Predict Risk of Relapse for Patients with Multiple Stages of Treatment of Depression

Zhi Nie[1,2], Pinghua Gong[2], Jieping Ye[2,3]
[1] Department of Computer Science and Engineering, Arizona State University, Tempe, AZ
[2] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI
[3] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI

ABSTRACT

Depression is a serious mood disorder afflicting millions of people around the globe. Medications of different types and with different effects on neural activity have been developed for its treatments during the past few decades. Due to the heterogeneity of the disorder, many patients cannot achieve symptomatic remission from a single clinical trial. Instead they need multiple clinical trials to achieve remission, resulting in a multiple stage treatment pattern. Furthermore those who indeed achieve symptom remission are still faced with substantial risk of relapse. One promising approach to predicting the risk of relapse is censored regression. Traditional censored regression typically applies only to situations in which the exact time of event of interest is known. However, follow-up studies that track the patients' relapse status can only provide an interval of time during which relapse occurs. The exact time of relapse is usually unknown. In this paper, we present a censored regression approach with a truncated l_1 loss function that can handle the uncertainty of relapse time. Based on this general loss function, we develop a gradient boosting algorithm and a stochastic dual coordinate ascent algorithm when the hypothesis in the loss function is represented as (1) an ensemble of decision trees and (2) a linear combination of covariates, respectively. As an extension of our linear model, a multi-stage linear approach is further proposed to harness the data collected from multiple stages of treatment. We evaluate the proposed algorithms using a real-world clinical trial dataset. Results show that our methods outperform the well-known Cox proportional hazard model. In addition, the risk factors identified by our multi-stage linear model not only corroborate findings from recent research but also yield some new insights into how to develop effective measures for prevention of relapse among patients after their initial remission from the acute treatment stage.

KDD '16, August 13-17, 2016, San Francisco, CA, USA

DOI: http://dx.doi.org/10.1145/2939672.2939870

Keywords

Major depressive disorder, relapse, censored regression, survival analysis

1. INTRODUCTION

Depression, clinically called Major Depressive Disorder (MDD), is a mood disorder that affects about one eighth of population in US [7] and an estimated 350 million people globally and is projected on track to be the second leading cause of disability in the world by the year 2020 [11]. Medications of different types, such as selective serotonin reuptake inhibitors (SSRIs) and serotonin norepinephrine reuptake inhibitors (SNRIs), which are based on different biological mechanisms and have different effects on neural activity have been developed and tested in a number of clinical trials during the past few decades. Typically, a clinical trial on one medication for depression lasts somewhere from one to four months during which the antidepressant under study is vigorously dosed to tolerance with the goal of symptom remission due to its implications of better daily function[19][8]. However, the interplay of multiple factors such as patients' gene expression profile [5], chronicity of depression [1], psychiatric and general heath comorbidities [16], intolerable side effects from medications, etc., makes many patients with MDD unlikely to respond to certain types of treatment. Thus they are unlikely to achieve remission with a single trial. For these patients, several sequential treatment stages are often necessary to obtain remission.

Another problem concerning the treatment of MDD is the potential risk of relapse among those who indeed achieve remission with one or several stages of treatment. Keller et al. [12] pointed out that there is a substantial probability of prompt relapse among patients without bipolar disorders who recovered from their first major depressive episode and should they relapse, they have an approximately 20% chance of remaining chronically depressed. Findings from [15] indicated that patients who achieved remission of MDD after treatment with citalopram still continued to experience residual symptoms which put them at a higher risk of relapse in a 12-month follow-up phase. Interestingly, the risk of relapse seems to be inversely proportional to the survival time after remission. For instance, Ramana et al. [18] found that all the relapses of subjects that participated in their study occurred within 10 months after they achieved remission. According to the results reported in [13], relapse occurs within four weeks for 12% of patients with remission. And it takes eight more weeks for the number to double. Thus it is of crucial importance to accurately predict the risk of relapse

of MDD patients after their remission, especially those who are likely to relapse shortly after their remission as there is evidence showing that putting remitted patients under continuation and maintenance therapy would greatly reduce their risk of relapse [23] [17].

An example of clinical trial to treat depression involving both multiple stages of treatment aimed at achieving symptom remission and a follow-up phase evaluating the long-term treatment outcome is the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) [1] trial. It involves over 4,000 outpatients with nonpsychotic MDD from a broad spectrum of social demography. It consists of four sequential acute treatment stages. Patients who did not achieve remission or suffered from intolerable side effects of medications in one treatment stage were encouraged to go to the next stage. Those who had achieved remission or shown significant symptomatic improvement could enter a 12-month naturalistic follow-up phase. In the follow-up phase, the measurements concerning patients' depressive symptoms were made on a monthly basis. However, in both acute treatment stages and follow-up, patients may miss certain scheduled clinic visits or drop out, resulting in different patterns of missingness.

One class of the widely used methods that can be potentially employed for building models to predict the risk of relapse is survival analysis, e.g., Cox proportional hazards model [4]. However, traditional methods for survival analysis usually assume that for uncensored cases (e.g., subjects that relapsed), the exact time when the event of interest (e.g., relapse) occurs is known, which is not the case in STAR*D or for data collected through clinical trials of depression in general. In STAR*D, for instance, if a subject does not relapse at the 5th month but is observed to have relapsed at the 6th month, we only know for certainty that the relapse occurs somewhere between the 5th month and the 6th month. Another issue with traditional survival analysis methods is that they apply only to the situation where all the subjects have the same covariates. They cannot be readily used for solving the problem in which certain patients have covariates from more treatment stages while others have less. An alternative is to build a sequence of dependent predictive models, one model for each time point with the dependency formulated through an explicit or implicit constraint that if a model built for a later time point predicts that the event of interest does not occur for a subject, the models built for all the earlier time points should predict the same. However, such methods usually result in very complicated optimization problems to solve or have to resort to some kind of approximation [3].

In this paper, we present a censored regression approach to predict the risk of relapse based on information collected from patients' acute treatment stages and their enrollment. Specifically, we employ a truncated l_1 loss to model the responses (patients' time of relapse) that are upper bounded and/or lower bounded. Based on this basic loss function, we consider the hypothesis that can be represented as (1) an ensemble of decision trees, and (2) a linear combination of covariates. For the hypothesis that takes the form of an ensemble of decision trees, we develop a gradient boosting approach to learn the base models and combination coefficients. When the hypothesis takes the form of a linear combination of covariates, we develop a stochastic dual coordinate ascent algorithm, which is the state-of-the-art method for solving large-scale machine learning problems with a convex loss function [24] with a guaranteed fast convergence rate [20]. Furthermore, we assume that the treatment stage varying covariates collected on patients around the same time before they entered follow-up study should share some commonalities in terms of their relative contribution to predicted risk of relapse and, at the same time covariates collected from different stages overall, contribute differently to the prediction. Based on this assumption, we propose a multi-stage linear approach that can simultaneously estimate multiple linear models for patients remitted after different numbers of treatment stages. Based on data collected from STAR*D trial, we generated several datasets by selecting different cut-off points. Our results show that our proposed methods consistently outperform the Cox model on all datasets.

Our major contributions in this paper are as follows: (1) We present a truncated l_1 loss based censored regression approach to deal with uncertainties of responses. (2) We develop an efficient gradient boosting algorithm and stochastic dual coordinate descent algorithm to solve the proposed formulation. (3) Based on the linear model, we further propose a multi-stage linear approach that can deal with covariates collected from different numbers of treatment stages. (4) We conduct experiments on both synthetic datasets and STAR*D to evaluate the effectiveness of our methods. (5) We identify risk factors which might provide some new insights into development of more effective therapies for prevention of relapse.

2. TRUNCATED L_1 LOSS FOR LEARNING THE RELAPSE TIME OF PATIENTS

In recent literature, different types of loss functions have been proposed to handle censored data under different application scenarios. For instance, in [10], the prediction of time of occurrence of stroke among subjects was modeled through the Huber loss which basically enforces the predicted time of the uncensored subjects to be the same as the observed time. In [21], the loss function for learning from censored targets was absolute deviation of predicted value away from its target interval. In our situation, the assessments of depressive status of all the patients were made on a monthly basis during the follow-up phase, which means that when a patient was regarded as having relapsed at the time of assessment, we only know for sure that relapse of the patient occurred at or before the time when the assessment was made. For this reason, the truncated l_1 loss, which includes the loss function used in [21] as a special case, is proposed to model the relapse risk of patients in the STAR*D cohort.

Suppose there are a total of n samples $D = \{x_1, \cdots, x_n\}$. Let $\mathcal{S}_l$ be the set of indices of samples whose responses are bounded from below by some values, i.e. $\mathcal{S}_l = \{i|a_i \geq l_i\}$ and $\mathcal{S}_u$ be the set of indices of samples whose responses are bounded from above by some values, i.e. $\mathcal{S}_u = \{i|a_i \leq u_i\}$ where a_i is the real unknown response of the i-th sample; l_i is its observed lower bound and u_i is its observed upper bound. The real unknown responses are both upper bounded and lower bounded for uncensored cases and only lower bounded for those censored cases.

[1] http://www.nimh.nih.gov/funding/clinical-research/practical/stard/index.shtml

We formulate our censored regression with truncated l_1 loss as the following optimization problem:

$$\begin{aligned}&\arg\min_{F(\mathbf{x})} L\left(F(\mathbf{x})|\tau,\mathcal{S}_l,\mathcal{S}_u,\mathbf{l},\mathbf{u},D\right)\\ =&\arg\min_{F(\mathbf{x})}\left[\frac{\tau}{n}\sum_{i\in\mathcal{S}_l}\left(l_i-F(\mathbf{x}_i)\right)_+ + \frac{1-\tau}{n}\sum_{i\in\mathcal{S}_u}\left(F(\mathbf{x}_i)-u_i\right)_+\right],\end{aligned} \tag{1}$$

where $(z)_+ = \max(0,z), \forall z\in\mathbb{R}$; $F(\mathbf{x}_i)$ gives an estimate of the response for the i-th sample $\mathbf{x}_i$; $\mathbf{l}$, $\mathbf{u}$ are vectors comprising of the lower and upper bounds for all the samples, respectively; $\tau\in(0,1)$ is a pre-specified constant that balances the tradeoff between the two terms.

Intuitively, if the response of an instance $\mathbf{x}_i$ is lower bounded by l_i, we would like the predicted response $F(\mathbf{x}_i)$ to be greater than or equal to l_i. Otherwise, a penalty would be imposed. This works similarly if its response is upper bounded. No penalty is incurred if $F(\mathbf{x}_i)\in[l_i,u_i]$.

Although in practice, the response of an instance $\mathbf{x}_i$ is both lower bounded and upper bounded if it is an uncensored case, in the following, to simplify our discussion, we simply replicate such instances with one for lower bound index set $\mathcal{S}_l$ and one for upper bound index set $\mathcal{S}_u$ such that $\mathcal{S}_l\cap\mathcal{S}_u=\emptyset$ and denote the new dataset as X. Furthermore, we use one vector $\mathbf{c}$ to denote the union of $\mathbf{l}$ and $\mathbf{u}$ such that $c_i=l_i$ if $i\in\mathcal{S}_l$ and $c_i=u_i$ if $i\in\mathcal{S}_u$. Thus, the resulting loss can be written as

$$\begin{aligned}&\arg\min_{F(\mathbf{x})} L\left(F(\mathbf{x})|\tau,\mathbf{c},\mathcal{S}_l,\mathcal{S}_u,X\right)\\ =&\arg\min_{F(\mathbf{x})}\left[\frac{\tau}{N}\sum_{i\in\mathcal{S}_l}\left(c_i-F(\mathbf{x}_i)\right)_+ + \frac{1-\tau}{N}\sum_{i\in\mathcal{S}_u}\left(F(\mathbf{x}_i)-c_i\right)_+\right],\end{aligned} \tag{2}$$

where $N=|\mathcal{S}_l|+|\mathcal{S}_u|$. Note that replicating instances serves only to decouple the index set $\mathcal{S}_l$ and $\mathcal{S}_u$ and has no effect on model $F(\mathbf{x})$ trained on the dataset.

To further simplify the problem (2), we introduce variables y_i's $(i=1,2,\cdots,N)$ which are defined as follows:

$$y_i=\begin{cases}1 & \text{if } i\in\mathcal{S}_l\ ;\\ -1 & \text{if } i\in\mathcal{S}_u.\end{cases} \tag{3}$$

Then the problem (2) could be reformulated as

$$\begin{aligned}&\arg\min_{F(\mathbf{x})} L\left(F(\mathbf{x})|\tau,\mathbf{c},\mathcal{S}_l,\mathcal{S}_u,X\right)\\ =&\arg\min_{F(\mathbf{x})}\tau\sum_{i\in\mathcal{S}_l}\left(c_i-F(\mathbf{x}_i)\right)_+ + (1-\tau)\sum_{i\in\mathcal{S}_u}\left(F(\mathbf{x}_i)-c_i\right)_+\\ =&\arg\min_{F(\mathbf{x})}\tau\sum_{i\in\mathcal{S}_l}\left[y_i\left(c_i-F(\mathbf{x}_i)\right)\right]_+\\ &+(1-\tau)\sum_{i\in\mathcal{S}_u}\left[y_i\left(c_i-F(\mathbf{x}_i)\right)\right]_+.\end{aligned} \tag{4}$$

Defining $\hat{c}_i,\hat{y}_i$ as

$$\hat{c}_i = \begin{cases}\tau y_i c_i & \text{if } i\in\mathcal{S}_l;\\ (1-\tau)y_i c_i & \text{if } i\in\mathcal{S}_u;\end{cases} \tag{5}$$

$$\hat{y}_i = \begin{cases}\tau y_i & \text{if } i\in\mathcal{S}_l;\\ (1-\tau)y_i & \text{if } i\in\mathcal{S}_u,\end{cases} \tag{6}$$

we can rewrite the problem (4) as

$$\arg\min_{F(\mathbf{x})}\sum_{i=1}^{N}\left(\hat{c}_i-\hat{y}_i F(\mathbf{x}_i)\right)_+. \tag{7}$$

3. A GRADIENT BOOSTING APPROACH TO CENSORED REGRESSION WITH TRUNCATED L_1 LOSS

In this section, we consider the case where $F(\mathbf{x})$ can be represented as an ensemble of regression trees. Namely, for an ensemble consisting of $M+1$ base learners:

$$F(\mathbf{x})=\sum_{i=0}^{M}\alpha_i f(\mathbf{x},\mathbf{a}_i),$$

where $f(\mathbf{x},\mathbf{a})$ represents the entire class of regression tree functions; each $\mathbf{a}_i$ represents a specific set of joint parameter values realizing a member of this function class and $\alpha_0,\cdots,\alpha_m$ are the combination coefficients .

The gradient boosting decision tree [6] iteratively adds new regression trees that fit the negative gradient of the loss function with respect to $F(\mathbf{x})$ at the most up-to-date estimate. Suppose that at the m-th step, we have in our ensemble a set of base learners $\{f(\mathbf{x};\mathbf{a}_i)\}_{i=0}^{m-1}$ each of which takes the form of a regression tree and the corresponding weights $\{\alpha_i\}_{i=0}^{m-1}$. Then, the current estimation for the response of the j-th sample is given by

$$F_{m-1}(\mathbf{x}_j)=\sum_{i=0}^{m-1}\alpha_i f(\mathbf{x}_j;\mathbf{a}_i).$$

The negative gradient of loss function with respect to $F(\mathbf{x}_j)$ at $F_{m-1}(\mathbf{x})$ is $\tilde{g}_j$ and

$$\begin{aligned}\tilde{g}_j &= -\left[\frac{\partial L\left(F(\mathbf{x}_j)|\tau,\mathbf{c},\mathcal{S}_l,\mathcal{S}_u,X\right)}{\partial F(\mathbf{x}_j)}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}\\ &= \begin{cases}\hat{y}_i & \text{if } \hat{y}_i F(\mathbf{x}_i)<\hat{c}_i;\\ 0 & \text{otherwise.}\end{cases}\end{aligned} \tag{8}$$

The new base learner (which, in our case, is a regression tree) parameterized by $\mathbf{a}_m$ to be added to the ensemble at the current step is typically obtained by solving the following optimization problem:

$$\mathbf{a_m}=\arg\min\sum_{i=1}^{N}\left(\tilde{g}_i-f(\mathbf{x}_i;\mathbf{a}_m)\right)^2.$$

For simplicity, let us denote $f(x_i,a_m)$ by $f_m(x_i)$. Once $\mathbf{a}_m$ is fixed, the optimal line search step size for $f_m(\mathbf{x})$ is obtained via

$$\begin{aligned}&\arg\min_{\rho>0} L_1\left(F(\mathbf{x})+\rho f_m(\mathbf{x})|\tau,\mathbf{c},\mathcal{S}_l,\mathcal{S}_u,X\right)\\ =&\arg\min_{\rho>0}\sum_{i=1}^{N}\left[\hat{c}_i-\hat{y}_i F_{m-1}(\mathbf{x}_i)-\hat{y}_i\rho f_m(\mathbf{x}_i)\right]_+\\ =&\arg\min_{\rho>0}\sum_{\hat{y}f_m(\mathbf{x}_i)>0}\hat{y}_i f_m(\mathbf{x}_i)\left[\frac{\hat{c}_i-\hat{y}_i F_{m-1}(\mathbf{x}_i)}{y_i f_m(\mathbf{x}_i)}-\rho\right]_+\\ &-\sum_{\hat{y}f_m(\mathbf{x}_i)<0}\hat{y}_i f_m(\mathbf{x}_i)\left[\rho-\frac{\hat{c}_i-\hat{y}_i F_{m-1}(\mathbf{x}_i)}{y_i f_m(\mathbf{x}_i)}\right]_+.\end{aligned} \tag{9}$$

Let

$$r_i = \left[\left(\hat{c}_i - \hat{y}_i F_{m-1}(\mathbf{x}_i)\right)/\left(y_i f_m(\mathbf{x}_i)\right)\right], i \in \{1, \cdots, N, f_m(\mathbf{x}_i) \neq 0\}.$$

Since (9) is a piece-wise linear function, it is straightforward that the optimal ρ is one of r_i's that satisfies $r_i > 0$, and could make (9) reach minimum.

When the number of instances N is very large, it is very time consuming to evaluate the function value for each r_i since each time of evaluation of the function value involves summing over N elements. However, since (9) is the difference of two monotonically decreasing functions, this repeated computation can be avoided by keeping track of the amount by which each function decreases each time when we increase ρ by a certain amount.

Algorithm 1 Gradient boosting approach for censored regression with truncated l_1 loss

1: **Input:** $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, $\mathcal{S}_l$, $\mathcal{S}_u$, $\mathbf{c}$, $\tau \in (0,1)$, M.
2: **Output:** $F_m(\mathbf{x})$.
3: $F_0(\mathbf{x}) \leftarrow \text{argmin}_{z \in \mathbb{R}} L(z|\tau, \mathbf{c}, \mathcal{S}_l, \mathcal{S}_u, X)$
4: **for** $m = 1, \cdots, M$ **do**
5: `Compute negative gradient` $\{\tilde{g}_i\}_{i=1}^N$ `by (8)`
6: $\mathbf{a}_m \leftarrow \arg\min \sum_{i=1}^N \left(\tilde{g}_i - f(\mathbf{x}_i; \mathbf{a}_m)\right)^2$
7: $\rho_t \leftarrow \arg\min_{\rho > 0} L\left(F_{m-1}(\mathbf{x}) + \rho f(\mathbf{x}; \mathbf{a}_m)|\tau, \mathbf{c}, \mathcal{S}_l, \mathcal{S}_u, X\right)$
8: $F_m(\mathbf{x}) \leftarrow F_{m-1}(\mathbf{x}) + \rho_m f(\mathbf{x}; \mathbf{a}_m)$
9: **end for**

The description of the algorithm for solving censored regression with the truncated l_1 loss based on the gradient boosting approach is shown in Algorithm 1. The time complexity of the algorithm is $O(pMNlogN)$, where p is the number of features, thus it may not be applicable for large-scale datasets. Next we introduce the linear model which can be applied to large-scale datasets.

4. A LINEAR MODEL FOR CENSORED REGRESSION WITH L_1 TRUNCATED LOSS

In this section, we consider the case where the hypothesis takes the form of a linear combination of the covariates. We formulate the problem as follows:

$$\min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} L\left(\mathbf{w}^T \mathbf{x}_i | \tau, \mathbf{c}, \mathcal{S}_l, \mathcal{S}_u, X\right) + \frac{1}{2}\lambda \|\mathbf{w}\|_2^2, \quad (10)$$

where $\lambda > 0$ is the regularization parameter. In the following, we simply denote the loss function for the sample x_i as $L\left(\mathbf{w}^T \mathbf{x}_i\right)$. To take into consideration the effects of bias, we can append ones at the end of $\mathbf{x}_i$'s.

By the definition of the loss function in (2), the above problem can be written as

$$\min_{\mathbf{w}} \frac{1}{N}\left[\tau \sum_{i \in \mathcal{S}_l} \left(c_i - \mathbf{w}^T \mathbf{x}_i\right)_+ + (1-\tau) \sum_{i \in \mathcal{S}_u} \left(\mathbf{w}^T \mathbf{x}_i - c_i\right)_+\right] + \frac{1}{2}\lambda \|\mathbf{w}\|_2^2. \quad (11)$$

Using the definition of auxiliary variables y_i, $\hat{c}_i$ defined in (3), (5) and further defining $\hat{\mathbf{x}}_i$ as:

$$\hat{\mathbf{x}}_i = \begin{cases} \tau y_i \mathbf{x}_i & \text{if } i \in \mathcal{S}_l; \\ (1-\tau) y_i \mathbf{x}_i & \text{if } i \in \mathcal{S}_u; \end{cases} \quad (12)$$

we can transform (11) into

$$\begin{aligned} &\min_{\mathbf{w}} \frac{1}{N}\left[\sum_{i \in \mathcal{S}_l} \tau \left[y_i\left(c_i - \mathbf{w}^T \mathbf{x}_i\right)\right]_+ \right. \\ &\left. + \sum_{i \in \mathcal{S}_u} (1-\tau)\left[y_i\left(c_i - \mathbf{w}^T \mathbf{x}_i\right)\right]_+\right] + \frac{1}{2}\lambda\|\mathbf{w}\|_2^2 \\ =&\min_{\mathbf{w}} \frac{1}{N}\sum_{i=1}^{N}[\hat{c}_i - \mathbf{w}^T \hat{\mathbf{x}}_i]_+ + \frac{1}{2}\lambda\|\mathbf{w}\|_2^2. \end{aligned} \quad (13)$$

The loss function above looks similar to the hinge loss used in SVM. However, the fact that $\hat{c}_i$ can be both positive and negative makes the popular packages such as LIBSVM and LIBLINEAR not applicable to solve this problem. An alternative is to transform the above problem into its dual form and use CVX to solve the resulting quadratic programming problem. However it is generally very slow and does not scale to large-scale datasets. Next, we show how to solve the problem with stochastic dual coordinate ascent which has been proven to achieve a fast convergence rate and can handle large-scale datasets.

Define $\phi_i : \mathbb{R} \to \mathbb{R}$ as $\phi_i(z) = [\hat{c}_i - z]_+$. Its convex conjugate $\phi_i^* : \mathbb{R} \to \mathbb{R}$ [2] is

$$\phi_i^*(u) \equiv \max_{z \in \mathbb{R}} zu - \phi_i(z). \quad (14)$$

By plugging the definition of $\phi_i(z)$ into (14), we have

$$\phi_i^*(u) = \begin{cases} -\hat{c}_i & \text{if } u \in [-1, 0]; \\ +\infty & \text{otherwise.} \end{cases} \quad (15)$$

For $\mathbf{w} \in \mathbb{R}^d$, the convex conjugate of the regularization term $g(\mathbf{w}) \equiv \frac{1}{2}\|\mathbf{w}\|_2^2$ is

$$g^*(\mathbf{v}) \equiv \max_{\mathbf{w} \in \mathbb{R}^d} \mathbf{w}^T \mathbf{v} - \frac{1}{2}\|\mathbf{w}\|_2^2 = \frac{1}{2}\|\mathbf{v}\|_2^2.$$

The stochastic dual coordinate ascent (SDCA)[20] solves the dual problem which can be formulated as

$$\max D(\boldsymbol{\alpha})$$

where

$$\begin{aligned} D(\boldsymbol{\alpha}) =& \frac{1}{N}\sum_{i=1}^{N} -\phi_i^*(-\alpha_i) - \lambda g^*\left(\frac{1}{\lambda N}\sum_{i=1}^{N} \hat{\mathbf{x}}_i \alpha_i\right) \\ =& \frac{1}{N}\sum_{i=1}^{N} \hat{c}_i \alpha_i - \frac{\lambda}{2}\|\frac{1}{\lambda N}\sum_{i=1}^{N} \hat{\mathbf{x}}_i \alpha_i\|_2^2 \\ s.t. \quad & \alpha_i \in [0,1], \quad i = 1, \cdots, N \end{aligned} \quad (16)$$

and

$$\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_N].$$

With $g(\cdot) = g^*(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, we can define $\mathbf{w}(\boldsymbol{\alpha}) = \nabla g^*(\mathbf{v}(\boldsymbol{\alpha}))$ where $\mathbf{v}(\boldsymbol{\alpha}) = \frac{1}{\lambda N}\sum_{i=1}^{N} \hat{\mathbf{x}}_i \alpha_i$. It is known that $\mathbf{w}^* = \mathbf{w}(\boldsymbol{\alpha}^*)$ where $\mathbf{w}^*$ and $\boldsymbol{\alpha}^*$ are the primal and dual optimal solutions with both the loss function and the regularization term in the primal problem being convex.

The SDCA method in each iteration randomly chooses one α_j $(1 \leq j \leq N)$ to update with the objective of increasing

the dual function value as much as possible. That is

$$\begin{aligned}\Delta\alpha_j = \underset{\Delta\alpha_j\in\mathbb{R}}{\operatorname{argmax}} & -\frac{1}{N}\phi_j^*(-(\alpha_j+\Delta\alpha_j)) \\ & -\lambda g^*\left(\mathbf{v}(\boldsymbol{\alpha})+\frac{1}{\lambda N}\hat{\mathbf{x}}_j\Delta\alpha_j\right) \\ = \underset{\Delta\alpha_j\in\mathbb{R}}{\operatorname{argmax}} & \frac{1}{N}\hat{c}_j\left(\alpha_j+\Delta\alpha_j\right)-\frac{\lambda}{2}\|\mathbf{v}(\boldsymbol{\alpha})+\frac{1}{\lambda N}\hat{\mathbf{x}}_j\Delta\alpha_j\|_2^2 \\ s.t. \quad & -\alpha_j \le \Delta\alpha_j \le 1-\alpha_j. \end{aligned} \tag{17}$$

By expanding the l_2-norm and discarding the terms unrelated to $\Delta\alpha_j$, we can further get

$$\begin{aligned}\Delta\alpha_j = \underset{\Delta\alpha_j\in\mathbb{R}}{\operatorname{argmax}} & \ \hat{c}_j\left(\alpha_j+\Delta\alpha_j\right) \\ & -\frac{1}{2\lambda N}\|\hat{\mathbf{x}}_j\|_2^2\left(\Delta\alpha_j\right)^2-\mathbf{v}(\boldsymbol{\alpha})^T\hat{x}_j\Delta\alpha_j \\ s.t. \quad & \alpha_j+\Delta\alpha_j\in[0,1]. \end{aligned} \tag{18}$$

Then, by letting

$$t_j = \frac{\lambda N\left(\hat{c}_j-\mathbf{v}(\boldsymbol{\alpha})^T\hat{\mathbf{x}}_j\right)}{\|\hat{\mathbf{x}}_j\|_2^2}, \tag{19}$$

we have

$$\Delta\alpha_j = \begin{cases} -\alpha_j, & \text{if } t_j \le -\alpha_j; \\ t_j, & \text{if } -\alpha_j \le t_j \le 1-\alpha_j; \\ 1-\alpha_j, & \text{if } t_j \ge 1-\alpha_j. \end{cases} \tag{20}$$

Algorithm 2 Proposed SDCA algorithm for truncated loss based censor regression with linear model

Input: $X=\{\mathbf{x}_0,\mathbf{x}_1,\ldots,\mathbf{x}_N\}$, S_l, S_u, $\mathbf{c}$, $\lambda\in\mathbb{R}^+$, T_0, T, $\tau\in(0,1)$.
Output: $\bar{\mathbf{w}},\bar{\boldsymbol{\alpha}}$
Initialize: $\boldsymbol{\alpha}^{(0)}=\mathbf{0},\mathbf{v}^{(0)}=\mathbf{0}$
for $i=1,\cdots,N$ **do**
  `Compute` $\hat{\mathbf{x}}_i,\hat{c}_i$ `by (12), (5) respectively`
end for
for $t=1,\cdots,T$ **do**
  `Randomly pick` $j\in[1,N]$
  `Compute` $\Delta\alpha_j$ `based on (20)`
  $\boldsymbol{\alpha}^{(t)}\leftarrow\boldsymbol{\alpha}^{(t-1)}+\Delta\alpha_j\mathbf{e}_j$
  $\mathbf{v}^{(t)}\leftarrow\mathbf{v}^{(t-1)}+\frac{1}{\lambda N}\hat{\mathbf{x}}_j\Delta\alpha_j$
end for
$\bar{\boldsymbol{\alpha}}=\frac{1}{T-T_0}\sum_{i=T_0+1}^{T}\boldsymbol{\alpha}^{(t-1)}$
$\bar{\mathbf{w}}=\frac{1}{T-T_0}\sum_{i=T_0+1}^{T}\mathbf{v}^{(t-1)}$

The algorithm for solving (13) is shown in Algorithm 2. In practice, the number of iterations T can be determined by the duality gap which is the difference between the function value of (13) and (16) at the attained primal and dual solution. The iteration can be terminated when this gap drops below a predefined threshold.

5. A MULTI-STAGE LINEAR MODEL FOR CENSORED REGRESSION

Based on the linear model introduced in the previous section, we propose in this section a mutli-stage linear model that can simultaneously estimate multiple linear models, one for each group of patients sharing the same number of treatment stages.

The key idea underlying our simultaneous estimation of models is that although patients may achieve remission from different stages of treatment, the stage-varying covariates collected on them around the same time before they entered follow-up study should share some commonalities in terms of their relative contribution to the prediction of relapse time and, at the same time covariates collected from different stages overall, contribute differently to the prediction. In this section, by assuming that commonalities shared among the patients remitted across different treatment stages take the form of a linear combination of covariates, we show how these commonalities can be exploited toward simultaneous learning of the prediction model for multiple stages of treatment.

5.1 Formulation

Suppose that patients in the clinical study of interest experience at most M stages of treatment before they achieve remission. Let all the covariates of the ith patient who remitted after m treatment stages be $\mathbf{x}_i=\left[\mathbf{x}_{i0}^T,\mathbf{x}_{i1}^T,\cdots,\mathbf{x}_{im}^T\right]^T$ $(1\le m\le M)$, where each of the $\mathbf{x}_{ik}^T$ $(0\le k\le m)$ is a column vector and $\mathbf{x}_{i0}^T$ represents covariates that are not related to treatment such as those recording the patient's demographic information and family medical history information; $\mathbf{x}_k^T$ $(k\ge 1)$ represents the covariates from the treatment stage $k-1$ stages away from their last treatment stage. For instance, $\mathbf{x}_{i2}^T$ represents the covariates from second to the last treatment stage. Let $\mathbf{w}=\left[\mathbf{w}_0^T,\mathbf{w}_1^T,\cdots,\mathbf{w}_M^T\right]^T$ where $\mathbf{w}_k^T$ represents the coefficients associated with $\mathbf{x}_k^T$. Note that in our approach, the patients across all the treatment stages share the same weight vector as long as they have the covariates that the specific segment of weights corresponds to. Also let $\boldsymbol{\alpha}=[\alpha_{10},\alpha_{11},\cdots\alpha_{k0},\cdots\alpha_{kk},\cdots\alpha_{M0},\cdots,\alpha_{MM}]$ be a vector of coefficients balancing the influence of different blocks of the covariates for patients that remit from each of the specific treatment stages. Let S_l^m and S_u^m be the set of indices of patients that have covariates from their last m treatment stages and have a lower bound and an upper bound in their relapse time, respectively.

Our proposed approach for simultaneously estimating censored regression models for patients remitting from multiple treatment stages can be formulated as follows:

$$\begin{aligned}\min_{\boldsymbol{\alpha},\mathbf{w}} \sum_{m=1}^{M}\Bigg[& \tau_m\sum_{i\in S_l^m}\left(l_i-\sum_{k=0}^{m}\alpha_{mk}\mathbf{w}_k^T\mathbf{x}_{ik}\right)_+ \\ & +(1-\tau_m)\sum_{i\in S_u^m}\left(\sum_{k=0}^{m}\alpha_{mk}\mathbf{w}_k^T\mathbf{x}_{ik}-u_i\right)_+\Bigg] \\ & +\lambda_1\sum_{m=1}^{M}\sum_{k=0}^{m}\alpha_{mk}^2+\lambda_2\sum_{k=0}^{M}\|\mathbf{w}_k\|^2. \end{aligned} \tag{21}$$

In making prediction, the coefficients for $\mathbf{x}_{ik}$ is $\alpha_{mk}\mathbf{w}_k$. All the covariates from the stage that is $k-1$ stages away from the last stage share the same $\mathbf{w}_k$ while for patients that experienced different number of stages, they have differ different α_{mk}.

5.2 Optimization

It is worth noting that although (21) is convex with respect to either $\boldsymbol{\alpha}$ or $\mathbf{w}$, it not jointly convex. The block coordinate descent algorithm that alternates between opti-

mizing over $\boldsymbol{\alpha}$ and optimizing over $\mathbf{w}$ is adopted to solve this problem. In the following, we show that each step of the optimizing over $\mathbf{w}$ and optimizing over $\boldsymbol{\alpha}$ can be transformed into a problem of the same form as (11).

Fixing $\mathbf{w}$*, solve* $\boldsymbol{\alpha}$

When $\mathbf{w}$ is fixed, the problem (21) can actually be decoupled into M separate problems. Let $r_{ik} = \mathbf{w}_k^T \mathbf{x}_{ik}$. The problem (21) becomes M separate optimization problems, each of which takes the form:

$$\min_{\alpha_{m0},\cdots,\alpha_{mm}} \tau_m \sum_{i\in S_l^m} \left(l_i - \sum_{k=0}^{m} \alpha_{mk} r_{ik} \right)_+ + (1-\tau_m) \sum_{i\in S_u^m} \left(\sum_{k=0}^{m} \alpha_{mk} r_{ik} - u_i \right)_+ + \lambda_1 \sum_{k=0}^{m} \alpha_{mk}^2. \quad (22)$$

This problem is actually the same as (11), thus the algorithm introduced in the previous section can be used to solve it.

Fixing $\boldsymbol{\alpha}$*, solve* $\mathbf{w}$

When $\boldsymbol{\alpha}$ is fixed, the problem (21) turns into

$$\min_{\mathbf{w}} \sum_{m=1}^{M} \left[\sum_{i\in S_u^m} \left(\sum_{k=0}^{m} \mathbf{w}_k^T (\alpha_{mk}(1-\tau_m)\mathbf{x}_{ik}) - (1-\tau_m)u_i \right)_+ + \sum_{i\in S_l^m} \left(\tau_m l_i - \sum_{k=0}^{m} \mathbf{w}_k^T (\tau_m \alpha_{mk} \mathbf{x}_{ik}) \right)_+ \right] + \lambda_2 \|\mathbf{w}\|^2. \quad (23)$$

For $i \in S_u^m$, denote $\mathbf{p}_{ik} = \alpha_{mk}(1-\tau_m)\mathbf{x}_{ik}$; $\mathbf{p}_i = [\mathbf{p}_{i0}, \cdots, \mathbf{p}_{im}]$ and $\hat{u}_i = (1-\tau_m)u_i$. For $i \in S_l^m$, denote $\mathbf{q}_{ik} = \alpha_{mk}(1-\tau_m)\mathbf{x}_{ik}$; $\mathbf{q}_i = [\mathbf{q}_{i0}, \cdots, \mathbf{q}_{im}]$ and $\hat{l}_i = \tau_m l_i$. Let d_k be the dimension of $\mathbf{x}_{ik}$ and $\mathbf{I}_{d_k}$ be an identity matrix of d_k rows and columns. Also let $\mathbf{I}_M = \text{diag}[\mathbf{I}_{d_0}, \cdots \mathbf{I}_{d_M}]$ which is also an identity matrix and $\mathbf{I}_m$ be its first $d_0 + \cdots + d_m$ rows. Then the problem above can be simplified as

$$\begin{aligned} &\min_{\mathbf{w}} \sum_{m=1}^{M} \left[\sum_{i\in S_u^m} \left(\sum_{k=0}^{m} \mathbf{w}_k^T \mathbf{p}_{ik} - \hat{u}_i \right)_+ + \sum_{i\in S_l^m} \left(\hat{l}_i - \sum_{k=0}^{m} \mathbf{w}_k^T \mathbf{q}_{ik} \right)_+ \right] + \lambda_2 \|\mathbf{w}\|^2 \\ &= \min_{\mathbf{w}} \sum_{m=1}^{M} \left[\sum_{i\in S_u^m} \left(\mathbf{w}^T \mathbf{I}_m^T \mathbf{p}_i - \hat{u}_i \right)_+ + \sum_{i\in S_l^m} \left(\hat{l}_i - \mathbf{w}^T \mathbf{I}_m^T \mathbf{q}_i \right)_+ \right] \\ &\quad + \lambda_2 \|\mathbf{w}\|^2. \end{aligned} \quad (24)$$

By introducing $\hat{\mathbf{p}}_i = \mathbf{I}_m^T \mathbf{p}_i$ and $\hat{\mathbf{q}}_i = \mathbf{I}_m^T \mathbf{q}_i$ and making use of the fact that $S_l = S_l^1 + \cdots + S_l^M$ and $S_u = S_u^1 + \cdots + S_u^M$, the above problem can be further written as

$$\min_{\mathbf{w}} \left[\sum_{i\in S_u} \left(\mathbf{w}^T \hat{\mathbf{p}}_i - \hat{u}_i \right)_+ + \sum_{i\in S_l} \left(\hat{l}_i - \mathbf{w}^T \hat{\mathbf{q}}_i \right)_+ \right] + \lambda_2 \|\mathbf{w}\|^2, \quad (25)$$

which also has the same form as (11).

Note that although in the STAR*D dataset, all the patients that achieved remission at a later stage have gone through all the previous treatment stages, this is not a prerequisite for our model. Our model only requires that the covariates of all the patients are aligned in the reverse order of the treatment stages leading to remission.

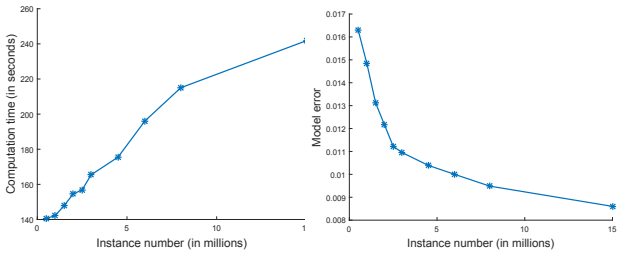
6. EXPERIMENTS

6.1 Simulated data

In this experiment, we use simulated data to answer the following two questions: (1) How well can our linear model scale to large datasets? (2) If there is an underlying linear relationship between the response and the covariates, to what extent that the linear model learned from minimizing the truncated l_1 loss recovers the true linear model when only an upper bound or a lower bound of the response of an instance is known?

6.1.1 *Data generation*

We first generate our datasets as follows: Each instance is randomly drawn from a 100-dimensional standard normal distribution. For each dataset we generate n such instances where n ranges from 50k to 15,000k. The ground truth linear model $\mathbf{w}^*$ is a 100 dimensional vector with each entry being 0.5. The ground truth response for each instance is $y_i = \mathbf{w}^{*T} \mathbf{x}_i + e_i$ where $e_i \sim \mathcal{N}(0,1)$; For each instance, we draw a random number b_i from $\mathcal{U}(0, \max_i |y|_i)$ and use $y_i + b_i$ as its upper bound or use $y_i - b_i$ as its lower bound. We assign a lower bound for half of the randomly chosen samples and an upper bound for the other half. In the simulation, we fix λ in (11) to be $1e-6$ and τ to be 0.5. The algorithm terminates when the duality gap is less than $1e-7$.



(a) Computational time (b) Model error

Figure 1: Change of the computational time and model error when the number of instances increases.

6.1.2 *Results*

The simulation was run on a machine with Intel Xeon(R) CPU E5-1620 v2 3.70GHz × 8 processor, 32 GB memory and Ubuntu 14.04 LTS system. The results are shown in Figure 1. The computational time grows approximately linearly with the number of instances. The model error, measured by $\sqrt{\|\hat{\mathbf{w}} - \mathbf{w}^*\|_2^2/100}$, decreases as the number of instances increases. Note that as the number of instances increases, the perturbation range $\max_i |y|_i$ also increases, bringing in more uncertainty to the response of each instance. It is interesting to observe that even though only an upper bound or a lower bound of responses is provided, with enough samples, the linear model learned from minimizing the truncated l_1

loss can still to some extent recover the true model under certain conditions.

6.2 Evaluation on STAR*D cohort

6.2.1 Dataset and preprocessing

STAR*D is a study designed to identify the most effective treatment or combination of treatments for patients diagnosed with nonpyschotic MDD. It lasted over a period of seven years and involved over 4,000 patients aging from 18-75 and has been so far the largest and longest study ever conducted for evaluating the effectiveness of treatments of depression. It consists of four treatment stages during each of which patients were treated with certain antidepressants and their depressive symptoms were evaluated every two to three weeks. Patients that could not achieve remission or suffered from intolerable side effects of medications in one treatment stage were encouraged to proceed to the subsequent stage. Those who did achieve remission or demonstrated significant symptomatic improvement were invited to enter a 12-month naturalistic follow-up phase where assessments of patients' depressive symptom severity were made on a monthly basis.

Due to subjects dropping out without relapse in the follow-up phase, we considered in our experiments three cut-off points in the follow-up - 10 months, 11 months, 12 months respectively, in order to, on one hand, keep our analysis in as much accordance with the original design of the study as possible, one the other hand, take into account the subjects that dropped out at later time points of the follow-up phase and see how our model performs in response to changes in the total number of right censored cases. For each cut-off point, we included into our analysis only the subjects that either have definitively relapsed at or before the chosen cut-off time point (uncensored cases) and the subjects whose relapse occurred later than the chosen cut-off time point (right censored cases). It is worth emphasizing here that we excluded from our analysis those who dropped out and did not relapse before the chosen cut-off time point so that risk of relapse for each sample is known.

As for the covariates, we included in our analysis those collected from the follow-up enrollment including demographics (DM), Eligibility (EL), Psychiatric Diagnostic Screening (PDS), etc. For each treatment stage, we took into consideration the covariates from patients' last observed record of Quick Inventory of Depressive Symptomatology - Clinician-rated (QIDS-C) and QIDS-SR (Self-rated), the baseline record of Interactive Voice Response (IVR) and Research Outcomes Assessments (ROA). According to [19], relapse is defined as an individual having an observed QIDS total score collected in IVR during the follow-up phase great than 10. All the subjects included in our analysis achieved remission from the acute treatment stages. The number of subjects that relapsed and did not relapse by the cut-off time point from each treatment stage is shown in Table 1. Note that, for the cut-off time points before the 12th month, "non-relapse" cases included all subjects that had definitively not relapsed until that point, even if they eventually relapsed at some time later than that point. The Kaplan-Meier survival curves for subjects with only one treatment stage and subjects with more than one treatment stage from different datasets are shown in Figure 2.

In training models, all the treatment stage varying covariates were aligned based on the reverse order of stages leading to remission. As is shown in Table 1 the number of subjects that remitted from stage 3 and stage 4 were too small, we omitted the covariates from the first treatment stage for those achieved remission from stage 3 and stage 4 as well as the covariates from the second treatment stage for those who achieved remission from stage 4. Missing values were imputed with the column mean. All the covariates that were included were normalized with zscore.

6.2.2 Feature selection

The dataset included a large number of covariates, ranging from demographic information, medical and psychiatric comorbidities to depressive symptom measurements. However, not all of them are related to the subjects' relapse status at the end of the follow-up phase or risk of relapse, which necessitates feature selection as one of the crucial steps to minimize overfitting and ensure the quality of predictive models. In this work, l_1 sparse logistic regression-based stability selection [14] was employed to perform feature selection on each of the datasets determined by different cut-off points. The fitting target for sparse logistic regression is the relapse status by the cut-off point associated with each dataset. Since we have covariates from different stages of treatment, we ran stability selection twice, one on covariates from enrollment and last treatment stage, the other on covariates from second to last treatment stage. The number of covariates selected was determined by the cross-validation.

6.2.3 Performance metrics

A commonly used metric for evaluating the performance of survival models is concordance index [22] [10] which is a generalization of Area Under ROC Curve (AUC) for continuous response and censored data. It measures the probability of concordance between the predicted response and the observed response. A high concordance index means that there is a high likelihood that for two randomly sampled individuals, the order of their predicted response matches the order of their observed response. In our context, concordance index can be regarded as a measure of the proportion of the pairs of subjects for whom the relative order of predicted time of relapse is concordant with the order of observed time of relapse among all the pairs of subjects whose observed time of relapse can be ordered. Suppose we have n samples in our testing set. The observed time of relapse for the i-th subject is t_i. Its predicted time of relapse is p_i. Let $\mathcal{A}$ be the set of pairs that can be ordered, that is

$$\mathcal{A} = \{< i, j > | t_i > t_j, i, j = 1, \cdots, n\}.$$

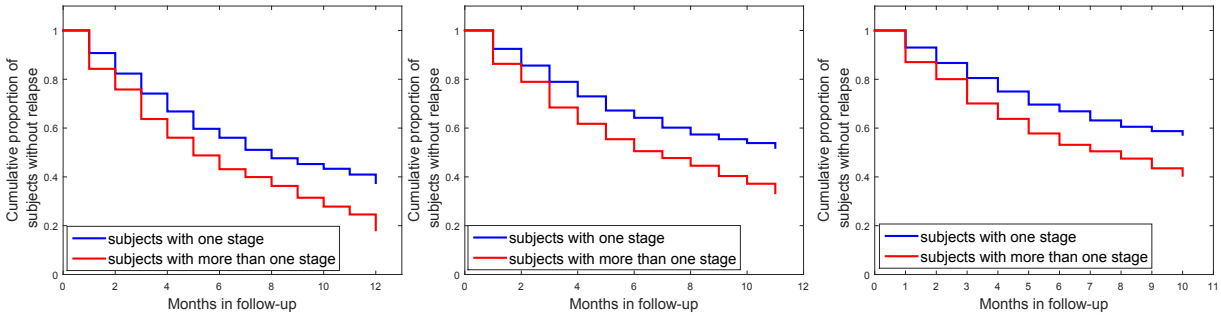
Then the concordance index can be defined as

$$\text{Concordance Index} = \frac{1}{|\mathcal{A}|} \sum_{<i,j> \in \mathcal{A}} \boldsymbol{I}(p_i > p_j),$$

where $\boldsymbol{I}(\cdot)$ is the indicator function.

6.2.4 Experimental setup and results

We can divide each of the datasets into two sets of samples, one with data from one treatment stage only, the other with data from at least two treatment stages. For each dataset, we randomly selected 80% of samples from those who relapsed and those who did not relapse in each set as training set and the rest 20% as the testing set. A five fold cross-validation was carried out on the training set to select parameters. This random split was repeated 10 times



(a) cut-off point at the 12th month (b) cut-off point at the 11th month (c) cut-off point at the 10th month

Figure 2: Kaplan-Meier survival curves of subjects in datasets determined by different cut-off points

Table 1: Data statistics of datasets determined by different cut-off points

	12th month			11th month			10th month		
Stage	Relapse	Non-relapse	Stage total	Relapse	Non-relapse	Stage total	Relapse	Non-relapse	Stage total
Stage 1	290	174	464	274	296	570	263	353	616
Stage 2	163	42	205	155	81	236	147	101	248
Stage 3	23	6	29	23	10	33	23	12	35
Stage 4	9	5	14	9	7	16	9	9	18
Total	485	227	712	461	394	855	442	475	917

and the mean performance on the test sets and the standard deviation were reported.

The linear model, gradient boosting model, and Cox model were all trained and tested on the covariates from enrollment and last treatment stage. The model parameters including the number of features selected for all the models, τ, λ for linear models, τ, the number of trees in the gradient boosting model were determined by cross validation. For the multi-stage linear model, the number of covariates selected from different stages were determined independently and the τ's were set to be the same.

Table 2: Performance of different methods on dataset determined by cut-off at the 12th month

Methods	`Concordance Index`
Multi-Stage Linear	0.7172 (0.0200)
Linear	0.7003 (0.0267)
Gradient Boosting	0.6804 (0.0279)
Cox	0.6800 (0.0310)

Table 3: Performance of different methods on dataset determined by cut-off at the 11th month

Methods	`Concordance Index`
Multi-Stage Linear	0.7423 (0.0172)
Linear	0.7181 (0.0188)
Gradient Boosting	0.7020 (0.0238)
Cox	0.6952 (0.0257)

The average concordance indices along with the standard deviation obtained by different methods on datasets determined by different cut-off time points are reported in Tables 2, 3, 4. Overall the performance on datasets with cut-off

Table 4: Performance of different methods on dataset determined by cut-off at the 10th month

Methods	`Concordance Index`
Multi-Stage Linear	0.7443 (0.0110)
Linear	0.7242 (0.0164)
Gradient Boosting	0.7012 (0.0165)
Cox	0.6993 (0.0232)

at the 10th and the 11th month is better, which is probably due to an increase in the number non-relapse cases that largely comes from subjects dropping out late in the follow-up phase. On all the datasets, our methods perform better than the Cox model. In particular, the multi-stage stage linear model produces the best performance, which, we think, can be accounted for by the fact that it can take into consideration the distributional difference of covariates from subjects remitted from different treatment stages. The performance of the gradient boosting approach is worse than that of the linear approach probably due to overfitting.

6.2.5 Identifying risk factors

One of the advantages of predicting relapse risk with linear models is that with all the covariates normalized to zero mean and the same variance, the magnitude of the coefficient associated with a covariate indicates its marginal contribution to the predicted risk, given all other covariates remaining unchanged. In this subsection, we use the multi-stage linear models obtained from previous subsection based on random splits of the dataset determined by the cut-off point at the 10th month to produce a ranking of the covariates. Although the models built on different random splits of data involve different numbers of covariates, the numbers clustered in small range, with the number of covariates selected

from Enrollment and last treatment stage varying from 70 to 90 and the number of covariates selected from the second to last treatment stage ranging from 10 to 20. Similar to the method used in [10] to cope with variance arising from cross-validation design, we averaged coefficients associated with each covariate over ten models and used the magnitude of the mean value minus their variance as a score to rank the coefficients.

Table 5 shows the top predictors of risk of relapse for subjects that experienced only one treatment stage. To our surprise, the "academic degree" comes on top of the list, suggesting that a higher academic degree is associated with a lower risk of relapse, according to the coding of this attribute and the sign of its coefficient. The residual symptoms, as mostly marked by "last observed", are also ranked high on the list, which corroborates the findings from [9] that the presence of residual symptoms such as depressed mood, hopelessness is associated with an earlier short-term relapse. Although residual sleep disturbance did not appear in our list of top predictors, which is consistent with the observation made in [15] that there is no significant different in Kaplan-Meier survival curves for those with and without the domain of residual sleep disturbance, we did find strong correlation between sleep onset insomnia at the baseline, which ranked second in our list of relapse risk factors. In addition, the subjective nature implied in some of the top-ranked predictors such as "Sad mood" and "impact of your family and friends" also help explain that mindfulness-based cognitive therapy [23] can reduce the risk of relapse of MDD patients in remission or recovery.

The top predictors of risk of relapse for subjects that experienced more than one treatment stage largely overlap with those in Table 5, thus we did not show a separate table here. However, it is worth mentioning that the scores of covariates for patients with more than one treatment stage are generally lower and more flat and among top predictors, predictors marked with "*" in Table 5 rise to a much higher place in the list. But there are some predictors with a relatively high magnitude of mean in their corresponding coefficients but did not make into the top list due to a great variance, which probably results from relatively scarce presence of the conditions specified in those predictors among the subjects under study, implying that they should be considered as potential high risk factors if they are found in subjects. Such predictors include "Visited emergency room in last three months", "Careless work due to emotional problem" from baseline IVR of second to last treatment stage, "Tired nearly every day past 2 weeks" from PDS and "Family history drug abuse" from PHX. Overall, from the top risk factors we identified, we can see that therapies focusing on improving subjects' outlook for the future, psychomotor functioning and negative thinking might be more effective in preventing the relapse among the patients that achieved remission from treatment with antidepressant.

7. CONCLUSION

In this paper, we proposed a censored regression approach for predicting the risk of relapse of patients after their initial remission from one or multiple stages of antidepressant treatment. Since the patients' relapse status was assessed once every month, we employed a truncated l_1 loss to model the response for which only a lower bound or an upper bound is observed. We considered the hypothesis in the loss function that can be represented as (1) an ensemble of regression trees; (2) a linear combination of covariates. We developed an efficient gradient boosting algorithm when the hypothesis takes the form of an ensemble of regression trees and a stochastic dual coordinate ascent algorithm when the hypothesis is a linear model. Furthermore, we extend the linear model to deal with covariates collected from multiple stages of treatment. Our experiments on synthetic data and STAR*D datasets demonstrate the efficiency and effectiveness of the proposed methods. In all cases, our multi-stage linear method achieves the best performance. In addition, the top risk factors identified by our multi-stage linear method are not only consistent with the findings from some of the recent research regarding relapse among patients with MDD who had initially achieved remission but also provided some insights into how to develop therapies for prevention of relapse.

Acknowledgement

The authors would like to thank Dr. Qingqin Li from Johnson & Johnson for some helpful discussions. This work is supported in part by grants from NIH (RF1AG051710) and NSF (IIS-0953662 and III-1421057).

8. REFERENCES

[1] J. Alpert and M. Fava. *Handbook of Chronic Depression: Diagnosis and Therapeutic Management*. Medical Psychiatry. Marcel Dekker Incorporated, 2014.

[2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, 2004.

[3] H. chin Lin, V. Baracos, R. Greiner, and C. nam J. Yu. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *NIPS*, pages 1845–1853. Curran Associates, Inc., 2011.

[4] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[5] H. A. Eyre, A. Eskin, S. F. Nelson, N. M. St. Cyr, P. Siddarth, B. T. Baune, and H. Lavretsky. Genomic predictors of remission to antidepressant treatment in geriatric depression using genome-wide expression analyses: a pilot study. *International Journal of Geriatric Psychiatry*, 2015.

[6] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000.

[7] B. N. Gaynes, D. Warden, M. H. Trivedi, S. R. Wisniewski, M. Fava, and A. J. Rush. What did star*d teach us? results from a large-scale, practical, clinical trial for patients with depression. *Psychiatric Services*, 60(11):1439–1445, 2009.

[8] J. E. Kelsey. Achieving remission in major depressive disorder: The first step to long-term recovery. *The Journal of the American Osteopathic Association*, 104:S6–S10, 2004.

[9] N. Kennedy and K. Foy. The impact of residual symptoms on outcome of major depression. *Current Psychiatry Reports*, 7(6):441–446.

Table 5: Top predictors from the multi-stage linear model for subjects with only one treatment stage

Predictor description	score	category
Academic degree	0.6630	Demographics
QIDS Sleep onset insomnia (baseline)	0.5074	IVR
Sum of QIDS sub-scores (baseline)	0.4327	IVR
American Indian or Alaskan Native	0.3481	Eligibility
QIDS Mood -sad (last observed)	0.3472	QIDS-SR
Impact of your family and friends	0.3436	Demographics
Worry obsessively about you'd act/speak violently	0.3206	PDS
How often have you missed taking meds (last observed)	0.2962	QIDS-SR
Are you currently employed	0.2712	IVR
QIDS Concentration/decision making (last observed)	0.2621	QIDS-SR
Feel hopeless about future for 2 years	0.2615	PDS
QIDS total score (last observed)	0.2470	QIDS-SR
QIDS Psychomotor agitation (last observed)	0.2385	QIDS-C
Stomach and intestinal problems*	0.2349	PDS
Do impulsive things*	0.2311	PDS
QIDS Weight (decrease) last 2 weeks (last observed)	0.2109	QIDS-SR
IDS Psychomotor slowing (baseline)*	0.1944	RA
IDS Outlook-future (baseline)*	0.1850	RA
QIDS Concentration/decision making (baseline)	0.1841	IVR
Number of answered QLESQ items	0.1719	IVR
Health been poor most of life*	0.1631	PDS
Depressed mood	0.1589	Eligibility
Worry when asking questions around others*	0.1588	PDS
Psychomotor agitation or retardation*	0.1543	Eligibility
On medical or psychiatric leave	0.1319	Demographics
Flashbacks of traumatic event	0.1215	PDS

[10] A. Khosla, Y. Cao, C. C.-Y. Lin, H.-K. Chiu, J. Hu, and H. Lee. An integrated machine learning approach to stroke prediction. In *KDD*, pages 183–192, 2010.

[11] M. Marcus, M. T. Yasamy, M. van Ommeren, D. Chisholm, S. Saxena, et al. Depression: A global public health concern. *WHO Department of Mental Health and Substance Abuse*, 1:6–8, 2012.

[12] K. MB, L. PW, L. CE, and K. GL. Predictors of relapse in major depressive disorder. *JAMA*, 250(24):3299–3304, 1983.

[13] K. MB, S. RW, L. PW, and W. N. Relapse in major depressive disorder: Analysis with the life table. *Archives of General Psychiatry*, 39(8):911–915, 1982.

[14] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[15] A. A. Nierenberg, M. M. Husain, M. H. Trivedi, M. Fava, D. Warden, S. R. Wisniewski, S. Miyahara, and A. J. Rush. Residual symptoms after remission of major depressive disorder with citalopram and risk of relapse: a star*d report. *Psychological Medicine*, 40(1):41–50, 2010.

[16] C. Otte. Incomplete remission in depression: role of psychiatric and somatic comorbidity. *Dialogues Clin Neurosci*, 10(4):453–460, 2008.

[17] E. S. Paykel. Continuation and maintenance therapy in depression. *British Medical Bulletin*, 57(1):145–159, 2001.

[18] R. Ramana, E. S. Paykel, Z. Cooper, H. Hayhurst, M. Saxty, and P. G. Surtees. Remission and relapse in major depression: a two-year prospective follow-up study. *Psychological Medicine*, 25(6):1161–1170, 1995.

[19] A. J. Rush, M. H. Trivedi, S. R. Wisniewski, A. A. Nierenberg, J. W. Stewart, D. Warden, G. Niederehe, M. E. Thase, P. W. Lavori, B. D. Lebowitz, P. J. McGrath, J. F. Rosenbaum, H. A. Sackeim, D. J. Kupfer, J. Luther, and M. Fava. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A star*d report. *American Journal of Psychiatry*, 163(11):1905–1917, 2006.

[20] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss. *J. Mach. Learn. Res.*, 14(1):567–599, Feb. 2013.

[21] P. K. Shivaswamy, W. Chu, and M. Jansche. A support vector approach to censored targets. In *ICDM*, pages 655 – 660, 2007.

[22] H. Steck, B. Krishnapuram, C. Dehing-oberije, P. Lambin, and V. C. Raykar. On ranking in survival analysis: Bounds on the concordance index. In *NIPS* , pages 1209–1216, 2008.

[23] J. D. Teasdale, Z. V. Segal, J. Mark, G. Williams, V. A. Ridgeway, J. M. Soulsby, M. A. Lau, J. D. Teasdale, and V. A. Ridgeway. Prevention of relapse/recurrence in major depression by mindfulness-based cognitive therapy. *Journal of Consulting and Clinical Psychology*, 68(4):615–623, 2000.

[24] K. Tran, S. Hosseini, L. Xiao, T. Finley, and M. Bilenko. Scaling up stochastic dual coordinate ascent. In *KDD*, pages 1185–1194, 2015.