Published in final edited form as:

Neuroimage. 2011 April 15; 55(4): 1519–1527. doi:10.1016/j.neuroimage.2010.12.028.

Penalized Least Squares Regression Methods and Applications to Neuroimaging

Florentina Bunea¹, Yiyuan She¹, Hernando Ombao², Assawin Gongvatana³, Kate Devlin³, and Ronald Cohen³

¹Florida State University, Department of Statistics

²Brown University, Center for Statistical Sciences

³Brown University, School of Medicine

Abstract

The goal of this paper is to review the most popular methods of predictor selection in regression models, to explain why some fail when the number P of explanatory variables exceeds the number N of participants, and to discuss alternative statistical methods that can be employed in this case. We focus on penalized least squares methods in regression models, and discuss in detail two such methods that are well established in the statistical literature, the LASSO and Elastic Net. We introduce bootstrap enhancements of these methods, the BE-LASSO and BE-Enet, that allow the user to attach a measure of uncertainty to each variable selected. Our work is motivated by a multimodal neuroimaging dataset that consists of morphometric measures (volumes at several anatomical regions of interest), white matter integrity measures from diffusion weighted data (fractional anisotropy, mean diffusivity, axial diffusivity and radial diffusivity) and clinical and demographic variables (age, education, alcohol and drug history). In this dataset, the number P of explanatory variables exceeds the number N of participants. We use the BE-LASSO and BE-Enet to provide the first statistical analysis that allows the assessment of neurocognitive performance from *high* dimensional neuroimaging and clinical predictors, including their interactions. The major novelty of this analysis is that biomarker selection and dimension reduction are accomplished with a view towards obtaining good predictions for the outcome of interest (i.e., the neurocognitive indices), unlike principal component analysis that are performed only on the predictors' space independently of the outcome of interest.

1 Introduction

Biomarker discovery has become a major focus of clinical research directed at a large number of different diseases, including neurological disorders, such as Alzheimer's disease and Stroke. The general strategy of these research efforts is to determine whether particular clinical or laboratory measures can be used to predict clinical outcome, including beneficial response to specific treatments. A common problem in this type of research is that there are

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

often many more potential clinical measures that need to be considered than there are clinical cases in the study, confounding statistical modeling efforts. This is often the case of neuroimaging studies, in which multi-modal neuroimaging data from a large number of brain regions may exist, and the goal is to use these data to predict a clinical outcome. Here, the number (P) of potential explanatory variables can be much larger than the number (N) of participants in the data set (i.e., $P \gg N$), and the application of standard statistical methodology becomes problematic. For instance, one common approach is to select first a manageable number of variables by performing stepwise forward selection or backward elimination, or a combination of the two, in regression models. However, it is well known that such procedures are greedy and may not recover the relevant set of predictors accurately. This drawback is particularly severe when P is large (Tibshirani, 1996).

Another common approach is to perform first dimension reduction on the space of predictors via principal components analysis (PCA), by selecting those components that explain most of the variance, and then fit a regression model of the response on the selected components. There are two major drawbacks of this approach: (i) The principal components are, by construction, linear combinations of the original predictors. Therefore, they may not necessarily yield easily interpretable results. Moreover, if one of the goals of the analysis is to identify a smaller set of predictors to be used in future studies, PCA cannot be used for this purpose: the principal components are, by definition, linear combinations of *all* P original predictors and thus require the collection of all the predictors. (ii) Another major issue, perhaps the most important, remains: most relevant components are selected without any regard for the outcome of interest. For instance, in our application, there is no guarantee that the selected PCs are those relevant for neurocognitive performance because no information on the neurocognitive indices is directly used in reducing the dimension of the space of explanatory variables.

In this article we discuss alternative approaches, based on recent statistical methods tailored specifically to the analysis of large dimensional ($P \gg N$) data. We focus on **penalized least squares** (PLS) methods. The PLS estimators minimize the residual sum of squares plus the penalty term. The aim of PLS is to select, from a large list of possible candidate models, the model that achieves the best trade-off between goodness of fit and model complexity. Here, models that use a smaller number of predictors are said to be of lower complexity than those that use a larger number of predictors. Following standard statistical terminology, models with a low number of predictors are also called sparse. Our goal in this paper is to discuss a number of specific methods that identify sparse models without sacrificing their goodness of fit and prediction accuracy. We will focus on computationally efficient methods. The type of penalty term one employs, convex or non-convex, is responsible for the computational complexity of the resulting PLS method. Since convex penalties have lower computational cost, we will focus on them in this article, with particular emphasis on the LASSO type penalty (Tibshirani, 1996) and its close variant, the Elastic Net (Zou and Hastie, 2005). One important aim of our work is to introduce some variants of these methods (Section 2.3) and to illustrate how they can be used for identifying neuroimaging and clinical biomarkers that best predict neurocognitive performance.

Penalized least squares methods are not new to neuroscience – although they have been previously employed to problems different than the one we treat here. For instance, one of the early applications to neuroimaging is the estimation of electrophysiological sources from electroencephalogram data. We refer the reader to Scherg and Voncramon (1986), Moshier and Leahy (1998), Uutela, et al. (1999), Hämäläinen and Ilmoniemi (1994), Silvaa, et al. (2004), Valdes-Sosa, et al. (2009) and Pascual-Marqui, et al. (1994, 2002), among many others. More recent applications of the PLS, in particular of the LASSO and E-Net methods, include Gunn et al. (2002), for modeling the dynamics of radiotracers, Carroll et al. (2009), for voxel selection in fMRI analysis, Vounou et al. (2010) in multivariate neuroimaging and genetics problems, Ryali et al. (2010) for classification problems.

The recent advent of the LASSO-type methods in many scientific areas and their increased usage by the neuroscience community at large motivated the structure of this work, with contributions outlined below:

- i. We present a detailed overview of the statistical properties of those penalized least squares (PLS) methods that can be used for variable selection and have high prediction accuracy. Our focus is on those methods that are computationally efficient when $P > N$. In this case, we discuss why the traditional methods based on testing are no longer reliable. Moreover, some of the more standard PLS methods, such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC) based methods, fail computationally when $P > 20$. We explain why LASSO-type methods provide a computationally attractive alternative, with sound theoretical foundation. This is presented in Sections 2.1 and 2.2. We also discuss some limitations of these methods in Section 2.3.
- ii. We introduce variants of the LASSO and Elastic Net: the Bootstrap Enhanced Lasso (BE-LASSO) and the Bootstrap Enhanced Elastic Net (BE-Enet). Since predictor selection, by any method, is accompanied by a certain amount of uncertainty, we propose to give a measure for it. We use each method, the LASSO and the Elastic Net, to select a subset of predictors. We then re-sample from the data, repeat the selection process and record the selected set. We repeat this a number of times and summarize the whole selection process by giving the percentage of times each predictor was selected – which we call the variable inclusion probability (VIP). Predictors with high VIP will be investigated further. Full details on the implementation of this procedure are given in Section 2.4.
- iii. We use the BE-LASSO and BE-Enet to provide the first statistical analysis that allows the assessment of neurocognitive performance from *high* dimensional neuroimaging and clinical predictors. The major novelty is that dimension reduction is performed with a view towards obtaining good predictions for the outcomes of interest (i.e., the neurocognitive indices). Prior analyses of the neurocognitive performance were either done in a lower dimensional setting, for instance via AIC/BIC, or involved dimension reduction techniques like PCA, with drawbacks discussed above.

Our analysis revealed that fractional anisotropy at the internal capsule appears as an important predictor. This finding is consistent with other studies (e.g., Shenkin et al., 2003)

that show that higher FA in the anterior limb of the internal capsule and improved response time in a visual target-detection task. However, it is interesting that the BE-LASSO and BE-Enet methods indicate the importance of the quadratic term which suggests a complex non-linear relationships between white matter injury and neurocognitive performance, which would likely not have been identified via the traditional regression methods. The data description, analysis and discussion of the results are the content of Section 3. We present an overall summary of this work in the Conclusions section.

2 Statistical Methodology

2.1 A Review: The LASSO and the Elastic Net Methods

The LASSO method is a particular penalized least squares method. The method consists in computing $\hat{\beta}$ that minimizes the following criterion

$$\sum_{i=1}^N [Y_i - \sum_{j=1}^P \beta_j X_{ij}]^2 + \lambda \sum_{j=1}^P |\beta_j|, \quad (1)$$

where Y_i is the dependent variable for subject i ; X_{ij} denotes a measurement on a predictor X_j for subject i ; and λ is the tuning parameter of the method whose role is to provide a balance between prediction accuracy and sparsity. In the context of neuroimaging, the independent variable Y_i is one of the neurocognitive assessments in the study (e.g., Grooved Pegboard, Trail Making A, Trail Making B, and so on) and the predictors X_{ij} 's include clinical, demographic, volumetric and diffusion tensor imaging variables. Moreover, to anticipate possible non-linearity in the relationship between neurocognitive assessment and the independent variables, X_{ij} 's can include quadratic terms of predictors (e.g., squared values of fractional anisotropy in the internal capsule) and interactions between predictors (e.g., fractional anisotropy in the internal capsule $\times$ mean diffusivity in the corpus callosum). We use the notation $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_j, \dots, \hat{\beta}_P)$ where $\hat{\beta}_j$ is the estimated coefficient of a generic predictor X_j .

Advantages of the LASSO method—The appeal and wide-spread usage of the LASSO for variable selection is justified by the following properties: (i) One remarkable property of the LASSO method is that it yields a sparse estimate $\hat{\beta}$, with some components $\hat{\beta}_j$ exactly equal to zero. The predictors X_j with zero coefficient estimates are discarded from the model, and only the rest are kept. This justifies the usage of the LASSO as a subset selection method; (ii) Another remarkable property of the LASSO is the scope of the method: it can be applied in the challenging situations when one has measured more predictors P than one has subjects N in the study. The theoretical validity of these estimates can be established via probabilistic arguments that are valid for any observed number of predictors P and sample size N . In particular, Bunea, Tsybakov and Wegkamp (2007(a) and (b)), Zhang and Huang (2008), and Meinshausen and Yu (2009) show how the LASSO estimators behave for any given N and P , and even when $P > N$ (iii) LASSO corresponds to a convex minimization procedure. Then, any algorithm that finds this minimum comes with the mathematical guarantee that it detects the global, overall, minimum. For this reason, there exist fast

algorithms for computing the LASSO estimates, for instance, the homotopy methods (Osborne et al. 2000), the LARS (Efron et al. 2004), interior point methods (Kim et al. 2007), and an iterative thresholding algorithm (Daubechies 2004, Friedman et al. 2007) among others. Friedman et al. (2007) provide an R package, glmnet, which is freely available online. We note that these algorithms are fast even when P is much larger than N , for instance P can be of the order N^3 or even larger.

Advantages of the Elastic Net method—When the predictors are highly correlated a further immediate improvement of the LASSO is the Elastic Net (Zou and Hastie, 2005). This corresponds to adding to the LASSO penalty a quadratic penalty term, designed to compensate for the correlation between predictors. Specifically, for given tuning parameters λ and μ , the Elastic Net (E-Net) estimator is the vector $\tilde{\beta}$ which minimizes the criterion

$$\sum_{i=1}^N [Y_i - \sum_{j=1}^P \beta_j X_{ij}]^2 + \lambda \sum_{j=1}^P |\beta_j| + \mu \sum_{j=1}^P \beta_j^2. \quad (2)$$

The minimization problem is still convex, and computationally optimal algorithms exist – see, e.g., the R package elasticnet. The E-Net estimator $\tilde{\beta}$ shares the sparsity properties of the LASSO estimator $\hat{\beta}$. When predictors have high correlations, $\tilde{\beta}$ leads to more accurate predictions of the response than $\hat{\beta}$. The theoretical properties of the E-Net estimate are also well understood, see e.g. Bunea (2008), who suggests caution in the choice of the tuning parameters of the method. In particular, if μ is too large relative to λ , then $\tilde{\beta}$ behaves essentially as the so called ridge regression estimator, $\bar{\beta}$, the minimizer of

$$\sum_{i=1}^N [Y_i - \sum_{j=1}^P \beta_j X_{ij}]^2 + \mu \sum_{j=1}^P \beta_j^2.$$

The ridge estimator $\bar{\beta}$ has no components exactly equal to zero and therefore cannot be used for model selection. However, if both λ and μ in (2) are chosen by cross validation over a carefully selected grid, as described in Section 2.4 below, the E-Net estimator $\tilde{\beta}$ will have the desirable sparsity property. The selection of the two tuning parameters is computationally more complex, but it is a price worth paying for models with highly correlated predictors. In practice, the E-Net estimator is a useful companion of the LASSO estimator and typically one fits both and chooses the one that yields the smallest mean squared error of the fit. We discuss the practical implementation of these estimators, together with our suggested modification in Section 2.4 below.

2.2 Comparison of the LASSO-type methods with other variable selection methods

Before we discuss the practical implementation of the LASSO based methods for our analysis, we first contrast them with other existing variable selection methods, with emphasis on the “Large P , small N ” case.

Variable selection methods based on testing are problematic when $P > N$ —It is well agreed upon in the statistical community that the standard variable selection methods based on hypotheses testing are not appropriate when $P > N$. To give a brief overview of these methods, we recall that the traditional selection methods based on hypothesis testing make use of p -values. To compute the p -values, the practitioner typically needs to : (i) either make exact distributional assumptions on the error terms in the model and perform what is referred to as exact tests — the F -test for comparison of two *nested* models and the t -test are examples of such tests; (ii) or use asymptotic results that lead to approximate distributions on which testing is based. To date, the latter is valid only when $P \leq N$. When $P > N$ such tests cannot be performed, as they involve computing least squares estimators, which are not defined in this case. Therefore, choosing among a large number of candidate models, possibly non-nested, poses a challenge. Even worse, even if approaches (i) and (ii) or their variants can be justified for particular cases, when $P > N$, the open problem of choosing the cut-off for multiple p -values remains: it is highly subjective what one declares as a significant predictor. There exist a number of corrections for the choice of the cut-off of the p -values, the so called False Discovery Rate (FDR) (Benjamini et al. 1995) type methods. However, for models with correlated predictors, they are only theoretically valid when $P \leq N$, sometimes much smaller, see, e.g. Bunea et al. (2006) and the references therein. Also, we add a note of caution on using ad-hoc procedures that would involve, generically, the following steps: (1) Fitting models that contain subsets of the original P variables that have size smaller than N and (2) Computing p -values to assess the importance of predictors in the arbitrarily chosen smaller models. Such procedures cannot be used reliably to select relevant variables. The relative ordering of the p -values depends crucially on which predictors were included in the original model on which the p -values were based. For instance, a generic predictor X_1 may appear significant in a model where only, say, two other predictors, X_2 and X_3 were included, but lose significance if X_4 , say, is added and the model is re-fitted. This is particularly visible if the predictors are correlated.

Other variable selection methods based on penalized least squares—Whereas the LASSO and E-Net penalties are convex, and can be computed efficiently even when P is very large, there exists a large family of the PLS methods that employ non-convex penalties, which may incur a much higher computational cost. An extreme case corresponds to the classic BIC (Schwarz 1976), AIC (Akaike 1978), Mallows' C_p (Mallow 1973) selection criteria and their variants. All these criteria involve fitting all possible 2^P regression models corresponding to all possible subsets of predictors. For each fitted model one computes a criterion that equals the mean squared error of the fitted model plus a term proportional with the number of predictors in that model. Then one selects the final model as the one with the smallest value of this criterion. This class of estimators have attractive theoretical properties see e.g. Bunea, Tsybakov and Wegkamp 2007(a) for an overview. The computation of these estimators involves either an exhaustive search over the space of all 2^P possible models, or approximate forward or backward stepwise procedures to produce a series of candidate models to minimize the criterion. Despite their success in selecting relevant predictors when P is much smaller than N , they fail computationally as soon as $P > 20$, and it has been shown that the problem is NP hard computationally: no combinatorial algorithm can solve this problem in polynomial time.

To address this major computational issue, a number of other non-convex penalties and optimization techniques have been introduced lately, see for instance Antoniadis and Fan (2001) for the SCAD penalty, Zou and Li (2008) for the LLA optimization, She (2009) for the hybrid hard-ridge penalty and the TISP, Zhang (2009) for firm shrinkage and MCP, and the references therein. All these newly developed methods can be implemented efficiently even if P is larger than N . These penalties and the resulting estimators are especially useful for models in which the predictors are very highly correlated. There are many subtle theoretical differences between model selection realized via convex penalties as the LASSO and the above mentioned methods corresponding to these non-convex penalties. However, there is also a major difference. If one employs a non-convex penalty, the criterion function can have many local minima. The research on choosing the best local minimum is still under development. In addition, the computational cost is higher than that of the convex methods as LASSO. It is for this reason that in this paper we only focus and describe further in detail the concrete implementation of the LASSO type methods.

2.3 Limitations of the LASSO and E-Net methods

The performance of the LASSO and E-Net methods depends crucially on the choice of the respective tuning parameters λ and μ of each method. It is known practically (see, e.g. Shi, Lee and Wahba (2007)) and theoretically (see, e.g. Bunea (2008) and Bunea and Barbu (2009)) that if the tuning parameters are chosen via cross-validation, as it will be described in detail in the next section, then the LASSO and E-Net will select a subset of predictors that will predict the response with high accuracy. However, this may not be the most parsimonious subset of predictors one can select: one may include, along with the predictors strongly associated with the response, some that are only weakly correlated, and may in fact be redundant. To understand how this can happen, consider the extreme example of having only two biomarkers, say X_1 and X_2 , associated with the response, although many more biomarkers, say 100, are available for the analysis. Then, if X_1 and X_2 are not highly correlated, the two methods will, with high probability, select X_1 and X_2 , but may also include in the model other irrelevant variables, say X_3 , X_7 and X_{15} . This will likely not inflate the overall prediction of the response by $\hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_7 X_7 + \hat{\beta}_{15} X_{15}$, as the LASSO estimated effect sizes, $\hat{\beta}_3$, $\hat{\beta}_7$ and $\hat{\beta}_{15}$, are typically very small. In fact, since some of these estimated coefficients may have negative signs, and if all predictors have positive values, it is possible that this prediction may be slightly better than predicting the response by using only the truly associated variables X_1 and X_2 . Therefore choosing tuning parameters in order to obtain best prediction may not yield the smallest useful model. This phenomenon is especially pronounced if the sample size is small and the predictors not associated with the response have medium to high correlation. Moreover, if the truly associated variables, X_1 and X_2 , are almost collinear then, by the construction of the LASSO estimator, only one of them will be selected, with high probability, if the sample size is small. The E-Net attempts to correct these drawbacks, but has only marginal success: if the truly associated predictors have medium correlation, then the E-Net will help include all of them in the model, but will still include extra variables, only fewer than the LASSO would.

Another important factor that may preclude the selection via LASSO or E-Net of predictors associated with the response is the effect size of these predictors, sometimes referred to as

the strength of association. It is known that only predictors with effect sizes above the noise level can be detected via these procedures: the noise level can be quantified exactly and is of order $\sqrt{\log(P)/N}$, in regression models, see e.g. Bunea (2008). Therefore, it may be the case that, in the example above, X_1 has a strong effect size, but the one of X_2 is below the noise level. In that case, either LASSO or the E-Net, or indeed any other method, will not select X_2 . Of course, with P kept fixed, if more data are collected, the detectable boundary $\sqrt{\log(P)/N}$ will be smaller, and even weakly associated variables such as X_2 could be detected. It is worth pointing out that this is not a drawback of penalized least squares methods, but a rather general phenomenon, shared by any model selection method, see e.g. Candès and Plan (2009).

In summary, if the LASSO and E-Net with parameters chosen via cross validation are applied to regression problems with *very highly correlated predictors, and the sample size is small relative to P* , then typically they will select a model that includes a few extra variables that are not associated with the response, and may also miss few relevant ones. Note that the latter happens only when those predictors have very weak association with the response, or if some predictors are almost collinear, in which case one may find this feature of the LASSO to be positive, as possibly redundant predictors, that measure essentially the same thing, are eliminated. It is worth mentioning that, if the predictors are only mildly correlated, there exist theoretical choices of the tuning parameters of these two methods that would yield models containing only the predictors truly associated with the response, see e.g. Bunea (2008). However, there is very little guidance as to how to choose these parameters in practice: cross validation will not work, as explained above, and the research on this topic is still open, see e.g. Wasserman and Roeder (2009) for a discussion.

For all these reasons we opted for the procedure outlined in the following section, where we complement the LASSO step by a bootstrap resampling step. This will allow us to measure the frequency with which each predictor is chosen by the LASSO procedure: predictors chosen more frequently will be investigated further. We describe the specific details below.

2.4 Proposed method: Bootstrap enhanced LASSO and E-Net

We describe our algorithm below.

- Step 1** [Model fitting] For each value of λ , perform the following procedure to obtain the (bias-corrected) LASSO estimate $\hat{\beta}(\lambda)$.
- 1a) [LASSO] Fit the LASSO on all *standardized* predictors.
 - 1b) [Bias correction] Fit a restricted OLS model on the selected predictors by the LASSO estimate from 1a), the corresponding estimate denoted by $\hat{\beta}(\lambda)$.
- Step 2** [Parameter tuning] Apply K -fold cross-validation to find the appropriate value λ_o (see Remark 3 below for a description). Record the corresponding selected predictors determined by the nonzero components of $\hat{\beta}(\lambda_o)$.

- Step 3** [Bootstrap] Draw B bootstrap samples from the data. For each bootstrap sample repeat Steps 1-2.
- Step 4** [Summarizing] Record the frequency with which each variable is selected by the LASSO in each bootstrap sample.

Remarks

1. All predictors are required to be mean-centered and then scaled such that the design matrix (without intercept) has all column-means equal to 0 and column-variances equal to 1. This allows for a fair comparison of the relative predictor importance across all explanatory variables.
2. The bias-correction step is based on the idea of ‘LARS-OLS hybrid’ (Efron 2004). The LASSO estimators have shrunk values, relative to the value of the ordinary least squares (OLS) estimators, and are therefore biased. The refitting step, after dimension reduction, corrects this.
3. To avoid overfitting we use K -fold cross-validation to tune the regularization parameter λ . To begin with, consider a fine grid G of possible values of the tuning parameter. For any $\lambda \in G$ we then repeat the following procedure. For a given value of K , typically 5 or 10, we randomly split the data into K subsets of approximately equal size. For each value of k , $1 \leq k \leq K$, run Step 1 on the data without the k -th subset. This step yields a sequence of estimates $\hat{\beta}^{(-k)}(\lambda)$. Then, for each $1 \leq k \leq K$, compute the prediction errors,

$$Err^{(-k)}(\lambda) := \sum_i (Y_i - \sum_{j=1}^P \hat{\beta}_j^{(-k)}(\lambda) X_{ij})^2; \text{ note that they are evaluated on the } k\text{-th subset of the data that had been left out for this purpose. Finally, compute}$$

the summarized cross-validation error curve $CVErr(\lambda) = \sum_{k=1}^K Err^{(-k)}(\lambda)$. The optimal parameter $\lambda_o \in G$ is then given by the value at which $CV-Err(\lambda)$ achieves its minimum.

Cross-validation is perhaps the most popular way for parameter tuning in the literature see, for instance, Tibshirani (1996), Zou and Hastie (2005), Gunn et al. (2002), Carroll et al. (2009), Ryali et al. (2010), although it may be computationally expensive for large-scale data. We comment on the computing time needed for the data analysis of our data in Section 3.3.

4. If a predictor is selected in more than 50% of the bootstrap samples, we declare it important and investigate its impact further. The bootstrap frequency associated with a certain predictor indicates to what extent the predictor is more likely to be included in the model, rather than be excluded, should we repeat the experiment a large number of times. We refer to this frequency as the **variable inclusion probability** (VIP). Of course, the threshold of 50% is user specified, we regard this value as a minimum requirement on investigating further a specific predictor. Since our goal is not to miss any possibly relevant predictors, we have opted for using the 50% conservative threshold.

It is worth mentioning that our bootstrap measure of the variable inclusion probability also has a Bayesian interpretation. The criteria (1) and (2) above, leading to the LASSO and E-Net estimators, can be regarded as the logarithm of the posterior probability distribution corresponding to a Gaussian likelihood with a Laplace prior, or mixed Laplace-Gaussian prior on β_j , respectively. Therefore, under this framework, our bootstrap frequencies can be regarded as estimates of the posterior probabilities $P(\beta_j \neq 0 | \text{Data})$, for each $1 \leq j \leq P$, and therefore of the posterior probability of including the j -th predictor in the model.

5. The Bootstrap-enhanced E-Net procedure for variable selection is similar. The only changes are: (i) replace the LASSO fitting by the E-Net in Step 1a); (ii) use a two-dimensional grid to cross-validate λ and η jointly in Step 2. The detailed procedure is not presented due to limited space.

3 Application to NeuroImaging data

3.1 Background

The need to consider a large number of potential biomarkers is apparent when one considers neurological conditions that produce diffuse brain disturbances that may not be attributable to a single clinical factor. Increasingly, multimodal neuroimaging is being employed in these efforts, resulting in large data sets of structural, functional, physiological, and metabolic measures that need to be considered relative to one another. This poses a methodological challenge, as the analysis of such data can involve a very large number of variables, especially when allowing for the nonlinear transformations of the variables and their interactions. This typically necessitates a selection of variables of interest based on previous literature prior to their inclusion in statistical models, at the expense of ignoring potentially significant variables. Penalized regression enables the analysis of a large number of explanatory variables even in cases where the number of explanatory variables exceed the sample size. To demonstrate such issues, we focus here on multimodal neuroimaging, cognitive and clinical data obtained from a cohort of HIV-infected individuals.

HIV-associated brain dysfunction—HIV infection is often accompanied by neurocognitive dysfunction, typically involving impairment in attention, speed of information processing, motor abilities, executive function, and learning and memory. Neuroimaging studies have demonstrated volume reduction in various brain structures, most notably the basal ganglia, and cerebral white matter abnormalities (Stout, et al., 1988 and McArthur, et al., 2005). To demonstrate the utility of the penalized least squares method, we applied this set of analytical methods on a sample of 62 HIV-infected individuals who were recruited as part of a larger longitudinal NIH-sponsored study of HIV-associated brain dysfunction. Brain volumetric measures and diffusion tensor imaging (DTI) measure of fractional anisotropy (FA), mean diffusivity (MD), axial diffusivity (AD), and radial diffusivity (RD) were analyzed to examine their relationship to HIV-associated neurocognitive deficits.

3.2 Data Example

Description of participants—Sixty-two HIV-infected (HIV+) participants were included in this study. Participants were excluded for history of 1) head injury with loss of consciousness > 10 minutes; 2) history of neurological conditions including dementia, seizure disorder, stroke, and opportunistic infection of the brain; 3) severe psychiatric illness that may impact brain function, e.g., schizophrenia; and 4) current active use of alcohol, cocaine, opiates, or illicit stimulants or sedatives. A significant proportion (39%) of participants had current HCV infection (HCV+). Table 1 shows demographic and HIV clinical information, along with alcohol and substance use history.

Neurocognitive Assessment—Five neurocognitive domains previously shown to be most affected in HIV infection were assessed including 1) attention/working memory (WAIS-III Digit Span and Letter-Number Sequencing), 2) speed of information processing (WAIS-III Digit Symbol and Symbol Search, Trail Making A), 3) psychomotor abilities (Grooved Pegboard), 4) executive function (Trail making B, Controlled Oral Word Association Test), and 5) learning and memory (Hopkins Verbal Learning Test -Revised, Brief Visuospatial Memory Test -Revised). All measures are widely used standardized neuropsychological tests with strong reliability and validity. Individual test scores were converted to demographically corrected T-score using the most updated normative data for each test.

Neuroimaging Data Acquisition—All neuroimaging was performed on one Siemens Tim Trio 3-tesla MRI imager located at Brown University MRI Research Facility. Diffusion weighted images (DWI) were acquired axially for the whole brain with TE/TR = 103/10060 ms, inplane resolution = 1.77mm × 1.77mm, slice thickness = 1.8mm, b-value = 1000 s/mm². One DWI was acquired in each of 64 diffusion gradient directions, in addition to 10 images with no diffusion encoding (b0). Structural MRI images were acquired sagittally using T1 MPRAGE with TE/TR = 3:06=2250 ms, isotropic 0.86-mm voxel size.

Brain Segmentation and Volumetric Measures—Automated brain segmentation was performed on the T1 MPRAGE image using tools from Freesurfer (Fischl, et al., 2004). This process produces volumetric measures of the cortical grey matter, white matter, caudate, putamen, pallidum, thalamus, hippocampus, amygdala, and corpus callosum subregions. For analysis of white matter integrity with DWI, two mask images were extracted, containing 1) cerebral white matter, and 2) subregions of the corpus callosum. Segmented T1 images were transformed to DWI space by applying transformation matrices derived from registering average b0 and T1 images. Means and standard deviations are reported in Table 4.

Diffusion Tensor Estimation—All image registrations were performed using FSL FLIRT tool (Smith, et al., 2004). The 10 b0 images were coregistered using rigid body registration to correct for movement. Registered images were then averaged. Each of the 64 DWIs were then registered to the average b0 image using affine registration to account for movement and eddy current artifacts. Each diffusion gradient direction was then rotated according to the corresponding affine transformation. Diffusion tensor estimation was performed using AFNI 3dDWItoDT tool (Cox, 1996), with the average b0 serving as the

normalization image, yielding the 3 principal eigenvectors and associated eigenvalues characterizing the diffusion ellipsoid. FA, MD, AD, and RD were then computed using standard formulas (Basser & Jones, 2002).

Regions of interest (ROI)—One isotropic 1 cm³ was placed on a standard T1 MNI152 template in each of the frontal and parietal lobes bilaterally. The mask image containing these 4 ROIs was registered to individual brains in DWI space by applying 1) affine transformations between MNI and T1 images, and 2) rigid-body transformations between T1 and average b0 image. Non-white matter voxels were removed using the white matter segmentation derived above. Corpus callosum ROIs derived above were divided into 3 new subregions: genu (anterior subregion from the original segmentation), body (middle 3 subregions), and splenium (posterior subregion). Internal capsule ROIs were placed on 5 consecutive axial slices on the transformed white matter segmentation images. Anterior and posterior regions of the internal capsules were demarcated by an imaginary line across the widest part of the genu on each slice. White matter segmentations of the ROIs as described above were further refined to include only voxels with FA > 0.3. Means and standard deviations of FA and MD values are reported in Table 4.

3.3 Description of the statistical modeling and approach

In our analysis, we fit separate regression models for each neurocognitive performance variable. The explanatory variables included all the clinical/demographic (denoted as **C**), brain volumetric (denoted as **M**) and DTI-derived measures (denoted **D**) along with all volumetrics × DTI interactions. We write

$$\text{neurocognitive} = \mathbf{C} + \mathbf{M} + \mathbf{D} + \text{higher order terms} + \text{Error}. \quad (3)$$

There are 31 “plain” variables of types **C**, **M**, and **D**. In addition to the plain variables, the models contained quadratic DTI and brain volumetric measures and the interactions between HIV stage, DTI variables, and volumetrics. The resulting Model (3) has a total of $P = 234$ predictors for each separate regression model. The number of unknown coefficients is larger than the total number of participants $N = 62$.

As discussed in Section 2.2, the number of predictors is too large to apply an exhaustive search over *all* possible sub-models. Greedy stepwise or sequential methods have to be used, to select among some of the possible models; typically the collection of these sub-models is decided upon in an ad-hoc manner. However, if one decides to pursue this option, in R, the leaps package contains a function regsubsets to generate a sequence of possible models of different sizes, by different means. Nevertheless, an issue still remains, if we want to select among these models using more traditional means: AIC, BIC, C_p will all pick the largest model (the one with the most number of explanatory variables). We refer to Chen and Chen (2008) for a rigorous theoretical justification. We will therefore use in what follows the bootstrap enhanced LASSO (BE-LASSO) and the bootstrap enhanced E-Net (BE-ENet) for selecting the best neuroimaging and clinical biomarkers.

Both methods were tuned by 5-fold cross-validation (CV). The CV running time for the E-Net method, which is the most computationally involved of the two methods, is approximately 12 minutes on an Intel Xeon 3.0GHz server with 8GB of RAM, which is acceptable. The mean squared errors (MSEs) of the two PLS methods are computed. We set aside 30% of the data as a separate test set, and used the remaining to fit the model and tune the regularization parameters. The median test error over the 13 neurocognitive assessments, computed on the test dataset, is 146:3 for the LASSO, and 105:1 for the E-Net. This improvement over the LASSO is due to the additional ridge penalty in the E-Net. Our experience is that even small μ values in (2) can lead to important prediction improvements.

We also fitted a random forest model (Breiman, 2001) to the data. Using bagging (an averaging technique) together with random selection of candidate predictors for splitting, the random forest algorithms construct a combination of trees. Since the thus constructed random forest captures various nonlinear relations and interactions among predictors in an automatic fashion, by construction, it predicts the response well. We used the randomForest package in R for our data and obtained a test error of 120:8. It is important to note that the output of such packages also contains a measure of variable importance. However, as noted in Hastie, Tibshirani, & Friedman (2008, page 593), although the random forest is a powerful predictive tool, it is not ideal if one is also interested in understanding which predictors should be included in the model. Therefore, it is not appropriate to estimate the probability of including a predictor in the model via the variable importance measure provided by random forest packages, as this may be misleading. Indeed, for our data set, random forest analysis did not select Education and hcv_current among the top important predictors. This finding is contrary to an established body of literature that have clearly demonstrated the strong association between education and neurocognitive assessments. Moreover, there is a growing body of literature that support the notion that HCV can injure the brain, as the HCV can infect cells in the central nervous system (CNS). As demonstrated by magnetic resonance spectroscopy, HCV-infected individuals have elevated choline-to-creatine ratios in the basal ganglia and white matter suggesting neuronal loss [see, e.g., Letendre (2008)]. Moreover, the E-Net, which is the method we recommend, has smaller test error (105:1) than the random forest (120.8), and therefore gives better prediction, *and also* can be reliably used for predictor selection.

As described earlier, for each response variable, we computed its bootstrap frequencies as estimates of that variable's inclusion probability. We then created a heatmap to display the inclusion probability of each predictor for the 13 neurocognitive assessments. In Figure 1, the predictors have been reordered based on the median of occurring frequencies for all neurocognitive assessments. The detailed important variables, with bootstrapping frequency > 50%, are given in Table 5 and Table 6.

3.4 Discussion of Results

The variables selected by both BE-LASSO and BE-ENet are similar, indicating that the results are consistent across the two methods. However, we note that the BE-ENet yielded models with more DTI and volumetric biomarkers than the BE-LASSO. Since the predictors in our fitted model are correlated, and since the BE-ENet is especially tailored for such

situations, we advocate its usage for further analyses of this type. The results of our analysis, via both methods, suggests that the models for most of the neurocognitive assessment are sparse – the optimal set of predictors is not large – although the actual number of most important variables differ across the different tests. This suggests redundancy in information across potential explanatory variables.

Among the demographic and clinical biomarkers, the most important predictors are co-infection with Hepatitis-C and education. This finding is consistent with results from other analyses (Gongvatana, et. al., 2010). After the biomarker selection procedure, we fitted regression models that contain as explanatory variables only those that were selected. As expected, performance level increases with education and Hepatitis-C co-infection is linked to lower performance in the neurocognitive function. The bootstrap-enhanced selection methods indicate that explanatory variables derived from neuroimaging data are good predictors for general neurocognitive assessments. However, the degree of importance of neuroimaging biomarkers vary across the different assessments. We note that fractional anisotropy at the corpus callosum and internal capsule and mean diffusivity at the corpus callosum appeared are important predictors for psychomotor skills and executive function, but not for learning and memory and attention/working memory domains. It is clear that while some neuroimaging variables appeared as linear, other explanatory variables enter as quadratic effects. This result underlines the potential impact of PLS in biomarker selection. Due to the “Large P small N ” constraints, it would not have been possible to fit models containing interactions and non-linear effects using standard variable selection techniques. By considering such more complex models, our analysis revealed that larger values of fractional anisotropy and lower values of mean diffusivity at the corpus callosum and the internal capsule are linked to higher performance in the neurocognitive assessment. These results confirm the importance of keeping the integrity of white matter in these areas as crucial to maintaining a high level of functioning in HIV+ patients.

It is also worth pointing out that some morphometric variables play a role in predicting certain neurocognitive functions in the BE-E-Net selection. They appear as interactions (MD at the posterior corpus callosum $\times$ genu volume; and MD at the posterior corpus callosum $\times$ putamen; MD at the posterior corpus callosum $\times$ cortex; and MD at the posterior corpus callosum $\times$ pallidum). Moreover, it enters non-linearly (quadratic effect of the corpus callosum volume) for predicting the Animal_T test.

4 Conclusion

Penalized least squares methods hold considerable promise in neuroimaging biomarker discovery. In this tutorial paper we have discussed in detail two such methods, the LASSO and Elastic Net. We contrasted them with more traditional variable selection methods, that typically fail when $P > N$. We discussed the merits and limitations of the two methods, both from a theoretical and from a computational perspective. We introduced the variants BE-LASSO and BE-Elastic net, that allow the researcher to attach a measure of certainty to each selected predictor. As a demonstration of the utility of these methods, we employed the bootstrap-enhanced LASSO and Elastic net methods in an analysis of clinical and neuroimaging data from a subset of HIV-infected individuals who were part of a larger

ongoing longitudinal study of HIV-effects on neurocognitive function. Using these methods, we were able to demonstrate the value of DTI measures of white matter integrity in predicting neurocognitive functioning among HIV-infected patients. DTI measures of FA and MD obtained from the corpus callosum and internal capsules were consistently retained along with other clinical factors as the strongest predictors of neurocognitive functioning across a number of specific tests of attention-executive, psychomotor and learning performance.

As an overall remark, we note that this general statistical approach is best used as a preliminary step in analyzing a data set having a large number of predictors. The methods discussed here are especially useful when there is no compelling a priori basis for selecting a small set of these variable to be utilized for further focused analyses, in which one could have employed more traditional statistical methods. Then, the bootstrap-enhanced PLS methods can be used reliable to select a reduced set of explanatory variables with high predictive power. Since the reduced set will be typically much smaller than the sample size, one can then conduct focused analyses and perform tests of hypotheses on the selected reduced model. Moreover, PLS can guide the researcher on the conduct of future clinical studies by collecting data only using the most useful predictors and on formulating more precise tests of hypotheses.

To conclude, we provide some new directions on statistical research in this area. Very often, researchers might be interested in analyzing several dependent (response) variables simultaneously rather than conducting separate individual regression models. For example, one might be interested in studying learning and memory as a single domain rather than analyzing each of the neurocognitive tests separately. To address this desiderata, we will develop dimension reduction methods for *multivariate response regression* models in our future research. By refining the method in Bunea, She and Wegkamp (2010), we will develop procedures that select the best biomarker predictors of a collection of, possibly correlated, outcome variables.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Antoniadis A, Fan J. Regularization of Wavelets Approximations. *JASA*. 2001; 96:939–967.
2. Basser PJ, Jones DK. Diffusion-tensor MRI: theory, experimental design and data analysis - a technical review. *NMR Biomed*. 2002; 15(7-8):456–467. [PubMed: 12489095]
3. Benjamini, Yoav, Hochberg, Yosef. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)*. 1995; 57:289–300.
4. Breiman L. Random forests. *Machine Learning*. 2001; 45:5–32.
5. Bunea F. Honest variable selection in linear and logistic models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics*. 2008; 2:1153–1194.
6. Bunea F, Barbu A. Dimension reduction and variable selection in case control studies via regularized likelihood optimization. *Electronic Journal of Statistics*. 2009; 3:1257–1287.

7. Bunea F, Tsybakov AB, Wegkamp MH. Aggregation for Gaussian regression. *The Annals of Statistics*. 2007; 35(4):1674–1697.
8. Bunea F, Tsybakov AB, Wegkamp MH. Sparsity oracle inequalities for the Lasso. *The Electronic Journal of Statistics*. 2007; 1:169–194.
9. Bunea F, Wegkamp MH, Auguste A. Consistent variable selection in high dimensional regression via multiple testing. *J Stat Planning and Inference*. 2006; 136:4349–4364.
10. Candès EJ, Plan Y. Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*. 2009; 37:2145–2177.
11. Carroll MK, Cecchi GA, Rish I, Garg R, Rao AR. Prediction and interpretation of distributed neural activity with sparse models. *Neuroimage*. 2009; 44:112–122. [PubMed: 18793733]
12. Chen J, Chen Z. Extended Bayesian information criterion for model selection with large model space. *Biometrika*. 2008; 95:759–771.
13. Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res*. 1996; 29(3):162–173. [PubMed: 8812068]
14. Daubechies I, Defrise M, De Mol C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*. 2004; 57:1413–1457.
15. Efron B, Hastie T, Johnstone I, Tibshirani R. Least Angle Regression. *Annals of Statistics*. 2004; 32:407–499.
16. Fischl B, Salat DH, van der Kouwe AJ, Makris N, Segonne F, Quinn BT. Sequence-independent segmentation of magnetic resonance images. *NeuroImage*. 2004; 23(Suppl 1):S69–84. [PubMed: 15501102]
17. Friedman J, Hastie T, Hofling H, Tibshirani R. Pathwise coordinate optimization. *Annals of Applied Statistics*. 2007:302–332.
18. Gongvatana A, Cohen R, Correia S, Devlin K, Miles J, Clark U, Westbrook M, Hana G, Kang H, Ombao H, Navia B, Laidlaw D, Tashima K. Impact of Hepatitis C and HIV Coinfection on Cerebral White Matter Integrity. *Neurology*, under review. 2010
19. Gunn R, Gunn S, Turkheimer F, Aston J, Cunningham V. Positron emission tomography compartmental models: A basis pursuit strategy for kinetic modeling. *Journal of Cerebral Blood Flow and Metabolism*. 2002; 22:635–652.
20. Hämmäläinen M, Ilmoniemi R. Interpreting measured magnetic fields of the brain: estimates of current distributions. *Med and Biol Eng Comput*. 1994; 32:35–42. [PubMed: 8182960]
21. Kim, Seung-Jean, Koh K, Lustig M, Boyd S, Gorinevsky D. An Interior-Point Method for Large-Scale L1-Regularized Least Squares. *IEEE Journal of Selected Topics in Signal Processing*. 2007; 1(4):606–617.
22. Letendre S. Neurocognitive effects of HCV, methamphetamine abuse, and HIV: multiple risks and mechanisms. *Annals of General Psychiatry*. 2008; 7(Suppl 1):S39.
23. McArthur JC, Brew BJ, Nath A. Neurological complications of HIV infection. *Lancet Neurology*. 2005; 4:543–555. [PubMed: 16109361]
24. Meinshausen N, Yu B. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*. 2009; 720:246–270.
25. Mosher J, Leahy R. Recursive MUSIC: A framework for EEG and MEG source localization. *IEEE Transactions on Biomedical Engineering*. 1998; 45:1342–1354. [PubMed: 9805833]
26. Osborne MR, Presnell B, Turlach BA. On the LASSO and its Dual. *J Comput Graph Statist*. 2000; 9(2):319–337.
27. Pascual-Marqui R. Standardized low-resolution brain electromagnetic tomography (sLORETA): Technical details. *Methods Find Exp Clin Pharmacol*. 2002; 24:512.
28. Pascual-Marqui R, Michel C, Lehmann D. Low-resolution electromagnetic tomography: A new method for localizing electrical-activity in the brain. *Int J Psychophysiol*. 1994; 18:49–65. [PubMed: 7876038]
29. Ryali S, Supekar K, Abrams DA, Menon V. Sparse logistic regression for whole-brain classification of fMRI data. *Neuroimage*. 2010; 18:752–764.

30. Scherg M, Voncramon D. Evoked dipole source potentials of the human auditory-cortex. *Electroencephalography Clinical Neurophysiology*. 1986; 65:344–360. [PubMed: 2427326]
31. Silva C, Maltez J, Trindade E, Arriaga A, Ducla-Soares E. Evaluation of L1 and L2 minimum norm performances on EEG localizations. *Clinical Neurophysiology*. 2004; 115:1657–1668. [PubMed: 15203067]
32. She Y. Thresholding-based Iterative Selection Procedures for Model Selection and Shrinkage. *The Electronic Journal of Statistics*. 2009; 3:384–415.
33. Shenkin S, Bastin M, MacGillivray T. Childhood and current cognitive function in healthy 80-year-olds: a DT-MRI study. *Neuroreport*. 2003; 14:345–49. [PubMed: 12634481]
34. Shi W, Lee KE, Wahba G. Detecting disease-causing genes by LASSOy-Patternsearch algorithm. *BMC Proceedings* 2007. 2007; 1(Suppl 1):S60.
35. Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*. 2004; 23(Suppl 1):S208–219. [PubMed: 15501092]
36. Stout J, Ellis RJ, Jernigan TL. Progressive cerebral volume loss in human immunodeficiency virus infection: a longitudinal volumetric magnetic resonance imaging study. HIV Neurobehavioral Research Center Group. *Archives of Neurology*. 1998; 55:161–168. [PubMed: 9482357]
37. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *JRSSB*. 1996; 58:267–288.
38. Uutela K, Hämeäläinen M, Somersalo E. Visualization of magnetoencephalographic data using minimum current estimates. *NeuroImage*. 1999; 10:173–180. [PubMed: 10417249]
39. Valdés-Sosa P, Vega-Hernandez M, Sanchez-Bornot J, Martınez-Montes E, Bobes M. EEG Source Imaging With Spatio-Temporal Tomographic Nonnegative Independent Component Analysis. *Human Brain Mapping*. 2009; 30:1898–1910. [PubMed: 19378278]
40. Vounou M, Nichols TE, Montana G. Alzheimer's Disease Neuroimaging Initiative. Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *Neuroimage*. 2010; 53:1147–1159. [PubMed: 20624472]
41. Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, to appear. 2009
42. Zhang C-H, Huang J. The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Annals of Statistics*. 2008; 36:1567–1594.
43. Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. *JRSSB*. 2005; 67:301–320.
44. Zou H, Li R. One-step Sparse Estimates in Nonconcave Penalized Likelihood Models. *Annals of Statistics*. 2008; 36:1509–1533. [PubMed: 19823597]
45. Wasserman L, Roeder K. High-dimensional variable selection. *Annals of Statistics*. 2009; 37:2178–2201. [PubMed: 19784398]

Figure 1.
Elastic net variable selection results (first 10 predictors only).

Table 1

Cognitive variables.

	Mean	SD
WAIS-III Symbol Search	48	7.45
WAIS-III Digit Symbol	45	8.40
WAIS-III Letter Number Sequencing	47	8.95
Grooved Pegboard worse hand	42	12.70
Trail Making A	48	10.20
Trail Making B	50	11.86
COWAT	48	9.66
Animal Fluency	52	8.80
HVLT-R total free recall	40	9.87
BVMT-R total free recall	41	15.03
HVLT-R delayed recall	39	12.82
BVMT-R delayed recall	44	17.66

Note: WAIS-III: Wechsler Adult Intelligence Scale-Third Edition, COWAT: Controlled Oral Word Association Test, HVLT-R: Hopkins Verbal Learning Test-Revised, BVMT-R: Brief Visuospatial Memory Test-Revised.

Table 2

Demographic and clinical characteristics.

Age (years)	44 ± 10.35
Education (years)	13 ± 2.04
% Male	68
% Caucasian	60
HIV infection duration (years)	12 ± 6.97
Nadir CD4 (cells/ μ l)	182 ± 171.03
Current CD4 (cells/ μ l)	431 ± 208.41
% with undetectable plasma HIV RNA	71
% with current hepatitis C	37
% with alcohol history	53
% with cocaine history	47
% with opiate history	16

Table 3

Brain volumetric measures.

Volume (cm ³)	Mean	SD
Cortical grey matter (left)	229,494	29,387.14
Cortical grey matter (right)	231,680	29,189.32
White matter (left)	213,474	30,473.23
White matter (right)	215,733	31,411.44
Caudate (left)	3,540	606.80
Caudate (right)	3,585	590.15
Putamen (left)	5,650	695.41
Putamen (right)	5,362	681.33
Pallidum (left)	1,781	270.66
Pallidum (right)	1,546	233.01
Thalamus (left)	6,524	925.35
Thalamus (right)	6,707	948.85
Hippocampus (left)	3,978	504.19
Hippocampus (right)	4,089	508.65
Amygdala (left)	1,646	235.22
Amygdala (right)	1,741	227.72
Corpus callosum: genu	895	157.38
Corpus callosum: body	1,308	300.39
Corpus callosum: splenium	929	144.49

Table 4

Diffusion tensor imaging measures of white matter integrity.

	FA Mean(SD) in mm ² /s	MD Mean(SD) in mm ² /s
Frontal white matter (left)	0.4255 (0.0236)	7.2053×10^{-04} (3.5022×10^{-05})
Frontal white matter (right)	0.4216 (0.0219)	7.2969×10^{-04} (3.3241×10^{-05})
Parietal white matter (left)	0.4597 (0.0196)	7.714×10^{-04} (4.2674×10^{-05})
Parietal white matter (right)	0.4479 (0.0206)	7.7381×10^{-04} (4.6688×10^{-05})
Anterior internal capsule (left)	0.5758 (0.0315)	6.8639×10^{-04} (3.3727×10^{-05})
Anterior internal capsule (right)	0.5670 (0.0255)	7.1386×10^{-04} (3.5269×10^{-05})
Posterior internal capsule (left)	0.6077 (0.0385)	6.9845×10^{-04} (3.9734×10^{-05})
Posterior internal capsule (right)	0.6315 (0.0338)	7.1987×10^{-04} (4.1889×10^{-05})
Corpus callosum: genu	0.6648 (0.0369)	8.4082×10^{-04} (6.6097×10^{-05})
Corpus callosum: body	0.6302 (0.0360)	9.7395×10^{-04} (8.6038×10^{-05})
Corpus callosum: splenium	0.7312 (0.0367)	8.0550×10^{-04} (4.8342×10^{-05})

Table 5

Predictors with bootstrapping frequency > 50% based on the bootstrapped enhanced LASSO.

Responses	Important predictors
HVLT_sum_T	kmsk_cocopi, Education
HVLT_delay_T	Education, hcv_current, kmsk_cocopi, kmsk_alc
BVMT_sum_T	kmsk_cocopi, Education, fa_cc1-squared
BVMT_delay_T	Education, kmsk_cocopi, fa_cc1-squared
WAIS_LNS_T	Education
WAIS_DigSym_T	kmsk_cocopi, kmsk_alc, fa_cc5-squared, md_cc234 $\times$ cortex
WAIS_SymSrch_T	kmsk_cocopi, md_cc1-squared, hcv_current
Gpeg_dom_T	fa_ic2-squared, kmsk_alc, hcv_current, md_cc234-squared, fa_cc1
Gpeg_nondom_T	fa_ic2-squared
Trail_A_T	kmsk_cocopi, md_cc234-squared, kmsk_alc
Trail_B_T	md_cc1-squared
COWAT_T	hcv_current
Animal_T	Education, md_ic1-squared

Table 6

Predictors with bootstrapping frequency > 50% based on the bootstrapped-enhanced E-Net.

Responses	Important predictors
HVLT_sum_T	kmsk_cocopi, Education, md_cc234 $\times$ cc_genu, kmsk_alc
HVLT_delay_T	Education, hcv_current, kmsk_cocopi, kmsk_alc
BVMT_sum_T	Education, kmsk_cocopi, fa_cc1-squared, hcv_current
BVMT_delay_T	Education, kmsk_cocopi, fa_cc1-squared, hcv_current, md_cc234 $\times$ putamen
WAIS_LNS_T	Education, fa_cc234-squared
WAIS_DigSym_T	kmsk_cocopi, kmsk_alc, md_cc5-squared, Education, fa_cc5-squared
WAIS_SymSrch_T	kmsk_cocopi, hcv_current, md_cc1-squared, Education
GPeg_dom_T	hcv_current, fa_ic2-squared, kmsk_alc, md_cc234-squared, fa_cc1
GPeg_nondom_T	fa_ic2-squared, kmsk_cocopi
Trail A_T	kmsk_cocopi, md_cc234-squared, kmsk_alc, hcv_current, md_cc234 $\times$ cortex, md cc234 $\times$ pallidum
Trail_B_T	md_cc1-squared, fa_ic2-squared
COWAT_T	hcv_current
Animal_T	Education, fa_ic2-squared, cc_body-squared