The Annals of Applied Statistics
2016, Vol. 10, No. 4, 2130–2156
DOI: 10.1214/16-AOAS967

COX REGRESSION WITH EXCLUSION FREQUENCY-BASED WEIGHTS TO IDENTIFY NEUROIMAGING MARKERS RELEVANT TO HUNTINGTON'S DISEASE ONSET

BY TANYA P. GARCIA[1] AND SAMUEL MÜLLER[2]

Texas A&M University and University of Sydney

Biomedical studies of neuroimaging and genomics collect large amounts of data on a small subset of subjects so as to not miss informative predictors. An important goal is identifying those predictors that provide better visualization of the data and that could serve as cost-effective measures for future clinical trials. Identifying such predictors is challenging, however, when the predictors are naturally interrelated and the response is a failure time prone to censoring. We propose to handle these challenges with a novel variable selection technique. Our approach casts the problem into several smaller dimensional settings and extracts from this intermediary step the relative importance of each predictor through data-driven weights called exclusion frequencies. The exclusion frequencies are used as weights in a weighted Lasso, and results yield low false discovery rates and a high geometric mean of sensitivity and specificity. We illustrate the method's advantages over existing ones in an extensive simulation study, and use the method to identify relevant neuroimaging markers associated with Huntington's disease onset.

1. Introduction. In studies of neuroimaging [Tabrizi et al. (2013)] and genomics [Witten and Tibshirani (2010)], the emergence of new technologies allows scientists to collect a copious amount of data on a small number of subjects. A major interest in such studies is identifying predictors truly relevant to a response of interest. These predictors are generally more cost-effective measures in future clinical trials, provide better understanding and visualization of the data, and improve predictions of a response. Identifying relevant predictors in large data sets is challenging especially when the response is a failure time prone to censoring. In this paper, we propose a new variable selection technique that casts the problem into several smaller dimensional settings and extracts from this intermediary step the relative importance of each predictor through data-driven weights called "exclusion frequencies" (formal definition in Section 2.2). We show that using the exclusion frequencies as weights in the adaptive Lasso [Zhang and Lu (2007)] im-

Received September 2015; revised July 2016.

[1]Supported in part by the Huntington's Disease Society of America Human Biology Project Fellowship.

[2]Supported in part by the Australian Research Council DP130100488.

Key words and phrases. Exclusion frequency, model selection, neuroimaging, proportional hazards model, weighted lasso.

proves variable selection accuracy, decreases false discovery rates [Storey (2003)], and has high geometric mean of sensitivity and specificity.

Our method is motivated by a neuroimaging study of Huntington's disease (HD) where the goal is to identify brain regions that impact age of motor onset. HD is a genetically inherited, neurodegenerative disease that results in motor and cognitive impairments, and eventual death. Its cause is an unstable expansion of the CAG trinucleotide repeat in the huntingtin gene [Huntington's Disease Collaborative Research Group (1993)], and research is ongoing to identify biomarkers for disease tracking and therapy development. A recent study, PREDICT-HD [Paulsen et al. (2008)], has collected numerous measures on brain regions to better understand HD development. Earlier analyses have identified some neuroimaging measures (e.g., striatal volume) that can discriminate individuals at risk for HD from healthy controls [Aylward et al. (2011), Tabrizi et al. (2012)]. These analyses, however, evaluated each measure individually, which can ignore correlation between variables, and did not account for potential censoring of the onset ages. In this paper, we propose a survival-based, variable selection procedure that handles censoring and accounts for interrelationships between neuroimaging measures by simultaneously assessing all variables. For PREDICT-HD, both the number of variables, p, and the sample size, n, are in the hundreds, but our method is general enough to deal with larger and smaller n's and p's.

Current variable selection methods for time-to-event outcomes extend from methods proposed for linear models. These include stepwise selection, best subset selection, bootstrap methods [Sauerbrei and Schumacher (1992)], and Bayesian approaches [Faraggi and Simon (1998), Ibrahim, Chen and MacEachern (1999)]. An approach favoring sparsity is one pioneered by Tibshirani (1997): a Lasso estimator applied to the Cox (1972) proportional hazards model. Fan and Li (2002) showed, however, that the Lasso estimator does not possess the oracle property, and instead proposed a smoothly clipped absolute deviation method (SCAD) for the Cox model which does. But the SCAD estimator can be difficult to solve and may be numerically unstable because it involves optimizing a nonconvex penalty. As an improvement, Zhang and Lu (2007) proposed an adaptive Lasso for the Cox model which involves weighing each predictor with the inverse of the maximizer of the log partial likelihood. The estimator possesses the oracle property and is numerically stable.

By design, weights in the adaptive Lasso for Cox models (and linear models) are based on learning from the full data: all n rows and p columns of the data matrix are used in an initial analysis to construct the weights. Other data-driven weights, however, have been shown to significantly improve variable selection accuracy. For example, Bergersen, Glad and Lyng (2011) showed that weights constructed from external genetic information to their study led to improved variable selection. Of course, external information is not always available and internal data can be used instead. For linear models, Garcia et al. (2013) and Garcia and

Müller (2014) showed improved variable selection using internal data. They constructed weights by regressing the response of interest on each predictor and taking as weights measures of significance of the predictor (i.e., t-statistic, p-value, q-value). They showed that using such weights in the adaptive Lasso improved the accuracy of the variables selected in a linear model. We propose to build on this idea in two ways. First, we will generate internal data weights for survival models. Second, we will construct weights by repeatedly learning from *subsets* of the columns of the design matrix to form exclusion frequencies defined as a variable's irrelevance to the model. Such frequencies are, in some sense, the intermediate and more general case of two "boundary" cases: the adaptive Lasso which uses the full data to construct feature weights, and the methods in Garcia et al. (2013) and Garcia and Müller (2014) which use one variable at a time.

Our motivation for exclusion frequencies stems from inclusion frequencies [Gong (1982, 1986)] which are measures of a predictor's relevance to a model. Inclusion frequencies are derived through repeatedly perturbing the data [Yu (2013)] and computing how often each predictor is selected after applying a variable selection procedure to the perturbed data. Well-established data perturbations involve resampling the n rows of the data matrix using the jacknife, cross-validation, or bootstrap. For example, Gong (1982) applied repeated forward logistic regression to the resampled data matrix after an initial screening of the most important variables. These ideas were further developed in Cox regression [Chen and George (1985), Sauerbrei and Schumacher (1992)] and Poisson regression [Buckland, Burnham and Augustin (1997)]. The notion has motivated other resampling procedures to gain a better understanding of selection method stability. For stability selection when $p < n$, Müller and Welsh (2010) proposed a bootstrap approach, and, when $p > n$, Meinshausen and Bühlmann (2010) proposed a leave-$\lceil n/2 \rceil$-out cross-validation resampling and Shah and Samworth (2013) proposed two consecutive cross-validation subsamples for improved resampling. In all aforementioned examples, selected "stable" variables are those which have inclusion frequencies exceeding a chosen (data-adaptive) threshold. However, such thresholding may ignore correlation between variables [Garcia et al. (2013)].

In contrast to the existing literature for inclusion frequencies, our idea is to compute and use exclusion (equivalently, inclusion) frequencies as follows:

1. Rather than perturbing the data by resampling the n *rows* of the data matrix, we will repeatedly subsample among the p *columns* of the design matrix. Our repeated column perturbations is novel to the variable selection literature, and it allows us to cast the problem into more familiar settings. Specifically, we will analyze only those p^* predictors in each subsample where p^* is much smaller than the sample size n. In these smaller dimensional problems, we will apply standard variable selection procedures (i.e., exhaustive best subset selection, stepwise selection) and compute how often each predictor is excluded (or included) in the procedure. The exclusion (inclusion) frequency weights are thus learned from these simpler, smaller dimensional problems.

2. Rather than select predictors individually based on those whose exclusion (inclusion) frequencies are less than (exceed) a prespecified threshold, predictors will be selected from a modified version of the adaptive Lasso applied to *all predictors*. Specifically, we will apply the adaptive Lasso such that its weights are the exclusion frequencies computed from the smaller dimensional settings. This technique allows us to avoid arbitrarily selecting a threshold where variations in its value can potentially result in different selections. Applying this new adaptive Lasso to all predictors simultaneously also helps to assess the impact of multiple predictors on the response.

We show that the adaptive Lasso with exclusion frequency weights yields more accurate model selections than when using other data-driven weights or approaches such as backward selection, Bolasso [Bach (2008)], and random Lasso [Wang et al. (2011)]. Due to the intractability of theoretical results, we demonstrate this accuracy empirically via simulation. Theoretical justification when subsampling columns of the design matrix might be achievable in future research, but we expect that this will require different methods to when subsampling rows [as in Meinshausen and Bühlmann (2010)]. Also, to the best of our knowledge, the Bolasso and random Lasso were implemented for the first time in the context of Cox regression in this article.

Section 2 describes the construction of exclusion frequency weights and its features. In Section 3, we demonstrate the effectiveness of these weights compared to seven alternative methods, five of which were considered for the first time in the context of Cox regression. In Section 4, we apply our method to PREDICT-HD to identify those brain measures that impact age of motor onset. Section 5 concludes the paper with generalizations of exclusion frequency weights to other regression models and penalization forms. An R implementation of the proposed procedure is available upon request.

2. Proposed exclusion frequency variable selection.

2.1. *Overview of the weighted Lasso.* For $i = 1, \ldots, n$, we denote the observed data as $(S_i, \delta_i, \mathbf{v}_i)$. Here, $S_i = \min(T_i, C_i)$, where T_i denotes the failure time and C_i the censoring time which is assumed independent of T_i. The censoring indicator is $\delta_i = I(T_i \leq C_i)$ and $\mathbf{v}_i$ are p-dimensional predictors. Without loss of generality, we assume the $\mathbf{v}_i$ predictors are standardized to have mean zero and sample variance one. For the PREDICT-HD study (Section 4), we have $n = 839$ genetically at-risk subjects, the failure time T corresponds to age of motor onset, and there are $p = 352$ predictors of volume and surface area measurements of different brain regions. Such a p is too large for exhaustive search algorithms, however, regularization methods can be used to economically parse through the large model space.

To model the relationship between failure times and predictors, we use the Cox (1972) proportional hazards model, though other survival models such as the additive hazards regression [Lin and Lv (2013)] could also be used. For a Cox model, regularization-based variable selection can be achieved using the adaptive Lasso [Zhang and Lu (2007)]. The solution is the minimizer of

$$Q^{\mathrm{cox}}(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\delta_i\left[\boldsymbol{\beta}^T\mathbf{v}_i - \log\left\{\sum_{j=1}^{n} I(S_j \geq S_i)\exp(\boldsymbol{\beta}^T\mathbf{v}_j)\right\}\right] + \lambda\sum_{k=1}^{p} w_k|\beta_k|, \tag{2.1}$$

with respect to $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, the coefficient vector. The first term in $Q^{\mathrm{cox}}(\boldsymbol{\beta})$ is the partial log-likelihood of the data; $\lambda > 0$ is a regularization parameter; and $w_k > 0$, $k = 1, \ldots, p$, are weights. Minimization of $Q^{\mathrm{cox}}(\boldsymbol{\beta})$ can be achieved by transforming the predictors as $v_{ik} \mapsto v_{ik}/w_k = v_{ik}^*$ and defining $\gamma_k = w_k\beta_k$. The minimizer $\widehat{\boldsymbol{\beta}}$ is found by minimizing

$$Q^{\mathrm{cox}}(\boldsymbol{\gamma}) = -\frac{1}{n}\sum_{i=1}^{n}\delta_i\left[\boldsymbol{\gamma}^T\mathbf{v}_i^* - \log\left\{\sum_{j=1}^{n} I(S_j \geq S_i)\exp(\boldsymbol{\gamma}^T\mathbf{v}_j^*)\right\}\right] + \lambda\sum_{k=1}^{p}|\gamma_k|,$$

with respect to $\widehat{\boldsymbol{\gamma}}$, and then transforming back via $\widehat{\beta}_k = \widehat{\gamma}_k/w_k$. Solving for $\widehat{\boldsymbol{\gamma}}$ is easily achieved using procedures for nonweighted, L_1 penalization [Zhang and Lu (2007), Simon et al. (2011)], which are available in software such as R.

2.1.1. *Use of weights.* A key advantage of the adaptive Lasso [Zhang and Lu (2007)] over the Lasso [Tibshirani (1997)] is its utility of weights which aid to yield numerical stability. For example, when using $w_k = 1/|\widetilde{\beta}_k|$, where $\widetilde{\beta}_k$ is the maximizer of the model's log partial likelihood, the oracle property is achieved since the term $|\beta_k|/|\widetilde{\beta}_k|$ converges to $I(\beta_k \neq 0)$.

Recent work from Bergersen, Glad and Lyng (2011), Garcia et al. (2013), and Garcia and Müller (2014) has shown that other data-driven weights can further improve variable selection accuracy. This results by defining weights that appropriately measure the predictors' relevance in relation to the response of interest as described next.

For ease in presentation, we refer collectively to all predictors as $\mathbf{v}$'s and denote the design matrix as $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p] \in \mathrm{mat}(n, p)$. For a fixed regularization parameter λ in (2.1), it is well known [Garcia and Müller (2014)] that those columns $\mathbf{v}_k$ with large weights will generally be excluded from the model (i.e., $\widehat{\beta}_k = 0$), whereas variables with small weights will generally be included in the model (i.e., $\widehat{\beta}_k \neq 0$). Thus, we aim to define small weights for relevant predictors and large weights for irrelevant ones.

To determine the relevance of each predictor, we consider a general situation where predictors belong to one of two groups: those that are indeed relevant to the

model (so-called "designed predictors"), and those that are subject to selection (so-called "candidate predictors"). Such a setting arises primarily in biological studies where certain phenomena warrant that some variables (the designed predictors) are *known* to act on the response, whereas the effect of others (the candidate predictors) is *unknown*. For example, when analyzing the age of onset of Huntington's disease, a designed predictor is the number of CAG repeats in the huntingtin gene, as it is *known* that longer repeats lead to earlier onset [Ross and Tabrizi (2010)]; candidate predictors are volume measures from different brain regions which are *unknown* a priori to have an effect on age of onset. See Garcia et al. (2013) and Garcia and Müller (2014) for other biological examples with designed and candidate predictors. For applications with no a priori information, the model would only have candidate predictors, and the weighting scheme for designed predictors below would be ignored.

For the general scenario, we denote the designed predictors as $\mathbf{z}$'s and candidate predictors as $\mathbf{x}$'s. Specifically, we let m_0 be the number of designed predictors denoted as $\mathbf{v}_1 := \mathbf{z}_1, \ldots, \mathbf{v}_{m_0} := \mathbf{z}_{m_0}$. We let m_1 be the number of candidate predictors denoted as $\mathbf{v}_{m_0+1} := \mathbf{x}_1, \ldots, \mathbf{v}_{m_0+m_1} := \mathbf{x}_{m_1}$. We have that $p = m_0 + m_1$ and p is potentially larger than n. When $m_0 = 0$ there are no designed predictors and all columns of $\mathbf{V}$ are subject to selection.

We will place very small weights on the designed predictors $\mathbf{z}$'s such as $w_1 = \cdots = w_{m_0} = 0$. This will ensure that the $\mathbf{z}$'s are in the final model. As the relevance of the $\mathbf{x}$'s to the model is not known a priori, we propose to define their weights based on information obtained in smaller dimensional settings (see Section 2.2). When $m_0 \neq 0$, the weights for the $\mathbf{x}$'s will also reflect their relevance to the failure times after accounting for the designed predictors $\mathbf{z}$'s. Weights that ignore the $\mathbf{z}$'s have been found to be inferior to unit weights on the $\mathbf{x}$'s; see the simulation study of Garcia et al. (2013).

Our proposed weights are based on exclusion frequencies constructed from two steps performed repeatedly. The first step involves partitioning the columns of the design matrix corresponding to the $\mathbf{x}$'s. The second step involves applying a standard variable selection method (e.g., stepwise regression) and tracking which predictors are not retained by the method. An overview of this algorithm is presented below with details given next (see Algorithm 1).

2.2. *Partitioning the candidate predictors.* Our first step in forming exclusion frequencies involves partitioning the candidate predictors into J different groups. While the choice of J depends on n, m_0, and p, we write J and not $J(n, m_0, p)$ for notational ease. When $n < p$, we aim that, in each partition, n exceeds the number of candidate predictors in the partition plus the number of designed predictors (m_0). Having this requirement allows us to apply nonregularized variable selection techniques to each partition and learn about the relevance of each candidate predictor. When p is already smaller than n but still very large, we show via simulation that partitioning has an advantage in these $p < n$ scenarios as well.

Algorithm 1 (**Exclusion frequency weighted Lasso algorithm**): Let $\mathbf{z}_1, \ldots, \mathbf{z}_{m_0}$ denote designed predictors known to be in the model, and $\mathcal{I} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{m_1}\}$ denote the predictors that are subject to selection. Choice of tuning parameters B, J, λ are discussed in Section 2.4.

Step 1: For $b = 1, \ldots, B$:

(a) Partition $\mathcal{I}$ into nearly equal-sized, disjoint groups $\mathcal{I}_1^{(b)}, \ldots, \mathcal{I}_J^{(b)}$ either with random partition (Section 2.2.1) or structured partition (Section 2.2.2).

(b) On each partition group, $j = 1, \ldots, J$, apply a stepwise regression to $(\mathbf{S}, \boldsymbol{\delta})$ on $\mathcal{I}_j^{(b)} \cup \{\mathbf{z}_1, \ldots, \mathbf{z}_{m_0}\}$, specifying that $\{\mathbf{z}_1, \ldots, \mathbf{z}_{m_0}\}$ are always in the final model. The final model is selected using the Bayesian information criterion (BIC). (Selection methods other than stepwise could be used, but we do not pursue these in this article.)

(c) The stepwise regressions will yield estimates $\widehat{\beta}_1^{(b)}, \ldots, \widehat{\beta}_{m_1}^{(b)}$ corresponding to the variables $\mathbf{x}_1, \ldots, \mathbf{x}_{m_1}$. Compute $E_k^{(b)} = I(\widehat{\beta}_k^{(b)} = 0)$ which tracks if candidate predictor $\mathbf{x}_k$ is excluded from the stepwise regression.

Step 2: Form the exclusion frequency weight for $\mathbf{x}_k$ as $w_{m_0+k} = \sum_{b=1}^{B} E_k^{(b)}/B$.

Step 3: For a fixed λ, solve the Cox regression weighted Lasso in (2.1) with weights $w_1 = \cdots = w_{m_0} = 0$ for $\mathbf{z}_1, \ldots, \mathbf{z}_{m_0}$ to ensure that these designed predictors are in the final model, and weights $w_{m_0+k} = \sum_{b=1}^{B} E_k^{(b)}/B$ for $\mathbf{x}_k$, $k = 1, \ldots, m_1$.

For example, in the Huntington's disease study where p is in the 100s and $p < n$, partitioning reduces the number of variables to the low 10s in each partition.

Two natural questions about partitioning emerge. First, how does one select J? Second, how does one form the partition? The first question is addressed in Section 2.4 where we discuss the selection of all tuning parameters. To address the second question, we propose two types of partitions: the first uses randomization (Section 2.2.1) and the second uses different structuring of the predictors (Section 2.2.2). For ease in notation, let the candidate predictors be collected in the set $\mathcal{I} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{m_1}\}$.

2.2.1. *Random partition.* Random partitioning involves randomly splitting $\mathcal{I}$ into J groups $\mathcal{I}_1, \ldots, \mathcal{I}_J$. To ensure that each group has roughly equal size, we require that the size of group $\mathcal{I}_j$, denoted as $r_j = |\mathcal{I}_j|$, satisfies

$$\sum_{j=1}^{J} r_j = m_1, \qquad \left\lfloor \frac{m_1}{J} \right\rfloor \le r_j \le \left\lceil \frac{m_1}{J} \right\rceil, \qquad r_j + m_0 < n,\ j = 1, \ldots, J.$$

The first requirement ensures that the partitions include all candidate predictors. The second requirement ensures that partition sizes are roughly equal. The last

requirement ensures that the total number of predictors (designed plus candidate ones) is less than sample size n.

Earlier work has shown random partitioning to be less robust than so-called structured partitioning in the variable selection literature [Müller and Welsh (2005, 2009)] when subsetting rows and not the columns as in this article. Therefore, we suggest different structured ways to partition the columns of the design matrix in the next subsection.

2.2.2. *Structured partition*. Structured partition is one directed by a measure of "similarity" among the predictors. We define "similarity" using a distance measure on the estimates from Cox ridge regression of $(\mathbf{S}, \boldsymbol{\delta})$ on $\{\mathbf{z}_1, \ldots, \mathbf{z}_{m_0}\} \cup \mathcal{I}$. Letting $\widetilde{\beta}_k$, $k = 1, \ldots, m_1$, denote the ridge regression estimates, we define the similarity of the predictors according to three measures: (i) those with similar magnitudes of $\widetilde{\beta}_k$; (ii) those whose estimated ridge estimates are in the same "k-means" clusters; and (iii) those whose estimated ridge estimates are in the same sample quantile group.

For each similarity measure, our idea is to initially partition $\mathcal{I}$ into J' groups, such that each partition $\mathcal{C}_{j'}$, $j' = 1, \ldots, J'$, has similarly behaving predictors. Then we will randomly sample from these J' partitions to form partitions $\mathcal{I}_1, \ldots, \mathcal{I}_J$ so that each $\mathcal{I}_j$, $j = 1, \ldots, J$, has predictors representative of the different similarity groups. This two-stage partitioning is easy to program (R code is available upon request) and, in our empirical experience, choosing $J' = J$ yields similar results as when $J' \neq J$.

We describe this two-stage partitioning in more detail below:

(a) *Sorted partition*: One structured partition uses the sorted magnitudes of the ridge parameter estimates. Let $|\widetilde{\beta}_{(1)}| \geq |\widetilde{\beta}_{(2)}| \geq \cdots \geq |\widetilde{\beta}_{(m_1)}|$ denote the ordered ridge estimates. First, we divide the predictors in $\mathcal{I}$ into J' clusters $\mathcal{C}_1, \ldots, \mathcal{C}_{J'}$ such that $\mathcal{C}_1$ contains the predictors associated with $|\widetilde{\beta}_{(1)}|, \ldots, |\widetilde{\beta}_{(k)}|$, $\mathcal{C}_2$ contains the predictors associated with $|\widetilde{\beta}_{(k+1)}|, \ldots, |\widetilde{\beta}_{(2k)}|$, and so on for $k = \lfloor m_1/J' \rfloor$ for the first $J' - 1$ partition groups; the remaining predictors are placed in the J'th partition. Second, we randomly select elements from each $\mathcal{C}_1, \ldots, \mathcal{C}_{J'}$ to form J disjoint groups $\mathcal{I}_1, \ldots, \mathcal{I}_J$. We aim to ensure that each $\mathcal{I}_j$ contains at least one element from each $\mathcal{C}_1, \ldots, \mathcal{C}_{J'}$, but this may not always be the case since the size of each $\mathcal{C}_j$ may be too small. However, because we repeat this procedure repeatedly and randomly, the collective set of partitions $\mathcal{I}_j$ will capture the behavior of the predictors.

(b) *Means and Quantile partitions*: An alternative partition involves means and quantiles of the ridge parameter estimates. First, we separate $\mathcal{I}$ into J'-clusters $\mathcal{C}_1, \ldots, \mathcal{C}_{J'}$ based on J'-means clustering or J'-quantiles of the estimates $\{\widetilde{\beta}_1, \ldots, \widetilde{\beta}_{m_1}\}$. Second, we randomly select elements from each $\mathcal{C}_1, \ldots, \mathcal{C}_{J'}$ to form J disjoint groups $\mathcal{I}_1, \ldots, \mathcal{I}_J$ such that each $\mathcal{I}_j$ contains at least one element from each $\mathcal{C}_1, \ldots, \mathcal{C}_{J'}$.

After forming the partition (either random or structured), we then form the exclusion frequencies by applying a variable selection method to each partition group and keeping track of which predictors are not selected. The procedure is described next.

2.3. *Forming and using exclusion frequencies*. On each partition group $j = 1, \ldots, J$, we apply a variable selection procedure to the set of predictors $\mathcal{I}_j \cup \{\mathbf{z}_1, \ldots, \mathbf{z}_{m_0}\}$ such that the predictors $\{\mathbf{z}_1, \ldots, \mathbf{z}_{m_0}\}$ are always in the final model. Given that n exceeds the number of predictors in $\mathcal{I}_j \cup \{\mathbf{z}_1, \ldots, \mathbf{z}_{m_0}\}$, nonregularization-based variable selection procedures can be used, such as stepwise, subset, forward, or backward regression. Among such methods, we propose to use stepwise regression and have the final model selected based on the Akaike information criterion (AIC) or Bayesian information criterion (BIC). In the simulation studies of Section 3, we show the results of both and explain our preference for BIC. There is a myriad of alternatives to using AIC or BIC, and for a recent review we refer to Müller and Welsh (2010). However, exploring what criterion or what other variable selection method is optimal is beyond the scope of this article.

After applying stepwise regression to each partition group, we will obtain estimates $\widehat{\beta}_1, \ldots, \widehat{\beta}_{m_1}$ corresponding to the variables $\{\mathbf{x}_1, \ldots, \mathbf{x}_{m_1}\}$. For $k = 1, \ldots, m_1$, we then define $E_k = I(\widehat{\beta}_k = 0)$, which tracks if candidate predictor $\mathbf{x}_k$ is *excluded* from the final model of the stepwise regressions. The procedure of partitioning the candidate predictors and computing the frequencies E_k is done repeatedly B-many times. The exclusion frequency for $\mathbf{x}_k$, $k = 1, \ldots, m_1$, is then $w_{m_0+k} = \sum_{b=1}^{B} E_k^{(b)}/B$.

Having formed the exclusion frequencies, we then use them as weights in equation (2.1). Specifically, for a fixed λ, we solve (2.1) with weights $w_1 = \cdots = w_{m_0} = 0$ for $\mathbf{z}_1, \ldots, \mathbf{z}_{m_0}$ to ensure that these designed predictors are in the final model, and weights $w_{m_0+k} = \sum_{b=1}^{B} E_k^{(b)}/B$ for the candidate predictor $\mathbf{x}_k$, $k = 1, \ldots, m_1$. Solving (2.1) can be quickly achieved using the glmnet package in R [Simon et al. (2011)].

2.4. *Selection of tuning parameters*. Our algorithm involves several tuning parameters: the number of partitions J (and J' for structured partition), how often partitions are formed (B), and the regularization parameter λ in equation (2.1).

We found that the exact choice of J minimally changed the performance in empirical studies. In Table 1, we report the false discovery rates [FDR, Storey (2003)] and the geometric mean of sensitivity and specificity $G \equiv (\text{specificity} \times \text{sensitivity})^{1/2}$ [Kubat, Holte and Matwin (1998)] of 500 simulated survival data sets with 50% censoring (see Section 3 for details of the design). Here, specificity is the proportion of irrelevant predictors that were not selected among irrelevant predictors, and sensitivity is the proportion of relevant predictors that are selected among relevant predictors. The ideal partition size will lead to low FDR (ideal is 0) and high G-score (ideal is 1).

TABLE 1

Effect of number of partitions when exclusion frequencies formed using J-quantile structured partition and stepwise regression with BIC applied to the partitions; 50% *censoring*, 500 *simulations*, $B = 100$ *replicates*. *Observed false discovery rate* (*FDR*; *low FDR is ideal*) *and G-score* (*balance between sensitivity and specificity*; *high G-score is ideal*) *when* $n = 120$, $p = 61$. *Covariates are either uncorrelated or correlated with* $\text{corr}(\mathbf{x}_i, \mathbf{x}_j) = \rho^{|i-j|}$

Method	Number of partitions	Covariates uncorrelated		Covariates correlated					
				$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
		FDR	G	FDR	G	FDR	G	FDR	G
J-quantile, BIC	$J = J' = 15$	0.09	0.97	0.08	1.00	0.12	0.97	0.16	0.71
	$J = J' = 10$	0.12	0.98	0.12	0.99	0.18	0.97	0.22	0.74
	$J = J' = 6$	0.03	0.98	0.01	1.00	0.01	0.96	0.01	0.69
	$J = J' = 4$	0.02	0.97	0.00	1.00	0.00	0.95	0.00	0.67

Overall, when $n = 120$ and $p = 61$ ($m_0 = 1, m_1 = 60$) and predictors are uncorrelated, the G-score (i) varies little and is between 0.96 and 0.98 whether $J = 4$, $J = 6$, $J = 10$, or $J = 15$; (ii) is unimodal over the range of considered values; and (iii) is largest for the two intermediate values $J = 6$ or $J = 10$. When the predictors are highly correlated, the G-score lowers slightly but remains stable for different J values. In general practice, we recommend choosing J so that there are 10 predictors on average in each partition. However, choosing the number of partitions is naturally driven by computational limitations. For example, one should choose J large enough so that there is a sufficiently large number of predictors in each partition and such that variable selection is computationally feasible when having to do this repeatedly on the resampled partitions. We encourage having enough predictors in each partition since we found that when there is only one predictor per partition group, then more irrelevant predictors are incorrectly selected than when having sufficiently large partition groups.

For B, we recommend $B = 100$ as a balance between learning appropriate data-driven exclusion frequency-based weights and computational demand. We found that using B-many partitions to form the exclusion frequencies has the additional advantage that highly correlated predictors are infrequently in the same partition group, which helps to avoid issues of multicollinearity.

Last, for the regularization parameter λ in (2.1), we recommend using the cross-validation procedure in the R glmnet package [Simon et al. (2011)].

3. Simulation studies.

3.1. *Simulation design*. We evaluated our proposed method and existing ones in a simulation study mimicking the Huntington's disease (HD) study in Section 4. The goal of the HD study is to determine which neuroimaging measures impact

age of motor onset after accounting for the known effect of a patient's disease progression level on onset age. A standard for measuring HD progression is through the CAG-Age-Product score [Zhang et al. (2011), CAP]: a product of the number of CAG repeats in the huntingtin gene and age at study entry. Higher CAP scores ($\geq$368) are associated with a patient having a high probability of experiencing HD onset, and lower scores ($\leq$290) are associated with a patient having a low probability of experiencing onset. Without loss of generality, we consider the CAP score and all neuroimaging predictors as standardized to have mean zero and variance 1.

To replicate the effect of CAP scores, we generated a uniformly distributed, designed predictor $\mathbf{z}$ (i.e., $m_0 = 1$). This predictor will act on the failure time (i.e., age of motor onset) and be correlated with other predictors to mimic the potential effect of disease progression on other neurological measures. To replicate the neuroimaging measures, we generated $\mathbf{x}_k, k = 1, \ldots, m_1$, such that the first 75% of $\mathbf{x}$'s depend on the designed predictor $\mathbf{z}$, and the remainder do not. Specifically, for $i = 1, \ldots, n$ and $k = 1, \ldots, 0.75m_1$, we set $x_{ik} = x^*_{ik} + z_i s_k$, and for $k = 0.75m_1 + 1, \ldots, m_1$, we set $x_{ik} = x^*_{ik}$. Here, s_k are independent uniform $(0.25, 0.5)$ random variables, and x^*_{ik} are generated to reflect different correlation structures between the $\mathbf{x}$'s; that is, we considered (i) x^*_{ik} are independent uniform $(0, 1)$ random variables so that $\mathbf{x}$'s are uncorrelated, and (ii) x^*_{ik} are normally distributed with mean zero and $\operatorname{corr}(x^*_{ik}, x_{i\ell}) = \rho^{|k-\ell|}$ for $\rho = 0.3, 0.5, 0.7$. Higher values of ρ incur larger correlations, and hence increase the problem complexity.

The relationship between the predictors $\mathbf{z}, \mathbf{x}_1, \ldots, \mathbf{x}_{m_1}$ and failure time T is then through the Cox model

$$T = \Lambda_0^{-1}\left[-\log(U)\exp\left\{-\left(\beta_0\mathbf{z} + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \sum_{k=4}^{m_1-1}\beta_k\mathbf{x}_k + \beta_{m_1}\mathbf{x}_{m_1}\right)\right\}\right],$$

where U is uniform(0,1) and the cumulative baseline hazard function is $\Lambda_0^{-1}(t) = t/\gamma$ for $\gamma > 0$. In all settings, we set $\boldsymbol{\beta} = (4.5, 3, -3, -3, \mathbf{0}^{\mathrm{T}}_{m_1-4}, 3)^{\mathrm{T}}$ with $\mathbf{0}_{m_1-4}$ an $(m_1 - 4)$-dimensional vector of zeros, and $\gamma = 0.025$. We also generated independent, uniformly distributed censoring times C to yield 50% censoring. Last, for n and m_1, we considered two settings: the first where $n = 40$ and $m_1 = 100$ (i.e., total number of predictors is $p = m_0 + m_1 = 101$), and the second where $n = 120$ and $m_1 = 60$ (i.e., total number of predictors is $p = m_0 + m_1 = 61$). The former has $n < p$ and the latter has $n > p$ and is reminiscent of the HD study (Section 4) where the number of candidate predictors is large but still less than the sample size.

In summary, we evaluated 8 data settings formed by the combination of the two choices for n, m_1 and four correlation structures for the $\mathbf{x}$'s. For each setting, we generated 500 independent data sets. Based on our data generation, the predictors $\mathbf{x}_1, \ldots, \mathbf{x}_{m_1}$ divide into four distinct categories:

Group 1. $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ depend on $\mathbf{z}$ and act on T;
Group 2. $\mathbf{x}_4, \ldots, \mathbf{x}_{0.75m_1}$ depend on $\mathbf{z}$ and do not act on T;
Group 3. $\mathbf{x}_{0.75m_1+1}, \ldots, \mathbf{x}_{m_1-1}$ neither depend on $\mathbf{z}$, nor act on T;
Group 4. $\mathbf{x}_{m_1}$ does not depend on $\mathbf{z}$, but acts on T.

The ideal variable selection procedure will thus largely select predictors in Groups 1 and 4, and disregard predictors in Groups 2 and 3.

3.2. *Methods evaluated.* We applied 12 different methods to the 8 data generation settings: 3 existing variable selection procedures in the literature, the weighted Lasso with 5 different exclusion frequency-based weights, and the weighted Lasso with 4 other data-driven weights.

3.2.1. *Existing methods.* We considered 3 existing variable selection methods: backward selection, Bolasso [Bach (2008)], and random Lasso [Wang et al. (2011)]. The latter two methods were originally presented for linear models, and we adapted them to a Cox model.

(I) *Backward selection*: (Only applicable when $n > p$). Models with $p-1$ variables are initially fitted (i.e., one variable is removed at a time), and the model with the smallest AIC or BIC is retained. Successive reduced models are fitted applying the same rule until all remaining variables are statistically significant, and the final model has the smallest AIC or BIC. Both AIC and BIC criteria are considered.

(II) *Bolasso*: The Bolasso [Bach (2008)] is a multi-step procedure as follows: (a) Bootstrap the samples B-many times (i.e., randomly select among the rows of the design matrix without replacement). (b) For each bootstrap sample, find the minimizer of equation (2.1) for a fixed λ and weights $w_{\mathbf{z}} = 0$ for $\mathbf{z}$ and $w_1 = \cdots = w_{m_1} = 1$ for $\mathbf{x}_1, \ldots, \mathbf{x}_{m_1}$. This ensures that $\mathbf{z}$ is in the final model. (c) Variable $\mathbf{x}_k$, $k = 1, \ldots, m_1$, is said to be in the final model if $\widehat{\beta}_k^{(b)} \neq 0$ for all $b = 1, \ldots, B$.

Based on the recommendations from Bach (2008), we set $B = 200$ and choose λ in Step (b) using cross-validation.

(III) *Random Lasso*: The random Lasso [Wang et al. (2011)] is also a multi-step procedure that involves two sets of bootstrap replicates. In the first set, importance measures for the coefficients are generated, and, in the second set, variables are selected as follows:

(a) Generate importance measures.
  (i) Bootstrap the samples B-many times (i.e., randomly select among the rows of the design matrix without replacement).
  (ii) For each bootstrap sample, randomly select q_1 candidate variables among $\{\mathbf{x}_1, \ldots, \mathbf{x}_{m_1}\}$, and apply the Cox Lasso, ensuring that $\mathbf{z}$ is always in the final model. Retain estimates $\widehat{\beta}_k^{(b)}$, where the estimate is zero if the corresponding variable was not selected by the Cox Lasso or not among the randomly selected q_1 variables.

(iii) Compute the importance measure for $\mathbf{x}_k$ as $L_k = |\sum_{b=1}^{B} \widehat{\beta}_k^{(b)}/B|$, $k = 1, \ldots, m_1$.

(b) Select variables.

(i) Draw another set of bootstrap samples B-many times (i.e., randomly select among the rows of the design matrix without replacement).

(ii) For each bootstrap sample, randomly select q_2 candidate variables among $\{\mathbf{x}_1, \ldots, \mathbf{x}_{m_1}\}$ proportional to the importance measures $L_1, \ldots, L_{m_1}$. Apply the weighted Cox Lasso with weights also proportional to the importance measures. Retain estimates $\widehat{\beta}_k^{(b)}$, $k = 1, \ldots, m_1$, where the estimate is zero if the corresponding variable was not selected by the weighted Cox Lasso or not among the randomly selected q_2 variables.

(iii) Compute the final estimator $\widetilde{\beta}_k$ as $\widetilde{\beta}_k = \sum_{b=1}^{B} \widehat{\beta}_k^{(b)}/B$. A variable is said to be selected if $|\widetilde{\beta}_k| > 1/n$.

As recommended by Wang et al. (2011), we evaluated different choices of q_1, q_2 similar to the partition sizes used for the exclusion frequency weights (see Table 3). The regularization parameters in the Cox Lasso of Steps (a)(ii) and (b)(ii) were chosen using cross-validation, and we set $B = 200$ in both bootstrap sets.

3.2.2. *Exclusion frequency weights*. We considered 5 different exclusion frequency-based weights constructed as follows:

(IV) *Random partition*, *AIC and BIC*: Variables in $\mathcal{I} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{m_1}\}$ are randomly partitioned into J groups according to the method described in Section 2.2.1. On each partition, a stepwise regression is applied and the selected model is the one yielding smallest AIC or BIC (both criteria are considered).

(V) *Sorted partition*, *AIC and BIC*: Process is the same as in (IV), except that the partitions of $\mathcal{I}$ are obtained by first organizing the ridge regression parameter estimates associated with $\mathbf{x}$'s in order of similar magnitude as described in Section 2.2.2, part (a).

(VI) *Means partition*, *AIC and BIC*: Process is the same as in (IV), except that the partitions of $\mathcal{I}$ are obtained by first dividing the $\mathbf{x}$'s by J'-means clustering of the ridge regression parameter estimates as described in Section 2.2.2, part (b). We take $J' = J$.

(VII) *Quantile partition*, *AIC and BIC*: Process is the same as in (IV), except that the partitions of $\mathcal{I}$ are obtained by first dividing the $\mathbf{x}$'s into J'-quantile groups of the ridge regression parameter estimates as described in Section 2.2.2, part (b). We take $J' = J$.

(VII) *Fixed partition*, *AIC and BIC*: We considered a so-called fixed partition not yet defined: We randomly partitioned the elements in $\mathcal{I}$ as in Section 2.2.1, but we ensure that each partition contains at least one of the truly relevant predictors (i.e., a predictor with nonzero coefficient). Such a structured partition is only for testing purposes because in a real application we do not know the truly relevant predictors.

Exclusion frequencies are formed based on $B = 100$ replicates. When $n = 40, m_1 = 100$, we set $J = J' = 10$, and when $n = 120, m_1 = 60$, we set $J = J' = 6$ so that each partition will have about 10 candidate predictors.

3.2.3. *Other data-driven weights*. Last, we considered the weighted Lasso with unit weights and three other data-driven weights similar to those studied by Garcia and Müller (2014) for linear models.

(IX) *Unit weights*: $w_{m_0+k} = 1, k = 1, \ldots, m_1$. The importance of all candidate predictors is considered equal, and the weighted Lasso may not be able to pick up the more relevant predictors.

(X) *p-Values*: $w_{m_0+k} = p_k$ on $\mathbf{x}_k$, $k = 1, \ldots, m_1$, where p_k are the p-values obtained from the individual Cox regressions of $\mathbf{S}$ on $(\mathbf{z}, \mathbf{x}_k)$ and it will always be clear from the context whether p is referring to a p-value or the number of variables. A statistically significant $\mathbf{x}_k$ tends to have a small p-value and a nonstatistically significant $\mathbf{x}_k$ has a large p-value. Weighing each $\mathbf{x}_k$ with its corresponding p-value will generally lead to including statistically significant $\mathbf{x}$'s in the final model.

(XI) *Benjamini–Hochberg (BH) Adjusted p-Values*: $w_{m_0+k} = p_k^{\mathrm{BH}}$ on $\mathbf{x}_k$, $k = 1, \ldots, m_1$, where p_k^{BH} are the Benjamini–Hochberg [Benjamini and Hochberg (1995)] adjusted p-values obtained from the individual Cox regressions of $\mathbf{S}$ on $(\mathbf{z}, \mathbf{x}_k)$. The BH adjusted p-value accounts for the multiplicity of the m_1 tests compared from the m_1 Cox regressions. Still, the impact of BH adjusted p-values is similar to that for p-values since a statistically significant $\mathbf{x}_k$ will have a small BH adjusted p-value even after the adjustment, and a statistically nonsignificant $\mathbf{x}_k$ will have a large BH adjusted p-value.

(XII) *q-Values*: $w_{m_0+k} = q_k$ on $\mathbf{x}_k$, $k = 1, \ldots, m_1$, where q_k are the q-values [Storey and Tibshirani (2003)] obtained from the individual Cox regressions of $\mathbf{S}$ on $(\mathbf{z}, \mathbf{x}_k)$. Q-values are a monotone transformation of p-values designed to control the false discovery rate. As with p-values and BH adjusted p-values, predictors with small q-values are generally relevant to the model, whereas predictors with large q-values are not.

3.3. *Simulation results*. All methods were evaluated using numerical and graphical measures. For numerical measures, we computed the observed false discovery rate (FDR) and the geometric mean of sensitivity and specificity G [Kubat, Holte and Matwin (1998)] as defined in Section 2.4. The ideal variable selection procedure will have small FDR (ideal is 0) and large G value (ideal is 1).

For graphical measures, we compared the observed average percentages of time variables in Groups 1, 2, 3, and 4 were selected compared to the ideal percentages. (Groups are defined in Section 3.1). Ideally, variables in Groups 1 and 4 are selected 100% of the time, and variables in Groups 2 and 3 are selected 0% of the time. Therefore, the best method has (nearly) null differences between the observed percentages and ideal percentages. In Figure 1 we display curves representing these differences. The curves are called difference curves, and the best

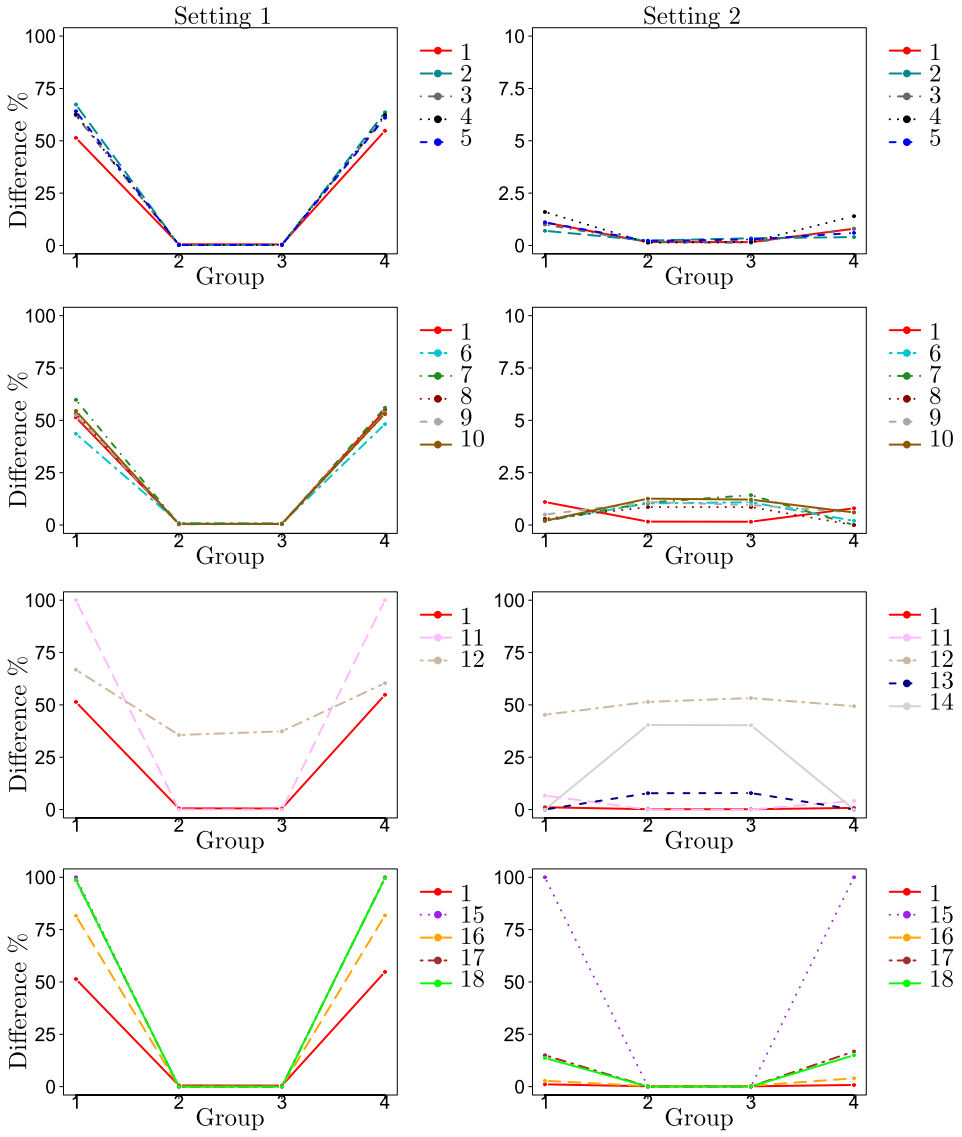

FIG. 1. *Difference curves between ideal and observed percentages of time variables in Groups* 1, 2, 3, *and* 4 *are selected when predictors are uncorrelated. Best method has null differences and thus lowest difference curve. Results shown for exclusion frequency-based weights when stepwise regression uses BIC criterion*; 50% *censoring*, 500 *simulations*, $n = 40$, $p = 101$, $J = J' = 10$ (*Setting* 1) *and* $n = 120$, $p = 61$, $J = J' = 6$ (*Setting* 2). *Curves shown are* 1: J*-quantile partition*, *BIC* (*preferred*); 2: J*-means partition*, *BIC*; 3: *Sorted partition*, *BIC*; 4: *Random partition*, *BIC*; 5: *Fixed partition*, *BIC*; 6: J*-quantile partition*, *AIC*; 7: J*-means partition*, *AIC*; 8: *Sorted partition*, *AIC*; 9: *Random partition*, *AIC*; 10: *Fixed partition*, *AIC*; 11: *Bolasso*; 12: *Random Lasso*; 13: *Backward selection*, *BIC*; 14: *Backward selection*, *AIC*; 15: *Unit weights*; 16: p*-values*; 17: *BH-adjusted* p*-values*; 18: q*-values*.

method corresponds to the lowest curve (i.e., null differences between the observed and ideal percentages in all groups).

We first discuss results when the **x**'s are uncorrelated and then discuss the impact of correlation in Section 3.3.3.

3.3.1. *Choice of exclusion frequency weights*. We first empirically explored the impact of random and structured partitions when forming the exclusion frequency weights in terms of variable selection performance.

Figure 1 (first row) displays difference curves from exclusion frequencies constructed by different partitioning forms with stepwise regression using a BIC criterion applied to the partitions. When $n > p$ (Setting 2), all partitioning forms yield equally good results: all difference curves are less than 2.5%. When $n < p$ (Setting 1), the difference curve corresponding to the J-quantile partitions is lowest. This suggests that exclusion frequencies formed from J-quantile partitions result in largely selecting the relevant predictors in Groups 1 and 4, and ignoring irrelevant ones in Groups 2 and 3. Its performance leads to the highest G-score and a lower false discovery rate among all methods considered (see Table 2). Therefore, among the partitioning forms, our empirical results suggest that J-quantile partitions are preferred.

A second interest of exclusion frequencies is whether the choice of AIC or BIC in the stepwise regression affects the results. It is well known that the AIC generally yields larger models compared to the BIC. As such, we expect AIC-based exclusion frequencies to have more correct selections in Groups 1 and 4, but more incorrect selections in Groups 2 and 3. This phenomenon is indeed observed in Figure 1 (row 2): differences between the ideal and observed percentages are lower in Groups 1 and 4, but the differences are higher in Groups 2 and 3 for all AIC-based curves particularly when $n > p$ (Setting 2). Having AIC-based exclusion frequencies make more incorrect selections in Groups 2 and 3 also produce higher false discovery rates. In Table 2, the AIC-based exclusion frequencies sometimes have false discovery rates 1.5 times larger than those from the BIC-based exclusion frequencies, and a minimal increase in the G-scores. To ensure a conservative solution and still obtain reasonable G-scores, we recommend using BIC-based exclusion frequency weights and, in particular, J-quantile BIC.

3.3.2. *Comparison with competing methods*. Our preferred exclusion frequency weight is formed from J-quantile partitions with BIC stepwise regression. We compared this preferred method to existing methods (Backward selection, Bolasso, random Lasso) and the weighted Lasso with other data-driven weights (unit weights, p-Values, BH adjusted p-Values, and q-values).

Figure 1 (row 3) shows difference curves for our preferred exclusion frequency method and Backward selection, Bolasso, and random Lasso. Whether $n < p$ (Setting 1) or $n > p$ (Setting 2), our preferred method has the lowest difference curve,

TABLE 2

Comparison of variable selection procedures in terms of observed false discovery rate (FDR; low FDR is ideal) and G-score (balance between sensitivity and specificity; high G-score is ideal) for $n = 40$, $p = 101$ (Setting 1) and $n = 120$, $p = 61$ (Setting 2). Covariates are uncorrelated, 50% censoring, 500 simulations. Exclusion frequency weights are formed from $B = 100$ replicates with variables partitioned into $J = J' = 10$ groups (Setting 1) or $J = J' = 6$ groups (Setting 2), and stepwise regression with AIC/BIC criterion applied to partitions. Ideal method is bold-faced and has small FDR and large G-score

		Setting 1		Setting 2	
	Method	**FDR**	**G**	**FDR**	**G**
Exclusion frequency weights	**J-quantile, BIC**	**0.28**	**0.39**	**0.03**	**0.98**
	J-means, BIC	0.16	0.31	0.05	0.98
	Sorted, BIC	0.18	0.33	0.03	0.98
	Random, BIC	0.17	0.33	0.03	0.96
	Fixed, BIC	0.15	0.32	0.04	0.98
	J-quantile, AIC	0.36	0.44	0.17	0.99
	J-means, AIC	0.25	0.3	0.18	1.00
	Sorted, AIC	0.27	0.39	0.14	0.99
	Random, AIC	0.27	0.39	0.17	0.99
	Fixed, AIC	0.25	0.38	0.19	0.99
Existing methods	Bolasso	—	0.00	0.00	0.87
	Random Lasso	0.97	0.30	0.95	0.39
	Backward selection, AIC	NA	NA	0.88	0.94
	Backward selection, BIC	NA	NA	0.60	0.99
Other weights	Unit weight	—	0.00	—	0.00
	p-values	0.07	0.21	0.05	0.93
	BH-adjusted p-values	0.00	0.04	0.01	0.72
	q-values	0.00	0.04	0.01	0.74

indicating that this method closely captures the ideal percentages in Groups 1, 2, 3, and 4.

Backward selection had a reasonably low difference curve when $n > p$ (Setting 2), but was not as competitive as our preferred method and is not even applicable when $n < p$ (Setting 1).

The Bolasso is most competitive to our preferred method particularly when $n > p$ (Setting 2): the Bolasso difference curve is low, but still higher than that from our preferred method. When $n > p$ (Setting 2), the Bolasso has an ideal false discovery rate of 0 and a G-score of 0.87, but the G-score is still lower than our preferred method's G-score of 0.98 (see Table 2, Setting 2). When $n < p$ (Setting 1), the Bolasso does not make any correct selections in Groups 1 and 4, and thus the difference curve is high (Figure 1, row 3) and the resulting G-score is 0 (see Table 2, Setting 1).

TABLE 3

Effect of number of candidate variables selected in the random Lasso; 50% *censoring*, 500 *simulations*, $B = 100$ *replicates*. *Observed false discovery rate* (*FDR*; *low FDR is ideal*) *and G-score* (*balance between sensitivity and specificity*; *high G-score is ideal*) *when* $n = 120$, $p = 61$. *Covariates are either uncorrelated or correlated with* $\text{corr}(\mathbf{x}_i, \mathbf{x}_j) = \rho^{|i-j|}$

Method	Number of variables	Covariates uncorrelated		Covariates correlated					
				$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
		FDR	G	FDR	G	FDR	G	FDR	G
Random Lasso	$q_1 = q_2 = 4$	0.95	0.40	0.95	0.41	0.95	0.39	0.95	0.40
	$q_1 = q_2 = 6$	0.95	0.39	0.95	0.40	0.95	0.39	0.95	0.39
	$q_1 = q_2 = 10$	0.95	0.39	0.95	0.39	0.95	0.39	0.95	0.39
	$q_1 = q_2 = 15$	0.95	0.36	0.95	0.37	0.95	0.37	0.95	0.38
	$q_1 = q_2 = 30$	0.95	0.35	0.95	0.36	0.95	0.36	0.95	0.36

The random Lasso performed poorly when $n < p$ (Setting 1) and $n > p$ (Setting 2). This does not imply that the random Lasso is an inferior method per se. Wang et al. (2011) proposed the random Lasso as a model selection method to cope with two situations, including when the number of truly nonzero regression coefficients is larger than n, which is not our simulation setting. For our simulation setting, the random Lasso chooses predictors in Groups 1, 2, 3, and 4 about 50% of the time on average. This behavior continued regardless of the choice of q_1, q_2 (i.e., the number of candidate predictors randomly selected to generate importance measures and select predictors). From Table 3, the false discovery rate and G-score are steady for different choices of q_1, q_2 when the predictors are correlated or not.

Our preferred method also outperformed the weighted Lasso with other data-driven weights. Table 2 shows that unit weights led to no predictors being selected, and thus weights that ignore the relevance of each predictor (i.e., unit weights) are inferior to weights that account for the relevance. An improvement to unit weights is the p-value, which is a meaningful measure of a predictor's relevance and one that led to the most correct selections among the other data-driven weights considered. P-value weights led to the highest G-score and low false discovery rate (Table 2), as well as the lowest difference curve (Figure 1, row 4). However, our preferred exclusion frequency method still remained preferable: it had higher G-score, lower false discovery rate, and lower difference curve. This suggests that exclusion frequencies capture the relevance of the predictors better than p-values or any of its transformations (i.e., BH-adjusted p-values or q-values).

3.3.3. *Impact of correlation*. When the predictors were correlated, all methods behaved similarly as when the predictors were uncorrelated but with some de-

terioration in performance. The J-quantile BIC exclusion frequency weights continued to perform optimally with the lowest difference curve (not shown) among the choice of exclusion frequency weights. In addition, while the number of partitions, J, J', did not alter the performance of this preferred exclusion frequency method, the increased correlation did lead to higher false discovery rates and lower G-scores as seen in Table 1. As expected, when the predictors were most correlated [i.e., $\text{corr}(\mathbf{x}_i, \mathbf{x}_j) = 0.7^{|i-j|}$], the G-scores were lowest and varying between 0.69 and 0.71. This contrasts considerably from the higher G-scores of 0.97 to 0.98 when the predictors are completely uncorrelated.

Compared to the competing methods of Bolasso, random Lasso, and Backward selection, the preferred exclusion frequency method continued to be competitive as the correlation between predictors increased. Introducing correlation actually resulted in Backward selection unable to converge even when $n > p$ (Setting 2). Correcting this issue would require fine-tuning the implemented convergence/singularity criteria, which is beyond the scope of the paper. We were thus only able to compare the preferred method to Bolasso and random Lasso in Figure 2. Overall, with larger correlation, difference curves for all methods became higher, reflecting the increased difficulty of variable selection under large correlation. Still, when $n < p$ (Setting 1) and $n > p$ (Setting 2), the difference curve for our preferred method was lowest with some overlap with the Bolasso difference curve in Setting 2.

Our method also had lower difference curves than those from the weighted Lasso with other data-driven weights. In summary, introducing correlation between predictors incurred a performance degradation in all methods, with the preferred exclusion frequency method performing optimally overall.

3.3.4. *Thresholding with exclusion frequency weights*. Up to now, we have used exclusion frequencies as weights to drive the selection of the weighted Lasso. One could also consider a variable selected if its corresponding exclusion frequency is less than a certain threshold (i.e., less than 0.15, say)—an approach similar in spirit to thresholding inclusion frequencies [Meinshausen and Bühlmann (2010), Shah and Samworth (2013)]. We report in Table 4 the variable selection results from thresholding the exclusion frequencies at 0.15. The results initially appear promising in that the G-scores were higher than when exclusion frequencies are used as weights in a weighted Lasso (Table 2). However, thresholding had false discovery rates at least double of those when using exclusion frequencies as weights in a weighted Lasso (compare Tables 2 and 4). Similar limitations of thresholding have been previously observed in a linear model context. Garcia et al. (2013) and Garcia and Müller (2014) showed that, for similar false discovery rates, the thresholding procedure often selected Groups 2 and 3 more often than did the weighted Lasso. This suggests then that thresholding is not competitive to using exclusion frequencies as weights in the weighted Lasso.

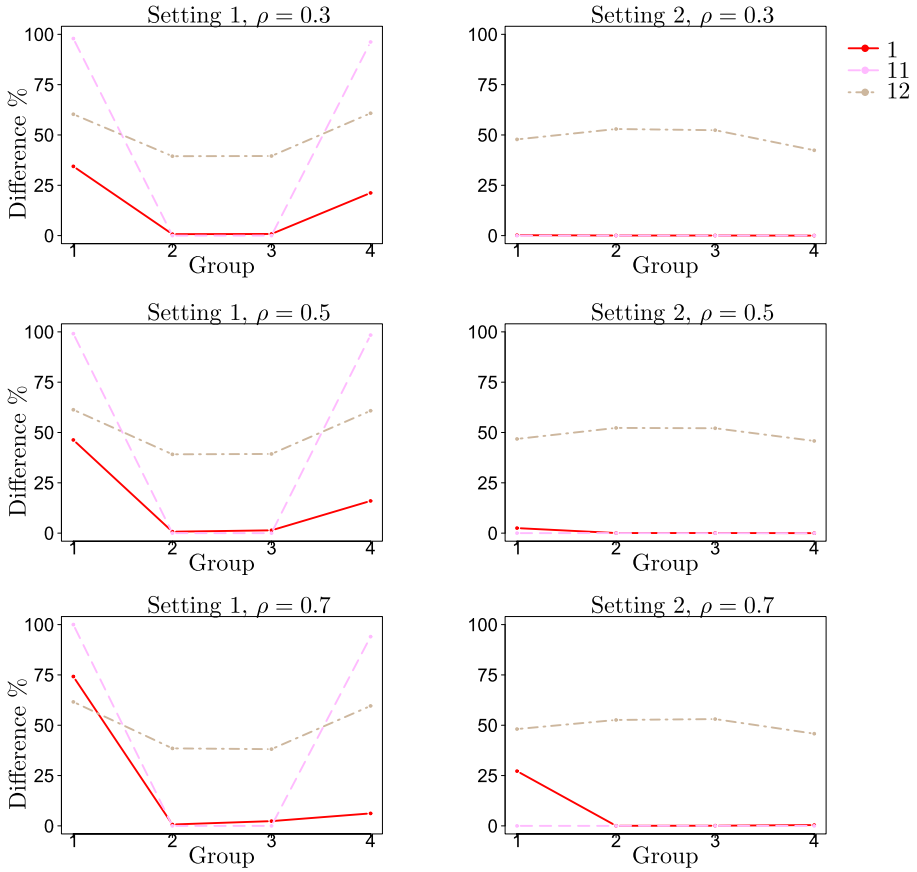

FIG. 2. *Difference curves between ideal and observed percentages of time variables in Groups* 1, 2, 3, *and* 4 *are selected when covariates are correlated with* $\text{corr}(\mathbf{x}_i, \mathbf{x}_j) = \rho^{|i-j|}$. *Best method has null differences and thus lowest difference curve. Results shown for existing variable selection methods compared to preferred exclusion frequency method*; 50% *censoring*, 500 *simulations*, $n = 40$, $p = 101$, $J = J' = 10$ (*Setting* 1) *and* $n = 120$, $p = 61$, $J = J' = 6$ (*Setting* 2). *Curves shown are* 1: J*-quantile partition*, *BIC* (*preferred*); 11: *Bolasso*; 12: *Random Lasso*.

4. Application to PREDICT-HD study of neuroimaging measures for Huntington's disease.

4.1. *Clinical research problem.* We applied our method to PREDICT-HD, a large, 12-year, observational study of potential neurobiological markers of Huntington's disease (HD). The study included $n = 839$ subjects at risk for developing HD (i.e., subjects had an expanded CAG trinucleotide repeat in the huntingtin gene). Among these 839 subjects, 13.9% experienced HD motor onset during the study period (i.e., they developed extrapyramidal signs unequivocally associated with HD), and 86.1% did not (i.e., onset ages were right censored).

TABLE 4

Results from thresholding exclusion frequencies at $\alpha = 0.15$; *predictors uncorrelated*, 50% *censoring*, 500 *simulations*. *Observed false discovery rate* (*FDR*; *low FDR is ideal*) *and G-score* (*balance between sensitivity and specificity*; *high G-score is ideal*) *for* $n = 40$, $p = 101$ (*Setting* 1) *and* $n = 120$, $p = 61$ (*Setting* 2). *Exclusion frequency weights are formed from* $B = 100$ *replicates with variables partitioned into* $J = J' = 10$ *groups* (*Setting* 1) *or* $J = J' = 6$ *groups* (*Setting* 2), *and stepwise regression with AIC/BIC criterion applied to partitions*

		Setting 1		Setting 2	
	Method	**FDR**	**G**	**FDR**	**G**
Thresholding exclusion frequency weights	J-quantile, BIC	0.59	0.60	0.19	0.99
	J-means, BIC	0.54	0.57	0.18	1.00
	Sorted, BIC	0.55	0.59	0.16	0.99
	Random, BIC	0.55	0.58	0.19	0.98
	Fixed, BIC	0.54	0.57	0.13	0.99
	J-quantile, AIC	0.73	0.69	0.56	0.99
	J-means, AIC	0.70	0.70	0.56	0.99
	Sorted, AIC	0.70	0.70	0.54	0.99
	Random, AIC	0.71	0.69	0.56	0.99
	Fixed, AIC	0.70	0.67	0.53	0.99

A key interest in PREDICT-HD is identifying brain regions associated with age of motor onset to better understand HD progression and define biomarkers for future clinical studies. In our modeling context, the outcome of interest is age of motor onset (T), and the predictors are the $m_1 = 352$ volume and surface area measurements of different brain regions taken at baseline. Because subjects enter the study at different disease phases, we included a designed predictor $\mathbf{z}$ to quantify the disease progression. Specifically, $\mathbf{z}$ is the CAG-Age-Product (CAP) score [Zhang et al. (2011)] defined by the product of CAG repeats and age at baseline. The CAP score has been shown to reliably quantify HD progression [Zhang et al. (2011)], and including it in the model ensures calibrating the subjects' disease severity.

Based on the empirical performance in Section 3.2, we applied five variable selection procedures. First, we applied the weighted Lasso with the best empirically performing exclusion frequency weights: those constructed using J-quantile partitions and a BIC criterion. We set $J = J' = 20$ so that each partition contained roughly 17 candidate brain measures, and exclusion frequencies were computed using $B = 100$ replicates. Second, we applied the competing existing methods: Backward selection, Bolasso, and Random Lasso. In Bolasso and Random Lasso, bootstrap replicates were set to $B = 200$ and all regularization parameters were chosen via cross-validation. Third, we applied the weighted Lasso with p-value weights as representative of the most competitive choice of other data-driven weights (Section 3.2.3). Implementation details of these methods are in Section 3.2.

4.2. *Results*. Of the 352 candidate brain measures, our preferred exclusion frequency-based method identified the following as being highly associated with age of motor onset: volumes for the left thalamus proper, left and right caudate nucleus, left and right ventral pallidum, right putamen, and the surface area of the right-hand side temporal transverse gyri.

Backward selection did not select any regions, as the method could not converge. Lack of convergence occurred because of the high correlation between neuroimaging predictors, and correcting this issue would require fine-tuning the implemented convergence/singularity criteria which is beyond the scope of the paper. Bolasso identified one region, the surface area of the right-hand side temporal transverse gyri. This single selection is reflective of Bolasso's conservative performance in that it minimizes the number of false positives but also may ignore relevant features (see Figure 1, row 3). In stark contrast, the Random Lasso selected 76, including the caudate nucleus and putamen. However, based on the empirical performance and high false discovery rates of the Random Lasso, the majority of the 76 regions identified may actually be chosen in error. Last, the weighted Lasso with p-value weights selected two additional regions beyond those identified by our preferred exclusion frequency-based method. The overlapping selections between the weighted Lasso with p-value weights and with the preferred exclusion frequency weights are expected particularly when $p < n$. From Figure 1 (row 4), the difference curves of these two methods are similar when $p < n$, suggesting similar selections will be made as so happens for PREDICT-HD.

The preferred method selecting the thalamus proper, caudate nucleus, ventral pallidum, and putamen is reasonable since these areas are part of the basal ganglia: a region linked to two abilities most affected by HD, motor movements, and cognition [Ross et al. (2014)]. The results agree with those found by Paulsen et al. (2010) and Younes et al. (2014) who also identified the volumes for the thalamus proper, caudate nucleus, and putamen to be associated with age of motor onset. Though Paulsen et al. (2010) did find other associative brain measures (total brain tissue, white matter, cerebral spinal fluid, and cortical grey matter), the discrepancies with our findings are expected since their studies used *estimated* ages of onset [Langbehn et al. (2004)], whereas we use actual ages of onset (or their censored values). In addition, Paulsen et al. (2010) used pairwise t-tests with Bonferroni-corrections to *individually* identify significantly different brain measures between prodromal patients and those not at-risk (i.e., healthy controls). This essentially resorts to thresholding p-values, which has been shown to make more false discoveries (Section 3.3.4).

The caudate nucleus and putamen have been repeatedly found to be associated with age of motor onset [Aylward et al. (2012), Younes et al. (2014)], which agrees with other studies that found atrophy in the caudate nucleus and putamen as the most prominent neuropathological changes in HD [Aylward et al. (2012), Wassef et al. (2015)]. In prodromal HD subjects alone [Hobbs et al. (2010)], significant decreases of these volume measures have been observed over time. Consequently,

volume measures of the caudate and putamen, collectively known as the dorsal striatum, have been proposed as a potential biomarker for HD [Aylward (2007)]. Our findings agree with this proposed suggestion, in addition to considering the three other regions: thalamus, pallidum, and temporal transverse gyri.

To the best of our knowledge, our study is the first to identify the pallidum and transverse gyri to be associated with age of motor onset. Georgiou-Karistianis et al. (2013) recently found that pallidum volumes significantly differentiated presymptomatic HD subjects from healthy controls, but the study did not consider the effects of pallidum volumes on age of motor onset. However, pallidum volumes affecting age of motor onset are feasible since reductions in pallidum volumes are associated with increased clinical severity [Jurgens et al. (2008)] and oculomotor problems [Hicks et al. (2010)], both of which contribute to increased motor impairment. The transverse temporal gyri is in the temporal lobe, a region that has never been specifically associated with disease progression in HD. Further replication and exploration of this region is needed to validate this find.

Our results suggest that potential biomarkers linked to age of motor onset are volume measures of the thalamus proper, caudate nucleus, ventral pallidum, putamen, and the surface area of the temporal transverse gyri. In some cases, our preferred method selected only the right or left portions of these regions, but this may be a consequence of using only L_1 penalization. To date, there are no clinical studies to suggest that only one-sided regions of the brain contribute to HD progression and onset. As such, a future work could be to simultaneously consider the right-left portions of brain measures by using combinations of L_2 and L_1 penalizations in the objective function in (2.1); the analysis would be along the lines of a sparse-group Lasso [Garcia et al. (2014), Simon et al. (2013)].

The analysis conducted was on the third version of PREDICT-HD data. As more data becomes available, other brain regions could be discovered because of larger sample sizes and more brain regions measured. Our results here still provide a meaningful avenue for future clinical investigations, and we will redo our analysis on future data to provide a more current answer.

5. Discussion. We present a novel variation of the weighted Lasso where weights are defined by exclusion frequencies, a data-driven weight formed by repeatedly partitioning the data matrix columns and computing how often each predictor is selected after applying a simple stepwise regression to the partition groups. Our method is shown to be useful for censored data, which is an important complication for which practitioners need appropriate analytic tools, especially when the objective is to identify features in a large p setting. In particular, we showed the utility of our method in a neuroimaging study of Huntington's disease (HD). Our method revealed that the thalamus proper, caudate nucleus, putamen, ventral pallidum, and temporal transverse gyri are associated with age of motor onset after controlling for HD progression through the CAP score. The first three

regions have been validated in earlier clinical studies as having an effect on motor onset, but the ventral pallidum and temporal tranverse gyri have never been identified as effects on motor onset. While these findings could be spurious effects (our method, like all variable selection procedures, are prone to false discoveries), they do stimulate further clinical research with the aim to confirm these two new potential association regions.

We developed our method in the context of a Cox model with L_1 penalization to appropriately handle the neuroimaging data from PREDICT-HD. Our method does extend to generalized linear regression and other penalty types. First, our method applies to other convex, nonlinear models. In this case, the partial log-likelihood in $Q^{\text{cox}}(\boldsymbol{\beta})$ from equation (2.1) is replaced by $-\log \mathcal{L}(\boldsymbol{\beta})$, where $\mathcal{L}(\boldsymbol{\beta})$ is a log-concave likelihood function for some model of interest. Second, our method can be adapted to handle L_2 or SCAD penalties by simply modifying the penalty term in $Q^{\text{cox}}(\boldsymbol{\beta})$.

Our focus for partitioning the columns of the design matrix was to assess the predictors in smaller dimensional problems. Partitioning the data into J partitions is also attractive when p is computationally too large, such as in ultrahigh-dimensional data. Although this article is not about introducing a new screening method, a possible screening method that could be explored in future work is as follows: (1) choosing J such that $p \gg p/J > n$ and that the number of variables in each partition group is of manageable size; (2) applying a regularization and variable selection method to each partition group repeatedly and obtaining exclusion frequency weights; (3) thresholding the variables by keeping those with smallest exclusion frequency for further analysis. Doing so would aid to decrease the ultrahigh-dimensional problem to a high-dimensional one with $p^* > n$ and the p^* predictors being the most relevant among the original p predictors. For this more manageable data set, one could then apply the methods discussed in this paper to the $p^* > n$ problem to identify those features most relevant to the response of interest. While such an approach may not be appropriate for the PREDICT-HD study which has a reasonable sample size and number of neuroimaging measures, this proposed screening could certainly be useful for larger neuroimaging studies with finer measurements, or in large genomic studies.

Last, a limitation of our method is its computational intensiveness in that we are repeatedly analyzing partitioned data. While this does incur a numerical burden, it is also beneficial in that the analyst can learn information from many small-dimensional problems. In our experience, the computational intensity increased with the number of partitions, but, in general, the computation time was similar to that of the Bolasso.

Acknowledgments. Both authors contributed equally to this work.

The authors thank Dr. Elizabeth Aylward (University of Washington) and Dr. Karen Marder (Columbia University) for their invaluable feedback on the neuroimaging results. The authors also thank the Editor and two anonymous referees

for their insightful and constructive feedback which greatly improved the quality of the paper.

Samples and/or data from the PREDICT-HD Study, which receives support from the National Institute of Neurological Disorders and Stroke and was collected by the PREDICT-HD investigators, were used in this study. The authors thank the PREDICT-HD investigators and coordinators who collected data and/or samples used in this study, as well as participants and their families who made this work possible.

REFERENCES

AYLWARD, E. H. (2007). Change in MRI striatal volumes as a biomarker in preclinical Huntington's disease. *Brain Res. Bull.* **72** 152–158.

AYLWARD, E. H., NOPOULOS, P. C., ROSS, C. A., LANGBEHN, D., PIERSON, R. K., MILLS, J. A., JOHNSON, H., MAGNOTTA, V., JUHL, A., PAULSEN, J. S. and THE PREDICT-HD INVESTIGATORS AND COORDINATORS OF THE HUNTINGTON STUDY GROUP (2011). Longitudinal change in regional brain volumes in prodromal Huntington disease. *J. Neurol. Neurosurg. Psychiatry* **82** 405–410.

AYLWARD, E. H., LIU, D., NOPOULOS, P. C., ROSS, C. A., PIERSON, R. K., MILLS, J. A., LONG, J. D., PAULSEN, J. S. and THE PREDICT-HD INVESTIGATORS, AND COORDINATORS OF THE HUNTINGTON STUDY GROUP (2012). Striatal volume contributes to the prediction of onset of Huntington disease in incident cases. *Biological Psychiatry* **71** 822–828. PMID: 21907324, PMCID, PMC3237730.

BACH, F. (2008). Bolasso: Model consistent Lasso estimation through the bootstrap. In *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland. 2008.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BERGERSEN, L. C., GLAD, I. K. and LYNG, H. (2011). Weighted lasso with data integration. *Stat. Appl. Genet. Mol. Biol.* **10** Art. 39, 31. MR2837183

BUCKLAND, S. T., BURNHAM, K. P. and AUGUSTIN, N. H. (1997). Model selection: An integral part of inference. *Biometrics* **53** 603–619.

CHEN, C. H. and GEORGE, S. L. (1985). The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model. *Stat. Med.* **4** 39–46.

COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. MR0341758

HUNTINGTON'S DISEASE COLLABORATIVE RESEARCH GROUP (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72** 971–983.

FAN, J. and LI, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30** 74–99. MR1892656

FARAGGI, D. and SIMON, R. (1998). Bayesian variable selection method for censored survival data. *Biometrics* **54** 1475–1485. MR1671590

GARCIA, T. P. and MÜLLER, S. (2014). Influence of measures of significance based weights in the weighted lasso. *J. Indian Soc. Agricultural Statist.* **68** 131–144. MR3242570

GARCIA, T. P., MÜLLER, S., CARROLL, R. J., DUNN, T. N., THOMAS, A. P., ADAMS, S. H., PILLAI, S. D. and WALZEM, R. L. (2013). Structured variable selection with q-values. *Biostatistics* **14** 695–707.

GARCIA, T. P., MÜLLER, S., CARROLL, R. J. and WALZEM, R. L. (2014). Identification of important regressor groups, subgroups and individuals via regularization methods: Application to gut microbiome data. *Bioinformatics* **30** 831–837.

GEORGIOU-KARISTIANIS, N., SCAHILL, R., TABRIZI, S. J., SQUITIERI, F. and AYLWARD, E. (2013). Structural MRI in Huntington's disease and recommendations for its potential use in clinical trials. *Neurosci. Biobehav. Rev.* **37** 480–490.

GONG, G. D. (1982). Cross-validation, the jacknife, and the bootstrap: Excess error estimation in forward logistic regression. Technical Report 192, Dept. of Statistics, Stanford Univ., 1–82.

GONG, G. (1986). Cross-validation, the jacknife, and the bootstrap: Excess error estimation in forward logistic regression. *J. Amer. Statist. Assoc.* **81** 108–113.

HICKS, S., ROSAS, H. D., BERNA, C., SCAHILL, R., DURMAS, E., ROOS, R. A. et al. (2010). PAW36 oculomotor deficits in presymptomatic and early Huntington's disease and their structural brain correlates. *J. Neurol. Neurosurg. Psychiatry* **81** e33.

HOBBS, N. Z., BARNES, J., FROST, C., HENLEY, S. M. D., WILD, E. J., MACDONALD, K., BARKER, R. A., SCAHILL, R. I., FOX, N. C. and TABRIZI, S. J. (2010). Onset and progression of pathologic atrophy in Huntington disease: A longitudinal MR imaging study. *Am. J. Neuroradiol.* **31** 1036–1041.

IBRAHIM, J. G., CHEN, M.-H. and MACEACHERN, S. N. (1999). Bayesian variable selection for proportional hazards models. *Canad. J. Statist.* **27** 701–717. MR1767142

JURGENS, C. K., VAN DE WIEL, L., VAN ES, A. C. G. M., GRIMBERGEN, Y. M., WITJES-ANE, M. N. W., VAN DER GROND, J. et al. (2008). Basal ganglia volume and clinical correlates in 'pre-clinical' Huntington's disease. *J. Neurol.* **255** 1785–1791.

KUBAT, M., HOLTE, R. C. and MATWIN, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Mach. Learn.* **30** 195–215.

LANGBEHN, D. R., BRINKMAN, R. R., FALUSH, D., PAULSEN, J. S., HAYDEN, M. R. and INTERNATIONAL HUNTINGTON'S DISEASE COLLABORATIVE GROUP (2004). A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clin. Genet.* **65** 267–277.

LIN, W. and LV, J. (2013). High-dimensional sparse additive hazards regression. *J. Amer. Statist. Assoc.* **108** 247–264. MR3174617

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. MR2758523

MÜLLER, S. and WELSH, A. H. (2005). Outlier robust model selection in linear regression. *J. Amer. Statist. Assoc.* **100** 1297–1310. MR2236443

MÜLLER, S. and WELSH, A. H. (2009). Robust model selection in generalized linear models. *Statist. Sinica* **19** 1155–1170. MR2536149

MÜLLER, S. and WELSH, A. H. (2010). On model selection curves. *Int. Stat. Rev.* **78** 240–256.

PAULSEN, J. S., LANGBEHN, D. R., STOUT, J. C., AYLWARD, E., ROSS, C. A., NANCE, M., GUTTMAN, M., JOHNSON, S., MCDONALD, M., BEGLINGER, L. J., DUFF, K., KAYSON, E., BIGLAN, K., SHOULSON, I., OAKES, D., HAYDEN, M. and COORDINATORS OF THE HUNTINGTON STUDY GROUP (2008). Detection of Huntington's disease decades before diagnosis: The Predict HD study. *J. Neurol. Neurosurg. Psychiatry* **79** 874–880.

PAULSEN, J. S., NOPOULOS, P. C., AYLWARD, E., ROSS, C. A., JOHNSON, H., MAGNOTTA, V. A., JUHL, A., PIERSON, R. K., MILLS, J., LANGBEHN, D. and NANCE, M. (2010). Striatal and white matter predictors of estimated diagnosis for Huntington disease. *Brain Res. Bull.* **82** 201–207.

ROSS, C. A. and TABRIZI, S. J. (2010). Huntington's disease: From molecular pathogenesis to clinical treatment. *Lancet Neurol.* **10** 83–98.

ROSS, C. A., PANTELYAT, A., KOGAN, J. and BRANDT, J. (2014). Determinants of functional disability in Huntington's disease: Role of cognitive and motor dysfunction. *Mov. Disord.* **29** 1351–1358.

SAUERBREI, W. and SCHUMACHER, M. (1992). A bootstrap resampling procedure for model building: Application to the Cox regression model. *Stat. Med.* **11** 2093–2109.

SHAH, R. D. and SAMWORTH, R. J. (2013). Variable selection with error control: Another look at stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 55–80. MR3008271

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **39** 1–13.

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.* **22** 231–245. MR3173712

STOREY, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.* **31** 2013–2035. MR2036398

STOREY, J. D. and TIBSHIRANI, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100** 9440–9445. MR1994856

TABRIZI, S. J., REILMANN, R., ROOS, R. A. C., DURR, A., LEAVITT, B., OWEN, G., JONES, R., JOHNSON, H., CRAUFURD, D., HICKS, S. L., KENNARD, C., LANDWEHRMEYER, B., STOUT, J. C., BOROWSKY, B., SCAHILL, R. I., FROST, C., LANGBEHN, D. R. and TRACK-HD INVESTIGATORS (2012). Potential endpoints for clinical trials in premanifest and early Huntington's disease in the TRACK-HD study: Analysis of 24 month observational data. *Lancet Neurol.* **11** 42–53.

TABRIZI, S. J., SCAHILL, R. I., OWEN, G., DURR, A., LEAVITT, B. R., ROOS, R. A., BOROWSKY, B., LANDWEHRMEYER, B., FROST, C., JOHNSON, H., CRAUFURD, D., REILMANN, R., STOUT, J. C., LANGBEHN, D. R. and TRACK-HD INVESTIGATORS (2013). Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: Analysis of 36-month observational data. *Lancet Neurol.* **12** 637–649.

TIBSHIRANI, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* **16** 385–395.

WANG, S., NAN, B., ROSSET, S. and ZHU, J. (2011). Random Lasso. *Ann. Appl. Stat.* **5** 468–485. MR2810406

WASSEF, S. N., WEMMIE, J., JOHNSON, C. P., JOHNSON, H., PAULSEN, J. S., LONG, J. D. and MAGNOTTA, V. A. (2015). T1ρ imaging in premanifest Huntington disease reveals changes associated with disease progression. *Mov. Disord.* **30** 1107–1114.

WITTEN, D. M. and TIBSHIRANI, R. (2010). Survival analysis with high-dimensional covariates. *Stat. Methods Med. Res.* **19** 29–51. MR2744491

YOUNES, L., RATNANATHER, J. T., BROWN, T., AYLWARD, E., NOPOULOS, P., JOHNSON, H., MAGNOTTA, V. A., PAULSEN, J. S., MARGOLIS, R. L., ALBIN, R. L., MILLER, M. I. and ROSS, C. A. (2014). Regionally selective atrophy of subcortical structures in prodromal HD as revealed by statistical shape analysis. *Hum. Brain Mapp.* **35** 792–809.

YU, B. (2013). Stability. *Bernoulli* **19** 1484–1500. MR3102560

ZHANG, H. H. and LU, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **94** 691–703. MR2410017

ZHANG, Y., LONG, J. D., MILLS, J. A., WARNER, J. H., LU, W., PAULSEN, J. S. and THE PREDICT-HD INVESTIGATORS OF THE HUNTINGTON STUDY GROUP, C. (2011). Indexing disease progression at study entry with individuals at-risk for Huntington disease. *Am. J. Med. Genet., Part B Neuropsychiatr. Genet.* **156B** 751–763.

DEPARTMENT OF EPIDEMIOLOGY AND BIOSTATISTICS
TEXAS A&M UNIVERSITY
TAMU 1266
COLLEGE STATION, TEXAS 77845
USA
E-MAIL: tpgarcia@sph.tamhsc.edu

SCHOOL OF MATHEMATICS AND STATISTICS
UNIVERSITY OF SYDNEY
NSW 2006
AUSTRALIA
E-MAIL: samuel.mueller@sydney.edu.au