Web-based Supplementary Materials for " Using the Optimal Robust Receiver Operating Characteristic (ROC) Curve for Predictive Genetic Tests" by Qing Lu, Nancy Obuchowski, Sungho Won, Xiaofeng Zhu, and Robert C. Elston

Web Appendix A: Forming the True Optimal ROC Curve

Suppose we are interested in combining data on *p* disease loci for disease prediction. Let the *j*-th $(j = 1, \cdots, p)$ disease locus have m_j genotypes, and let $R_j = (r_{j1}, \cdots, r_{jm_j})$ and $F_j = (f_{j1}, \cdots, f_{jm_j})$ denote the corresponding relative risks and genotype frequencies for these m_j genotypes. We can then calculate the conditional probability given disease of the k_j th $(k_j = 1, \cdots, m_j)$ monogenic genotype at the *j*-th locus as

$$P(g_{jk_j} \mid S = 1) = \frac{r_{jk_j} f_{jk_j}}{\sum_{k_j=1}^{m_j} r_{jk_j} f_{jk_j}}. \qquad (1)$$

Then by applying equations (2) – (4) in section 3.1, we form the true optimal ROC curve and calculate its AUC.

What we described in the text was based on the *p* disease loci exhibiting no interaction. If some of the *p* loci interact, the method can be easily extended to handle such a situation. Without loss of generality, we consider two loci, j and j', that interact with each other. The joint probabilities of the two-locus genotypes can be calculated from the single locus genotype frequencies, on assuming linkage equilibrium, as $P(G_l) = P(g_{jk_j}, g_{j'k_{j'}}) = f_{jk_j} \cdot f_{j'k_{j'}}$. If the two loci are in linkage disequilibrium, we can obtain the joint probabilities of the genotypes from the haplotype frequencies. Then, given the underlying interaction model, we cluster together the two-locus genotypes that share the same risk of disease when estimating the cluster frequencies. For instance, if the underlying interaction model is the threshold model (Marchini, Donnelly, and

Cardon, 2005), we would cluster all possible two-locus genotypes into two groups: (1) a single high risk group, which we denote H, for all individuals having at least one of the disease-susceptibility alleles at each of the two loci, with frequency $f_H = \sum_{G_l \in H} P(G_l)$; and (2) a common low risk group, which we denote C, for all other individuals, with frequency $f_C = \sum_{G_l \in C} P(G_l)$. Let $r_{H/C}$ denote the risk of the high risk group relative to that of the low risk group. By treating these interacting loci as comprising one set of genotypes and applying equations (2) – (4), we can also incorporate interacting loci with this approach.

Web Appendix B: A Simple Example Illustrating the Method Described in Section 3.2

Suppose we have two disease susceptibility loci, A and B, that were found to be significantly associated with disease in previous association studies. We wish to investigate their potential role in disease prediction. Assuming each locus has three genotypes (i.e., genotypes AA, $A\overline{A}$ and $\overline{A}\overline{A}$ for locus A, and BB, $B\overline{B}$ and $\overline{B}\overline{B}$ for locus B), from which we could have nine possible multi-locus genotypes, we estimate the likelihood ratios (LRs) for each of the nine multi-locus genotype using the entire dataset. We then rank the nine multi-locus genotypes in descending order of their LRs, from the highest rank to the lowest rank and plot the optimal ROC curve. This represents the full model with the largest number of multi-locus genotype clusters (each comprising one genotype). To reduce the model complexity, we adopt the backward clustering algorithm to gradually combine the multi-locus genotype clusters with each other. In the first step of the backward clustering process, we consider the following six possibilities (models) to combine pairs of multi-locus genotypes based on two particular one-locus genotypes:

i. $AA + A\overline{A}$

ii. $AA + \overline{A}\overline{A}$

iii. $A\overline{A} + \overline{A}\overline{A}$

iv. $BB + B\overline{B}$

v. $BB + \overline{B}\overline{B}$

vi. $B\overline{B} + \overline{B}\overline{B}$

Note that each of these six possible clusterings results in six clusters. For each one of them, we rank the six clusters by their LRs, and form the corresponding optimal ROC curve using the entire dataset. The candidate model at this step is chosen from the above six models based on having the highest AUC value. Note that because the models ii. and v. are not normally considered biologically plausible, we might exclude them from the algorithm, and only consider the remaining four models. Assuming model i. is chosen as the candidate model, i.e. we have pooled together the AA and $A\overline{A}$ multi-locus genotypes, we consider the following four models in the next step:

i. $AA + A\overline{A} + \overline{A}\overline{A}$

ii. $BB + B\overline{B}$

iii. $B\overline{B} + \overline{B}\overline{B}$

iv. $BB + \overline{B}\overline{B}$

Using the same strategy, we can choose the candidate model for step two. Here we might also exclude model iv. because it is not biologically plausible. If model i is chosen, then locus A is dropped from consideration. We repeat the clustering process until all multi-locus genotypes fall into one group at step T (in this example $T = 4$). By repeating the clustering process, we obtain as a maximum $T + 1$ candidate models, $G^{(0)}, G^{(1)}, \cdots, G^{(T)}$, with respectively nine, six, three(four), two, and one multi-locus genotype clusters. We use the number of multi-locus genotype clusters to measure the model's complexity, and conduct ten-fold cross validation to choose the most parsimonious model with an appropriate model complexity.

In ten-fold cross validation, we randomly partition the entire dataset into ten subsets. Nine subsets are used for the purpose of training and one subset is used for the purpose of validation. We repeat the backward clustering algorithm described above in each of the ten training datasets to obtain candidate models and in each case calculate the prediction AUC in the corresponding validation dataset, the latter being averaged over the ten values. Letting $nc^{(m)}$ denoted the number of multi-locus genotype clusters with a maximum average prediction AUC in the cross validation, the corresponding candidate model $G^{(m)}$ with $nc^{(m)}$ multi-locus genotype clusters is chosen as the most parsimonious model. Based on the selected model, we estimate the fitted AUC value using the entire dataset. In order to obtain the prediction AUC, an independent dataset is required.

For a dataset with a smaller sample size and a large number of disease susceptibility loci, occasionally in the cross validation process some of the multi-locus genotypes might only be present in the validation dataset and not in the training dataset. Assume, for example, $\{AA \ BB\}$ is only presents in the validation dataset, so that we are unable to infer where to place $\{AA \ BB\}$ from any model built in the training dataset. Instead of treating it as missing, we adopt the same strategy of gradually clustering the multi-locus genotypes until, after a few steps, $\{A\bar{A} \ BB\}$ and $\{\bar{A}\bar{A} \ BB\}$ are clustered together in the training dataset. Because the statistic (e.g., the LR) associated with the cluster $\{A\bar{A} \ BB + \bar{A}\bar{A} \ BB\}$ is the same as for $\{BB\}$, we can remove locus A and use $\{BB\}$ to represent $\{A\bar{A} \ BB + \bar{A}\bar{A} \ BB\}$ in the training dataset. We apply this process to the validation dataset, and use the LR of $\{BB\}$ estimated from the training dataset to infer the order of $\{AA \ BB\}$ in the validation dataset.

The backward clustering algorithm applies naturally to both disease and marker selection. For instance, if the chosen model is i. in step one, this would imply that SNP A follow a model in which A is dominant and $\overline{A}$ is recessive. In another case, as noted above, if the final chosen is i. in step two, this would imply that SNP A is a noise locus and it is removed by the algorithm.

Web Appendix C: Simulation III

We simulated three independent diallelic SNP loci with the disease susceptibility allele frequencies 0.4, 0.3 and 0.2, respectively. We assumed that the disease prevalence ρ is equal to 0.05 and at each locus the rarer allele follows a recessive model with respective relative risks 4, 3 and 2. Sampling 1000 cases and 1000 controls from the simulated population data, we investigated the logistic regression model, the classification tree and the optimal robust ROC curve method described in section 3.2. Among all approaches, the logistic regression that assumed a recessive model (i.e., the right mode of inheritance) gave the most accurate AUC estimates and the logistic regression that assumed a dominant model seriously underestimated the AUC values. The optimal robust ROC curve performs better than the logistic regression model that assumes a multiplicative model in terms of MSE (Web Table 1).

REFERENCES

1. Baker SG. (2000). Identifying combinations of cancer markers for further study as triggers of early intervention. *Biometrics* 56 (4):1082-1087.

2. Marchini J, Donnelly P, and Cardon LR. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics* 37 (4):413-417.

TABLES AND FIGURES

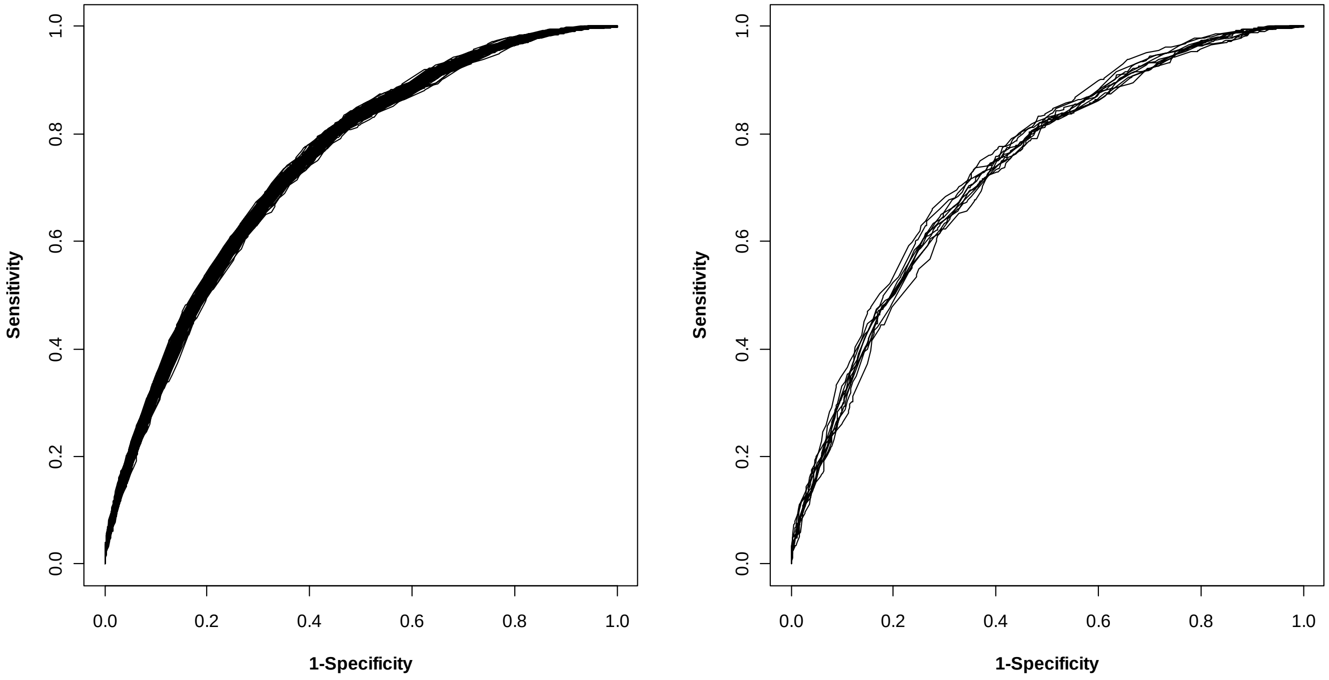
Web Table 1. AUC estimates and standard deviation (SD) of the unordered optimal (UNORD) and jagged ordered optimal (JAGGED) methods (Baker, 2000), based on 1000 repeated split samples from the Wellcome Trust dataset.

	Training Dataset		Validation Dataset	
	AUC	SD	AUC	SD
UNORD	0.7639	0.0045	0.7153	0.0095
JAGGED	0.7618	0.0046	0.7171	0.0095

Web Table 2. Comparison among of AUC obtained by the optimal robust ROC curve method (OPT-ROC), logistic regression with backward selection (Mul LOG-REG, Dom LOG-REG, and Rec LOG-REG), and classification tree (CLA-TREE)

Cases:Controls	1000:1000		
	BIAS	SD	MSE
OPT-ROC	0.0074	0.0133	0.00023
Mul LOG-REG	-0.0144	0.0126	0.00037
Dom LOG-REG	-0.1024	0.0129	0.01065
Rec LOG-REG	0.0005	0.0104	0.00011
CLA-TREE	-0.0116	0.0104	0.00024

Web Figure 1. ROC curves, obtained using the multiplicative logistic regression model, from 100 repeated split samples. The left panel shows the ROC curves estimated from the training dataset and the right panel shows the ROC curves estimated from the validation dataset.


Sensitivity
1-Specificity
Sensitivity
1-Specificity