Research Article

Statistics in Medicine

Received 8 June 2010, Accepted 15 February 2011 Published online 15 April 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4246

Assessment of a disease screener by hierarchical all-subset selection using area under the receiver operating characteristic curves

Yuanjia Wang,[a*†] Huaihou Chen,[a] Theresa Schwartz,[b] Naihua Duan,[b,c,d] Angela Parcesepe[b] and Roberto Lewis-Fernández[b,c]

In many clinical settings, a commonly encountered problem is to assess the accuracy of a screening test for early detection of a disease. In this article, we develop hierarchical all-subset variable selection methods to assess and improve a psychosis screening test designed to detect psychotic patients in primary care clinics. We select items from an existing screener to achieve best prediction accuracy based on a gold standard psychosis status diagnosis. The existing screener has a hierarchical structure: the questions fall into five domains, and there is a root question followed by several stem questions in each domain. The statistical question lies in how to implement the hierarchical structure in the screening items when performing variable selection such that when a stem question is selected in the screener, its root question should also be selected. We develop an all-subset variable selection procedure that takes into account the hierarchical structure in a questionnaire. By enforcing a hierarchical rule, we reduce the dimensionality of the search space, thereby allowing for fast all-subset selection, which is usually computationally prohibitive. To focus on prediction performance of a selected model, we use area under the ROC curve as the criterion to rank all admissible models. We compare the procedure to a logistic regression-based approach and a stepwise regression that ignores the hierarchical structure. We use the procedure to construct a psychosis screening test to be used at a primary care clinic that will optimally screen low-income, Latino psychotic patients for further specialty referral.

Keywords: disease diagnostic test; early screening; hierarchical variable selection; best subset selection

1. Introduction

In many clinical settings, a commonly encountered problem is to assess the accuracy of a screening test for early detection of a disease. Subjects screened positive will be referred for more definitive diagnostic tests. For example, cervical cancer screening with Pap smear and breast cancer screening with mammogram are routinely performed in clinics. Primary care settings are increasingly important in the early detection and treatment of psychopathology. Some screening instruments, however, such as the Psychosis Screening Questionnaire (PSQ), have good psychometric properties in primary care when used in the cultural groups in which they were originally designed [1], but limited psychometric properties when transposed to other cultural groups, such as low-income minorities [2]. For example, the PSQ yielded a screening rate of 14 per cent for psychotic disorder in a low-income Latino primary

[a]*Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, U.S.A.*
[b]*New York State Psychiatric Institute, New York, NY 10032, U.S.A.*
[c]*Department of Psychiatry, Columbia University, New York, NY 10032, U.S.A.*
[d]*Department of Biostatistics, Columbia University, New York, NY 10032, U.S.A.*
[*]*Correspondence to: Yuanjia Wang, Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, U.S.A.*
[†]*E-mail: yw2016@columbia.edu*

care sample, which is clearly inconsistent with the clinical review of this patient population [2]. Without accurate screening methods, racial and ethnic minority patients in primary care suffering from psychopathology are particularly subject to misdiagnosis, stemming from both under- and over-diagnosis. The result is frequently invalid diagnoses and inadequate treatment that call for improved screening practices in primary clinics.

A study using the PSQ as a screener for current psychosis among low-income Latino primary care patients was conducted at the Columbia University Medical Center (CUMC, Lewis-Fernández [2]). The goal of the study was to develop an improved screener by selecting PSQ items that would best predict the presence of a psychotic disorder as diagnosed by the gold standard of the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID, First *et al.* [3]). Based on the questions selected from the existing PSQ, a patient visiting a primary care clinic classified as positive will be referred to a psychiatrist for further evaluation, while a subject classified as negative will not be referred.

Constructing a new screener by selecting questions from an existing questionnaire is a variable selection problem. The PSQ is a screener with a hierarchical structure. There are 12 questions grouped into five domains. In each domain, there is a root question followed by several stem questions. The stem questions will be asked after a subject has answered the root question. If one ignores the hierarchical structure in the PSQ, it is possible that one would select a model that contains a stem question but not its root question. However, such a model is not admissible. The reasons to use a nested questionnaire are two folds. First, to start with questions that were deliberately vague as a gentle introduction, so as to avoid 'over-incisive' initial questions about psychosis and thereby decrease subjects' desire to continue. Second, to streamline the instrument so that subjects who responded negatively to the initial question could be given a negative answer to later items. The questions thus act as a funnel, with gentle initial questions getting progressively sharper. The questionnaire requires both kinds of questions in order to entice subjects in, but then reduce the number of likely false positives, which is known to be high for psychosis screeners.

To account for the hierarchical structure between variables in a variable selection procedure, several penalized likelihood-based structured variable selection approaches have been proposed. Huang *et al.* [4] applied a group bridge penalty to perform both the group level and within-group individual level variable selection. Wang *et al.* [5] proposed penalized Cox regression to impose a hierarchical structure among variables by re-parametrization. Zhao *et al.* [6] developed a structured variable selection approach by constructing appropriate penalty functions. Yuan and Lin [7] proposed methods for variable selection with variables assuming certain group structures. However, the loss functions used in these methods are all squared loss or likelihood functions, which may not be the optimal choice for evaluating prediction performance of a diagnostic test [8].

To assess accuracy of medical tests, true positive and false negative rates are two popular indices. From a continuous test score Y, one can define a binary test under a threshold c as: $Y \geqslant c$: positive, and $Y < c$: negative. The receiver operating characteristic (ROC) curve is the entire collection of possible true positives and false positives with different thresholds. A summary index of medical test performance can then be defined as the area under the ROC curve (the AUC of an ROC). An interesting interpretation of the AUC of an ROC is the probability that test results from a randomly selected pair of diseased and non-diseased subjects are correctly ordered, that is, the diseased subject has a higher score than the non-diseased subject. Since in our application the prediction performance of the screener used for diagnosing psychosis at a primary care clinic is of interest, we will use AUC of an ROC under a given model as our model selection criterion. There are no existing methods that can handle the hierarchical structure and use the AUC as the objective function to select models. Methods and algorithms developed for other methods (e.g. [5, 7]) are not straightforward to implement, partially because the empirical AUC is not a continuous function which can be difficult to optimize.

In this work, we apply an all-subset variable selection procedure that takes into account the hierarchical structure among variables. The procedure exhausts all possible admissible models that enforce the hierarchical rule that whenever a stem question is selected, its root question is also selected. We take advantage of the hierarchical constraints of the model search space to reduce dimensionality, thereby allowing for fast all-subset selection which is usually computationally prohibitive. To focus on the prediction performance of a selected model, we use the area under an ROC curve as the criterion to rank all admissible models. We compare the procedure to a logistic regression-based approach and a stepwise regression that ignores the hierarchical structure. Finally, we use the procedure to construct a

psychosis screening test at a primary care clinic that will optimally screen low-income Latino patients for further specialty referral.

2. Methods

2.1. Unstructured model selection with unrelated variables

First we illustrate the search procedure with unrelated variables. A traditional model selection procedure includes forward, backward, stepwise and all-subset selection. Popular criteria used for model selection include Marlows's C_p, AIC, BIC, and so on. These criteria can be viewed as a loss function plus a penalty function. For example, the loss function for AIC and BIC is the logarithm of the likelihood function and for Marlow's C_p, it is the mean squared error.

Statistical approaches available to select PSQ questions that best match the SCID diagnosis include association-based methods, for example, logistic regression of SCID diagnosis on PSQ items, and ROC-based methods. For a continuous test, one may use a decision rule of the test score being greater than a threshold to classify a subject as positive. The ROC curve plots the true positive rate versus the false positive rate for all possible thresholds. The area under an ROC curve is used to summarize the performance of a diagnostic test [8, 9]. Pepe [8] showed that when the prediction is of interest, ROC-based methods can outperform traditional association-based methods in terms of classification accuracy because the likelihood of a logistic regression may not be a useful index in measuring the prediction performance. For example, a variable with a large odds ratio has a large association, but might contribute little in terms of prediction. Therefore, we propose to use area under the ROC curve as a measure to select candidate models because it directly addresses prediction performance.

When there is no structure among the items in a questionnaire, a traditional step-wise or all-subset selection can be carried out. For each given model, a subset of PSQ questions will be included to construct a test score. When the questions are unrelated, since each question refers to the presence of a disease symptom, an answer of 'Yes' to each question will have a score of one and the final score for each subject is the sum of the scores for all questions, i.e. the number of positive symptoms. The area under the ROC curve for this given model will then be computed. Using this criterion, one can rank all the models and select the one with the highest area under the curve.

2.2. All-subset selection with structured screener

The 12 questions in the PSQ are grouped into five parts to represent psychosis symptoms in five domains (e.g. hypomania, thought interference, persecution, strange experiences, and hallucinosis). Given that the screeners are administered in a structured way, it is desirable to incorporate this structure into our analysis. To this end, we would impose the following constraint during model selection:

$C1$: Stem questions are considered only if the corresponding root question is included.

For example, with this method, there are five possible options for PSQ group 1 which includes questions Q1, Q1a, and Q1b: (NULL: None of the three questions), (Q1 only), (Q1 and Q1a), (Q1 and Q1b), or (Q1 and Q1a and Q1b). Other options such as (Q1a only), (Q1b only), and (Q1a and Q1b) are excluded because they violate the hierarchical structure ($C1$). Similarly, the possible options for items in group 2, which include questions Q2 and Q2a, would consist of the following three options: (NULL), (Q2 only), or (Q2 and Q2a). We summarize in Table I all possible options in each domain, obeying the hierarchical structure. The entire search space for all admissible hierarchical models consists of combinations of selecting one legitimate option from each of the five domains (Table I). In total, there are 675 ($=5\times3\times5\times3\times3$) admissible models. This is in contrast to the search space for 12 unrelated variables that contains $2^{12}-1=4095$ models. It is evident that the rule ($C1$) enforces constraints on the search space, thereby reducing its dimensionality.

Our search algorithm will perform all-subset selection among the legitimate models, i.e. we restrict the search space to all models that satisfy the hierarchical rule ($C1$).

2.3. Implementing a scoring system for a structured screener

In this section, we develop a scoring system to incorporate the structured nature of the counting instrument into the scoring rule. For example, if the final selected model includes questions: Q1, Q1b, Q2, Q4, Q5, and Q5a, we would follow the structured scoring procedure used for the PSQ which scores

Table I. Legitimate models.

Group	Questions involved	Legitimate cases
G1	Q1, Q1a, Q1b	NULL, {Q1}, {Q1 Q1a}, {Q1 Q1b}, {Q1 Q1a Q1b}
G2	Q2, Q2a	NULL, {Q2}, {Q2 Q2a}
G3	Q3, Q3a, Q3b	NULL, {Q3}, {Q3 Q3a}, {Q3 Q3b}, {Q3 Q3a Q3b}
G4	Q4, Q4a	NULL, {Q4}, {Q4 Q4a}
G5	Q5, Q5a	NULL, {Q5}, {Q5 Q5a}

Table II. Hypothetical patients and scores.

Questions	Q1, Q1a, Q1b	Q2, Q2a	Q3, Q3a, Q3b	Q4, Q4a	Q5, Q5a
Psychotic	No No No	Yes Yes	Yes Yes No	No No	Yes Yes
Score	0	1	0	0	1
Non-psychotic	No No No	Yes No	No No No	Yes No	No No
Score	0	0	0	0	0

the instrument by domain. Following this method, we would score Q1 and Q1b as one item. Therefore, the respondent would need to endorse both Q1 and Q1b in order to score positively on domain 1. Similarly, the respondent would need to endorse both Q5 and Q5a in order to score positively on domain 5. Thus, this candidate instrument would have five possible scores: 0, 1, 2, 3, and 4, depending on the number of domains endorsed among the four candidate domains Q1, Q2, Q4, and Q5 included in this candidate instrument.

We present two hypothetical subjects to illustrate the linear scoring rule with the full PSQ including all items. First, we present the hypothetical example of a subject who would screen positive for psychosis using the PSQ. Assume that a score of 1 or above on the PSQ qualifies as a positive screen for psychosis. The subjects' answers to all items are listed in Table II. This person's final score would be 2, since he scored positive on items 2 and 5. Note that this person did not score positively on item 3 since he did not positively endorse item 3b. To score positively on any one item (i.e. item 3), one must endorse both the root question (i.e. Question 3) and all related stem questions (i.e. Questions 3a and 3b).

Next we provide a hypothetical example of a subject who would screen negative for psychosis using the PSQ. This person's final score would be 0, as he did not score positive on any items. Note that this person did not score positively on items 2 or 4. For both of these items, while the root questions (Questions 2 and 4) were positively endorsed, the related stem questions (Questions 2a and 4a) were not positively endorsed.

In general, for a model with p domains selected, let $X_1, \ldots, X_p$ denote the score for each domain under the above scoring rule. The variables X_k are binary variables depending on whether the answers to all individual items in a domain are positively endorsed. A linearly combined score for all domains can be defined as

$$L(X)=\sum_{k=1}^{p} X_k. \tag{1}$$

When $L(X)>c_0$ where c_0 is a threshold, a subject is classified as likely to be psychotic and needs further specialty referral. Since the linear decision rules are scale invariant and location-shift invariant in computing the empirical AUC of an ROC as in (3), a decision rule of $L(X)>c_0$ is equivalent to a decision rule of $\theta_0+\theta_1 L(X)>c_1$. It follows that the rule (1) is equivalent to $L(X)=\sum_{k=1}^{p} cX_k$ for an arbitrary c. In other words, the combined test (1) assumes that the weight for the individual score from each domain in the linear combination is identical. This assumption is reasonable given the clinical insights in designing the widely accepted SCID diagnostic test in DSM-IV [3], which scores symptoms by domain and adds up all domain-wise positive scores.

The linear scoring rule described in (1) assumes that the individual score from each domain has the same weight using clinical insights. When such clinical knowledge is not available to specify weights, it may be desirable to estimate weights for each domain from the data through a logistic regression. To be specific, let Y denote the gold standard SCID diagnosis and let $X_1, \ldots, X_p$ denote the answers

to a domain defined before for the rule (1). A logistic regression specifies

$$\text{logit Pr}(Y=1|X)=\beta_0+\beta_1X_1\cdots+\beta_pX_p.$$

A linear score can be constructed by the parameters estimated from the logistic regression. To be specific, let $\alpha_k=\beta_k/\beta_1,\ k=2,\ldots,p$, then we form a linear rule,

$$L_\alpha(X)=X_1+\alpha_2X_2+\cdots+\alpha_pX_p, \tag{2}$$

and when $L_\alpha(X)>c_0$ where c_0 is a threshold, a subject is classified as likely to be psychotic and needs specialty referral. Note that the combined score does not include an intercept and that the coefficient for the first variable is fixed as one. This is again due to the linear decision rules being scale invariant and location-shift invariant in computing the AUC. The parameters in a logistic regression are estimated by maximum likelihood estimation. Usually the AIC, BIC, or the area under the ROC curve can be used to select the best model. To focus on prediction performance, we use the AUC of the ROC curve to rank the models.

It is worth pointing out that $L_\alpha(X)$ differs from $L(X)$ in (1) in that for a given model with a known set of variables, the weights in $L_\alpha(X)$ are unknown and estimated from data using a logistic regression, while under (1), the weights, and therefore the combined test scores, are known.

2.4. Algorithms to fit the model

Here we describe an exhaustive algorithm to select the best model. We first describe the Algorithm A for fitting the model under the scoring system (1) followed by the Algorithm B under system (2).

Algorithm A

Step 1. Define legitimate models: We create a binary variable indicator that takes a value of one if a variable is selected. Define all the models satisfying the constraint ($C1$), i.e. if a model has any of the stem questions it must include the root question, by a sequence of zero–one indicator variables. Following this, the variable indicators for each domain should start with one; otherwise, none of the questions in this group will be selected. For example, if a candidate root question is followed by two candidate stem questions, the proper variable indicators are (1, 0, 0), (1,1,0), (1,0,1), (1,1,1), and (0, 0, 0). Sequences such as (0,1,0) or (0,1,1) are not allowed. Selecting the variables with indicator equal to one leads to a legitimate model. For each legitimate model, we calculate the score using the method in Section 2.3. In other words, the score for each domain is the product of all individual indicator variables in the domain.

Step 2. For each legitimate model, compute its AUC of the ROC curve to assess the predictivity of the selected items for the gold standard DSM-IV psychosis status. Suppose there are n_D psychotic subjects diagnosed by DMS-IV and n_H non-psychotic subjects. Denote S_i^+, $i=1,\ldots,n_D$, and S_j^+, $j=1,\ldots,n_H$, as the corresponding scores for psychotic and non-psychotic subjects, respectively. The AUC of the empirical ROC can be estimated using the Mann–Whitney U-statistic

$$\widehat{\text{AUC}}=\frac{\sum_{i=1}^{n_D}\sum_{j=1}^{n_H}I\{S_i^+>S_j^+\}}{n_Dn_H}, \tag{3}$$

where $I(\cdot)$ is an indicator function. We rank all models by $\widehat{\text{AUC}}$.

Step 3. For each legitimate model defined in Step 1, repeat Step 2 to obtain the empirical AUC. Select the final model as the one with the highest AUC.

Next we describe the algorithm using logistic regression to estimate weights for (2). To enforce the hierarchical structure in the model, we still restrict the search space to the legitimate models described in step 1 of Algorithm A. In step 2, we will use parameters estimated from the logistic regression to form the linear scoring rule. A complication arises in computing AUC to rank different models for the logistic regression-based approach. It is well known that if one uses the data to obtain the parameters of a model and then uses the same data to estimate measures of prediction performance (i.e. prediction error or AUC of an ROC), the estimated classification performance will be overly optimistic [10, 11]. An honest measure of prediction performance should be assessed using independent data. However, such data are not usually available in practice. We use a random partitioning-based approach.

Algorithm B

Step 1. Define the legitimate models as in Step 1 of Algorithm A.

Step 2a. Partition the data into a training set of size $n_1=2n/3$ and a testing set of size $n_2=n/3$ [12].

Step 2b. For each given model defined in Step 1, fit the logistic model on the training set under constraint ($C1$) and form a linear score. Compute the AUC of the ROC curve of the linear test using the testing set.

Step 2c. To avoid getting a higher AUC just by a lucky partition, we repeat the random partition B times (for example $B=100$), and use the mean AUC of ROC as the criterion to rank all models.

Step 3. For each legitimate model defined in Step 1, repeat Steps 2a–2c to obtain the empirical AUC. Select the final model as the one with the highest AUC.

Note that the random partitioning to compute AUC described in Algorithm B is not implemented for the symptom counting-based method in Algorithm A. This is because, for a given model, the classification rule in (1) is known with all weights β_k being equal; therefore, no parameter is estimated from the data. Here, we are interested in computing an honest estimate of AUC for each given model in a given step of a variable selection procedure, and use these AUCs to rank legitimate models. The random partitioning is used to correct the potential upward bias caused by using the same data to estimate the coefficients β_k and to perform prediction. Essentially, in Algorithm B, given a model that contains a known set of variables but unknown weights, we carry out Steps 2a–2c to compute a corrected AUC. The corrected AUCs for all legitimate models are then used to choose the final model. In contrast, for symptom counting-based method, the coefficients in (1) are known for a given model containing known variables and no correction is needed if one is interested in using a good estimate of AUC to rank models.

We point out that if one is interested in assessing the predictive performance of the variable selection procedure itself, then without random partitioning the AUC computed in Algorithm A for the simple symptom counting would be biased by not adjusting for the model selection [13]. In other words, the estimated AUCs here can be interpreted as measuring prediction ability for a given model, but not as measuring predictivity of the variable selection procedure. Finally, to correct a potential overfitting bias in the model selection itself for the logistic regression-based method, one would have to embed the proposed all-subset selection into another level of random partitioning. In other words, one needs to randomly partition the data into two sets, apply all steps in Algorithm B on the training set, obtain the AUCs on the testing set, and average over repetitions. The computational burden of such a procedure is heavy because there are two levels of random partitioning, one used to correct for bias inherited in the model selection procedure itself and the other used to correct for bias in estimating β_k.

3. Simulation study

We carried out the following simulation study to evaluate the proposed methods. We constructed simulation model based on the real data. We generated nine predictors that belong to three groups with three variables in each group. Variables between groups are independent, whereas variables within a group are correlated. We specify the first variable in each group as the root variable and the following two variables as the stem variables. Specifically, the predictors (X_1, X_{1a}, X_{1b}) were first generated from a multivariate normal distribution with a mean of zero and an AR-1 correlation structure ($\rho=0.2$), and then dichotomized as 1 if a predictor was greater than or equal to 0 and otherwise 0. The variables in the other two groups were similarly generated. We assume that (X_1, X_{1a}, X_2, X_3) are the effective predictors. The outcomes were generated from the model

$$g(P(Y=1))=-1+2\times X_1\times X_{1a}+2\times X_2+2\times X_3, \tag{4}$$

where g is a link function (a logistic or a probit function). We considered four different scenarios. In scenarios I and II, we used the logistic and probit link, respectively, with a sample size $n=100$. In scenarios III and IV, we used the same simulation setting but doubled the sample size so that $n=200$. For each scenario, there were $5\times5\times5=125$ legitimate models. In each simulation, we ranked all the legitimate models according to their AUCs and chose the model with the highest AUC. Two methods were applied: the symptom counting-based selection and logistic regression. For the logistic regression, we estimated the AUC with 100 random partitions.

Table III summarizes the simulation results based on 200 replications. To compare sparsity of the fitted models by different methods, we summarize measures of model complexity and other features as usually reported in the literature (e.g. Zou and Li [14]). In Table III, the column indexed as 'C' is the mean number of effective variables correctly selected in the model. The column indexed as 'IC' is the mean number of noise variables incorrectly included in the model. The column 'Under-fit' is

Table III. Simulation results based on 200 replications: model (4).

	No. of zeros			Proportion of		
Method	AUC	C	IC	Under-fit	Correct-fit	Over-fit
Scenario I						
Symptom counting	0.820	3.815	0.315	0.160	0.610	0.230
Logistic regression	0.816	3.725	0.425	0.255	0.490	0.255
Scenario II						
Symptom counting	0.901	3.925	0.150	0.065	0.820	0.115
Logistic regression	0.898	3.745	0.235	0.250	0.605	0.145
Scenario III						
Symptom counting	0.815	3.935	0.130	0.060	0.845	0.095
Logistic regression	0.815	3.745	0.175	0.255	0.650	0.095
Scenario IV						
Symptom counting	0.904	3.968	0.048	0.011	0.936	0.047
Logistic regression	0.899	3.777	0.117	0.202	0.697	0.096

the proportion of models that miss some of the non-noise variables, the column 'Correct-fit' is the proportion of models that correctly select the exact subset of the non-null variables, and 'Over-fit' is the proportion of models that include some noise variables.

From Table III, we can see that the symptom counting-based selection and logistic regression both had a low 'under-fit' and a high 'correct-fit' proportion. In all the scenarios, the symptom counting-based selection performed better than logistic regression judging by its higher proportions of selecting the correct model, which is shown in the column 'Correct-fit'. Note that symptom counting uses the simple summation of all positive domains to form a linear test based on clinical knowledge without estimating a coefficient for each individual item as in a logistic regression. The proportion of correctly fitted models improves when the sample size is increased. It is interesting to see that the mean AUCs based on the 200 replications of the symptom counting selection and logistic regression are very close.

To evaluate the relative performance of the two approaches when the coefficients for the individual items are not all the same, we conducted a second set of simulations. The outcomes were generated from the following model:

$$g(P(Y=1))=-1+2\times X_1\times X_{1a}+1\times X_2+1.5\times X_3, \tag{5}$$

where g is again a logistic link function in scenarios I and III and a probit function in scenarios II and IV. All other parameters were the same as in the first set of simulations. Table IV summarizes the results of this simulation. Under model (5), symptom counting assumes a misspecified model with equal weight for each domain and is therefore expected to perform worse. We can see from Table IV that the logistic regression-based approach has a higher AUC, although the difference is small. In terms of sparsity, the logistic regression-based approach clearly outperforms symptom counting. For example, the proportion of 'Correct-fit' is higher for the logistic regression and the difference can be up to 15 per cent. These simulations suggest that the AUC of the final selected model is relatively insensitive to the model misspecification and the approach that is taken, while the sparsity of the selected model is less robust to model misspecification. We also conducted sensitivity analysis on the partition size through simulations with a $\frac{1}{2}n$ and $\frac{1}{2}n$ random partitioning. Compared with the original simulations with sizes $\frac{2}{3}n$ and $\frac{1}{3}n$, the AUCs and the sparseness of the fitted models are similar (results not shown).

4. Data analysis

Bisoffi *et al.* [15] discussed using ROC curves to evaluate screening questionnaires. Here, we applied the hierarchical search methods developed in Section 2 to design a new screening tool using the PSQ data. One hundred eighty patients being seen for primary care visits were randomly selected from a parent study of 1005 patients who had been systematically sampled from the patient population at the Associates of Internal Medicine clinic at CUMC. Out of the subsample of 180 patients, we located and fully interviewed 77. These 77 subjects were selected to undergo a PSQ screening and their SCID diagnosis was obtained by a trained clinical interviewer. SCID interviews and PSQ screens were

Table IV. Simulation results based on 200 replications: model (5).

Method	No. of zeros			Proportion of		
	AUC	C	IC	Under-fit	Correct-fit	Over-fit
Scenario I						
Symptom counting	0.761	3.760	0.575	0.220	0.460	0.320
Logistic regression	0.765	3.830	0.550	0.160	0.485	0.355
Scenario II						
Symptom counting	0.839	3.840	0.350	0.160	0.585	0.255
Logistic regression	0.859	3.925	0.270	0.070	0.730	0.200
Scenario III						
Symptom counting	0.761	3.890	0.290	0.110	0.675	0.215
Logistic regression	0.774	3.980	0.180	0.020	0.815	0.165
Scenario IV						
Symptom counting	0.842	3.930	0.170	0.070	0.800	0.130
Logistic regression	0.867	3.990	0.045	0.010	0.945	0.045

Table V. Top 10 models from hierarchical all-subset selection by symptom counting.*

Rank	AUC	Q1	Q1a	Q1b	Q2	Q2a	Q3	Q3a	Q3b	Q4	Q4a	Q5	Q5a
1	0.930	1	0	0	0	0	1	1	1	0	0	1	0
2	0.928	1	0	0	1	0	1	1	1	0	0	1	0
3	0.927	1	0	0	0	0	1	0	1	1	1	1	0
4	0.927	1	0	0	0	0	1	1	1	1	1	1	0
5	0.926	1	0	0	0	0	1	1	1	1	0	1	0
6	0.926	1	1	0	0	0	1	0	1	1	1	1	1
7	0.926	1	1	0	0	0	1	1	1	1	1	1	1
8	0.926	1	0	1	0	0	1	1	1	0	0	1	0
9	0.924	1	1	0	0	0	1	0	1	1	1	1	0
10	0.924	1	1	0	0	0	1	1	1	1	1	1	0

*Ranked by AUCs.

Table VI. Top 10 models from hierarchical all-subset selection by logistic regression*.

Rank	AUC†	Q1	Q1a	Q1b	Q2	Q2a	Q3	Q3a	Q3b	Q4	Q4a	Q5	Q5a
1	0.920	1	0	0	0	0	1	0	1	0	0	1	1
2	0.920	1	0	0	0	0	1	1	1	0	0	1	1
3	0.919	1	0	0	0	0	1	0	1	0	0	1	0
4	0.919	1	0	0	0	0	1	1	1	0	0	1	0
5	0.918	1	1	0	0	0	1	0	1	0	0	1	1
6	0.918	1	1	0	0	0	1	1	1	0	0	1	1
7	0.910	1	1	0	0	0	1	0	1	1	1	1	1
8	0.910	1	1	0	0	0	1	1	1	1	1	1	1
9	0.910	1	1	0	0	0	1	0	1	0	0	1	0
10	0.910	1	1	0	0	0	1	1	1	0	0	1	0

*Ranked by AUCs.
†Mean AUC estimated from 1000 random partitions.

conducted in either English or Spanish, based on respondents' preference. All respondents were self-identified as Latino. Over half (55.3 per cent) of the respondents spoke only Spanish and an additional third (32.9 per cent) spoke Spanish more fluently than English.

With the hierarchical constraint ($C1$), there were 675 legitimate models. Since the AUC of some top-ranking models may be very close, to decide on a final screening test, we take into account both the AUC and the parsimony of a model. We show the top 10 models were ranked based on their AUCs in Table V. Note that model 1 dominates models 2, 4, 5, 7, 8, and 10 in the sense that its items are a subset of items in the other models and it has a higher AUC. In other words, model 1 uses the least number of items to achieve a higher accuracy in terms of AUC compared with models 2, 4, 5, 7, 8, and 10. Model 3 dominates models 6 and 9 in the same sense. Since parsimony is preferred, we choose

a final model between 1 and 3. Furthermore, since model 1 has a higher AUC and 1 fewer item than model 3, our final model is model 1 with questions Q1, Q3 Q3a Q3b, and Q5.

We compare these results with a stepwise search that uses AUC as a model selection criterion but does not enforce the hierarchical structure among the variables. To be specific, at the first step in a stepwise procedure, we add the variable that has the largest AUC in predicting the SCID diagnosis without considering the hierarchical structure. A stem question by itself can be chosen to enter the model if it has the largest AUC. At the subsequent steps, we choose a variable such that the combination of this variable and the ones chosen at the previous step has the largest AUC. We stop when no addition of variables will increase the AUC. The final model contains variables Q1, Q3b, Q5, and Q5a. It is evident that this model does not satisfy the constraint ($C1$) because the root question for Q3b is not selected.

As a sensitivity analysis, we also carried out a logistic regression analysis based on all-subset selection introduced in Section 2. Table VI summarizes the top 10 models selected by logistic regression with 1000 random partitions to compute the AUC for each legitimate model. It is easy to see that all other models contain model 3. In other words, model 3 is the most parsimonious model of the top 10 candidate models. Note that the difference in the AUC for model 3 and model 1 is small, and model 3 contains the least number of items. Owing to parsimony being a desirable feature of our questionnaire, we select model 3 to be the final model from the logistic regression-based analysis. This model contains questions Q1, Q3, Q3b, and Q5.

The two hierarchical search methods lead to slightly different results. The direct counting of symptoms in each domain selects one more stem question (Q3a) compared with the logistic regression, and all the other selected questions are the same. The symptom counting search has slightly higher AUCs than those using logistic regression. However, the AUC computed from the logistic regression using the random partition can be more variable than the AUC obtained from the direct counting due to the sample splitting.

5. Discussion

Here we propose an all-subset model selection procedure for designing a questionnaire that takes into account the hierarchical structure among the variables. The goal of the variable selection is to search for the best items in a questionnaire (PSQ items) to predict a gold standard test diagnosis (SCID-based psychotic disorder). Since prediction performance is our main interest, we used the area under the ROC curve as the model selection criterion. Our scoring system of the items is a domain symptom count that also takes into account the hierarchical structure among the variables in a domain. For a given model, the rule of combining the scores from all items measured from an individual is a simple summation where each domain receives a weight of one. Therefore, for each given model, the parameters in the linear combination of constructing combined scores are known.

The all-subset selection can be cumbersome when the number of variables is large even with the hierarchical constraint. To facilitate an automatic variable selection with hierarchical structure, the procedures proposed in Zhao *et al.* [6] or Huang *et al.* [4] based on penalizing a loss function may be considered. In our application, predicting a gold standard test is of main interest. Therefore, a natural loss function would be the empirical AUC of an ROC curve defined in (3). However as noted in Pepe *et al.* [16], maximizing the empirical AUC of an ROC curve of a combined test with coefficients estimated from data is a numerically difficult problem. Further research along these lines deserves some attention.

Our criterion used to rank models is the empirical AUC of an ROC curve. However, this index does not take into account the number of variables used to construct the combined score. In our data analysis example, we used parsimony as a desirable feature to assist decision on the final model. It would be desirable to develop an AIC- or BIC-like criterion that automatically accounts for the number of variables used to construct the combined score as well as to accommodate the prediction goal.

Acknowledgements

The research of Yuanjia Wang's is supported by US NIH grant AG031113-01A2. The research of Roberto Lewis-Fernández is supported by the National Alliance for Research on Schizophrenia and Depression (NARSAD),

by the Bureau of Cultural Competence of the New York State Office of Mental Health, and by institutional funds from the New York State Psychiatric Institute.

References

1. Bebbington P, Nayani T. The psychosis screening questionnaire. *International Journal of Methods in Psychiatric Research* 1995; **5**(1):11–19.
2. Lewis-Fernández R. Assessing psychosis screeners among underserved urban primary care patients. *Fifteenth Annual Scientific Symposium*, New York, NY, National Alliance for Research on Schizophrenia and Depression (NARSAD), Great Neck, NY, 17–18 October 2003.
3. First MB, Spitzer RL, Gibbon M, Williams J. Structured clinical interview for DSM-IV axis I disorders, Patient Edition (SCID-I/P, version 2.0, 9/98 revision). Biometrics Research Department, New York State Research Institute, 1998.
4. Huang J, Ma S, Xie H, Zhang C. A group bridge approach for variable selection. *Biometrika* 2009; **96**:339–355.
5. Wang S, Nan B, Zhou N, Zhu J. Hierarchically penalized Cox regression with grouped variables. *Biometrika* 2009; **96**:307–322.
6. Zhao P, Rocha G, Yu B. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics* 2009; **37**:3468–3497.
7. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, *Series B* 2006; **68**:49–67.
8. Pepe MS. Evaluating technologies for classification and prediction in medicine. *Statistics in Medicine* 2005; **24**: 3687–3696.
9. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford, 2003.
10. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* 1983; **78**:316–331.
11. Efron B. The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association* 2004; **99**:619–632.
12. Ma S, Huang J. Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics* 2005; **21**:4356–4362.
13. Ye J. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association* 1998; **93**(441):120–131.
14. Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics* 2008; **36**:1509–1533.
15. Bisoffi G, Mazzi MA, Dunn G. Evaluating screening questionnaires using Receiver Operating Characteristics (ROC) curves from two-phase (double) samples. *International Journal of Methods in Psychiatric Research* 2000; **9**:121–133.
16. Pepe MS, Cai T, Longton G. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* 2006; **62**:221–229.