirms nor refutes their role.

It is worth noting that the proposed method relies on the Laplace approximation. In the context of a generalized linear mixed model (GLMM), this method is called penalized quasi-likelihood (PQL) [50] and is the most commonly used method due to its convenient computation [51]. Breslow and Lin [52] show that the PQL estimators of regression parameters and variance component are subject to bias when applied to correlated-binary or count data, which is not our case. Their numerical studies suggest that the biases are minimal for $0 \leq \theta \leq 0.25$, in which the estimate of θ in our study falls. Therefore, the results in this study are expected to not suffer from the bias of PQL.

Sutradhar [53] proposes a generalized quasi-likelihood (GQL) approach that produces consistent as well as more efficient estimates as shown in [54]. However, the GQL does not require any estimates for the random effects b, which are our main interest. Some other alternative estimation methods involve numerical techniques, including a Bayesian approach [55], MCEM algorithm [56], Gauss-Hermite quadrature (GHO) [57], and Quasi-Monte Carlo (QMC) [58]. One common drawback of this approach is that those methods are time consuming [59]. In future research, one could consider to penalize the variance of the random effects θ instead of penalizing the random effects b. This idea was discussed in [51] for GLMM and it can be generalized to survival analysis.

It is important to note that we do not lose information in discriminating patients at different risk scores by extending the assumption of the random effects distribution. In our real lung cancer dataset, the median survival times of patients in the 10th, 50th, and 90th percentile of risk scores remain consistent, regardless of whether the distribution of random effects is as proposed, or any of its special cases. This indicates that the proposed model's ability to identify genomic regions associated with the survival is mainly due to the model's formulation.

Lastly, the proposed methodology can be implemented in other omics data

where the genomewide information contain spatial dependencies. For example, we speculate that the methodology may be fully implemented in the case genetic data (based on single nucleotide polymorphisms) or epigenetic data (DNA methylation), whenever survival analysis is needed. In a case where spatial dependencies are not substantial, such as gene expression data [60], the methodology can still be implemented. However, its novel advantage may not be able to be fully seen. These are currently part of our current active research.

7. Conclusion

In addition to the clinical phenotypes, copy number alterations are informative for predicting cancer patients' survival. This study identifies important genomic regions that additionally associated with cancer patients' survival. This is done by addressing the characteristics of copy number alteration data that are highly correlated. The extended Cox PH model we proposed enables simultaneous prediction and identification of important genomic regions, while taking into account the dependencies between the genomic regions.

Acknowledgements

Funding: The first author (KA) was supported by the Saudi Arabian Ministry of Education, Prince Sattam Bin Abdulaziz University, Saudi Arabia.

References

[1] R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, et al., Global variation in copy number in the human genome, Nature 444 (7118) (2006) 444–454.

[2] J. L. Freeman, G. H. Perry, L. Feuk, R. Redon, S. A. McCarroll, D. M. Altshuler, H. Aburatani, K. W. Jones, C. Tyler-Smith, M. E. Hurles, et al., Copy number variation: new insights in genome diversity, Genome Research 16 (8) (2006) 949–961.

[3] A. B. Olshen, E. S. Venkatraman, R. Lucito, M. Wigler, Circular binary segmentation for the analysis of array-based DNA copy number data, Biostatistics 5 (4) (2004) 557–572. arXiv:http://biostatistics.oxfordjournals.org/content/5/4/5 doi:10.1093/biostatistics/kxh008.
URL http://biostatistics.oxfordjournals.org/content/5/4/557.abstract

[4] J. Huang, A. Gusnanto, K. O'Sullivan, J. Staaf, Å. Borg, Y. Pawitan, Robust smooth segmentation approach for array CGH data analysis, Bioinformatics 23 (18) (2007) 2463–2469.

[5] A. J. Holland, D. W. Cleveland, Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis, Nature reviews Molecular cell biology 10 (7) (2009) 478–487.

[6] D. R. Cox, Regression models and life tables, Journal of the Royal Statistical Society B 34 (2) (1972) 187–220.

[7] P. J. Verweij, H. C. Van Houwelingen, Penalized likelihood in Cox regression, Statistics in Medicine 13 (23-24) (1994) 2427–2436.

[8] R. Tibshirani, et al., The lasso method for variable selection in the Cox model, Statistics in Medicine 16 (4) (1997) 385–395.

[9] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society, Series B 67 (2005) 301–320.

[10] D. A. Engler, Y. Li, Survival analysis with large dimensional covariates: an application in microarray studies, Statistical Applications in Genetics and Molecular Biology 8 (1) (2009) 14.

[11] L. Waldron, M. Pintilie, M.-S. Tsao, F. A. Shepherd, C. Huttenhower, I. Jurisica, Optimized application of penalized regression methods to diverse genomic data, Bioinformatics 27 (24) (2011) 3399–3406.

[12] J. J. Goeman, L_1 penalized estimation in the Cox proportional hazards model, Biometrical Journal 52 (1) (2010) 70–84.

[13] O. Belvedere, S. Berri, R. Chalkley, C. Conway, F. Barbone, F. Pisa, K. MacLennan, C. Daly, M. Alsop, J. Morgan, et al., A computational index derived from whole-genome copy number analysis is a novel tool for prognosis in early stage lung squamous cell carcinoma, Genomics 99 (1) (2012) 18–24.

[14] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics 25 (14) (2009) 1754–1760.

[15] A. Gusnanto, C. C. Taylor, I. Nafisah, H. M. Wood, P. Rabbitts, S. Berri, Estimating optimal window size for analysis of low-coverage next-generation sequence data, Bioinformatics 30 (2014) 1823–1829.

[16] A. Gusnanto, H. M. Wood, Y. Pawitan, P. Rabbitts, S. Berri, Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data, Bioinformatics 28 (1) (2012) 40–47.

[17] P. Micke, K. Edlund, L. Holmberg, H. G. Kultima, L. Mansouri, S. Ekman, M. Bergqvist, L. Scheibenflug, K. Lamberg, G. Myrdal, A. Berglund, A. Andersson, M. Lambe, F. Nyberg, A. Thomas, A. Isaksson, J. Botling, Gene copy number aberrations are associated with survival in histologic subgroups of non-small cell lung Journal of Thoracic Oncology 6 (11) (2011) 1833 – 1840. doi:https://doi.org/10.1097/JTO.0b013e3182295917.
URL http://www.sciencedirect.com/science/article/pii/S1556086415322450

[18] A. M. Bowcock, Dna copy number changes as diagnostic tools for lung cancer, Thorax 69 (5) (2014) 496–497. arXiv:https://thorax.bmj.com/content/69/5/496.full.pdf, doi:10.1136/thoraxjnl-2013-204681.
URL https://thorax.bmj.com/content/69/5/496

[19] N. Zhang, M. Wang, P. Zhang, T. Huang, Classification of cancers based on copy number variation landscapes, Biochimica et Biophysica Acta (BBA) - General Subjects

1860 (11, Part B) (2016) 2750–2755, systems Genetics - Deciphering the Complex Disease with a Systems Approach. doi:https://doi.org/10.1016/j.bbagen.2016.06.003.
URL https://www.sciencedirect.com/science/article/pii/S0304416516302082

[20] J. Li, Q. Xu, M. Wu, T. Huang, Y. Wang, Pan-cancer classification based on self-normalizing neural networks and feature selection, Frontiers in Bioengineering and Biotechnology 8 (2020) 766. doi:10.3389/fbioe.2020.00766.
URL https://www.frontiersin.org/article/10.3389/fbioe.2020.00766

[21] A. Gusnanto, P. Tcherveniakov, F. Shuweihdi, M. Samman, P. Rabbitts, H. M. Wood, Stratifying tumour subtypes based on copy number alteration profiles using next-generation sequence dat Bioinformatics 31 (16) (2015) 2713–2720. arXiv:https://academic.oup.com/bioinformatics/article-pdf/31/16/2713/17084814/btv191.pdf doi:10.1093/bioinformatics/btv191.
URL https://doi.org/10.1093/bioinformatics/btv191

[22] X. Pan, X. Hu, Y.-H. Zhang, L. Chen, L. Zhu, S. Wan, T. Huang, Y.-D. Cai, Identification of the copy number variant biomarkers for breast cancer subtypes, Molecular Genetics and Genomics 294 (2019) 95–110.

[23] S. Zhang, X. Pan, T. Zeng, W. Guo, Z. Gan, Y.-H. Zhang, L. Chen, Y. Zhang, T. Huang, Y.-D. Cai, Copy number variation pattern for discriminating macrod2 states of colorectal cancer subtypes, Frontiers in Bioengineering and Biotechnology 7 (2019) 407. doi:10.3389/fbioe.2019.00407.
URL https://www.frontiersin.org/article/10.3389/fbioe.2019.00407

[24] D. Cox, D. Oakes, Analysis of Survival Data, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis, 1984.
URL https://books.google.co.uk/books?id=Y4pdM2soP4IC

[25] D. Kleinbaum, M. Klein, Survival Analysis: A Self-Learning Text, Third Edition, Statistics for Biology and Health, Springer New York, 2011.
URL https://books.google.co.uk/books?id=7Vg5wTSdwuMC

[26] S. Ripatti, J. Palmgren, Estimation of multivariate frailty models using penalized partial likelihood, Biometrics 56 (4) (2000) 1016–1022.

[27] T. Therneau, T. Grambsch, P. Grambsch, Modeling Survival Data: Extending the Cox Model, Statistics for Biology and Health, Springer, 2000.
URL https://books.google.co.uk/books?id=9kY4XRuUMUsC

[28] Y. Pawitan, In all likelihood: statistical modelling and inference using likelihood, OUP Oxford, 2013.

[29] W. Zhou, M. Yin, H. Cui, N. Wang, L. Zhao, L. Yuan, X. Yang, X. Ding, F. Men, X. Ma, et al., Identification of potential therapeutic target genes and mechanisms in non-small-cell lung carcinoma in non-smoking women based on bioinformatics analysis, European Review for Medical and Pharmacological Sciences 19 (2015) 3375–3384.

[30] R. J. Gray, Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis, Journal of the American Statistical Association 87 (420) (1992) 942–951.

[31] P. J. Verweij, H. C. Van Houwelingen, Cross-validation in survival analysis, Statistics in Medicine 12 (24) (1993) 2305–2314.

[32] N. Breslow, Covariance analysis of censored survival data., Biometrics 30 (1) (1974) 89–99.

[33] D. Cox, E. J. Snell, A general definition of residuals, Journal of the Royal Statistical Society. Series B (Methodological) 30 (2) (1968) 248–275.
URL http://www.jstor.org/stable/2984505

[34] D. Lee, Y. Lee, Y. Pawitan, W. Lee, Sparse partial least-squares regression for high-throughput survival data analysis, Statistics in Medicine 32 (30) (2013) 5340–5352.

[35] S. Nygård, Ø. Borgan, O. C. Lingjærde, H. L. Størvold, Partial least squares Cox regression for genome-wide data, Lifetime Data Analysis 14 (2) (2008) 179–195.

[36] H. M. Bøvelstad, S. Nygård, H. L. Størvold, M. Aldrin, Ø. Borgan, A. Frigessi, O. C. Lingjærde, Predicting survival from microarray data comparative study, Bioinformatics 23 (16) (2007) 2080–2087.

[37] H. Chun, S. Keleş, Sparse partial least squares regression for simultaneous dimension reduction and variable selection, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72 (1) (2010) 3–25.

[38] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, et al., Regularization paths for Cox's proportional hazards model via coordinate descent, Journal of Statistical Software 39 (5) (2011) 1–13.

[39] A. Flacco, V. Ludovini, F. Bianconi, M. Ragusa, G. Bellezza, F. R. Tofanetti, L. Pistola, A. Siggillino, J. Vannucci, L. Cagini, et al., MYC and Human Telomerase Gene (TERC) copy number gain in early-stage non–small cell lung cancer, American Journal of Clinical Oncology 38 (2) (2015) 152–158.

[40] K. Antoniou, K. Samara, I. Lasithiotaki, G. Margaritopoulos, G. Soufla, I. Lambiri, I. Giannarakis, I. Drositis, D. Spandidos, N. Siafakas, Differential telomerase expression in idiopathic pulmonary fibrosis and non-small cell lung cancer, Oncology reports 30 (6) (2013) 2617–2624.

[41] G. Pelosi, B. Del Curto, M. Trubia, A. G. Nicholson, M. Manzotti, G. Veronesi, L. Spaggiari, P. Maisonneuve, F. Pasini, A. Terzi, et al., 3q26 amplification and polysomy of chromosome 3 in squamous cell lesions of the

lung: a fluorescence in situ hybridization study, Clinical Cancer Research 13 (7) (2007) 1995–2004.

[42] L. E. Buckingham, J. S. Coon, L. E. Morrison, K. K. Jacobson, S. S. Jewell, K. A. Kaiser, A. M. Mauer, T. Muzzafar, C. Polowy, S. Basu, et al., The prognostic value of chromosome 7 polysomy in non-small cell lung cancer patients treated with gefitinib, Journal of Thoracic Oncology 2 (5) (2007) 414–422.

[43] K. Kitada, T. Yamasaki, The complicated copy number alterations in chromosome 7 of a lung cancer cell line is explained by a model based on repeated breakage-fusion-bridge cycles, Cancer Genetics and Cytogenetics 185 (1) (2008) 11–19.

[44] J. S. Lee, S. Pathak, V. Hopwood, B. Tomasovic, T. D. Mullins, F. L. Baker, G. Spitzer, J. A. Neidhart, Involvement of chromosome 7 in primary lung tumor and nonmalignant normal lung tissue, Cancer Research 47 (23) (1987) 6349–6352.

[45] M. Schemionek, O. Herrmann, M. M. Reher, N. Chatain, C. Schubert, I. Costa, S. Haenzelmann, E. Gusmao, S. Kintsler, T. Braunschweig, et al., MTSS1 is a critical epigenetically regulated tumor suppressor in CML, Leukemia 30 (4) (2016) 823–832.

[46] X. Yue, Z. Zhang, X. Liang, L. Gao, X. Zhang, D. Zhao, X. Liu, H. Ma, M. Guo, B. T. Spear, et al., Zinc fingers and homeoboxes 2 inhibits hepatocellular carcinoma cell proliferation and represses expression of Cyclins A and E, Gastroenterology 142 (7) (2012) 1559–1570.

[47] L.-D. Li, H.-F. Sun, X.-X. Liu, S.-P. Gao, H.-L. Jiang, X. Hu, W. Jin, Down-regulation of NDUFB9 promotes breast cancer cell proliferation, metastasis by mediating mitochondrial metabolism, PloS One 10 (12) (2015) e0144441.

[48] C. Wu, S. Qiu, L. Lu, J. Zou, W.-f. Li, O. Wang, H. Zhao, H. Wang, J. Tang, L. Chen, et al., RSPO2–LGR5 signaling has tumour-suppressive activity in colorectal cancer, Nature Communications 5.

[49] W. Lee, A. Belkhiri, A. C. Lockhart, N. Merchant, H. Glaeser, E. I. Harris, M. K. Washington, E. M. Brunt, A. Zaika, R. B. Kim, et al., Overexpression of OATP1B3 confers apoptotic resistance in colon cancer, Cancer Research 68 (24) (2008) 10315–10323.

[50] N. E. Breslow, D. G. Clayton, Approximate inference in generalized linear mixed models, Journal of the American statistical Association 88 (421) (1993) 9–25.

[51] J. Pan, C. Huang, Random effects selection in generalized linear mixed models via shrinkage penalty function, Statistics and Computing 24 (5) (2014) 725–738.

[52] N. E. Breslow, X. Lin, Bias correction in generalised linear mixed models with a single component of dispersion, Biometrika (1995) 81–91.

[53] B. C. Sutradhar, On exact quasilikelihood inference in generalized linear mixed models, Sankhyā: The Indian Journal of Statistics (2004) 263–291.

[54] M. R. I. Chowdhury, B. Sutradhar, Generalized quasi-likelihood versus hierarchical likelihood inferences in generalized linear mixed models for count data, Sankhyā: The Indian Journal of Statistics, Series B (2008-) (2009) 55–78.

[55] M. R. Karim, S. L. Zeger, Generalized linear models with random effects; salamander mating revisited, Biometrics (1992) 631–644.

[56] J. G. Booth, J. P. Hobert, Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61 (1) (1999) 265–285.

[57] J. Pan, R. Thompson, Gauss-hermite quadrature approximation for estimation in generalised linear mixed models, Computational Statistics 18 (1) (2003) 57–78.

[58] J. Pan, R. Thompson, Quasi-monte carlo estimation in generalized linear mixed models, Computational Statistics & Data Analysis 51 (12) (2007) 5765–5775.

[59] P. Newcombe, H. R. Ali, F. Blows, E. Provenzano, P. Pharoah, C. Caldas, S. Richardson, Weibull regression with bayesian variable selection to identify prognostic tumour markers of breast cancer survival, Statistical methods in medical research (2014) 0962280214548748.

[60] Y. Pawitan, J. Bjöhle, S. Wedren, K. Humphreys, L. Skoog, F. Huang, L. Amler, P. Shaw, P. Hall, J. Bergh, Gene expression profiling for prognosis using Cox regression, Statistics in Medicine 23 (11) (2004) 1767–1780.