Published in final edited form as:

Biometrics. 2014 March ; 70(1): 153–163. doi:10.1111/biom.12113.

Variable Selection and Prediction Using a Nested, Matched Case-Control Study: Application to Hospital Acquired Pneumonia in Stroke Patients

Jing Qian^{1,2,*}, Seyedmehdi Payabvash³, André Kemmling³, Michael H. Lev³, Lee H. Schwamm⁴, and Rebecca A. Betensky²

¹Division of Biostatistics and Epidemiology, University of Massachusetts, Amherst, MA, 01003

²Department of Biostatistics, Harvard School of Public Health, Boston, MA, 02115

³Department of Radiology, Massachusetts General Hospital, Boston, MA 02114

⁴Department of Neurology, Massachusetts General Hospital, Boston, MA 02114

Summary

Matched case-control designs are commonly used in epidemiologic studies for increased efficiency. These designs have recently been introduced to the setting of modern imaging and genomic studies, which are characterized by high-dimensional covariates. However, appropriate statistical analyses that adjust for the matching have not been widely adopted. A matched case-control study of 430 acute ischemic stroke patients was conducted at Massachusetts General Hospital (MGH) in order to identify specific brain regions of acute infarction that are associated with hospital acquired pneumonia (HAP) in these patients. There are 138 brain regions in which infarction was measured, which introduce nearly 10,000 two-way interactions, and challenge the statistical analysis. We investigate penalized conditional and unconditional logistic regression approaches to this variable selection problem that properly differentiate between selection of main effects and of interactions, and that acknowledge the matching. This neuroimaging study was nested within a larger prospective study of HAP in 1915 stroke patients at MGH, which recorded clinical variables, but did not include neuroimaging. We demonstrate how the larger study, in conjunction with the nested, matched study, affords us the capability to derive a score for prediction of HAP in future stroke patients based on imaging and clinical features. We evaluate the proposed methods in simulation studies and we apply them to the MGH HAP study.

Keywords

AUC; Cerebral infarction; Conditional logistic regression; Elastic net; Lasso; Penalized likelihood; ROC analysis

1. Introduction

Hospital acquired pneumonia (HAP) is a major complication after stroke, and is associated with higher mortality, larger neurological deficits, longer hospitalization and increased costs for medical care (Davenport et al., 1996; Katzan et al., 1998). A study was conducted in a cohort of 1915 acute ischemic stroke patients admitted to Massachusetts General Hospital

*qian@schoolph.umass.edu.

SUPPLEMENTARY MATERIALS Web Appendices A, B, C, D and E referenced in Sections 5 and 6 are available with this paper at the *Biometrics* website on Wiley Online Library.

(MGH) between June 2004 and March 2008, for which the outcome of interest (HAP), demographic information and clinical variables were obtained for all participants. The prevalence of HAP among this cohort was 12%. The goal of the study was twofold: (1) to identify specific brain regions of acute infarction that are associated with HAP in acute ischemic stroke patients and (2) to derive a prediction rule based on imaging and clinical variables for early assessment of the risk of HAP in new patients. However, as it would have been prohibitively expensive and time-consuming to conduct neuroimaging studies for the entire cohort, a nested matched case-control study design was conducted. In this substudy of acute stroke patients, 215 with HAP were matched with 215 without HAP, based on gender, the National Institutes of Health stroke scale (NIHSS) and age. Neuroimaging analysis was undertaken for these matched pairs. The resultant 138 imaging variables, along with their nearly 10,000 two-way interactions, which are important to consider given spatial correlations in the brain, present a challenging high-dimensional variable selection problem. Another challenge is how to optimally leverage the information in the prospective cohort to use alongside that in the nested study, for development of a powerful prediction rule for future patients.

Many high-dimensional variable selection methods within the context of likelihood-based estimation have been developed, including penalized regression approaches, such as lasso (Tibshirani, 1996) and elastic net (Zou and Hastie, 2005); an overview is provided by Fan and Lv (2010). Additional issues arise when interaction terms are also of interest, including substantially increased dimensionality and desired differential weighting of main effects and interaction effects. Wu et al. (2009) developed a lasso penalized logistic regression model with a cyclic coordinate ascent algorithm to select both main and interaction predictors in the context of genome-wide association studies. Wu et al. (2010) proposed an alternative lasso penalization procedure called “screen and clean” for the purpose of identifying and then “cleaning” interactions in case-control genome-wide association studies. Neither of these proposals addressed matched study designs.

Matched case-control designs are commonly used in epidemiologic studies for increased efficiency by enforcing similar distributions of confounding variables for cases and controls (Rothman et al., 2008). These designs have recently been introduced to imaging and genomic studies, which are characterized by high-dimensional covariates. However, with few exceptions, statistical analyses that undertake high-dimensional feature selection do not appropriately adjust for the matching (e.g., Davatzikos et al., 2008; Duchesne et al., 2009). We identified only three papers that do acknowledge the underlying matched designs. Tan, Thomassen, and Kruse (2007) proposed a two-stage variable selection procedure for matched case-control microarray data, where in the first stage, important features are identified using a modified paired t -test statistic, and in the second stage, a support vector machine classifier is built using these selected features. The second stage classification ignores the matched design, which may lead to biased results (Breslow and Day, 1980). Adewale, Dinu, and Yasui (2010) developed two variants of boosting for classification of matched-pair or correlated binary responses with high-dimensional predictors. One is based on the generic functional gradient descent boosting algorithm and employs a loss function that handles correlation in binary data. The other is a likelihood-based boosting via generalized linear mixed models that handles the correlation via a random effect. This approach entails strong modeling assumptions. Balasubramanian et al. (2013) proposed a random forest penalized conditional logistic regression algorithm, which adjusts for the matched case-control design through the conditional likelihood and incorporates certain attractive features of the random forest machinery to assess variable importance in a high dimensional feature space.

When potential risk factors are expensive, case-control studies are often nested within larger cohort studies and the measurements are taken only from the subjects in the nested studies. This design is cost-effective and efficient (Lin and Ying, 1993; Woodward, 2005), and is widely used in large epidemiology cohort studies such as the Nurses' Health Study (Tworoger et al., 2011). They are even more important in the context of very high dimensional and expensive imaging and genetic variables, such as the MGH HAP study.

With nested, matched case-control data, we also seek to derive a prediction rule for the outcome (e.g., HAP) in future patients. It is well known that if a logistic regression model holds for a prospective sample, then it holds as well for a retrospective case-control sample, albeit with a different log-odds parameter (Mantel, 1973; Prentice and Pyke, 1979). The difference in the log-odds parameters between the retrospective and prospective models is a function of the sampling probabilities of the cases and controls. Thus, if the sampling probabilities are known, a risk prediction model for future subjects can be constructed simply by fitting a logistic regression model to the retrospective case-control sample and adjusting the intercept. With matched case-control data, the adjustment must be conditional on the matching variables. The papers that propose techniques for high dimensional variable selection based on matched retrospective studies do not address the problem of prediction.

In this paper, we propose and assess penalized variable selection strategies to select important main effects and two-way interaction effects generated by the imaging data in the MGH HAP matched case-control study. Furthermore, we extend the results of Mantel (1973) and Prentice and Pyke (1979) that elucidated the link between the prospective and retrospective logistic regression models to our setting of the nested, matched case-control study. In Section 2, we specify population-level and sampling models for HAP, and describe two procedures for our dual goals of variable selection and prediction. In Section 3, we develop a two-stage procedure for variable selection based on a penalized conditional likelihood for the nested, matched case-control study. We then derive a prediction rule for future subjects based on the selected imaging variables and the matching variables by leveraging the larger study from which the case-control study was sampled. In Section 4, we develop a simultaneous procedure for variable selection and prospective modeling. We report the results of simulation studies in Section 5, and we apply the proposed methods to the MGH HAP study in Section 6. We end with a discussion in Section 7.

2. Notation, models and sampling

Let Y be binary response variable ($Y = 1$ or 0 for presence or absence of HAP), $\mathbf{X}_{main} = (X_1, \dots, X_p)^T$ the vector of predictors, $\mathbf{X}_{int} = (X_i X_j, 1 \leq i < j \leq p)^T$ the vector of two-way interactions of the predictors in $\mathbf{X}_{main}$, and $\mathbf{Z} = (Z_1, \dots, Z_q)^T$ the vector of matching variables. We assume that a logistic regression model holds on the population level,

$$\text{logit}\{\Pr(Y=1|\mathbf{X}, \mathbf{Z})\} = \alpha + \beta^T \mathbf{X} + \gamma^T \mathbf{Z}, \quad (1)$$

where $\mathbf{X} = (\mathbf{X}_{main}^T, \mathbf{X}_{int}^T)^T$, $\beta = (\beta_1, \dots, \beta_p, \beta_{1.2}, \beta_{1.3}, \dots, \beta_{(p-1).p})^T$, $\gamma = (\gamma_1, \dots, \gamma_q)^T$. In high dimensional settings, it is likely that several of the predictors are not associated with the response, and thus several components of β equal 0. In addition, among the set of $\mathbf{X}_{main}$ with nonzero coefficients, only a subset of their two-way interactions have nonzero coefficients.

As for the MGH HAP study, N subjects are prospectively enrolled into the study, and the response variable, Y , and strongly associated clinical variables, $\mathbf{Z}$, are recorded for each subject. Due to limited resources, and with the goal of efficiency, the predictors of interest, $\mathbf{X}_{main}$, are not measured for all N subjects. Instead, a nested, matched case-control study is

conducted. In particular, n pairs of subjects from this cohort are sampled, such that each pair contains one case ($Y = 1$) and one control ($Y = 0$) which are matched on the important clinical variables, $\mathbf{Z}$. The observed data for this substudy are $\{(Y_{kj}, \mathbf{X}_{kj}, \mathbf{Z}_k), k = 1, \dots, n, j = 1, 2\}$, with the index $j = 1$ indicating case and $j = 2$ indicating control. The standard conditional likelihood approach for analysis of these data enables estimation of β , but not of α or γ , which would be needed for prediction of HAP in future patients.

Let $V = 1$ indicate that a subject is sampled into the matched case-control study and $V = 0$ otherwise. We assume that given $(Y, \mathbf{Z})$, the sampling probability does not depend on $\mathbf{X}$, i.e., $\Pr(V = 1|Y, \mathbf{X}, \mathbf{Z}) = \Pr(V = 1|Y, \mathbf{Z})$, which is usually satisfied in practice. Under the one-to- m matched case-control design, in which $\Pr(Y = 1|V = 1, \mathbf{Z}) = \Pr(Y = 0|V = 1, \mathbf{Z})/m$, it follows that

$$\frac{\Pr(Y=1|\mathbf{X}, \mathbf{Z}, V=1)}{\Pr(Y=0|\mathbf{X}, \mathbf{Z}, V=1)} = \frac{\Pr(Y=1|\mathbf{X}, \mathbf{Z})}{\Pr(Y=0|\mathbf{X}, \mathbf{Z})} \times \frac{1}{m} \times \frac{\Pr(Y=0|\mathbf{Z})}{\Pr(Y=1|\mathbf{Z})}.$$

Therefore, it follows from (1) that the probability of HAP among those stroke patients sampled for the case-control study is

$$\begin{aligned} \text{logit} \{ \Pr(Y=1|\mathbf{X}, \mathbf{Z}, V=1) \} &= \alpha + f(\mathbf{Z}) + \beta^T \mathbf{X} + \gamma^T \mathbf{Z}, \\ \text{where } f(\mathbf{Z}) &= \log \left\{ \frac{\Pr(Y=0|\mathbf{Z})}{\Pr(Y=1|\mathbf{Z})} \right\} - \log m. \end{aligned} \quad (2)$$

Note that $f(\mathbf{Z})$ is estimable from the prospective study from which the matched case-control study was sampled; we provide details in Section 6.1. Also, $f(\mathbf{Z})$ is not of use for the prediction of HAP in future stroke patients. Rather, it enables us to estimate α and γ using the nested matched study, and these parameters are necessary for prediction for future patients.

This sampling model (2) suggests two approaches that will achieve the dual goals of variable selection and prediction. A two-stage approach entails estimation of β using the conditional likelihood for the matched pairs, and then use of $\hat{f}(\mathbf{Z}) + \hat{\beta}^T \mathbf{X}$ as an offset in an *unconditional* logistic regression model for estimation of α and γ . The second approach does not utilize conditional logistic regression, but rather fits an unconditional logistic regression model for the matched case-control data, using $\hat{f}(\mathbf{Z})$ as an offset. This enables simultaneous estimation of α , β , and γ . The advantage of the first approach is that estimation of β is not compromised by variability in $\hat{f}(\mathbf{Z})$ or by instability induced by simultaneous estimation of α and γ . The advantage of the second approach is that it is a single stage approach. Without loss of generality, we assume one-to-one matching in the remainder of this paper.

3. Two-stage Procedure for Variable Selection and Prediction

For the two-stage approach, we first conduct variable selection, and estimate β using the conditional likelihood appropriate for the matched case control study, based on (1):

$$L_c(\beta) = \prod_{k=1}^n \left\{ \frac{\exp(\beta^T \mathbf{X}_{k1})}{\exp(\beta^T \mathbf{X}_{k1}) + \exp(\beta^T \mathbf{X}_{k2})} \right\}. \quad (3)$$

3.1 Variable selection

Stepwise selection is commonly used for variable selection. However, it is well known to overfit the data. Logic regression (Ruczinski et al., 2003) was designed for selection of binary predictors, though can be extended to continuous ones through construction of categorical variables with associated indicators. It finds optimal Boolean combinations of predictors within the context of any likelihood function, but has downside of being computationally expensive, even when a relatively efficient simulated annealing algorithm is adopted. Thus, we focus on penalized likelihood methods. We compare the performance of our proposed methods to stepwise and logic regression for the MGH HAP study in Section 6.

We define the penalized log likelihood as

$$\log L_C(\beta) - \sum_{i=1}^p g_{\lambda_1}(|\beta_i|) - \sum_{i=1}^p \sum_{j>i}^p g_{\lambda_2}(|\beta_{i,j}|),$$

where $\log L_C(\beta)$ is the log conditional likelihood given in (3), $g_{\lambda_1}(\cdot)$ is a penalty function for the main effects and $g_{\lambda_2}(\cdot)$ is a penalty function for the two-way interaction effects. These functions are indexed by tuning parameters $(\lambda_1, \lambda_2) \geq 0$. The Lasso penalty defines $g_{\lambda}(\theta) = \lambda|\theta|$, and the ridge defines it as $\lambda|\theta|^2$. The lasso penalty serves to shrink the regression coefficients toward zero; the larger the value of λ , the greater the amount of shrinkage. As λ increases, the lasso regression will set some of the coefficients to be exactly zero and obtain a subset of covariates with nonzero regression coefficients. The ridge penalty serves to shrink the regression coefficients toward zero and each other by imposing a penalty on their size, but does not reduce the number of covariates with nonzero coefficients. The elastic net procedure combines the lasso and the ridge, and implements variable selection, like the lasso, and also shrinks together the coefficients of strongly correlated predictors, like the ridge. Thus, strongly correlated predictors tend to be in or out of the selected model together. This is potentially advantageous over the lasso when some true predictors are highly correlated, as are some of the imaging variables, and it is of scientific interest to identify them all. When variable selection is an important aim, both lasso and elastic net penalization are of interest. In studies with predefined groupings of variables that are known to operate jointly on the outcome, the group lasso (Meier et al., 2008) could be used.

Due to the large number of imaging variables (138), the much larger number of two-way interactions between them, and our desire to weight the main effects and interaction effects differentially, we consider three different penalization strategies, termed “Pen1”, “Pen2” and “Pen3” when based on the lasso, and “EN1”, “EN2” and “EN3” when based on elastic net:

Pen1/EN1 : This strategy does not include any interactions. We fit penalized conditional logistic regression models with main effects, $\mathbf{X}_{main}$, only.

Pen2/EN2 : This strategy considers interactions of pre-selected main effects. First, we fit univariate conditional logistic regression models with each of the predictors, $\mathbf{X}_{main}$, and select the predictors whose p -values are less than a pre-selected threshold. Second, we fit conditional multiple logistic regression models with a penalty on the two-way interactions of the selected main predictors.

Pen3/EN3 : This is a variation on Pen2/EN2 that involves joint selection of main effects using PEN1/EN1, followed by penalized selection of interactions of selected main effects. We also considered simultaneous selection of main effects and interactions, with two different penalty parameters to allow for differential weighting of the two

types of terms, but encountered numerical instabilities in this approach that made it infeasible.

3.2 Cross-validation

We employ ten-fold cross-validation to select the optimal penalty parameters, and to evaluate the performance of proposed variable selection strategies (using the corresponding optimal penalty parameters). We randomly split the original n pairs of case-control observations into 10 groups. Using nine out of ten groups as the training dataset, we implement each of the three proposed procedures. We then apply the final fitted regression models to the group that was omitted from the training set and calculated the (unpenalized) conditional log likelihood for that subset of pairs. We repeat this procedure ten times so that each group of pairs served as the validation set once. The sum of all of the conditional log likelihood contributions from the validation groups is the cross-validation score. The procedure with the highest score at its optimal penalty parameter has the best performance. Variable selection is completed by maximizing the penalized likelihood at the optimal penalty parameter for the entire dataset.

3.3 Estimation of population parameters for prediction

We estimate $f(Z)$ using the entire cohort (see 6.1) and then, inserting $\hat{f}(Z)$ and $\hat{\beta}X$ as offsets, we fit the unconditional logistic regression model (2) to the matched case control data. This provides us with estimates of the population parameters (α, γ) . Prediction for future patients will be based on the linear score function, $\hat{\alpha} + \hat{\beta}^T X + \hat{\gamma}^T Z$.

4. Simultaneous Variable Selection and Prediction Procedure

As seen in the derived regression model for the sampled data given in (2), if we have available an external estimate for $f(Z)$, as we do in the case of a nested study, we can directly fit an unconditional logistic regression model to the nested matched case-control data for simultaneous estimation of α, β , and γ using the unconditional pseudo-likelihood:

$$\prod_{k=1}^n \left[\frac{\exp \left\{ \alpha + \hat{f}(Z_k) + \beta^T X_{k1} + \gamma^T Z_k \right\}}{1 + \exp \left\{ \alpha + \hat{f}(Z_k) + \beta^T X_{k1} + \gamma^T Z_k \right\}} \right] \left[\frac{1}{1 + \exp \left\{ \alpha + \hat{f}(Z_k) + \beta^T X_{k2} + \gamma^T Z_k \right\}} \right]. \quad (4)$$

We apply the variable selection strategies outlined in Section 3.1 to this unconditional pseudo-likelihood. We apply the penalization with cross-validation to β only, and not to (α, γ) . Penalized pseudo-likelihood approaches have been used successfully for variable selection in other contexts, as well (e.g., Cai et al., 2005); we evaluate the effect of plugging in $\hat{f}(Z_k)$ in simulations.

5. Simulation Studies

We conducted three sets of simulation studies to compare the performances of our three variable selection strategies with the lasso and elastic net penalties, and to compare the two-stage and simultaneous procedures. We also compared the performance of the proposed methods with stepwise regression. We used the R package “penalized”.

In all simulation scenarios we first generated a prospective sample from a population logistic model, and then selected the nested matched case-control study from it. In the first set of simulations, we included a moderate number of predictors (50) and explored various association structures among the non-null predictors and the null predictors and different effect sizes of regression coefficients. Our main interest was to investigate the impact of

these associations and effect sizes on the performance of the various variable selection procedures and on their subsequent predictive performances. In the second set of simulations, we mimicked the MGH HAP study in numbers of predictors, associated block correlation structure, and matching variables. In the third set of simulations, the number of main predictor candidates is larger than in Setup II, and the predictors are continuous rather than binary, which illustrates the applicability of our method to other settings. In Web Appendix D, we illustrate minimal impact of use of the pseudo-likelihood (4) on prediction performance for the simultaneous procedure under simulation Setup II.

5.1 Simulation I: moderate number of predictors

In this simulation, we considered 50 Bernoulli predictors, of which, five (X_4, X_6, X_8, X_9 and X_{10}) are associated with the outcome, Y . Within this simulation, we evaluated four scenarios that illustrated the impact of different association structures between (X_4, X_6, X_8), (X_9, X_{10}) and Z , as well as different association structures between (X_1, X_2, X_3) and (X_4, X_6, X_8). The matching variable Z follows a uniform distribution on integers 1, 2, ..., 10. The population model is as in (1). In scenarios (i) and (ii), $\alpha = -1$, $\beta_4 = -0.4$, $\beta_6 = \beta_8 = 0.4$, $\beta_9 = -0.4$, $\beta_{10} = 0.25$ and $\beta_i = 0$ for all other covariates. We further assumed three non-null interactions, $X_4 \times X_6$, $X_4 \times X_8$ and $X_6 \times X_8$, with coefficients, $\beta_{4,6} = -0.4$, $\beta_{4,8} = \beta_{6,8} = 0.4$, and set the regression coefficient of Z to be $\gamma = 0.05$. In scenario (iii), $\alpha = -2$, $\beta_4 = -0.5$, $\beta_6 = \beta_8 = 0.5$, $\beta_9 = \beta_{10} = 0.5$ and $\beta_i = 0$ for all other covariates, and we further included the same three non-null interactions with coefficients $\beta_{4,6} = \beta_{6,8} = -1.75$, $\beta_{4,8} = 1.5$, and set $\gamma = 0.15$. In scenario (iv), $\alpha = -2$, $\beta_4 = -0.75$, $\beta_6 = 1.25$, $\beta_8 = 1$, $\beta_9 = 0.75$, $\beta_{10} = 1$ and $\beta_i = 0$ for all other covariates, and the three non-null interactions with coefficients $\beta_{4,6} = -0.75$, $\beta_{4,8} = \beta_{6,8} = 1.5$, and $\gamma = 0.30$. Based on these models, we generated a prospective sample of 2500 subjects, and then selected a nested, matched case-control study, consisting of 250 pairs. We provide additional details in Web Appendix A.

We applied the proposed procedures to each of 500 generated nested case-control studies. Table 1 summarizes the variable selection and prediction results for both procedures. The AUC was calculated by applying the estimated prediction rule from the nested study to an independent validation data set of sample size 2000, which followed the same distribution as the cohort data from which the matched case-control data were sampled. We also list the “true” AUC, which we obtained by fitting the true population logistic regression model to the independent validation data set. It represents the optimal AUC and serves as a benchmark. It is notable that the best of the three procedures with regard to variable selection, as measured by the cross validated log likelihood, almost always corresponds to the best with regard to the ultimate prediction performance, as measured by AUC. The AUC values for the two-stage and simultaneous procedures are comparable.

The four scenarios in Table 1 have AUC's ranging from 0.66 to 0.85. Scenarios (i) and (ii) have the same model for Y , but different association structures among the covariates. Scenario (iii) has the strongest interaction effects for the outcome model, and Scenario (iv) has the strongest main effects with moderate interaction effects. For both the two-stage and simultaneous procedures, the Pen1 strategy, which selects main effects only, is best for Scenarios (i,ii,iv), which have relatively weak interaction effects. In contrast, the EN2 strategy, which selects some interaction terms, is best for Scenario (iii), which has the strongest interaction effects. In summary, the performance of the selection procedure is robust to association structure and is influenced by the relative magnitudes of the main versus interaction effects.

5.2 Simulation II: large number of predictors mimicking stroke imaging study

In this simulation, we mimicked the MGH HAP study with 100 correlated predictors, and three matching variables with marginal distributions that correspond to those of age, gender, and NIHSS from the MGH study, and with correlation between gender and NIHSS, also as in the study. For simplicity, we assumed age to be independent of gender and NIHSS.

We modeled age (Z_1) as multinomial on the integers [18,100], gender (Z_2) as Bernoulli, and NIHSS (Z_3) as multinomial on the integers [0,36], all with probabilities as in the MGH study. We took the X 's to have marginal Bernoulli distributions, with p ranging from 0.14 to 0.36, and we assumed a blockwise correlation structure for $\mathbf{X}_{main}$. We set $\alpha = -3.5$, $\beta_6 = \beta_7 = \beta_8 = \beta_{10} = 0.5$, $\beta_9 = 1.0$, $\beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = 0.5$, and $\beta_i = 0.0$ for all other predictors. The nonzero regression coefficients for the interactions are $\beta_{6,9} = -3$, $\beta_{6,10} = \beta_{6,11} = -1$, $\beta_{7,9} = 1.2$, $\beta_{7,12} = -3$, $\beta_{8,10} = -2$, $\beta_{8,12} = 0.25$, $\beta_{8,13} = -3$, $\beta_{11,12} = -3$, $\beta_{11,13} = \beta_{12,13} = -2$, $\beta_{14,15} = -2.5$. The regression coefficients for the matching variables are $\gamma_1 = 0.003$, $\gamma_2 = 0.1$ and $\gamma_3 = 0.05$. Under this parameter configuration, the prevalence rate for HAP in the population is about 12%, as in MGH study. We generated cohort studies of 2500 subjects, and from them, we retrospectively selected matched case-control studies of 250 pairs of subjects. We provide additional details in Web Appendix B.

The penalization method Pen2 exhibits the best performance in variable selection and prediction (Table 2). We have included averaged 95% bootstrap confidence intervals for the AUC's from the independent samples for the simultaneous procedure under simulation Setup II. We did not include these for all simulations due to the computational demand of conducting a bootstrap within a simulation. These are suggestive that the AUC's from the different strategies are not formally significantly different from each other. However, the slightly larger AUC for Pen2 (0.727) versus Pen1 (0.693) provides informal justification for use of the cross-validated log likelihood as a metric for variable selection strategy, as it does seem to give a boost, albeit small, in AUC in an independent, prospective sample. In applications, it would be worth additionally considering the difference in the sets of variables selected by each procedure and their associated costs; it may be that Pen2 does not add any cost over Pen1, for example, but rather simply includes some interaction terms from among main effects that are already included in Pen1.

Figure 1 depicts the frequency of selection based on the two-stage implementation of procedure Pen 2 for each of the 100 main effect predictors and 12 non-null interactions, across the 500 replications. The tick marks on the x -axis in Figure 1(a) indicate the true main effects. It is apparent that the variables that are selected with highest frequency are the true, non-null, main effects. Two variables, X_{17} and X_{19} , which are not true main effects but are selected with high frequency, are correlated with true main effects $X_6, X_7, \dots, X_{10}$. The frequency of selection of the non-null interaction effects depends on their magnitude, the magnitude of their associated main effects and the correlation of the associated main effects with other main effects. These factors lead to relatively infrequent selection of $X_{14} \times X_{15}$. Among the remaining 4938 null interactions, 64.9% are never selected across the 500 replication studies, 18.7% are selected once, 5.7% are selected three times, and the remaining 10.7% are selected four or more times. The results for the simultaneous procedure are similar.

Table 3 lists additional summary statistics for the variable selection performance from this simulation. Consistent with the results of Table 2, Pen2 and EN2 outperform the other strategies in true and false positive rates for selection of main effects and interactions. The moderate true positive rates (34-61%), in conjunction with the near-optimal AUC of 72-73% suggests that there is redundancy in variables with regard to prediction, and that the variable selection strategies are optimizing prediction at the expense of variable selection.

Interestingly, the elastic net is not performing any better in variable selection than the lasso even in the presence of some correlated variables.

Finally, we assessed the accuracy of estimation of $f(\mathbf{Z})$ in the context of the simultaneous procedure. We investigated the impact of window size on the estimation of $f(\mathbf{Z})$ by choosing window sizes that lead to at least 5, or 10, or 15 subjects in a window. We also examined the impact of this estimation on the independent sample AUC. Detailed results are summarized in Web Appendix D.

5.3 Simulation III: large number of continuous predictors

In this simulation, the number of main predictor candidates is larger than that of Setup II, and they are continuous rather than dichotomous. In these regards, Setup III demonstrates the applicability of our method to common biomarker study settings, such as a matched case-control study conducted by the High Risk Plaque Initiative to discover proteomic and metabolic biomarkers for cardiovascular events (Balasubramanian et al., 2013).

We considered 150 main predictor candidates, of which 10 are non-null, 9 non-null interactions and 2 matching variables. A cohort of 4000 subjects was generated, and a nested, matched case-control study consisting of 400 pairs was selected. The prevalence rate is 25%. Matching variable Z_1 is multinomial on the integers $[1, 10]$, and Z_2 has Bernoulli distribution with success probability associated with Z_1 . We took the X 's to have marginal standard normal distributions, and we assumed a blockwise correlation structure for $\mathbf{X}_{main}$. Variables associated with Y include X_6 to X_{10} , which are correlated with each other, and X_{11} to X_{15} , which are independent of each other. Additional details are found in Web Appendix C.

The lower section of Table 2 summarizes the results for the two-stage and simultaneous procedures. The strategy that penalizes interaction effects of pre-selected main effects (PEN2/EN2) exhibits the best performance in variable selection and prediction. In additional simulations (not shown) we varied the cohort size and nested case-control study size and EN2 and Pen2 maintained their optimal performance. As expected, the AUC's increased relative to the true values with increasing sample size.

6. Application to MGH HAP in Stroke Study

There is inconsistent evidence about the potential association between lesion location and post-stroke hospital acquired pneumonia (HAP) (Upadya et al., 2004; Hilker et al., 2003). The MGH HAP study was undertaken to identify specific brain regions of acute infarction that are linked to HAP in acute ischemic stroke patients and to formulate a prediction rule that could be used for early evaluation of the risk of HAP in newly admitted stroke patients.

6.1 Variable selection

The potential predictors include both clinical and neuroimaging variables. Clinical variables include dyslipidemia, smoking history, coronary artery disease, diabetes mellitus, atrial fibrillation and hypertension. Raw images based on either non-contrast computed tomography (CT) scans or diffusion weighted magnetic resonance imaging (MRI-DWI) were obtained soon after onset of stroke symptoms. After co-registering raw images using specialized software, the infarct lesion maps were subsegmented into 69 pairs of mirrored cortical and subcortical regions based on "Harvard-Oxford cortical structural" and "JHU DTI-based white-matter" atlases. The regional percentage of infarcted tissue for each patient was then determined for each of the 138 standardized regions. We dichotomized the percentage of infarction at each brain region using its median. We categorized infarct volume according to the tertiles of its distribution. To stabilize the numerical calculations,

we excluded the 16 brain regions for which fewer than 5% of cases or 5% of controls displayed positive infarction.

We compared the performance of the proposed lasso and elastic net based strategies with stepwise regression and logic regression. The elastic net potentially advantageous over the lasso when some true predictors are highly correlated, as are some of the imaging variables, and it is of scientific interest to identify them all.

For variable selection using the simultaneous procedure, we need to estimate the offset $f(\mathbf{Z})$ defined in Section 2. This estimate is also necessary to estimate the prediction model based on the two-stage procedure. We have two choices for estimation of this quantity: we may estimate it directly, or we may decompose it according to Bayes' theorem and estimate the components of it separately. We illustrate the latter approach, as it is preferable when there are external estimates available for any of the components, or if any of the components may be particularly well-estimated using the prospective parent study. By Bayes' theorem, $\Pr(Y = 1 | \mathbf{Z} = \mathbf{z}) = \Pr(Y = 1) \Pr(\mathbf{Z} = \mathbf{z} | Y = 1) / \Pr(\mathbf{Z} = \mathbf{z})$. The prevalence rate of HAP in acute stroke patients, $\Pr(Y = 1)$, is estimated to be 12.2%. Estimation of $\Pr(\mathbf{Z} = \mathbf{z} | Y = 1) / \Pr(\mathbf{Z} = \mathbf{z})$ is based on the available disease status and matching variable information for the 1851 patients from among the 1915 in the cohort study who are not missing any of these variables. Among the 1851 patients, age ranges from 11 years to 103 years, and NIHSS ranges from 0 to 36. To estimate $\Pr(\mathbf{Z} = \mathbf{z}_i | Y = 1) / \Pr(\mathbf{Z} = \mathbf{z}_i)$ empirically, we create a nearest neighbor window for the i th observed subject in cohort data as $\text{Age}_i \pm \varepsilon_{i1}$, $\text{NIHSS}_i \pm \varepsilon_{i2}$ and count the number of subjects with and without HAP falling within the window. Details of algorithm are provided in Web Appendix D. This nonparametric approach to estimation of $f(\mathbf{Z})$ may suffer the "curse of dimensionality" when $\mathbf{Z}$ is high dimensional. In this situation, we recommend direct estimation of $f(\mathbf{Z})$, i.e., via $P(Y | \mathbf{Z})$, but based on a model, such as logistic regression.

Table 4 summarizes 10-fold cross-validation results for the different variable selection methods, under both procedures. Stepwise conditional logistic regression has considerably worse performance than other methods. Pen3 exhibits the highest cross-validated log-likelihood among Pen1, Pen2 and Pen3. We additionally evaluated Pen3 with the elastic net penalty function replacing the lasso. We considered a sequence of different L_2 penalties (λ_2) along with the internal cross-validation procedure for the L_1 component of the elastic net. We selected the minimum λ_2 that yielded a cross-validated log-likelihood that was within 3% of the maximum, to satisfy the accompanying goal of a parsimonious model. Among all applications of the elastic net, the use of the elastic net penalty function on the main effects and the lasso penalty on the interactions performed the best.

Based on our simulation results, it is reasonable to choose the variable selection method with the highest cross-validated (conditional) log-likelihood, as it typically leads to the best prediction accuracy. As seen in Table 5, among the 10 main effects chosen by the two-stage and simultaneous procedures with lasso penalty, eight are shared in common. Region j26 is selected by the two-stage procedure only and h72 is selected by the simultaneous procedure only. One possible reason that the two-stage procedure does not select h72 is that h72 is highly correlated with h32, which was selected by the two-stage procedure. Similarly, the simultaneous procedure does not select j26, which is highly correlated with j34, and was selected by the simultaneous procedure. We list the 20 selected interaction effects selected by Pen3 under the two-stage procedure and the 24 selected interaction effects selected under the simultaneous procedure in the Web Appendix E. Fifteen of these terms are common to both procedures. As expected, the elastic net strategy selects more main effects than the lasso, and all of the main effects selected by lasso are also selected by the elastic net. The

elastic net strategy under the two-stage procedure and simultaneous procedure selected 15 and 31 interactions respectively; these are listed in the Web Appendix E.

Several of the brain regions selected have plausible associations with the risk of pneumonia. As a brainstem motor control center, infarctions in the cerebral peduncle (j10) would be expected to impair motor control of swallowing and increase the risk of pneumonia. Notably, functional MRI studies of healthy individuals have shown activation of the middle and inferior frontal gyri, the cingulate cortex (adjacent to fornix and stria terminalis), the insular cortex, the superior and transverse temporal gyri during different swallowing activation tasks (Martin et al., 2001).

Figure 2 displays a heatmap of the phi coefficient (a correlation coefficient for binary data) matrix plot for the imaging variables and infarct volume. Black represents perfect positive or negative correlation, and white represents no correlation. The darker the grey, the higher the association. A block correlation structure is apparent, likely reflecting high local spatial correlations in the brain. The bars on the left depict the univariate log odds ratios and associated p -values for each variable relative to the HAP outcome based on a conditional logistic regression model that accounts for the matching: black depicts the highest positive or negative association $\log OR = 1.07$ ($OR = 2.9$), white depicts no association $\log OR = 0$ ($OR = 1.0$); black depicts the smallest p -value=0.001 and white depicts the largest p -value=1. The bold ticks on the axes of the correlation matrix mark the variables that were selected in the two-stage lasso version of Pen3. It is apparent that our proposed lasso procedure selects variables with high univariate association and does not select redundant variables. The bold ticks below the x -axis labels mark the variables that were selected by the elastic net method embedded in the two-stage procedure. As expected, the elastic net selects additional main effects, which are correlated with those selected by lasso. The heatmap plot for simultaneous procedure is similar and is not shown here.

Finally, using the elastic net procedure, EN3, we fit an unconditional logistic regression model that includes $\hat{f}(Z)$ as the offset and we simultaneously estimated α , β and γ . The estimated coefficients are listed in Appendix E and could be used for prediction of HAP for a future patient.

7. Discussion

We have proposed two variable selection and prediction procedures for high dimensional data from a matched case-control study that is nested within a larger prospective study. In our context, the predictor variables of interest are available only for the nested study participants. To appropriately account for the matching, which is growing in popularity for high dimensional imaging and genomic studies, we take a likelihood-based approach. We further employ penalization to stabilize the estimation and to facilitate variable selection. Being likelihood based, our procedures can be extended to other clustered designs. Although we have not considered interactions between the matching variables and the predictor variables, these could be included as well. Our ability to go beyond variable selection to prediction hinges on the availability of the parent prospective study, from which we can estimate the proper adjustment to the log odds parameter to obtain the correct population outcome model.

We have compared the performances of our proposed two-stage and simultaneous procedures in simulation studies and for the MGH HAP study. While the variable selection and prediction performances are similar for the two procedures, the simultaneous procedure demonstrated marked computational efficiency over the two-stage procedure. In our simulation studies, the computational time of the two-stage procedure ranged from 13 to 35

times the computational time of the simultaneous procedure. The potential disadvantage of the simultaneous procedure is that it requires correct modeling of the predictor variables as well as the matching variables for variable selection, whereas the two-stage procedure requires only correct modeling of the predictor variables for variable selection. This is because the two-stage procedure removes the matching variables from the predictor variable selection step through its use of the conditional likelihood.

Wu et al. (2010) proposed a lasso penalization procedure called “screen and clean” for the purpose of identifying interactions in case-control genome-wide association studies. This procedure is similar to our Pen2 method, however, it additionally includes a final step to “clean” the selected interactions using an independent source of data. It will be useful to extend this procedure to accommodate matched study designs and for prediction when both an independent matched case-control study and a parent prospective study are available.

In conclusion, we have developed proposals for the dual goals of variable selection and prediction in the setting of high dimensional predictors that are too expensive to measure on all members of a prospective study, necessitating use of a nested, matched case-control study. This scenario arises commonly in the context of imaging and genomic variables. We have exhibited our approach on the MGH study of hospital acquired pneumonia among acute stroke patients. Based on this study, we have identified nine non-redundant, biologically plausible brain regions, together with volume, and a subset of their interactions. This may lead to selective prophylactic treatment and ultimately serve to minimize this detrimental outcome in this vulnerable population.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful to the editor, the associate editor, and two anonymous referees for their valuable comments. This research was supported in part by the Harvard NeuroDiscovery Center, the Harvard Catalyst (8UL1TR000170), NIH grant R01-CA075971.

References

- Adewale AJ, Dinu I, Yasui Y. Boosting for correlated binary classification. *Journal of computational and graphical statistics*. 2010; 19:140–153.
- Balasubramanian R, Houseman E, Coull B, Lev M, Schwamm L, Betensky R. Variable importance in matched case-control studies in settings of high dimensional data. *Journal of the Royal Statistical Society*. 2013 in press. Series C
- Breslow, NE.; Day, NE. *The Analysis of Case-control Studies*. Vol. Vol. 1. World Health Organization; 1980. *Statistical Methods in Cancer Research*.
- Cai J, Fan J, Li R, Zhou H. Variable selection for multivariate failure time data. *Biometrika*. 2005; 92:303–316. [PubMed: 19458784]
- Davatzikos C, Resnick S, Wu X, Parnpi P, Clark C. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. *NeuroImage*. 2008; 41:1220–1227. [PubMed: 18474436]
- Davenport R, Dennis M, Wellwood I, Warlow C. Complications after acute stroke. *Stroke*. 1996; 27:415–420. [PubMed: 8610305]
- Duchesne S, Caroli A, Geroldi C, Collins DL, Frisoni GB. Relating one-year cognitive change in mild cognitive impairment to baseline MRI features. *NeuroImage*. 2009; 47:1363–1370. [PubMed: 19371783]

- Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*. 2010; 20:101–148. [PubMed: 21572976]
- Hilker R, Poetter C, Findeisen N, Sobesky J, Jacobs A, Neveling M, et al. Nosocomial pneumonia after acute stroke: implications for neurological intensive care medicine. *Stroke*. 2003; 34:975–981. [PubMed: 12637700]
- Katzan IL, Dawson NV, Thomas CL, Votruba ME, Cebul RD. The cost of pneumonia after acute stroke. *Neurology*. 1998; 68:1938–1943. [PubMed: 17536051]
- Lin DY, Ying Z. Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association*. 1993; 88:1341–1349.
- Mantel N. Synthetic retrospective studies and related topics. *Biometrics*. 1973; 29:479–486. [PubMed: 4793136]
- Martin R, Goodyear B, Gati J, Menon R. Cerebral cortical representation of automatic and volitional swallowing in humans. *Journal of Neurophysiology*. 2001; 85:938–950. [PubMed: 11160524]
- Meier L, van de Geer S, Bühlmann P. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B*. 2008; 70:53–71.
- Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979; 66:403–412.
- Rothman, KJ.; Greenland, S.; Lash, TL. *Modern Epidemiology*. 3rd edition. Lippincott Williams & Wilkins; 2008.
- Ruczinski I, Kooperberg C, Leblanc M. Logic regression. *Journal of Computational and Graphical Statistics*. 2003; 12:475–511.
- Tan Q, Thomassen M, Kruse TA. Feature selection for predicting tumor metastases in microarray experiments using paired design. *Cancer Informatics*. 2007; 3:213–218. [PubMed: 19455244]
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B: Methodological*. 1996; 58:267–288.
- Twohoger S, Rosner B, Willett W, Hankinson S. The combined influence of multiple sex and growth hormones on risk of postmenopausal breast cancer: a nested case-control study. *Breast Cancer Research*. 2011; 13(5):R99. [PubMed: 22017816]
- Upadya A, Thorevska N, Sena K, Manthous C, Amoaeng-Adjepong Y. Predictors and consequences of pneumonia in critically ill patients with stroke. *The Journal of Critical Care*. 2004; 19:16–22.
- Woodward, M. *Epidemiology: Study Design and Data Analysis*. 2nd edition. Chapman & Hall/CRC; 2005.
- Wu J, Devlin B, Ringquist S, Trucco M, Roeder K. Screen and clean: A tool for identifying interactions in genome-wide association studies. *Genetic Epidemiology*. 2010; 34:275–285. [PubMed: 20088021]
- Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*. 2009; 25(6):714–721. [PubMed: 19176549]
- Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*. 2005; 67:301–320.

Figure 1. Plot of frequency that each variable is selected for setup II based on Pen2 of two-stage procedure: (a) 100 main effects, and tick marks on the x -axis indicate 10 true main effects, which are $X_6, X_7, \dots, X_{15}$. (b) 12 true interaction effects.

Figure 2.

Heatmap of the correlations among the imaging variables and infarct volume (black is perfect positive or negative correlation, white is no correlation; the darker the grey, the higher the association). The bars on the left depict the univariate log odds ratios and p -values for each imaging variable relative to the HAP outcome based on a conditional logistic regression model that accounts for the matching: black depicts the highest positive or negative association $\log OR = 1.07$ ($OR = 2.9$), white depicts no association $\log OR = 0$ ($OR = 1.0$); black depicts the smallest p -value=0.001 and white depicts the largest p -value=1. The bold ticks on the axes of the correlation matrix mark the variables that were selected in the two-stage lasso version of Pen3. The bold ticks below the x -axis labels mark the variables that were selected by the elastic net method embedded in the two-stage procedure. The last two ticks correspond to two indicator variables of infarct volume.

Table 1

Summary of variable selection and prediction results for simulation setups I.

	"true"-AUC	Stepwise	Pen1	Pen2	Pen3	EN1	EN2	EN3
two-stage procedure								
<i>scenario (i)</i>								
CV-LK		-186.91	-165.83	-169.52	-183.27	-166.08	-196.49	-183.29
AUC	0.658	0.598	0.616	0.610	0.6051	0.616	0.607	0.605
<i>scenario (ii)</i>								
CV-LK		-178.53	-155.21	-159.25	-170.11	-155.20	-159.37	-170.13
AUC	0.688	0.633	0.662	0.654	0.645	0.660	0.654	0.644
<i>scenario (iii)</i>								
CV-LK		-187.86	-170.08	-162.75	-180.50	-170.03	-162.68	-180.10
AUC	0.727	0.592	0.611	0.639	0.624	0.611	0.639	0.624
<i>scenario (iv)</i>								
CV-LK		-155.16	-104.35	-111.24	-136.87	-104.88	-111.28	-137.00
AUC	0.854	0.796	0.825	0.819	0.800	0.825	0.819	0.800
simultaneous procedure								
<i>scenario (i)</i>								
CV-LK		-352.43	-336.91	-342.76	-351.71	-355.03	-347.45	-351.87
AUC	0.658	0.602	0.617	0.611	0.603	0.616	0.608	0.602
<i>scenario (ii)</i>								
CV-LK		-343.55	-327.52	-331.87	-340.74	-349.82	-337.63	-340.91
AUC	0.688	0.639	0.663	0.656	0.644	0.662	0.653	0.644
<i>scenario (iii)</i>								
CV-LK		-353.84	-338.77	-335.55	-344.09	-338.80	-334.83	-344.10
AUC	0.727	0.597	0.612	0.640	0.631	0.612	0.641	0.631
<i>scenario (iv)</i>								
CV-LK		-282.21	-268.14	-271.43	-283.92	-268.46	-271.45	-283.93
AUC	0.854	0.812	0.828	0.823	0.810	0.828	0.823	0.810

NOTE: CV-LK: cross-validated conditional log-likelihood (for two-stage procedure) or cross-validated log-likelihood (for simultaneous procedure); AUC: area under the ROC curve for independent validation dataset.

Table 2

Summary of variable selection and prediction results for simulation setups II and III.

	"true" AUC	Stepwise	Pen1	Pen2	Pen3	EN1	EN2	EN3
<i>Simulation Setup II</i>								
two-stage procedure								
CV-LK		-785.39	-149.02	-145.20	-157.26	-151.82	-147.25	-158.67
AUC	0.826	0.632	0.692	0.718	0.683	0.692	0.718	0.684
simultaneous procedure								
CV-LK		-378.95	-325.85	-317.16	-343.80	-325.97	-319.11	-338.19
AUC	0.826	0.654	0.693	0.727	0.696	0.693	0.727	0.698
95% CI		(0.617, 0.690)	(0.656, 0.729)	(0.692, 0.760)	(0.660, 0.731)	(0.656, 0.729)	(0.692, 0.760)	(0.662, 0.733)
<i>Simulation Setup III</i>								
two-stage procedure								
CV-LK		-508.80	-249.46	-228.61	-233.27	-249.46	-228.59	-233.26
AUC	0.815	0.622	0.676	0.719	0.708	0.676	0.719	0.708
simultaneous procedure								
CV-LK		-595.04	-527.20	-502.45	-506.77	-527.20	-501.74	-506.74
AUC	0.815	0.638	0.677	0.723	0.714	0.677	0.723	0.714

NOTE: CV-LK: cross-validated conditional log-likelihood (for two-stage procedure) or cross-validated log-likelihood (for simultaneous procedure); AUC: area under the ROC curve for independent validation dataset. For simultaneous procedure under Setup II, we also give the 95% averaged bootstrap percentile confidence interval for AUC based on 3000 bootstrap samples.

Table 3

Simulation setup II: Summary of variable selection.

Methods	Stepwise	Pen1	Pen2	Pen3	EN1	EN2	EN3
two-stage procedure							
<i>main effects</i>							
average # of effects found	29.63	19.17	12.27	19.48	19.18	12.27	19.78
average # of true effects found	5.78	5.71	6.10	5.77	5.71	6.10	5.76
TPR	0.578	0.571	0.610	0.577	0.571	0.610	0.576
FPR	0.265	0.150	0.069	0.152	0.150	0.069	0.156
<i>interaction effects</i>							
average # of effects found	–	–	14.09	9.85	–	14.09	9.80
average # of true effects found	–	–	4.10	2.23	–	4.10	2.20
TPR	–	–	0.342	0.186	–	0.342	0.183
FPR	–	–	0.002	0.002	–	0.002	0.002
simultaneous procedure							
<i>main effects</i>							
average # of effects found	24.29	20.23	12.27	20.23	20.24	12.27	20.27
average # of true effects found	5.95	5.93	6.10	5.93	5.93	6.10	5.93
TPR	0.595	0.593	0.610	0.593	0.593	0.610	0.593
FPR	0.204	0.159	0.069	0.159	0.159	0.069	0.159
<i>interaction effects</i>							
average # of effects found	–	–	17.09	18.93	–	17.12	18.94
average # of true effects found	–	–	4.60	3.40	–	4.60	3.40
TPR	–	–	0.384	0.284	–	0.384	0.283
FPR	–	–	0.003	0.003	–	0.003	0.003

NOTE: TPR, true positive rate; FPR, false positive rate. The total number of main effects is 100, whereas the number of true main effects is 10. The total number of two-way interaction effects is 4950, whereas the number of true interactions is 12.

Table 4

MGH stroke imaging data: 10-fold cross-validated (conditional) log-likelihood for different variable selection methods (excluded imaging variables with nonzero values less than 5% in either cases or controls).

	Stepwise	Logic Reg	Pen1	Pen2	Pen3	Elastic Net
Two-stage procedure	-1230.30	-149.50	-145.83	-128.80	-120.07	-104.02
Simultaneous procedure	-383.87	-276.43	-302.24	-277.12	-268.36	-245.50

Table 5

Names of selected main effects by two procedures with lasso or elastic net penalty.

code	name	two-stage procedure		simultaneous procedure	
		(lasso)	(EN)	(lasso)	(EN)
CAD	Coronary artery disease (Y/N)				✓
j10	Cerebral peduncle R	✓	✓	✓	✓
j11	Anterior limb of internal capsule L				✓
j19	Superior corona radiata L				✓
j23	Posterior thalamic radiation L		✓		
j24	Posterior thalamic radiation R		✓		
j26	Sagittal stratum R	✓	✓		✓
j34	Fornix (cres) / Stria terminalis R	✓	✓	✓	✓
j40	Uncinate fasciculus R		✓		

code	name	two-stage procedure		simultaneous procedure	
		(lasso)	(EN)	(lasso)	(EN)
h5	Superior Frontal Gyrus L	✓	✓	✓	✓
h8	Middle Frontal Gyrus R	✓	✓	✓	✓
h21	Middle Temporal Gyrus - anterior division L				✓
h28	Inferior Temporal Gyrus - anterior division R				✓
h25	Middle Temporal Gyrus - temporooccipital part L		✓		
h32	Inferior Temporal Gyrus - temporooccipital part R	✓	✓	✓	✓
h43	Lateral Occipital Cortex - superioir division L				✓
h45	Lateral Occipital Cortex - inferior division L		✓		
h46	Lateral Occipital Cortex - inferior division R		✓		✓

code	name	two-stage procedure		simultaneous procedure	
		(lasso)	(EN)	(lasso)	(EN)
h51	Juxtapositional Lobule Cortex L	✓	✓	✓	✓
h72	Lingual Gyrus R		✓	✓	✓
h74	Temporal Fusiform Cortex - anterior division R		✓		✓
vol1	Volume great than or equal to its 33% percentile	✓	✓	✓	✓
vol2	Volume great than or equal to its 67% percentile	✓	✓	✓	✓

NOTE: EN, elastic net; L, left side of brain; R, right side of brain.