Cox Regression with Correlation based Regularization for Electronic Health Records

Bhanukiran Vinzamuri and Chandan K. Reddy
Department of Computer Science
Wayne State University, Detroit, Michigan USA
bhanukiranv@wayne.edu, reddy@cs.wayne.edu

***Abstract*—Survival Regression models play a vital role in analyzing time-to-event data in many practical applications ranging from engineering to economics to healthcare. These models are ideal for prediction in complex data problems where the response is a time-to-event variable. An event is defined as the occurrence of a specific event of interest such as a chronic health condition. Cox regression is one of the most popular survival regression model used in such applications. However, these models have the tendency to overfit the data which is not desirable for healthcare applications because it limits their generalization to other hospital scenarios. In this paper, we address these challenges for the cox regression model. We combine two unique correlation based regularizers with cox regression to handle correlated and grouped features which are commonly seen in many practical problems. The proposed optimization problems are solved efficiently using cyclic coordinate descent and Alternate Direction Method of Multipliers algorithms. We conduct experimental analysis on the performance of these algorithms over several synthetic datasets and electronic health records (EHR) data about heart failure diagnosed patients from a hospital. We demonstrate through our experiments that these regularizers effectively enhance the ability of cox regression to handle correlated features. In addition, we extensively compare our results with other regularized linear and logistic regression algorithms. We validate the goodness of the features selected by these regularized cox regression models using the biomedical literature and different feature selection algorithms.**

***Keywords*-cox regression; regularization; feature selection; healthcare**

I. Introduction

Survival regression analysis [13], [11] on time-to-event data is an important component in analyzing complex datasets in many practical applications ranging from engineering to economics to healthcare. Time-to-event data analysis deals with prediction of the time to a particular event directly. An event here is defined as the occurrence of a specific interest point. We consider a real-world healthcare application problem to demonstrate the meaning of the time-to-event data. Large number of patients get readmitted at different time points for chronic conditions such as heart failure. It is critical to build robust regression methods that can predict the probability of readmission for a particular patient. Traditional linear and logistic regression models have a limited purview in this domain. These models require certain constraints to be satisfied before they can be applied on such time-to-event clinical data.

We explain the shortcomings of the linear and logistic regression models using the example of readmission risk prediction for heart failure. For this problem, these models generally require a risk stratification scheme on the class label consisting of predefined categories such as low risk, intermediate risk and high risk. However, it is observed that one cannot generalize this stratification scheme to all hospitals. Another inherent flaw of these traditional models is that they do not provide any insight on how different are the patients from intermediate risk category to low risk and high risk categories. Survival regression models can inherently overcome such limitations.

Survival models are more effective than linear and logistic regression models as they directly model the probability of occurrence of an event for each patient in contrast to assigning a nominal label to the patient. They are more adept at handling the *non symmetry* in time to failure data for different patients than linear or logistic learners. The predictions obtained by using these models provide the healthcare physician with a more thorough understanding on the expected probability of readmission for a given patient.

Cox regression [4] is one of the most popularly used survival regression models. The unique formulation of cox regression and it's proportional hazards assumption makes it ideal for many practical applications. However, it is observed that cox regression models tend to overfit the data; thus, limiting their generalization ability to future unseen data. Regularization [3], [15] is a method used effectively in statistics and machine learning to improve the generalization ability of a learner.

Regularized cox regression models such as the lasso cox regression (LASSO-COX) and elastic net cox regression (EN-COX) have been studied in the literature[22], [20]. These cox regression models provide sparsity and good generalization ability. However, the LASSO-COX algorithm does not work effectively in the presence of correlation in the data, and EN-COX is only partially effective at handling structured sparsity in the data.

Structured sparsity in high dimensional data is difficult to capture using LASSO-COX and EN-COX. In order to mitigate these defects of regularized cox regression models; we propose a framework which combines cox regression with novel regularization functions to capture correlation and grouping of features effectively. The major

contributions of our work are as follows

1) Propose a Kernel elastic net regularized Cox regression (KEN-COX) algorithm which uses a novel kernel elastic net penalty term as a regularization factor. Our experimental results demonstrate that this method is more effective than the standard elastic net at handling correlated features. The novel pairwise feature similarity regularizer in this method is obtained using kernel functions.
2) Propose a Graph-based OSCAR (Octagonal Shrinkage and Clustering Algorithm for Regression) regularized Cox regression method (OSCAR-COX). This method is effective since it can capture structured sparsity. It exploits the graph structure of the features in the dataset to capture the grouping of the features.
3) Demonstrate the improved discriminative ability of KEN-COX and OSCAR-COX using standard evaluation metrics. We also compare their performance to the state-of-the-art linear and logistic regression models using several synthetic and healthcare datasets of varying diversity.
4) Demonstrate the non redundancy of the features selected by KEN-COX and OSCAR-COX in comparison to the state-of-the-art feature selection algorithms. In addition, we use the parsimonious models from KEN-COX and OSCAR-COX to identify important biomarkers for the heart failure readmission from EHR data. We validate the biomarkers identified using the clinical literature.

In this paper, we propose two algorithms which integrate novel regularizers in the cox regression framework to build a system which is effective and has good generalization ability. Experimental analysis conducted over real EHR data with more than 8,000 patients from a large hospital shows that our proposed algorithms can obtain models with lower mean squared error (MSE) values compared to the cox regression model and it's variants. Further experimental results obtained on synthetic datasets demonstrate the good generalization ability of the proposed algorithms in comparison to different linear and logistic regression algorithms.

This paper is organized as follows: In Section II, we introduce the basic survival regression framework, the concept of censoring in healthcare applications and discuss the cox regression model in detail. In Section III, we provide the details of the two algorithms proposed in this paper. We discuss the structure of the embedded optimization of both these approaches in detail. We show the experimental results in detail for the proposed algorithms in Section IV. In Section V, we discuss the related work on regularized cox regression models. In Section VI, we conclude our discussion and discuss the possible future directions.

II. Survival Regression Models

In this section, we explain the basic components of a survival regression model. We begin by explaining the patient readmission cycle briefly and then explaining the concepts associated with survival regression. We will then discuss these important concepts in a survival regression framework using the patient readmission cycle in a hospital.

The patient readmission cycle consists of the different stages a patient goes through from the initial admission to the next readmission. The different kinds of information obtained from the patient beginning from the admission to discharge includes demographics, comorbidities, medications, procedures and pharmacy claims. All these constitute an EHR for that particular hospitalization of the patient. We now describe the basic concepts, input and output components in the survival regression framework for the particular case of 30 day readmission [10] where a patient is readmitted within 30 days of discharge from a hospital.

A. Concepts in Survival regression

1) **Event of interest**: A patient is readmitted within 30 days.
2) **Time to event**: The time from entry into a study (date hospitalized) until the subject has experienced the event of interest (30 day readmission).
3) **Right Censoring**: A patient who does not experience the event of interest for the duration of the study is said to be right censored. The survival time for this patient is considered to be at least as long as the duration of the study. Another case of right censoring is when the patient drops out of the study before the end of the study and he does not experience the event of interest. In this paper, we use the term censoring instead of right censoring with a slight abuse of the original term.
4) **Survival function**: The survival function $s(t)$ gives the probability of surviving (or not experiencing the event of interest) until time t.
5) **Hazard function** :The hazard function $h(t)$ gives the potential that the event of interest will occur per time unit given that an individual has survived up to a given specified time t.
6) **Input:** Input data X, Censored Times C and Time to event T .
   Output: Survival function $s(t)$, Regression coefficient vector $\hat{\beta}$.

B. Importance of censoring in healthcare applications

Censoring is an important part of clinical data analysis as explained in the patient readmission cycle. Traditional regression methods such as linear and logistic regression cannot handle survival times which are typically positive numbers and hence they would need to be transformed in a way to apply these methods. Standard linear and logistic regression methods cannot handle censoring directly which

Table I
AN EXAMPLE TO DEMONSTRATE RIGHT CENSORING WITH C=12 DAYS

Patient ID	T	Event	δ	Interpretation
122	2	HF Readmission	1	Patient readmitted 2 days after discharge
61	12	End of Study	0	Patient not readmitted even 12 days after discharge
45	6	Drop from Study	0	Lost follow up of patient 6 days after discharge
21	4	HF Readmission	1	Patient readmitted 4 days after discharge

enforces the requirement of survival regression methods for EHR data.

We explain the application of right censoring through a simple example. We consider a simple EHR dataset in Table I consisting of 4 instances. In this example, the time is measured in days. The censored times C is set to 12 days for all the patients.

One can observe that instances with patient ID 122 and 21 are not censored and hence δ is set to 1 with the survival time equivalent to the time to event of interest (T). Instances with patient ID 61 and 45 are censored with δ set to 0. In this manner, censoring is applied on the instances in the dataset.

With a brief understanding of the survival regression framework we now introduce some notations that will help in comprehending the cox regression framework in Table II. Given a dataset X which consists of n data points. Let x_i denote the i^{th} feature vector. Let $T = \{t_1 < t_2 < t_3 < \ldots < t_k\}$ represent the set of sorted k unique time to failure values. δ_i represents the censoring status for the i^{th} patient. δ_i=1 represents a failure and δ_i=0 represents a censored instance.

Table II
NOTATIONS USED IN THIS PAPER

Name	Description
X	n x m matrix of feature vectors.
T	k x 1 vector of sorted unique failure times.
R_i	set of all patients j at risk at time t_i ($y_j > t_i$).
C_k	set of all times for which patient k is still at risk. ($t_i < y_k$)
d_i	number of patients readmitted within time t_i .
δ	n x 1 vector of censored status.
$\hat{\beta}$	m x 1 regression coefficient vector
W	n x n symmetric weight matrix
Z	n x 1 pseudo response vector

C. Cox Regression Model

The Cox regression model [4], is a survival regression model which provides useful and easy to interpret information regarding the relationship of the hazard function to the predictors. It is by far the most popular survival regression model and is implemented in a large number of statistical software packages.

The Cox regression model is a semi parametric method of estimation. This means that we specify a model for the effect of the covariates but we do not specify a model for the baseline hazard function. This implies that we can estimate the regression coefficient vector $\hat{\beta}$ without having any knowledge of the baseline hazard function. The regression coefficient vector can then be used to estimate the baseline hazard and survival functions.

Notation: In the equations used throughout this paper we will be using the following notations. A represents a matrix and $A(i,:)$ represents the i^{th} row vector in the matrix. $A(:,j)$ represents the j^{th} column vector of the matrix. $A(i,j)$ represents the i, j^{th} element of the matrix.

In Equation (1), we provide the formulation for the estimation of $\hat{\beta}$ and Equation (4) provides the formulae for computing the survival function and hazard function in cox regression. Equation (1) provides the steps associated in estimating $\hat{\beta}$ in a weighted least squares framework. A more detailed derivation of the formulae for the weight matrix W and the pseudo response vector Z which are provided here [20].

$$\hat{\beta} = \min_{\beta} \frac{1}{n} \sum_{k=1}^{n} (W(k,k)((Z(k,:) - X(k,:)\beta)^2) \quad (1)$$

$$W(k,k) = \sum_{i \in C_k} [\frac{e^{\tilde{\eta}(k,:)} \sum_{j \in R_i} e^{\tilde{\eta}(j,:)} - (e^{\tilde{\eta}(k,:)})^2}{(\sum_{j \in R_i} e^{\tilde{\eta}(j,:)})^2}] \quad (2)$$

$$Z(k,:) = \tilde{\eta}(k,:) + \frac{1}{W(k,k)}[\delta_k - \sum_{i \in C_k} (\frac{e^{\tilde{\eta}(k,:)}}{\sum_{j \in R_i} e^{\tilde{\eta}(j,:)}})] \quad (3)$$

Using the value of $\hat{\beta}$ obtained from Equation (1) we now compute the probability of survival for a given patient with feature vector x_i at time t. $\tilde{\eta}$ is set to $X\hat{\beta}$. To compute the survival probability, we compute the baseline hazard function $h_0(t)$ at a given time t. We then compute the survival probability $s(t|x_i)$ using the baseline hazard function and the values of i^{th} patient feature vector x_i.

$$\begin{aligned} h_0(t) &= \sum_{t_i \leq t} \frac{\delta_i}{\sum_{j \in R_i} exp(\hat{\beta}^T x_j)} \\ s_0(t) &= exp(-h_0(t)) \\ s(t|x_i) &= s_0(t)^{exp(\hat{\beta}^T x_i)} \end{aligned} \quad (4)$$

In Algorithm (1), we present the basic steps involved in the cox regression framework. We formulate cox regression as a weighted least squares problem.

Algorithm 1 Cox Regression

Require: Input X, Censored Times C, Failure Times T

1: **repeat**
2: Initialize $\tilde{\beta}$
3: Compute W and Z using Equations (2) and (3)
4: Solve for $\hat{\beta}$ in Equation (1) using a convex solver
5: $\tilde{\beta} = \hat{\beta}$
6: **until** Convergence of $\hat{\beta}$
7: Output $\hat{\beta}$, hazard and survival functions for each time t

III. Cox Regression with Correlation based Regularization

In this section, we describe the algorithms developed by combining two novel correlation regularizers with cox regression. We denote a regularizer as $P_\alpha(\beta)$ where β is the regression coefficient vector. Generally, most regularizers considered here are convex loss functions because of their desirable properties. The motivation for applying convex functions in the clinical domain arises from the success achieved by using convex non-smooth functions such as the L_1 and the $L_{2,1}$ norms for different biomedical applications [15], [3]. Their properties of sparsity and group sparsity have proven to be very effective for such applications. Our novel regularizers are functions which use the L_1, L_2, and L_∞ norms. Regularizers also need a parameter which governs their importance in the framework. In Equation (5), λ is called the regularization parameter. The quadratic optimization problem in Equation (5) can be solved using methods such as conjugate gradient descent and cyclic coordinate descent.

$$\begin{aligned} L(\beta) &= \frac{1}{n}\sum_{k=1}^{n}(W(k,k)((Z(k,:) - X(k,:)\beta)^2) \\ \hat{\beta} &= \min_{\beta} L(\beta) + \lambda P_\alpha(\beta) \end{aligned} \tag{5}$$

A. KEN-COX

In this subsection, we describe the kernel elastic net cox regression (KEN-COX) algorithm. The goal of this algorithm is to compensate for the drawbacks of the elastic net regularized cox regression (EN-COX) which is only partially effective at handling correlated attributes in EHR data [20]. We propose to modify this formulation by introducing kernels and proposing the kernel elastic net (KEN-COX) regularized cox regression. Kernel functions [5] can be used to find out the pairwise similarity between a set of features. In our formulation, we use the individual features as the input and build a kernel similarity matrix over the set of features (columns of the original data). The motivation behind using this is to introduce the pairwise feature similarity in the cox regression formulation.

$$\hat{\beta} = \min_{\beta} L(\beta) + \lambda(\alpha \parallel \beta \parallel_1) + \lambda(1-\alpha)\beta^T K\beta \tag{6}$$

$$K(x_i, x_j) = exp(\frac{- \parallel x_i - x_j \parallel_2^2}{2\sigma^2})$$

Equation (6) provides the formulation of KEN-COX regression. K is a symmetric RBF kernel matrix. The formulation of KEN-COX here consists of one smooth and one non smooth L_1 term. This optimization problem can be solved using the cyclic coordinate descent procedure. In Line 6 of Algorithm 2, the cyclic coordinate descent step for KEN-COX is provided. In this equation, S refers to the soft thresholding function. In this step, the p^{th} coordinate of the new coefficient vector $\hat{\beta}$ is calculated at each iteration by cyclically setting all the remaining p-1 coordinates constant from $\tilde{\beta}$ and keeping the p^{th} coordinate alone as the variable for minimization.

Algorithm 2 KEN-COX

Require: Input Data X, Censored times C, Time to event T, Regularization parameter λ, Elastic Net parameter α

1: Initialize $\tilde{\beta} = zeros(m, 1)$
2: **repeat**
3: Compute W and Z using Equations (2) and (3)
4: **repeat**
5: p=1:1:m
6: $\hat{\beta}(p,:) = \frac{S(\frac{1}{n}\sum_{k=1}^{n} W(k,k)X(k,p)[Z(k,:) - \sum_{j\neq p} X(k,j)\tilde{\beta}(j,:)], \lambda\alpha)}{\frac{1}{n}\sum_{k=1}^{n} W(k,k)X(k,p)^2 + \lambda\alpha K(p,p)}$
7: **until** $p \neq m$
8: $\tilde{\beta} = \hat{\beta}$
9: **until** Convergence of $\hat{\beta}$
10: Output $\hat{\beta}$

B. OSCAR-COX

Grouping of features in high dimensional EHR data is a property where a set of features are correlated with similar strength to the prediction label (response). An unbiased model must deal with such attributes on the same scale and assign similar prediction coefficients to them. In this algorithm, we incorporate the OSCAR (Octagonal Shrinkage and Clustering Algorithm for Regression) regularizer into the cox regression framework [1].

OSCAR performs variable selection for regression with many highly correlated predictors. The advantage of using this penalty over other penalties such as the elastic net and LASSO is that this method promotes equality of coefficients which are similarly related to the response. OSCAR obtains the advantages of both individual sparsity due to the L_1 norm and the group sparsity because of the pairwise L_∞ norm. It can select features and form different groups of features. In this way, OSCAR also does supervised clustering of the features.

In this paper, we use the modified Graph OSCAR

(GOSCAR) regularizer [25], [24] in the cox regression formulation. The formulation of the GOSCAR penalty is given in Equation (7). In this formulation, T is the sparse symmetric (m x m) egde set matrix obtained after building a graph using the set of features alone as individual nodes. The sparse adjacency graph is built using the Euclidean distance among the features.

In this manner, a pairwise feature regularizer is added to the cox regression formulation. OSCAR has proven to be more effective than the elastic net in handling correlation among variables and hence is more suited for EHR data. For the sake of brevity, we refer to the *GOSCAR-COX* algorithm as *OSCAR-COX* throughout this paper.

$$\hat{\beta} \quad = \quad \min_{\beta} L(\beta) + \lambda_1(\parallel \beta \parallel_1) + \lambda_2(\parallel T\beta \parallel_1) \qquad (7)$$

In contrast to the KEN-COX algorithm, the formulation in OSCAR-COX is completely non smooth. This problem can be solved using the Alternate Direction Method of Multipliers (ADMM) method effectively [2]. The ADMM method has proven to have a very fast convergence rate and is particularly useful for our application.

$$\hat{\beta} = \min_{\beta,q,p} L(\beta) + \lambda_1 \parallel q \parallel_1 + \lambda_2 \parallel p \parallel_1$$
$$\text{s. t. } \beta - q = 0, T\beta - p = 0$$

In Equation (8), we provide the steps involved in solving OSCAR-COX using the ADMM procedure. In the first step we convert Equation (7) into a form that is suitable for applying ADMM optimization. We then express the augmented lagrangian L_ρ where ρ is the augmented lagrangian parameter. μ and v are the lagrange multiplier vectors. $\tilde{\beta}$, $\tilde{q}$ and $\tilde{p}$ are the initial values to begin the optimization routine. We provide equations for estimating $\hat{\beta}$, $\hat{q}$ and $\hat{p}$. We use the *conjugate gradient descent* method for solving the unconstrained minimization problems to determine $\hat{\beta}$, $\hat{q}$ and $\hat{p}$ during each iteration.

$$L_\rho(\beta, q, p) = L(\beta) + \lambda_1 \parallel q \parallel_1 + \lambda_2 \parallel p \parallel_1 \qquad (8)$$
$$+ \mu^T(\beta - q) + v^T(T\beta - p)$$
$$+ \frac{\rho}{2} \parallel \beta - q \parallel^2 + \frac{\rho}{2} \parallel T\beta - p \parallel^2$$

$$\hat{\beta} = \min_{\beta} L_\rho(\beta, \tilde{q}, \tilde{p}) \qquad (9)$$

$$\hat{q} = \min_{q} L_\rho(\hat{\beta}, q, \tilde{p}) \qquad (10)$$

$$\hat{p} = \min_{p} L_\rho(\hat{\beta}, \hat{q}, p) \qquad (11)$$

C. Discussion

The difference between the KEN-COX and OSCAR-COX algorithm lies in their uniqueness in handling correlated variables in the EHR data. *KEN-COX* uses a kernel based pairwise regularizer in the elastic net formulation to supplement the original elastic net algorithm to handle correlated variables effectively. In this algorithm, the choice of the kernel function is important but it does not cause great variation in the performance. *KEN-COX* uses the L_1 and L_2 norms in it's regularizer. In *OSCAR-COX*, we use the L_1 norm and a pairwise L_∞ norm term. The pairwise L_∞ function encourages similar coefficient values for correlated variables. This helps us understand that both these algorithms handle correlated variables in their own unique ways.

Algorithm 3 OSCAR-COX

Require: Input Data X, Censored Times C, Time to event T, Regularization Param λ_1 Regularization Param λ_2, AugLag Param ρ

Initialize $\tilde{\beta} = zeros(m, 1)$
repeat
  Compute W and Z using Equations (2) and (3)
  Compute the adjacency matrix from the features X.
  Compute the sparse edgeset matrix T from the adjacency matrix
  Compute $\hat{\beta}$ using Equation (9)
  Compute $\hat{q}$ and $\hat{p}$ from Equations (10), (11)
  $\tilde{\beta} = \hat{\beta}$
until Convergence of $\hat{\beta}$
Output $\hat{\beta}$

IV. Experimental Results

In this section, we discuss the experimental results obtained by using the proposed *KEN-COX* and *OSCAR-COX* regression algorithms on both real world EHR and synthetic datasets. We evaluate the goodness of these algorithms in terms of non redundancy in feature selection , goodness of fit using mean squared error (MSE) , R^2 coefficients and the AUC metric. We compare the goodness of fit of these algorithms against cox regression and it's popular variants. We assess the performance of these algorithms on synthetic survival analysis datasets we generated. We also compare the AUC values for these algorithms against state of the art regularized least-squares and logistic loss regression.

In addition, we also conduct a study on the redundancy of the features selected by the proposed algorithms on EHR datasets. We run and tabulate the results obtained from using popular feature selection algorithms [28] implemented in the ASU feature selection repository[1]. In this experiment, we also conduct a study on the biomarkers obtained by using these algorithms and validate those biomarkers using existing clinical literature.

A. Experimental Setup

We generate synthetic datasets by setting the pairwise correlation ρ between any pair of covariates to 0.2. We

[1] http://featureselection.asu.edu

generate the feature vectors using this correlation and a normal distribution $N(0,1)$. We generate feature vectors of different dimensionality to construct five synthetic datasets. For each of these synthetic datasets we generate the failure times T using a Weibull distribution with γ set to 1.5. The Weibull distribution is used here to generate positive responses (failure times) to suit the constraints of synthetic survival data. Censoring times C for each dataset was set to the average of all the failure times in that dataset. In Table III, further description on the dimensionality of these synthetic datasets is provided.

We use the electronic health records (EHR) retrieved from the Henry Ford Health System at Detroit, Michigan. We considered 18,701 hospitalization records for 8,692 patients admitted for heart failure over a duration of 10 years. We generated individual longitudinal EHR datasets from this raw data file. Each *DS1-DS5* represents the set of records for the first to fifth readmission for these patients. We summarize some of the important steps in the process of generating this longitudinal data

We create binary variables from the procedures and medications list which indicate the presence or absence of that particular procedure or medication for the patient. For the labs, we apply the logarithm transformation to make the data follow a normal distribution. For each distinct lab variable we compute the maximum, minimum and average values and create separate variables for each of them. We also create a new feature which signifies the % of *abnormal* labs for a patient.

In Table III, we provide the details about the number of records in these datasets. The variation in the number of columns for these EHR datasets arises from the difference in the number of common lab tests, procedures and medications administered to the patients during their different readmissions. In this longitudinal data, we observe the phenomenon that as the readmission index increases the number of patients readmitted decreases.

The *KEN-COX* and *OSCAR-COX* algorithms were implemented in the *MATLAB* environment. The regularization parameters for both of these algorithms were determined using 5 fold cross validation. In the *OSCAR-COX* algorithm the augmented lagrangian parameter ρ was set to 5 and we use the same values for both the regularization parameters λ_1 and λ_2. In the *KEN-COX* algorithm, the elastic net parameter α was varied from 0.4 to 0.7 with increments of 0.05. We observed that an α value of 0.6 gave us the best performance.

B. Goodness of Fit

In this subsection, we compare the proposed *KEN-COX* and *OSCAR-COX* regression with the elastic net regularized cox regression (*EN-COX)* and cox regression algorithms. We precede this by providing a brief description of the EN-COX optimization formulation.

Table III
DESCRIPTION OF DATASETS

Dataset	# Features	# Instances
Syn1	**15**	**5000**
Syn2	**50**	**5000**
Syn3	**100**	**100**
Syn4	**1000**	**100**
Syn5	**500**	**50**
DS1	**732**	**5675**
DS2	**709**	**4379**
DS3	**668**	**3543**
DS4	**658**	**2826**
DS5	**609**	**2278**

Elastic Net COX (EN-COX) Regression: [20] The formulation of the loss function to be minimized in the elastic net regularized cox regression is given in Equation (12). This optimization problem is solved using the cyclic coordinate descent procedure. We implemented this algorithm using MATLAB.

$$\hat{\beta} = \min_{\beta} L(\beta) + \lambda P_\alpha(\beta) \tag{12}$$
$$P_\alpha(\beta) = \alpha \parallel \beta \parallel_1 + (1-\alpha) \parallel \beta \parallel_2^2$$

In Table IV, we provide the mean squared error (MSE) and R^2 coefficients to assess the goodness of fit. The Martingale residual for a given patient x_i is defined as $\delta_i - exp(x_i^T \hat{\beta}) h_0(t_i)$. The mean squared error (MSE) is calculated as the mean of the squared martingale residuals for all the patients. The R^2 coefficient is calculated using the formula $R^2 = 1 - exp(-\frac{2}{n}(l(\hat{\beta}) - l(0)))$. In this formula, l represents the partial log likelihood function from cox regression.

For each EHR dataset, we consider the case of 30 day readmission to determine the censoring status (δ). The reason for choosing a 30 day readmission scheme for censoring is because this duration is clinically relevant in practice. The 30 day time period post discharge from a hospitalization is considered to be the most important time period where a relapse is expected [10]. Hence, the censoring times C is set to 30 for all the instances and then δ is determined as shown in the example in Section II. Each of the algorithms in Table IV is run on the dataset through 5 fold cross validation to report the MSE and R^2 coefficient values in this format.

In Table IV, we highlight the algorithm with the best fit using bold. A model is considered to be a good one if the mean squared error (MSE) is low and the R^2 coefficient is high. The value of the R^2 coefficient ranges from 0 to 1. We observe that for all the five datasets our proposed algorithms provide a better fit compared to standard cox regression and EN-COX algorithms. This demonstrates the effectiveness of our approach in real world clinical settings.

Table IV
COMPARISON OF MSE AND R^2 VALUES OF KEN-COX AND OSCAR-COX WITH EN-COX AND COX

Dataset	OSCAR-COX		KEN-COX		EN-COX		COX	
	MSE	R^2	**MSE**	R^2	**MSE**	R^2	**MSE**	R^2
DS1	**2.85**	**0.41**	2.96	0.29	2.9825	0.21	2.96	0.3
DS2	**2.78**	**0.40**	2.97	0.31	2.988	0.24	3.41	0.19
DS3	**2.27**	**0.35**	2.95	0.29	3.05	0.24	3.22	0.22
DS4	**2.03**	**0.48**	2.79	0.36	3.03	0.23	3.14	0.15
DS5	2.9	0.29	**2.8**	**0.32**	3.10	0.20	3.25	0.17

C. Redundancy in Features

In the proposed regularized cox regression algorithms, we use sparsity inducing norms with specific mathematical structure to handle correlation among attributes. Due to the sparsity induced, these methods also perform feature selection implicitly. We compare the goodness of the features selected by these methods against state of the art feature selection methods. The metric we use for comparing is the redundancy of features given in Equation 13. In this equation, ρ_{ij} is the Pearson correlation coefficient , F is the set of features selected by the corresponding parsimonious model and m is the number of features present in the dataset.

$$Redundancy = \frac{1}{m(m-1)} \sum_{f_i, f_j \in F i>j} \rho_{ij} \qquad (13)$$

We compare the redundancy scores of our algorithms with prominent feature selection methods in the literature.

1) *SPEC* [27]: Spectral Feature Selection uses a kernel function and determines the Laplacian matrix for the data. It's formulation also consists of the degree and affinity matrices of the dataset. Using these components a scoring function is evaluated for each feature.
2) *mRmR* [18]: Minimum redundancy and maximum relevance feature selection minimizes the correlation and mutual information between features and maximizes the correlation between features and class label.
3) *Fisher Score* [6]: Fisher score gives higher rank to those features that have similar values in the samples from the same class and have different values in the samples from different classes.
4) *Relief-F* [12]: Relief-F is a more recent extension of Relief to handle multi-class problems. It basically computes the values on the corresponding feature of the nearest points to the instance in consideration with the same and different class label respectively. In this manner, it selects the features with good discriminative ability.

In Table V, we compute the redundancy scores using Relief-F, Fisher Score, SPEC and mRmR against the features in the parsimonious models of *KEN-COX* and *OSCAR-COX*. We observe that our methods provide the least redundancy scores for 3 out of 5 datasets considered. The redundancy scores in the case of the other two datasets are also competitive. This suggests that our methods retain the best set of explanatory features in the dataset for prediction.

Table V
COMPARISON IN THE REDUNDANCY OF FEATURES SELECTED BY KEN-COX AND OSCAR-COX AGAINST STANDARD FEATURE SELECTION ALGORITHMS

Method	DS1	DS2	DS3	DS4	DS5
Relief-F	0.05	**0.039**	0.07	0.055	0.06
Fisher Score	0.042	0.043	**0.05**	0.067	0.07
SPEC	0.042	0.045	**0.05**	0.054	0.06
mRmR	0.04	0.046	0.054	0.051	0.058
KEN-COX	**0.039**	0.047	0.051	0.048	0.052
OSCAR-COX	0.042	0.045	0.052	**0.043**	**0.045**

D. Comparison with linear and logistic learners

In this experiment, we compare the performance of *KEN-COX* and *OSCAR-COX* against different set of regularized linear and logistic regression algorithms. For this experiment, we segregate the original data and build two different datasets. We consider both the time-to-event data for the survival regression and the data with nominal labels for the linear and logistic learners. We generate nominal labels for the data using a simple binary labelling scheme. To construct the nominal dataset we assign a label 0 if the time to event is > 30 and a label 1 if the time to event is $<= 30$.

For the regression models constructed for the time to event synthetic data the discrimination ability is evaluated using a metric called the *survivalROC* [9]. This metric considers censored survival data and predicts the survival probability for the subjects in the dataset.

$$FLA = \| Y - X\beta \|_2^2 + \lambda_1 \| \beta \|_1 + \lambda_2 \sum_{j=1}^{p} \| \beta_j - \beta_{j-1} \|_1$$
$$TLA = \| Y - X\beta \|_2^2 + \lambda \| XDiag(\beta) \|_* \qquad (14)$$

To compare the goodness of the models we apply the LASSO [21], ELASTIC NET (EN) [30], FUSED-LASSO (FLA) [23] and TRACE-LASSO (TLA) [8] regression models for the given synthetic datasets. The motivation to use EN, FLA and TLA models is due to their inherent ability

Table VI
COMPARISON OF AUC VALUES FOR KEN-COX AND OSCAR-COX AGAINST REGULARIZED LINEAR AND LOGISTIC LEARNERS

Dataset	OSCAR-COX	KEN-COX	LASSO	EN	FLA	TLA	SPLOG	ENLOG
Syn1	0.81	**0.8472**	0.64	0.68	0.80	0.7978	0.7454	0.7701
Syn2	0.8814	0.8605	0.76	0.784	**0.8918**	0.8449	0.7342	0.8101
Syn3	**0.8909**	0.8412	0.76	0.762	0.874	0.8179	0.7071	0.67
Syn4	**0.881**	0.8605	0.72	0.69	0.8533	0.8446	0.68	0.63
Syn5	**0.8875**	0.859	0.71	0.75	0.8575	0.8553	0.64	0.7

to handle correlations effectively and provide better performance than the LASSO. The FLA imposes both sparsity and smoothness constraints in the model by imposing individual sparsity and sparsity in the differences of the coefficient values. TLA imposes the nuclear norm regularization in the least squares regression framework. We report the AUC values in Table VI using 5 fold cross validation. We report the AUC values for *KEN-COX* and *OSCAR-COX* using *survivalROC*. The regularization parameter for each of these algorithms was obtained through cross validation. The maximum number of iterations for convergence was set to 100.

The TLA algorithm [8] is implemented as described in the original paper. SLEP [16] package is used to implement the linear regression regularized lasso and elastic net algorithms along with the FLA algorithm. Similarly, loss function in SLEP is modified to implement the sparse logistic regression (SPLOG) [14] and elastic net penalized logistic loss regression (ENLOG) algorithms respectively.

In Table VI, the first and second best performing algorithms are marked in bold. We observe that for 4 out of 5 datasets either *KEN-COX* or *OSCAR-COX* is the best performing algorithm. The better performance of these algorithms can be attributed to their inherent capability of handling the non symmetry in time to event data. The improved discriminative ability in correlated feature spaces over both FLA and TLA algorithms is due to the novelty of the regularizers. *KEN-COX* and *OSCAR-COX* effectively use the kernel and graph based structure to exploit correlation and grouping of features more effectively.

E. Biomarker identification from Heart Failure EHR data

Biomarkers are important indicators (variables) of the progression of a disease in real world clinical setting. In this section, we provide a comparative analysis of the biomarkers obtained by applying our methods on the real EHR data. We begin by explaining how we created the baseline to evaluate the biomarkers obtained.

Baseline generation: In a recent clinical review [19], the authors conducted a survey over medical journal articles to determine the important variables for predicting readmission risk for heart failure. The survey statistics included capturing the % of studies where the clinical variable was included in the model, % of studies where the variable was included and found to be statistically associated with readmission risk and other related measurements.

In Table VII, the second column represents the % of studies which reported a statistical association between the candidate variable and heart failure readmission risk. We use this number as the baseline and sort the important biomarkers in the descending order of their statistical association. For each important biomarker, we use a ✔ mark to represent that this variable is selected in the parsimonious feature model and ✖ to mark it's absence from the model. The sparse regression models we consider in this experiment are those of KEN-COX, OSCAR-COX and LASSO. We also consider the top 11 variables with highest absolute regression coefficient values from the Cox regression model.

Table VII
STATISTICAL ASSOCIATION BETWEEN BIOMARKERS AND HEART FAILURE READMISSION

Variable	Assoc	LASSO	COX	OSCAR-COX	KEN-COX
HGB	**0.81**	✖	✔	✔	✔
ckd[2]	**0.75**	✖	✔	✔	✔
diabetes	**0.71**	✖	✖	✔	✔
hypertension	**0.70**	✖	✔	✔	✔
BUN/CREAT	**0.66**	✔	✔	✔	✔
age	**0.66**	✔	✖	✖	✖
cad[3]	**0.61**	✖	✖	✔	✔
heart_failure	**0.60**	✔	✖	✔	✔
afib[4]	**0.60**	✖	✖	✔	✔
HAP	**0.57**	✖	✖	✔	✔
pvd[5]	**0.56**	✖	✖	✔	✔

We observe that both *KEN-COX* and *OSCAR-COX* identify 10 out of 11 important baseline biomarkers and use them in their model. Cox ranks only 4 out of these 11 biomarkers in it's top ranked feature list. LASSO also identifies only 3 out of the 11 important biomarkers. This proves that our methods identify clinically relevant variables from the entire set and retain those variables in their parsimonious models effectively.

V. RELATED WORK

In this section, we discuss the relevant work conducted in the field of regularized cox regression models and their

[2]chronic kidney disease
[3]coronary artery disease
[4]atrial fibrillation
[5]peripheral vascular disease

applications to healthcare.

In the literature, cox regression has been combined with regularizers such as the LASSO [21] which uses the L_1 norm regularization and encourages sparsity in the regression coefficient values. In [22], the LASSO penalty was used along with the cox scaled partial log likelihood function to obtain the LASSO-COX algorithm. This algorithm was used to identify important predictors in lung cancer and liver data. LASSO-COX proved to be a strong competitor to naive feature selection approaches.

In [7], the SCAD penalty was used as a regularizer in the cox regression framework. It was observed that this method had the quality of behaving like an oracle and can obtain the most important predictors with the optimal regularization parameters. A generalized formulation of the SCAD penalty has spawned the inception of many different sparsity inducing norms.

In [26], [29], the adaptive LASSO penalty was used as a regularizer in the cox regression framework resulting in a robust regression algorithm. This improved over LASSO-COX by considering a weighted L_1 norm in the formulation. The weights were also determined using the regression coefficients from the LASSO-COX regression. This solution obtained good sparsity and was also proclaimed to have the oracle property.

The elastic net uses a convex combination of the L_1 and squared L_2 norm (ridge) penalty to obtain both sparsity and handle correlated feature spaces [30]. Using prostate cancer data, the authors have shown that elastic net outperformed other competing methods including the LASSO. In [20], the elastic net penalty was used as a regularizer in the cox regression framework to propose a elastic net cox (EN-COX) algorithm.

In [17], the problem of diabetes risk prediction was tackled using real patient data consisting of around 200,000 patients. For the risk prediction, the authors used methods such as LASSO-COX and cox regression coupled with strong feature selection mechanisms. They also applied different variants of cox and other machine learning techniques such as k-nearest neighbour method to obtain highly discriminative models. The problem with this approach is that they do not capture the clinical data semantics such as variable correlations and structured sparsity. In contrast our algorithms *KEN-COX* and *OSCAR-COX* specifically address these challenges and combine the effectiveness of cox with novel regularizers to build algorithms which are effective and non-redundant.

VI. Conclusion and Future Work

In this paper, we combine cox regression with two novel regularizers to propose the *KEN-COX* and *OSCAR-COX* algorithms respectively. The motivation behind choosing these regularizers is to handle correlation and structured sparsity in high dimensional EHR data effectively. We solve both these problems using scalable optimization procedures to obtain faster convergence.

We conduct different kinds of experiments to evaluate the discriminative ability and the feature selection quality for these two algorithms over a wide range of synthetic and real EHR datasets. Experimental results suggest that our algorithms provide a good fit for the data points and the AUC is better compared to state of the art linear and logistic regularized learners. Our algorithms also improve over the original cox and elastic net cox regression. This suggests the effectiveness of our approach.

The directions for future work include building cox regression models which can deal with multiple outputs at the same time and combine cox with multi-task learning and multi-output regression models. Through this we can explore the true power of these models in dealing with longitudinal clinical data.

Acknowledgements

This work was supported in part by the U.S. National Science Foundation grants IIS-1231742 and IIS-1242304. The authors would like to thank Dr. David Lanfear from Henry Ford Health System, Detroit, Michigan for providing the electronic health records for our analysis.

References

[1] H. D. Bondell and B. J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123, 2008.

[2] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[3] X. Chen, W. Pan, J. T. Kwok, and J. G. Carbonell. Accelerated gradient method for multi-task sparse learning problem. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pages 746–751. IEEE, 2009.

[4] D. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

[5] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[6] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern classification and scene analysis 2nd ed. 2001.

[7] J. Fan and R. Li. Variable selection for cox's proportional hazards model and frailty model. *The Annals of Statistics*, 30(1):74–99, 2002.

[8] E. Grave, G. Obozinski, and F. Bach. Trace lasso: a trace norm regularization for correlated designs. *arXiv preprint arXiv:1109.1990*, 2011.

[9] P. J. Heagerty and Y. Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105, 2005.

[10] A. F. Hernandez, M. A. Greiner, G. C. Fonarow, B. G. Hammill, P. A. Heidenreich, C. W. Yancy, E. D. Peterson, and L. H. Curtis. Relationship between early physician follow-up and 30-day readmission among medicare beneficiaries hospitalized for heart failure. *JAMA: The Journal of the American Medical Association*, 303(17):1716–1722, 2010.

[11] D. W. Hosmer Jr, S. Lemeshow, and S. May. *Applied survival analysis: regression modeling of time to event data*, volume 618. Wiley-Interscience, 2011.

[12] K. Kira and L. A. Rendell. A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning*, pages 249–256. Morgan Kaufmann Publishers Inc., 1992.

[13] J. P. Klein and M. L. Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer, 2003.

[14] J. Liu, J. Chen, and J. Ye. Large-scale sparse logistic regression. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 547–556. ACM, 2009.

[15] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient l 2, 1-norm minimization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 339–348. UAI Press, 2009.

[16] J. Liu, S. Ji, and J. Ye. SLEP: Sparse learning with efficient projections, 2009.

[17] H. Neuvirth, M. Ozery-Flato, J. Hu, J. Laserson, M. S. Kohn, S. Ebadollahi, and M. Rosen-Zvi. Toward personalized care management of patients at risk: the diabetes case study. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 395–403. ACM, 2011.

[18] H. Peng, C. Ding, F. Long, et al. Minimum redundancy maximum relevance feature selection. *IEEE Intelligent Systems*, 20(6):70–71, 2005.

[19] J. S. Ross, G. K. Mulvey, B. Stauffer, V. Patlolla, S. M. Bernheim, P. S. Keenan, and H. M. Krumholz. Statistical models and patient predictors of readmission for heart failure: a systematic review. *Archives of internal medicine*, 168(13):1371, 2008.

[20] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011.

[21] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[22] R. Tibshirani et al. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

[23] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[24] S. Yang, L. Yuan, Y.-C. Lai, X. Shen, P. Wonka, and J. Ye. Feature grouping and selection over an undirected graph. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 922–930. ACM, 2012.

[25] J. Ye and J. Liu. Sparse methods for biomedical data. *ACM SIGKDD Explorations Newsletter*, 14(1):4–15, 2012.

[26] H. H. Zhang and W. Lu. Adaptive lasso for cox's proportional hazards model. *Biometrika*, 94(3):691–703, 2007.

[27] Z. Zhao and H. Liu. Spectral feature selection for supervised and unsupervised learning. In *Proceedings of the 24th international conference on Machine learning*, pages 1151–1157. ACM, 2007.

[28] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H. Liu. Advancing feature selection research. *ASU Feature Selection Repository*, 2010.

[29] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[30] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.