Improved biomarker performance for the detection of hepatocellular carcinoma by inclusion of clinical parameters

Mengjun Wang[1], Timothy M. Block[1], Jorge Marrero[2], Adrian M. Di Bisceglie[3], Karthik Devarajan[4], and Anand Mehta[1]

[1]Drexel University College of Medicine, Doylestown, PA, 18901. [2]Division of Gastroenterology, University of Michigan, 3912 Taubman Center, Ann Arbor, MI 48109-0362. [3]Saint Louis University School of Medicine, 1402 S. Grand FDT 12th Floor, St. Louis, Mo 63104 [4]Fox Chase Cancer Center, Philadelphia, PA, 18901.

Abstract- **We have previously identified several biomarkers of hepatocellular carcinoma (HCC). The levels of three of these biomarkers were analyzed individually and in combination with the currently used marker, alpha fetoprotein (AFP), for the ability to distinguish between a diagnosis of cirrhosis (n=113) and HCC (n=164). We have utilized several novel biostatistical tools, along with the inclusion of clinical factors such as age and gender, to determine if improved algorithms could be used to increase the probability of detection of cancer. Using several of these methods, we are able to detect HCC in the background of cirrhosis with an AUC of at least 0.95. The use of clinical factors in combination with biomarker values to detect HCC is discussed.**

Keywords- Hepatocellular Carcinoma, biomarkers, Hepatitis B virus, logistic regression, penalized logistic regression, classification and regression trees.

I   INTRODUCTION

Infection with hepatitis B virus (HBV) and/or hepatitis C virus (HCV) is the major etiology of hepatocellular cancer (HCC)[1-4]. The progression of into liver cancer is primarily monitored by serum levels of the oncofetal glycoprotein, alpha-fetoprotein (AFP) or the core fucosylated glycoform of AFP (AFP-L3). However, AFP can be produced under many circumstances [5-7], and is not present in all with HCC. Therefore the use of AFP as a the sole screen for HCC has been questioned [8] and more sensitive serum biomarkers for HCC are desired.

Using fucose specific lectins to identify the proteins that become fucosylated with liver disease, we have identified glycoproteins that contained increased fucosylation with HCC [9]. In the current study we have analyzed the performance of several biomarkers in 113 patients with cirrhosis and 164 patients with cirrhosis plus HCC. In an effort to maximize the detection of patients with cancer, we utilized several novel biostatistical tools to determine if improved algorithms could be used to increase the probability of cancer detection. This included combining marker values with clinical factors such as age and gender to improve diagnosis. Using

several of these methods, we are able to detect HCC in the background of cirrhosis with a predictive probability of at least 0.95, which was much greater than any marker when used alone. The potential of using this combination of markers and clinical values is discussed.

II   MATERIALS AND METHODS

Patients: Serum samples were obtained from Saint Louis University School of Medicine and the University of Michigan. In both cases the Institutional Review Board approved the study protocol and written informed consent was obtained from each subject. Patient and clinical information is presented in [10].

Lectin FLISA and analysis of GP73: Analysis of fucosylated A1AT and kininogen was performed as described in [10]. GP73 was analyzed by immunoblotting as described in [11].

Statistical Methods: Univariate statistical analyses were performed using the Fisher's Exact test for categorical variables and the Mann-Whitney test for continuous variables. Univariate logistic regression analyses were also performed for each individual biomarker separately. A variety of methods were used in multivariable analyses for associating the incidence of HCC with biomarker levels and clinical/demographic variables such as age and gender. Specifically, three different but related methods were investigated in this approach – logistic regression (LR), penalized logistic regression (PLR) and Classification and Regression Trees (CART). All tests were two-sided and used a Type I Error of 0.05 to determine statistical significance.

PLR is a variant of logistic regression based on a quadratic penalty that is ideal for associating discrete factors and continuous variables such as gender, age and biomarker levels with a binary response such as HCC incidence[12]. In PLR, we maximize the log-likelihood subject to a size constraint on the L_2-norm of the coefficients (excluding intercept)[12]. This penalized likelihood can be written as $L(\beta) = -l(\beta) + (\lambda/2) \|\beta\|^2_2$. Here, l indicates the binomial log-likelihood, β is the parameter vector and λ is a positive constant. PLR is well suited for modeling a large number of variables. Variable selection can be done using a forward stepwise approach. Different values of the penalty parameter λ were considered in our approach. PLR is implemented in the open-source R package stepPLR [13].

CART [14] is based on decision trees and is non-parametric. A decision tree is a logical model represented as a binary tree that shows how the value of a response variable can be predicted by using the values of a set of variables. If the response variable is binary such as whether a patient developed HCC or not, then a

classification tree is generated that predicts the probability of developing HCC. The unified CART framework embeds recursive binary partitioning into the theory of permutation tests and is implemented in the open-source R package PARTY [13].

In order to evaluate the performance of combining multiple biomarkers and/or clinical variables, values of multiple biomarkers were inputted into the model from the appropriate method, and in each case the output (predicted value) was between 0 and 1, with 0 being cirrhosis and 1 being cancer. A cut-off of 0.5 was used for the predicted probability p and patients were classified as being HCC positive when $p>=0.5$, otherwise they were classified as cirrhotic ($p<0.5$). To determine the optimal cutoff value for each biomarker or a combination of biomarkers and/or clinical variables, Receiver Operating Characteristic (ROC) curves were constructed using all possible cutoffs for each method. Sensitivity and specificity (along with 95% confidence intervals (CI)) were used to characterize the precision of binary predictions from LR, PLR and CART. Area under the ROC curves (along with 95% CI), prediction accuracy (ACC), positive predictive value (PPV) and negative predictive value (NPV) were used to characterize the predictive value of models from these methods. Model selection within each method was done using the Akaike Information Criterion (AIC) wherever appropriate. In addition, the performance of each model was evaluated using leave-one-out cross validation (LOOCV) and three-fold cross validation (3CV).

Using results from LOOCV, an ROC curve and its AUC (with 95% CI) was computed based on the predicted probabilities. This is the cross-validated AUC. In order to evaluate the performance of each model on independent data in the absence of a validation set, 3CV was used. Here, the dataset is divided randomly into three equal parts. The combined data from two parts are used to fit a model using a particular method, and this model is used to predict the HCC status of each observation in the left out part. This process is repeated for 200 random partitions of the dataset and the mean AUC (and its 95% CI) was computed.

III   RESULTS AND INTERPRETATION

Univariate analyses revealed a significant association between gender and the incidence of HCC (Fisher's Exact test p-value=0.036). The odds of HCC in males was 1.75 times higher than that in females (95% CI: (1.01,3.02)). There was also a statistically significant association between age and incidence of HCC (Mann-Whitney test p-value < 0.0001). The median age of patients with HCC was 58 years compared to 51 years for those with cirrhosis. There was no significant difference in gender or in HCC incidence between the two sites.

Data obtained across two sites were used in the analyses. In order to adjust for any potential differences in biomarker levels obtained at different sites, a dichotomous, nominal variable Site (indicating the site where the data was obtained for each observation) was incorporated into the modeling as a covariate. No statistically significant effect due to site was observed in any of the models or methods applied. For each statistical method used, four different models were considered based on the inclusion of age and gender in multivariable analysis. These are listed in Table 1. The stratified dataset consisting of males only (with or without age) was of particular importance due to the known higher incidence of HCC in male patients [2]. Whenever age was included in a multivariable model using any method, it was found to be statistically significant (data not shown).

In addition, no statistically significant interactions were identified in multivariable LR analyses. Results from multivariable analyses (presented in Tables 1 & 2, Figures 1-4) were compared with those from univariate LR (data not shown) applied to each individual biomarker separately. Univariate LR models performed uniformly worse than multivariable models that utilized multiple biomarkers using *any* of the three methods. It turns out that the best performing univariate model (GP73) produced a model-based AUC of 0.87 (95% CI (0.84, 0.91)) and ACC of 0.78, a result that fell far short of those of multivariable models, and thus emphasized the need for including multiple biomarkers and additional confounding clinical variables into the model.

Using multivariable models, there was a significant improvement in the predictive performance of each method when age was included in the model after adjusting for gender differences compared to the model excluding age (Table 1, Figures 2 & 3). In particular, PLR and CART showed improvements in AUC (ACC) of 2.53% (2.27%) and 5% (4.16%), respectively. This difference was more pronounced when the stratified dataset consisting of only males was used in the analysis (Table 1, Figures 1 & 4). In this case, all three methods showed improvements in AUC and ACC in excess of 4%, with PLR ($\lambda=1$) showing an increase in ACC of nearly 5%. In fact, a marked improvement is observed in the predictive performance of each method based on this stratified dataset independent of whether age is included in the model. However, the inclusion of age results in the best predictive model across all methods considered (Table 1). It should also be noted that the inclusion of age results in a substantial decrease in AIC for LR and PLR (all choices of λ) both for the stratified male only dataset and when gender differences are accounted for in the model (Table 1).

Furthermore, PPV and NPV capture other critical aspects of the performance of a model. For our application, PPV represents the proportion of patients *correctly* predicted to have HCC while NPV represents the proportion of patients *correctly* predicted to have cirrhosis. A high PPV means that the model only rarely classifies a HCC patient as having cirrhosis, and is

therefore a desirable characteristic in a model. Table 1 lists the best performing models and methods in terms of PPV and NPV. Once again, models that included age generally showed a higher PPV or NPV compared to those that did not. For the stratified male only data, LR and PLR (λ=0.1) resulted in a 3.57% increase in PPV with the inclusion of age while PLR (λ=10) and CART increased NPV by nearly 9% and 6.6%, respectively. For all three methods, the highest PPV (91.96%) was achieved for the stratified male only data. When gender effect was adjusted for in the model, PLR (λ=10) resulted in the maximum increase in NPV of 4.7% due to the inclusion of age.

Interpretation of CART results: Multivariable CART analysis of the complete dataset revealed that age and levels of GP73, AFP and Kininogen were significantly associated with increased incidence of HCC after controlling for site and gender. A complex interplay between the various biomarkers and age was observed. Similarly, multivariable CART analysis of the stratified (male only) data revealed that age and levels of GP73, AFP and Kininogen were significantly associated with increased incidence of HCC in males after controlling for site. Higher levels of the markers GP73 (>3.8), AFP (>1.3) and Kininogen (>1.7) were significantly associated with increased incidence of HCC (p<0.001 in all cases). These correspond to node pairs (1,9), (3,7) and (4,6), respectively, in Figure 4. Moreover, older men were identified to have a significantly higher incidence of HCC (GP73<=3.8, age>60, p=0.014 and GP73>3.8, age>48, p<0.001) corresponding to node pairs (2,8) and (9,11), respectively, in Figure 4). The highest incidence of HCC was observed in the subgroup of men with GP73>3.8 and aged over 48 years (74 patients, 72/74(97.29%) are HCC) while the lowest incidence of HCC was observed in the subgroup of men with Kininogen<=1.7 (as well as AFP<=1.3 and GP73<=3.8) under 60 years of age (44 patients, 0/44(0%) are HCC).

Predictive Performance of Multivariable Models using Cross-validation: While model based metrics such as AUC, ACC, PPV and NPV provide a measure of the predictive performance of a model, equivalent versions of these quantities based on cross-validation are based on blinded, independent datasets and therefore provide the true predictive performance of the model. Table 2 presents the AUC (with 95% CI) and ACC for each model and method used based on LOOCV and 3CV. A notable improvement in AUC is observed in models that include age across all three methods for the stratified male only data. The median value of this increase is around 5% for AUC based on LOOCV and around 4% for AUC based on 3CV. When gender is accounted for in the model, the inclusion of age also results in an improvement in AUC of about 2.5% for PLR (median value over 3 choices of λ, based on both LOOCV and 3CV) and a substantial 6% increase for CART using LOOCV. On the other hand,

LR did not contribute to a significant increase in AUC (0.45 and 0.16 based on LOOCV and 3CV, respectively. In terms of prediction accuracy, significant improvement in ACC is observed in models that include age for LR and PLR for the stratified male only data. The median value of this increase is around 5.5% for ACC based on LOOCV and around 5.25% for ACC based on 3CV. When gender is accounted for in the model, the inclusion of age also results in improvements of 3.41% and 3.09% for PLR (median value over 3 choices of λ) based on both LOOCV and 3CV, respectively. These improvements are relatively smaller for LR (1.27% and 2.79%) while the performance of CART is seen to vary between models and cross-validation methods. It is clear that multivariable models using methods considered here outperform corresponding univariate models based on each individual biomarker. Moreover there is strong evidence, overall, that the predictive performance of PLR is superior to that of LR and CART.

IV   SUMMARY AND DISCUSSION

In this paper, we demonstrated the usefulness of incorporating multiple biomarkers and relevant clinical variables into a statistical model for predicting the incidence of HCC. Specifically, we investigated the predictive performance of three different yet related methods, namely LR, PLR and CART, in distinguishing HCC patients from cirrhotic patients. While all three approaches provided overall improvement compared to the use of single biomarkers, our results suggested that important differences exist between these methods. For example, PLR and CART provided more significant improvements in various aspects of predictive performance compared to traditional LR. One novel aspect of our approach has been the application of CART for analyzing and interpreting biomarker data for HCC. The non-parametric approach in CART is a useful alternative to traditional parametric methods like LR and PLR. CART automatically incorporates interactions between multiple biomarkers and/or clinical variables. It provided potentially useful cut-offs for biomarkers and clinical variables alike that indicated a statistically significant association with increased HCC incidence. In that sense, CART can be seen as a complementary approach to LR and PLR and it sets the stage for further evaluation and validation of the clinical significance of these results in future, larger studies. An important finding in this study is the marked improvement in predictive performance due to the inclusion of clinical factors such as age and gender. This improvement was seen to be independent of the method used in the analysis. One of the goals in this study has been to identify a model predictive of HCC in males due to its known higher risk in this subgroup. It turned out that models based on the stratified male only subset showed the best predictive performance overall.

One possible avenue for future research on this topic would be the application of a method that borrows strength from the binary recursive partitioning approach in CART as well as the parametric approach in LR [14]. It will form the basis of our future investigation [14]. In addition, the inclusion of other clinical factors such as Alanine transaminase, Aspartate transaminase and Alkaline phosphatase levels may be able to increase performance even further. This is currently under investigation.

Table 1: Model-based performance measures (Multivariable Models)

Method	AUC (95% CI)	ACC	PPV	NPV	AIC
LR (with age, males only)	0.97 (0.94-0.99)	88.4	92.0	82.6	93.2
LR (without age, males only)	0.92 (0.89-0.96)	84.0	88.4	76.8	133.3
LR (with age, gender)	0.96 (0.94-0.98)	88.3	91.0	84.3	150.6
LR (without age, gender)	0.95 (0.93-0.97)	87.9	89.6	86.2	187.6
PLR (λ=0.1) (male only,with age)	0.97 (0.95-0.99)	88.4	92.0	82.6	93.3
PLR (λ=1) (male only, with age)	0.97 (0.94-0.99)	88.4	91.2	83.6	96.0
PLR (λ=10) (male only, with age)	0.95 (0.93-.98)	88.4	90.5	84.6	111.9
PLR (λ=0.1) (male only,without age)	0.92 (0.87-0.967)	84.0	88.4	76.8	131.2
PLR (λ=1) (male only, without age)	0.92 (0.89-0.960)	83.4	88.3	75.7	132.3
PLR (λ=10) (male only, without age)	0.91 (0.87-0.95)	83.5	88.3	75.7	143.9
PLR (λ=0.1) (gender,with age)	0.96 (0.94-0.98)	88.3	91.0	84.3	148.6
PLR (λ=1) (gender, with age)	0.958 (0.936-0.981)	88.3	91.0	84.3	149.86
PLR (λ=10) (gender, with age)	0.95 (0.93-0.98)	88.3	91.6	83.6	165.4
PLR (λ=0.1) (gender,without age)	0.93 (0.90-0.96)	86.0	89.6	90.9	183.6
PLR (λ=1) (gender, without age)	0.93 (0.90-0.96)	86.0	90.13	80.4	184.4
PLR (λ=10) (gender, without age)	0.93 (0.90-0.96)	84.8	89.9	78.9	197.4
CART (with age,male only)	0.930 (0.90-0.97)	88.4	92.0	82.6	NA
CART (without age,male only)	0.89 (0.85-0.94)	85.1	91.5	76	NA
CART (with age,gender)	0.96 (0.94-0.99)	91.3	91.3	85.3	NA
CART (without age,gender)	0.91(0.88-0.95)	87.1	90.9	82.0	NA

Table 2: Performance measures based on cross-validation (Multivariable Models)

Method	LOOCV AUC (95% CI)	LOOCV ACC	3CV AUC (95% CI)	3CV ACC (SD)
LR (with age, males only)	0.97(0.94-0.99)	87.3	0.97 (0.94-0.99)	87.3 (4.0)
LR (without age, males only)	0.92 (0.89-0.96)	81.8	0.92 (0.89-0.96)	81.8 (4.3)
LR (with age, gender)	0.96 (0.94-0.98)	87.9	0.96 (0.94-0.98)	88.0 (3.1)
LR (without age, gender)	0.95 (0.93-0.97)	86.6	0.95 (0.93-0.97)	85.2 (3.1)
PLR (λ=0.1) (male only, with age)	0.97 (0.95-0.99)	87.3	0.97 (0.95-0.99)	86.7 (3.4)
PLR (λ=1) (male only, with age)	0.97 (0.94-0.99)	87.3	0.97 (0.94-0.99)	87.8 (8.4)
PLR (λ=10) (male only, with age)	0.95 (0.93-0.98)	87.3	0.95 (0.93-0.98)	86.4 (3.9)
PLR (λ=0.1) (male only, without age)	0.92 (0.89-0.96)	81.8	0.92 (0.89-0.96)	81.73 (4.31)
PLR (λ=1) (male only, without age)	0.92 (0.89-0.96)	82.3	0.92 (0.89-0.96)	81.96 (4.40)
PLR (λ=10) (male only, without age)	0.91 (0.87-0.95)	81.2	0.91 (0.87-0.95)	81.8 (4.4)
PLR (λ=0.1) (gender, with age)	0.96 (0.94-0.98)	87.9	0.96 (0.94-0.98)	87.2 (3.1)
PLR (λ=1) (gender, with age)	0.96 (0.94-0.98)	87.9	0.96 (0.94-0.98)	88.0 (2.7)
PLR (λ=10) (gender, with age)	0.95 (0.93-0.98)	86.7	0.95 (0.93-0.97)	86.3 (3.1)
PLR (λ=0.1) (gender, without age)	0.93 (0.90-0.96)	84.5	0.93(0.90-0.96)	84.1 (3.40)
PLR (λ=1) (gender, without age)	0.93 (0.90-0.96)	84.1	0.93 (0.94-0.96)	84.4 (3.40)
PLR (λ=10) (gender, without age)	0.93(0.90-0.96)	84.1	0.93 (0.90-0.96)	83.9 (3.4)
CART (with age, male only)	0.93(0.90-0.97)	83.4	0.93 (0.90-0.97)	81.5 (5.0)
CART (without age, male only)	0.89 (0.85-0.94)	84.5	0.89 (0.85-0.94)	79.3 (5.0)
CART (with age, gender)	0.96 (0.94-0.99)	82.2	0.96 (0.94-0.99)	81.9 (3.5)
CART (without age, gender)	0.91 (0.88-0.95)	79.5	0.91 (0.88-0.95)	83.6 (3.9)

References

1. Di Bisceglie, A.M., *Hepatocellular carcinoma: molecular biology of its growth and relationship to hepatitis B virus infection*. Med Clin North Am, 1989. **73**(4): p. 985-97.

2. Block, T.M., A.S. Mehta, C.J. Fimmel, and R. Jordan, *Molecular viral oncology of hepatocellular carcinoma*. Oncogene, 2003. **22**(33): p. 5093-107.

3. Marrero, J.A., *Hepatocellular carcinoma*. Curr Opin Gastroenterol, 2006. **22**(3): p. 248-53.

4. Sallie, R. and A.M. Di Bisceglie, *Viral hepatitis and hepatocellular carcinoma*. Gastroenterol Clin North Am, 1994. **23**(3): p. 567-79.

5. Alpert, M.E., J. Uriel, and B. de Nechaud, *alpha fetogloblin in the diagnosis of human hepatoma*. N Engl J Med, 1968. **278**: p. 984-6.

6. Ruoslahti, E., M. Salaspuro, H. Pihko, L. Andersson, and M. Seppala, *Serum alpha-fetoprotein: diagnostic significance in liver disease*. Br Med J, 1974. **2**(918): p. 527-9.

7. Di Bisceglie, A.M. and J.H. Hoofnagle, *Elevations in serum alpha-fetoprotein levels in patients with chronic hepatitis B*. Cancer, 1989. **64**(10): p. 2117-20.

8. Sherman, M., *Hepatocellular carcinoma: epidemiology, risk factors, and screening*. Semin Liver Dis, 2005. **25**(2): p. 143-54.

9. Comunale, M.A., Lowman, M., Long, R.E., Krakover, J., Philip, R., Seeholzer, S., Evans, A. A., Hann, H.W.L, Block, T.M, Mehta, A.S, *Proteomic analysis of serum associated fucosylated glycoproteins in the development of primary hepatocellular carcinoma*. Journal of Proteome Research., 2006. **6**(5): p. 308-315.

10. Wang, M., R.E. Long, M.A. Comunale, O. Junaidi, J. Marrero, A.M. Di Bisceglie, T.M. Block, and A.S. Mehta, *Novel fucosylated biomarkers for the early detection of hepatocellular carcinoma*. Cancer Epidemiol Biomarkers Prev, 2009. **18**(6): p. 1914-21.

11. Marrero, J.A., P.R. Romano, O. Nikolaeva, L. Steel, A. Mehta, C.J. Fimmel, M.A. Comunale, A. D'Amelio, A.S. Lok, and T.M. Block, *GP73, a resident Golgi glycoprotein, is a novel serum marker for hepatocellular carcinoma*. J Hepatol, 2005. **43**(6): p. 1007-12.

12. Park, M.Y. and T. Hastie, *Penalized logistic regression for detecting gene interactions*. Biostatistics, 2008. **9**(1): p. 30-50.

13. Team, R.D.C., *R: A language and environment for statistical computing.* , in *R Foundation for Statistical Computing* 2005. p. URL http://www.R-project.org.

14. Hothorn, T., K. Hornik, and A. Zeileis, *Unbiased recursive partitioning: A conditional inference framework*. Journal of Computational and Graphical Statistics, 2006. **15**(3).
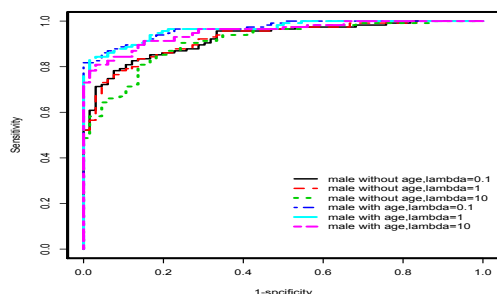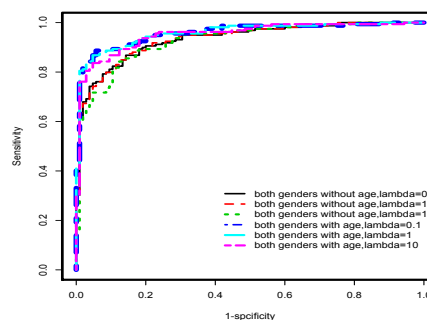
Figure 2. ROC curves for penalized logistic regression models incorporating gender. See Tables 1&2 and text for AUCs and further discussion
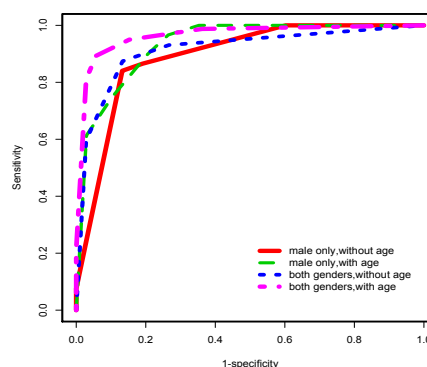

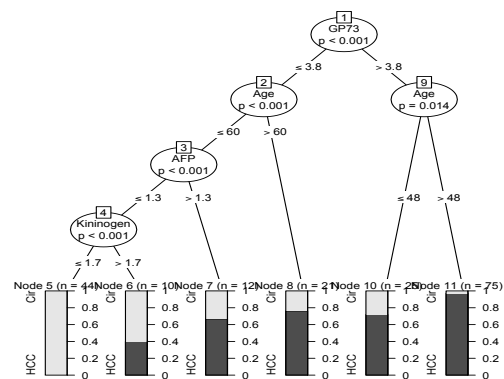Figure 3. ROC curves for CART analysis. See Tables 1&2 for more information.

Figure 4. CART analysis in male patients. See text and Tables 1& 2 for more details.

Figure 1. ROC curves for penalized logistic regression models in males only. See Tables 1-2 and text for AUCs and more discussion.