Published in final edited form as:

Biometrics. 2014 December ; 70(4): 1042–1051. doi:10.1111/biom.12217.

A Serial Risk Score Approach to Disease Classification that Accounts for Accuracy and Cost

Dat Huynh^{1,*}, Oliver Laeyendecker^{2,**}, and Ron Brookmeyer^{3,***}

¹Department of Biostatistics, University of California, Los Angeles

²National Institute of Allergy and Infectious Diseases, National Institutes of Health, and Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD

³Department of Biostatistics, University of California, Los Angeles

Summary

The performance of diagnostic tests for disease classification is often measured by accuracy (e.g. sensitivity or specificity); however, costs of the diagnostic test are a concern as well. Combinations of multiple diagnostic tests may improve accuracy, but incur additional costs. Here we consider serial testing approaches that maintain accuracy while controlling costs of the diagnostic tests. We present a serial risk score classification approach. The basic idea is to sequentially test with additional diagnostic tests just until persons are classified. In this way, it is not necessary to test all persons with all tests. The methods are studied in simulations and compared with logistic regression. We applied the methods to data from HIV cohort studies to identify HIV infected individuals who are recently infected (< 1 year) by testing with assays for multiple biomarkers. We find that the serial risk score classification approach can maintain accuracy while achieving a reduction in cost compared to testing all individuals with all assays.

Keywords

Biomarkers; classification; diagnostic tests; HIV

1. Introduction

Diagnostic testing for classifying a person into one of two states (e.g. disease or no disease) can be costly, especially if multiple diagnostic tests are required to make accurate diagnoses. While increasing the number of biomarkers may increase accuracy, that comes at the price of increased costs of the assays. The trade-off between accuracy and costs in diagnostic

* dath@ucla.edu

** olaeyen1@jhmi.edu

*** rbrookmeyer@ucla.edu

Supplementary Materials

A Web Appendix, with R-code to implement the serial risk score classification method is available with this paper at the Biometrics website on the Wiley Online Library.

Conflict of interest: none declared.

classification is of concern particularly in countries with limited resources where adequate screening is costly (Parpia et al., 2010).

The cost of diagnosis can be reduced by using a *serial testing algorithm* that classifies some subjects with fewer tests. A serial testing algorithm is a classification method where tests are performed and evaluated in a specified order, one at a time. The costs associated with diagnosis can be reduced if classification can occur before all tests are performed.

To make this concrete, consider the following simple example. Suppose there are 2 diagnostic tests for a disease, B_1 and B_2 (e.g. assays for two different biomarkers). These tests can be combined into a classification rule such as ‘*either positive*’ or ‘*both positive*’ where the former refers to a positive diagnosis of disease if either test is positive and the latter referring to positive diagnosis if both tests are positive. If we use an ‘*either positive*’ rule, and B_1 is positive, there is no need to test with B_2 . Similarly, if we use a ‘*both positive*’ strategy and B_1 is negative, there again is no need to test with B_2 .

Serial testing algorithms have been used to reduce the cost of testing to identify HIV infected persons. For example, one study in the Ivory Coast showed that a serial testing algorithm to identify HIV infected persons could reduce assay costs while maintaining accuracy (Nkengasong et al., 1999).

Our work was motivated by the problem of classifying HIV infected persons as to whether their infection occurred recently or not (e.g., within the previous year or not). The problem of accurately identifying recently occurring infections is important for several reasons. It has been shown that individuals who are recently infected have higher rates of HIV transmission (Wawer et al., 2005). It is important to identify sexual partners of those recently infected so they can seek testing and care. Treatment decisions may also depend on the recency of infections (Sáez-Cirión et al., 2013). Additionally, knowledge of recency of infection is useful in studies of risk factors for HIV infection and may help identify subgroups of the population where HIV incidence is growing rapidly.

There are a number of different laboratory assays that are informative about the recency of infection, including the BED capture immunoassay (BED-CEIA), BioRad-Avidity assay (avidity), CD4 cell count and HIV viral load assays (Laeyendecker et al., 2013). As these assays measure different aspects of the immune and clinical status of the individual, it is useful to consider them in combination; however, some of the assays such as viral load and the CD4 cell count are expensive due to the cost of the laboratory assays and logistical issues for performing certain assays in the field.

Statistical methods for disease classification using multiple diagnostic tests have been extensively developed (Pepe, 2003). A fundamental approach is based on estimating risk scores, $S(\mathbf{X}) = P(Y = +1|\mathbf{X})$ where $\mathbf{X}$ is a vector of predictors (e.g., diagnostic tests or assays for biomarkers) and $Y = +1$ if a person has the disease and $Y = -1$ otherwise (Pepe, 2003). Logistic regression is frequently used for risk score modeling. In this approach, a person is classified as positive with disease if $S(\mathbf{x}) > c$ and otherwise negative without disease, where c are discrimination thresholds chosen in an optimal way to minimize loss functions.

An advantage of risk score modeling is that it uses all the biomarker (or diagnostic test) information, and can account for biomarker data that is either continuous or categorical; however, it does require that all the diagnostic tests be performed on all persons which can be expensive. On the other hand, serial testing algorithms can be cost effective because not all persons are necessarily evaluated with all diagnostic tests.

In this paper, we develop a hybrid approach called serial risk score classification. The approach merges the strengths of risk score modeling with serial testing algorithms. The basic idea is to use a serial testing strategy, which involve a series of stages in which persons not yet classified are subject to additional diagnostic testing. Risk score modeling is used for classification using the diagnostic tests performed up to that stage. Persons that do not have sufficient evidence for classification proceed to the next stage. The biomarker data can be either continuous or categorical. The cost savings results from the fact that not all diagnostic tests are performed on all persons.

In Section 2.2, we develop the methodology for serial risk score classification. In Section 2.3, we discuss a bootstrap method to adjust performance measures for bias from overfitting. In Section 3, the method is evaluated by simulation and compared to logistic regression. In Section 4, the method is applied to the HIV problem of determining if HIV infected persons were infected within the preceding year or not. The results are discussed in Section 5.

2. Approach

2.1 Overview of Serial Risk Score Classification

In this section, we outline a strategy that combines aspects of a risk score modeling approach and a serial testing approach to use all available information while minimizing costs. We call the method serial risk score classification. Our approach introduces several stages of classification where a person may either be classified as positive for disease, negative for disease, or neutral (i.e. undecided) when there is not enough evidence for either a positive or negative classification. Neutral classification was originally suggested to keep misclassification rates under a specified level (Rao, 1947) and has also been described in engineering literature as a *reject*, an event when a recognition system withholds its recognition decision (Chow, 1970). Most recently, Jeske et al. (2007) formalized a framework for a bayesian neutral zone classifier using posterior class probabilities to define a neutral zone.

A general description of our strategy is as follows. Suppose we have a total of M different diagnostic tests that can be used for disease classification. At the first stage of diagnostic testing, each person is evaluated with m_1 diagnostic tests. The m_1 diagnostic tests are a subset of the M diagnostic tests. Based on those m_1 test results, we classify each person into one of three categories: positive for disease ($\hat{Y} = +1$), negative for disease ($\hat{Y} = -1$), or neutral (undecided; $\hat{Y} = 0$). Persons classified as neutral proceed to the second stage. At the second stage, an additional diagnostic test is performed on persons who were classified neutral at the first stage. We then reclassify those neutral persons from the first stage as either positive, negative, or still neutral using all $m_1 + 1$ diagnostic test results available at the second stage (i.e, the m_1 tests from the first stage together with the additional test

performed at the second stage). We proceed similarly through each stage. In general, at the i^{th} stage, we reclassify those persons classified as neutral at the $(i - 1)$ stage using all the m_i diagnostic tests that are available (i.e., m_{i-1} tests from the preceding $(i - 1)$ stages plus the additional test performed at the i^{th} stage). The classification at the i^{th} stage is based on risk scores, determined, for example, from logistic regression. Risk score thresholds are determined at each stage, such that if the score is below (above) a lower (upper) threshold, the person is classified as negative (positive) for disease, and otherwise the person is classified as neutral. We allow a maximum of K stages where $K \leq M$. All persons are classified at or before the K^{th} stage because we do not allow for neutral classification at the last K^{th} stage. The serial risk score classification approach can be less costly than testing all persons with all M diagnostic tests because some persons are classified as positive or negative before undergoing all M diagnostic tests.

Figure 1 is an illustration of our strategy with 4 stages and 4 diagnostic tests which are labeled z_1, z_2, z_3 and z_4 . At the first stage, a single diagnostic test z_1 is performed ($m_1 = 1$). Risk scores $S(z_1)$ are estimated from a risk score model (e.g., logistic regression) and persons are classified as negative if their scores are below a lower threshold c_{1l} , as positive if their risk scores are above an upper threshold c_{1u} , or as neutral if their scores lie between the lower and upper thresholds. At the second stage, an additional diagnostic test z_2 is performed only on those persons who were classified as neutral at the first stage. These persons are assigned updated risk scores $S(z_1, z_2)$ from a risk score model using both diagnostic tests z_1 and z_2 and they are then reclassified (using these updated risk scores) based on updated thresholds c_{2l} and c_{2u} . Similarly, at the third stage, an additional diagnostic test z_3 is performed on those persons still neutral at the second stage. At the fourth and final stage, the last diagnostic test z_4 is performed on those persons neutral at the third stage, and a classification (positive or negative) is finally made for these persons.

2.2 Methods and Estimation

In this section, we detail and formalize the approach outlined in 2.1. We assume we have available a training dataset that includes Y and all M diagnostic test results for each person in the dataset. In this section, we describe the methods for determining the risk score thresholds and the sequential order to add particular diagnostic tests at each successive stage.

Suppose we have a training data set of N subjects, of whom n_+ have disease ($Y = +1$) and n_- do not have disease ($Y = -1$). The total number of diagnostic tests is M which are named $z_1, \dots, z_M$. Each of the M diagnostic tests are performed on each person. Let $\mathbf{X}_i$ be a vector of m_i diagnostic tests available at the i^{th} stage for risk score modeling. The vector $\mathbf{X}_i$ includes all the diagnostic tests in $\mathbf{X}_{i-1}$ plus one additional diagnostic test that is performed at stage i , and thus $m_i = m_{i-1} + 1$.

To be clear, the tests included in earlier stages are available for analysis at later stages, that is, the diagnostic tests in $\mathbf{X}_i$ are a subset of the diagnostic tests in $\mathbf{X}_j$ for $i < j$. For example, in figure 1, $\mathbf{X}_1 = (z_1)$, $\mathbf{X}_2 = (z_1, z_2)$, $\mathbf{X}_3 = (z_1, z_2, z_3)$, and $\mathbf{X}_4 = (z_1, z_2, z_3, z_4)$. The $\mathbf{X}_i$ vectors are not fixed or pre-determined. Rather, we use our methods described below to determine the

sequential order of tests to perform at each stage that balances accuracy and cost considerations.

At the i th stage, let $\hat{S}_i(\mathbf{X}_i)$ be a model for the risk score, $P(Y = +1|\mathbf{X}_i)$. For example, when using a logistic regression model, we have $\hat{S}_i(\mathbf{X}_i) = 1 / \{1 + \exp(-\mathbf{X}_i^T \boldsymbol{\beta}_i)\}$ where $\boldsymbol{\beta}_i$ are estimated regression coefficients.

For risk score models $\hat{S}_i(\mathbf{X}_i)$, $1 \leq i \leq k-1$, we divide the risk score space into “positive”, “negative”, and “neutral” regions using the lower and upper thresholds c_{il} and c_{iu} , where c_{il} , c_{iu} , and both lie between 0 and 1, so that

$$\hat{Y}_i = \begin{cases} -1, & \text{if } \hat{S}_i(\mathbf{X}_i) < c_{il} \\ 0 & \text{if } c_{il} \leq \hat{S}_i(\mathbf{X}_i) < c_{iu} \\ +1, & \text{if } \hat{S}_i(\mathbf{X}_i) \geq c_{iu} \end{cases}$$

where 0 indicates neutral or no classification in stage i . At the last stage, we do not assign a neutral zone and use a single threshold c_k to divide the risk score space, i.e.

$$\hat{Y}_k = \begin{cases} -1, & \text{if } \hat{S}_k(\mathbf{X}_k) < c_k \\ +1, & \text{if } \hat{S}_k(\mathbf{X}_k) \geq c_k \end{cases}.$$

For a person assigned to the neutral zone in the first $j-1$ stages, the final classification of that individual is $\hat{Y}_j$.

We now present a formal definition for a *serial risk score classification algorithm*. It is a classification algorithm where the prediction $\hat{Y}$ ($\hat{Y} = 1$ if classified positive for disease and -1 otherwise) is based on a logical rule r (or algorithm) of the following form (where $\wedge$ is a logical AND operator and $\vee$ is a logical OR operator):

$$\begin{aligned} r = & \{ \hat{S}_1(\mathbf{X}_1) \geq c_{1u} \} \\ & \vee [\{ \hat{S}_2(\mathbf{X}_2) \geq c_{2u} \} \wedge \{ \hat{S}_1(\mathbf{X}_1) \geq c_{1l} \}] \\ & \vee [\{ \hat{S}_3(\mathbf{X}_3) \geq c_{3u} \} \wedge \{ \hat{S}_1(\mathbf{X}_1) \geq c_{1l} \} \wedge \{ \hat{S}_2(\mathbf{X}_2) \geq c_{2l} \}] \\ & \vdots \\ & \vee [\{ \hat{S}_k(\mathbf{X}_k) \geq c_k \} \wedge \{ \hat{S}_1(\mathbf{X}_1) \geq c_{1l} \} \wedge \cdots \wedge \{ \hat{S}_{k-1}(\mathbf{X}_{k-1}) \geq c_{(k-1)l} \}] \end{aligned} \quad (1)$$

where

$$\hat{Y} = \begin{cases} +1, & \text{if } r = \text{true} \\ -1, & \text{if } r = \text{false}. \end{cases}$$

In equation 1, the first line refers to persons classified as positive at the end of stage 1; the second line refers to persons neutral at the end of stage 1 but subsequently classified as positive at the end of stage 2; and so on. The set of algorithms given by (1) define a space of

algorithms $\mathbf{R}$ which is generated by the various sequential orders of adding diagnostic tests and choices for the risk score thresholds.

To choose algorithms from this expansive space, we develop optimality criteria that balance accuracy and cost. We begin by defining an accuracy loss function; here we use $L(r, p) = pFP(r) + FN(r)$ where p is the relative contribution to the loss of a false positive (FP) compared to a false negative (FN) as discussed in Etzioni et al. (2003). Our methods could also be applied to other loss functions. We define

$$FP(r) = \sum_{\{i|Y_i=-1\}} I(\hat{Y}_i=+1)$$

and

$$FN(r) = \sum_{\{i|Y_i=1\}} I(\hat{Y}_i=-1).$$

For a specified value of p , we could simply find the minimal value, $L_{min}(p)$, of the loss function that is achievable with any algorithms in the space of algorithms $\mathbf{R}$. However we also want to balance accuracy with cost considerations. Associated with each stage is additional costs from diagnostic testing of those persons still neutral at the preceding stage.

The total cost of an algorithm is defined by the number of individuals who are tested with a diagnostic test and the cost of the diagnostic test itself. Let b_i be the cost for a diagnostic test z_i . Let n_{zi} be the number of individuals tested with z_i in an algorithm. For example, suppose an algorithm uses diagnostic test z_i in the first stage, then in this case, $n_{zi} = N$. We define the total cost of an algorithm by:

$$\text{Total Cost} = \sum_{i=1}^M b_i n_{zi}.$$

The algorithm that corresponds to $L_{min}(p)$ may not have the lowest total cost. To help achieve balance between both cost and accuracy considerations, we adopt the following strategy. We find the collection of algorithms that have accuracy loss functions “close” to the minimal value. Among these algorithms that are “nearly” equivalent with regard to accuracy, we find the one with minimal total cost. To quantify the meaning of “close” with respect to accuracy, we introduce a tolerance t on the accuracy loss function. Consider the set of algorithms $\mathbf{R}(t, p)$ that give values of the loss function within a percentage $t \times 100\%$ of $L_{min}(p)$ so that

$$\mathbf{R}(t, p) = \{r: L_{min}(p) \leq L(r, p) < (1+t) \cdot L_{min}(p)\}.$$

That is, $\mathbf{R}(t, p)$ is a set of algorithms that are nearly equivalent with respect to accuracy as calibrated by the tolerance t . We define an optimal algorithm r_p as an algorithm in the set

$\mathbf{R}(t, p)$ with minimal total cost. It is possible that multiple algorithms satisfy this optimality criteria. For a chosen set of $p \in \mathbf{P}$, these optimal algorithms form a subset of $\mathbf{R}$,

$\mathbf{R}_p(t) = \{r_p : \exists p \in \mathbf{P}, r_p = \arg \min_{cost} \mathbf{R}(t, p)\}$. The set $\mathbf{R}_p(t)$ is used to construct a ROC curve. As t decreases more weight is given to accuracy compared to cost for determining the optimal algorithm.

The space of algorithms $\mathbf{R}$ includes an algorithm of performing all M diagnostic tests at the first stage and not allowing any person to be classified as neutral, that is, $m_1 = M$ and $c_{1l} = c_{1u}$. Such an algorithm is equivalent to a risk score model using all diagnostic tests. Thus, if the tolerance t is set to 0, then the serial risk score classification algorithm is guaranteed to achieve at least the same accuracy as standard logistic regression. The reason is that logistic regression can be viewed as a special case of the serial risk score approach when $t = 0$. In fact, the serial risk score classification algorithm can be both more accurate and less expensive than standard logistic regression. It should also be noted that even if t is greater than 0, it is still possible that the serial risk score approach will yield greater accuracy than logistic regression.

2.3 Bootstrapping misclassification rates and cost

One of the objectives when generating a classification algorithm is characterizing the prediction ability of the algorithm with a future unobserved subject. We use two measures to characterize the accuracy of an algorithm r_p : sensitivity, $Se = P(\hat{Y} = 1 | Y = 1)$ and specificity, $Sp = P(\hat{Y} = -1 | Y = -1)$. When the same data is used to both construct and evaluate the algorithm, these measures of accuracy tend to be over optimistic (Efron, 1983). The reason is that the classifiers are chosen specifically to minimize the loss function with respect to the training data. In this section, we describe a bootstrapping method to estimate this optimism, or bias, and how we adjust for it.

In the following discussion, as it is understood that Se and Sp will depend on a particular p and optimal algorithm r_p , we omit this dependency for brevity. Define the *apparent* sensitivity as:

$$\overline{Se} = 1 - \frac{FN}{n_+}.$$

The *apparent* specificity is defined as:

$$\overline{Sp} = 1 - \frac{FP}{n_-}.$$

We define the biases in the apparent sensitivity and specificity respectively by $e_{se} = \overline{Se} - Se$ with expectation ε_{se} , and $e_{sp} = \overline{Sp} - Sp$ with expectation ε_{sp} . If ε_{se} was known, then $\widehat{Se} = \overline{Se} - \varepsilon_{se}$ would be an unbiased estimate for Se . Similarly if ε_{se} was known, $\widehat{Sp} = \overline{Sp} - \varepsilon_{sp}$ is an unbiased estimate for Sp as well.

Our strategy is to estimate ε_{se} and ε_{sp} using a bootstrap method. Once an estimate is found, we can adjust our apparent estimates by the errors to reduce the bias incurred from overfitting. The diagnostic costs can be adjusted by the bootstrap error in the same fashion. Let $\overline{Cost}$ be the apparent total cost of an algorithm found from (2.2). The error of the total cost is $e_{cost} = Cost - \overline{Cost}$ where $Cost$ is the true cost for an optimal algorithm with some p . If we let ε_{cost} be the expectation of this error, then the adjusted cost $\widehat{Cost} = \overline{Cost} + \varepsilon_{cost}$.

For a training data set $T = \{Y, \mathbf{X}\}$ where $\mathbf{X}$ is a $N \times M$ matrix of diagnostic tests and a chosen p , the bootstrap procedure is described below.

1. Draw a bootstrap sample T^* from T with replacement.
2. Use the serial risk score classification method to find the optimal algorithm r_p^* from T^* .
3. Use the optimal algorithm r_p^* applied to the bootstrap sample, T^* , to calculate the apparent sensitivity, specificity and cost, which we denote respectively by, $\overline{Se}(r_p^*|T^*), \overline{Sp}(r_p^*|T^*), \overline{Cost}(r_p^*|T^*)$.
4. Use the optimal algorithm r_p^* applied to the original training data set, T , to calculate the sensitivity, specificity, and cost, which we denote respectively by, $\overline{Se}(r_p^*|T), \overline{Sp}(r_p^*|T), \overline{Cost}(r_p^*|T)$.
5. Calculate
$$e_{se}^* = \overline{Se}(r_p^*|T^*) - \overline{Se}(r_p^*|T), e_{sp}^* = \overline{Sp}(r_p^*|T^*) - \overline{Sp}(r_p^*|T), e_{cost}^* = \overline{Cost}(r_p^*|T^*) - \overline{Cost}(r_p^*|T)$$
.
6. Repeat 1 – 5 for B times. Estimate
$$\hat{\varepsilon}_{se} = 1/B \sum_{b=1}^B e_{se_b}^*, \hat{\varepsilon}_{sp} = 1/B \sum_{b=1}^B e_{sp_b}^*, \hat{\varepsilon}_{cost} = 1/B \sum_{b=1}^B e_{cost_b}^*.$$

Following the bootstrap procedure, we estimate the adjusted sensitivity $\widehat{Se} = \overline{Se} - \hat{\varepsilon}_{se}$, specificity $\widehat{Sp} = \overline{Sp} - \hat{\varepsilon}_{sp}$, and cost, $\widehat{Cost} = \overline{Cost} + \varepsilon_{cost}$.

The bootstrap samples allow the construction of a $(1 - 2\alpha)\%$ confidence interval for Se and Sp . For each p , we find the interval $(\widehat{Se}_\alpha, \widehat{Se}_{1-\alpha})$ where $\widehat{Se}_\alpha$ and $\widehat{Se}_{1-\alpha}$ are the α and $1 - \alpha$ percentiles of the empirical distribution of the bootstrapped sensitivity (Efron and Tibshirani, 1993). The percentile-based bootstrap confidence interval is constructed similarly for Sp .

3. Simulation Results

We evaluated the performance of the serial risk score classification method and bootstrap procedures through a series of simulations. Each simulation was comprised of 1000 simulated datasets. In each dataset, we generated data from 500 persons with two diagnostic tests z_1 and z_2 from a truncated bivariate normal distribution. We considered the situations when the correlation coefficient for the two was $\rho = 0.5$ and $\rho = 0.0$. The cost of z_2 was 10 times that of z_1 . We modeled the outcome Y using a logistic regression model. We

considered 2 cases for the coefficients in the logistic regression model: (I) the coefficients β_1 and β_2 are chosen so that a 1 standard deviation increase in z_1 doubles the risk of $Y = 1$ and 1 standard deviation increase in z_2 triples the risk; and (II) where 1 standard deviation increase in z_2 triples the risk and $\beta_1 = 0$. Bootstrap adjusted estimates for the serial risk score algorithm were computed using 500 bootstrap samples. We also evaluated the performance characteristics of the serial risk score classification method as determined by our bootstrapping procedure by generating an independent data set of 500 values for each simulation and classified each person using the serial risk score classification algorithm. The results for correlated and uncorrelated diagnostic tests are shown in tables 1 and 2 respectively. We performed the simulations for two different tolerance values ($t = 0.05$ and $t = 0.10$). The tolerance t calibrates the tradeoff between accuracy and cost. Increasing t puts more priority on decreasing costs at the expense of accuracy. In each simulation, we compared the serial risk score classification method to logistic regression with backward elimination (non-significant terms at the $\alpha = 0.05$ level were eliminated).

Table 1 gives the results with correlated diagnostic tests ($\rho = 0.5$). We find in case (I) that the serial risk score method has slightly less accuracy performance as measured by the AUC (area under curve) compared to logistic regression, but the cost savings are substantial. This makes intuitive sense because some information from one diagnostic test about disease risk is included in the other, and thus the less expensive diagnostic test can take the place of the more expensive one to some extent. We find in case (II), smaller cost savings compared to case (I) which is explained because $\beta_1 = 0$. Some significant cost savings is still achieved in case (II) because of the correlation between the two diagnostic tests and as such the less expensive diagnostic test can be used as a proxy for the significantly more expensive test.

Table 2 gives the results with uncorrelated diagnostic tests ($\rho = 0.0$). In case (I), we find a markedly reduced cost savings with uncorrelated (Table 2) compared to correlated diagnostic tests (Table 1) with a minor reduction in accuracy. Because the diagnostic tests are no longer correlated, both tests are more often necessary for correct classification. In case (II), the cost savings are further reduced, and only very minor cost savings are achieved. It should be noted that the cost savings we report in this simulation are dependent on the disease prevalence and different results will be obtained with different prevalence rates.

The bootstrap adjusted SRS estimates of the performance measures appear in good agreement with the estimates found when the algorithms are applied to the independent data sets. We believe this demonstrates the bias reduction from the bootstrap procedure is effective. We also performed the simulation reported in Tables 1 and 2 with only 100 bootstraps and obtained very similar results as that reported here with 500 bootstraps. When the serial risk score algorithm is applied to data where the diagnostic tests have no discrimination ability, we find an AUC of .50 as expected.

4. Application to HIV study

We used our methods to develop cost effective algorithms for classifying HIV infected persons as to whether they were infected < 1 year or ≥ 1 year using a dataset that has been

previously described (Brookmeyer et al., 2013; Laeyendecker et al., 2013). The dataset consisted of 1782 samples from three cohort studies: a vaccine preparedness study (HIVNET001, men and women with different risk factors for HIV infection, (Celum et al., 2001)), a cohort study of intravenous drug users (the AIDS Linked to Intravenous Experience (ALIVE) cohort, (Vlahov et al., 1991)), and a cohort study of men who have sex with men (the Multicenter AIDS Cohort Study (MACS), (Kaslow et al., 1987)). The data set included estimated dates of infection for each individual based on midpoints of the seroconversion interval and those dates are considered correct in this analysis. Each sample was tested with four different assays: BED-CEIA, avidity, CD4 and viral load. The laboratory costs for performing the BED-CEIA, avidity, CD4, and viral load assays were, relative to the BED assay, 1,2,5 and 10 respectively.

Because of the computational constraints of brute force searching, we limited the risk score thresholds $\{c_{1l} \cdots c_{(k-1)l}\}$, $\{c_{1u} \cdots c_{(k-1)u}\}$, and c_k to using only the deciles of the risk score space at each stage. This approximate algorithm set $\mathbf{R}_{approx}$ was used to find the $\mathbf{R}_p(t)$, an optimal algorithm set where the tolerance $t = 0.10$. We employed the bootstrap procedure described in Section 2.3 to compute adjusted estimates for sensitivity and specificity and cost using 500 bootstrap samples. We have compared results using 100 bootstraps and 500 bootstraps and similar results were obtained.

The ROC curve presented in figure 2, shows the ROC curves for the adjusted and unadjusted error rates from algorithms generated by the serial risk score classification algorithm. Each point on the ROC curve represents a different algorithm corresponding to a particular p selected by minimization of the loss $L(p)$. For each point, the 95% lower and upper confidence limits of $\widehat{S_e}$ and $\widehat{S_p}$ are plotted to form the confidence intervals shown on the graph.

In table 3, the adjusted and unadjusted sensitivities and specificities for selected optimal algorithms from the ROC curve are presented and compared with classification from logistic regression. As expected, the adjusted bootstrap estimates have reduced the upward bias of the apparent estimates for sensitivity and specificity. For each of these algorithms, the cost savings for the values of p shown is decreased by at least 62% versus a classification method using all biomarkers (e.g. logistic regression). Furthermore, the serial risk score algorithm performs equal to or better than logistic regression in both sensitivity and specificity for most values of p .

In table 4, we present an example optimal algorithm chosen when $p = .3$. In each column, the risk score model coefficients for $\widehat{S}(X_i)$ for classification stage i are displayed. The decision rules for recent classification are also displayed under the corresponding risk score model coefficients. We also show the cumulative percentage that are classified in each stage as well as cumulative percentage of total cost in each stage. From the table, we can see that the majority of the observations are classified with only 2 tests.

5. Discussion

Serial risk score classification is a hybrid method for disease classification that combines risk score modeling with serial testing algorithms. This approach uses all information available at each stage to make classifications before proceeding to the next stage and incurring additional costs associated with additional diagnostic tests. Using this methodology, we were able to realize at least a 62% reduction in costs for identifying recently HIV infected persons (<1 year) using up to 4 diagnostic tests compared to testing all persons with all 4 diagnostic tests using standard risk score modeling based on logistic regression. The algorithms created by serial risk score classification were nearly as accurate and sometimes more accurate than logistic regression. That finding may seem surprising because logistic regression is using data on all biomarkers from each individual while the serial risk score approach is generally using only a subset of that data. The explanation for the finding is that the serial risk score approach performs binary regressions at each stage of the algorithm and as such more parameters are being estimated than with logistic regression. Furthermore, as discussed in McIntosh and Pepe (2002) simply adding a biomarker as a linear term in a model may not produce the optimal combination of biomarkers. Indeed, these authors suggest risk score estimation within sub-regions of the biomarker space as a practical approach for approximating optimal combination rules.

The methods we have described could be extended in a number of directions. It would be useful to incorporate covariates into the risk score modeling performed at each stage. For example, in our HIV application, it may be helpful to incorporate covariates such as HIV transmission group (e.g., intravenous drug use, men who have sex with men) and age. It would also be useful to account for errors in the disease classification Y . For example, in our HIV application, the calendar date of infection was based on midpoint imputation of the interval of seroconversion which would introduce error into Y . Methods have been suggested for incorporating uncertainties in the dates of infection in HIV studies, and their extensions to our methods would be useful (Brookmeyer et al., 2013; Sweeting et al., 2010).

A challenge with the risk score classification approach we have proposed is the computational intensity of the optimization. Each additional biomarker exponentially increases the computation time as well as memory requirements to store the generated algorithms. The bootstrapping adjustment procedure further complicates the method, as each bootstrap sample requires a new set of generated algorithms. Fortunately, the approach is capable of being parallelized in a computation cluster.

The implementation of serial risk score classification described in the application was limited in a number of ways. The number of risk score thresholds for consideration at each stage had been limited to deciles of the risk score space. It would be useful to examine the sensitivity of the accuracy to this choice. We might consider introducing refinements to the risk score model such as interactions or polynomial terms. We used logistic regression models to estimate the risk score, but alternative models could be used such as a probit or a bayesian probability classifier (Kim, 2013). There may also be restrictions in the ordering of the assays that may require consideration. For example, it may be logistically important to perform the CD4 assay first, because of the difficulty associated with cryo preserving viable

samples (Brookmeyer et al., 2013; Laeyendecker et al., 2013). In addition, there are considerations in choosing the tolerance of the loss function. Using a larger tolerance may lower costs further but may diminish accuracy. A smaller tolerance may increase costs, but increase accuracy. Approaches for choosing the tolerance could be studied. Finally, McIntosh and Pepe (2002), have discussed modeling considerations for combining diagnostic tests which could be incorporated into our risk score estimation within each stage.

An alternative approach to creating serial testing algorithms may be through the use of logic regression (Ruczinski et al., 2003). This method is used to define a logical classification rules with binary predictors. These rules can be translated into a serial testing algorithm. Logic regression can be adapted to minimization of cost by looking at permutations of equivalent logic rules that assign the same classification to the same predictors, but at different costs. One limitation with logic regression when compared to serial risk score is that binary predictors will need to be determined a priori. This may require discretization of continuous measures from diagnostic tests in advance.

The method we have presented describes an approach to develop algorithms for diagnostic testing that mediates the trade-off between cost and accuracy. The method has shown to be accurate but with considerably less cost than logistic regression in many situations of interest. We believe the serial risk score classifier can be a useful approach for screening populations in resource-limited settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by National Institutes of Health grant R01-AI095068 and in part by the Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health.

The authors thank the following investigators for providing or generating the data analyzed in paper including data from the MACS, ALIVE and HIVNET001 studies: Dr. Susan H. Eshleman, Caroline E. Mullis, Dr. Matthew M. Cousins, and Drs. Thomas C. Quinn, Deborah Donnell, Connie Celum, Susan P. Buchbinder, George R. Seage III, Lisa P. Jacobson, Joseph B. Margolick, Joelle Brown, Gregory D. Kirk, Shruti H. Mehta, and Jeanne C. Keruly.

The HIV Network for Prevention Trials (HIVNET) 001 Study was funded by the HIVNET and sponsored by the NIAID; the AIDS Link to Intravenous Experience (ALIVE) Study was funded by the NIDA; and the Multicenter AIDS Cohort Study (MACS) was funded by the NIAID, with additional supplemental funding from the National Cancer Institute and National Heart, Lung, and Blood Institute.

References

- Brookmeyer R, Konikoff J, Laeyendecker O, Eshleman SH. Estimation of HIV incidence using multiple biomarkers. *American Journal of Epidemiology*. 2013; 177:264–272. [PubMed: 23302151]
- Brookmeyer R, Laeyendecker O, Donnell D, Eshleman SH. Cross-sectional HIV incidence estimation in HIV prevention research. *JAIDS Journal of Acquired Immune Deficiency Syndromes*. 2013; 63:S233–S239. [PubMed: 23764641]
- Celum CL, Buchbinder SP, Donnell D, Douglas JM, Mayer K, Koblin B, Marmor M, Bozeman S, Grant RM, Flores J, et al. Early human immunodeficiency virus (HIV) infection in the HIV network for prevention trials vaccine preparedness cohort: risk behaviors, symptoms, and early plasma and genital tract virus load. *Journal of Infectious Diseases*. 2001; 183:23–35. [PubMed: 11106536]

- Chow C. On optimum recognition error and reject tradeoff. *Information Theory, IEEE Transactions on*. 1970; 16:41–46.
- Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*. 1983; 78:316–331.
- Efron, B.; Tibshirani, R. *An introduction to the bootstrap*. Vol. 57. CRC press; 1993.
- Etzioni R, Kooperberg C, Pepe M, Smith R, Gann P. Combining biomarkers to detect disease with application to prostate cancer. *Biostatistics*. 2003; 4:523–538. [PubMed: 14557109]
- Jeske DR, Liu Z, Bent E, Borneman J. Classification rules that include neutral zones and their application to microbial community profiling. *Communications in Statistics Theory and Methods*. 2007; 36:1965–1980.
- Kaslow RA, Ostrow DG, Detels R, Phair JP, Polk BF, Rinaldo CR, et al. The multicenter AIDS cohort study: rationale, organization, and selected characteristics of the participants. *American Journal of Epidemiology*. 1987; 126:310–318. [PubMed: 3300281]
- Kim, H. Sequential neutral zone classifier; Unpublished paper presented at Quality and Productivity Research Conference 2013; 2013.
- Laeyendecker O, Brookmeyer R, Cousins MM, Mullis CE, Konikoff J, Donnell D, Celum C, Buchbinder SP, Seage GR, Kirk GD, et al. HIV incidence determination in the United States: a multiassay approach. *Journal of Infectious Diseases*. 2013; 207:232–239. [PubMed: 23129760]
- McIntosh MW, Pepe MS. Combining several screening tests: Optimality of the risk score. *Biometrics*. 2002; 58:657–664. [PubMed: 12230001]
- Nkengasong JN, Maurice C, Koblavi S, Kalou M, Yavo D, Maran M, Bile C, Nguessan K, Kouadio J, Bony S, Wiktor SZ, Greenberg AE. Evaluation of HIV serial and parallel serologic testing algorithms in Abidjan, Cote d'Ivoire. *AIDS*. 1999; 13:109–117. [PubMed: 10207552]
- Parpia ZA, Elghanian R, Nabatiyan A, Hardie DR, Kelso DM. p24 antigen rapid test for diagnosis of acute pediatric HIV infection. *JAIDS Journal of Acquired Immune Deficiency Syndromes*. 2010; 55:413. [PubMed: 20811289]
- Pepe, MS. *The statistical evaluation of medical tests for classification and prediction*. Oxford, United Kingdom: Oxford University Press; 2003.
- Rao RC. The problem of classification and distance between two populations. *Nature*. 1947; 159:30–31. [PubMed: 20279071]
- Ruczinski I, Kooperberg C, LeBlanc M. Logic regression. *Journal of Computational and Graphical Statistics*. 2003; 12:475–511.
- Sáez-Cirión A, Bacchus C, Hocqueloux L, Avettand-Fenoel V, Girault I, Lecroux C, Potard V, Versmisse P, Melard A, Prazuck T, et al. Post-treatment HIV-1 controllers with a long-term virological remission after the interruption of early initiated antiretroviral therapy ANRS VISCONTI study. *PLoS pathogens*. 2013; 9:e1003211. [PubMed: 23516360]
- Sweeting MJ, De Angelis D, Parry J, Suligoi B. Estimating the distribution of the window period for recent HIV infections: a comparison of statistical methods. *Statistics in medicine*. 2010; 29:3194–3202. [PubMed: 21170913]
- Vlahov D, Anthony J, Munoz A, Margolick J, Nelson K, Celentano D, Solomon L, Polk B. The ALIVE study, a longitudinal study of HIV-1 infection in intravenous drug users: description of methods and characteristics of participants. *NIDA research monograph*. 1991; 109:75. [PubMed: 1661376]
- Wawer MJ, Gray RH, Sewankambo NK, Serwadda D, Li X, Laeyendecker O, Kiwanuka N, Kigozi G, Kiddugavu M, Lutalo T, et al. Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda. *Journal of Infectious Diseases*. 2005; 191:1403–1409. [PubMed: 15809897]

Figure 1.

Illustration of serial risk score classifier with a maximum of $K = 4$ stages. At each level of the classifier, a risk score model is estimated and thresholds are defined. The threshold values determine whether an observation continues onto further testing or is classified as positive or negative. Four diagnostic tests are shown which are labeled z_1, z_2, z_3 and z_4 . The notational shortcut $\hat{S}_i(\cdot)$ is used here in place of the $\hat{S}_i(z_1, \dots, z_i)$ for conciseness of display.

Figure 2.

ROC curves generated using serial risk score classification procedure. The bootstrap adjusted ROC curve shown as a solid black line was generated from the algorithm specified in Section 2.3. The adjusted ROC curve represents the expected performance of the algorithm. The unadjusted ROC curve is constructed from the unadjusted (apparent) error rates given by $\overline{S_p}$ and $\overline{S_e}$. The 95% confidence interval curves are constructed using a percentile bootstrap confidence interval.

Table 1

Simulation results with two correlated diagnostic tests ($\rho = 0.5$) comparing the serial risk score classification algorithm (SRS) to logistic regression. The cost of the second diagnostic test is 10 times that of the first. Shown is the mean sensitivity (sens.), specificity (spec.) and the average cost savings (ACS) of bootstrap adjusted SRS compared to logistic regression with variable selection over 1000 simulations. The performance characteristics of SRS was validated with independent datasets in column labeled SRS validation.

	Logistic Regression				SRS Estimates				SRS Validation			
	p	Sens.	Spec.		Sens.	Spec.	ACS		Sens.	Spec.		
(I) $\beta_1 = .69, \beta_2 = .37$	$t = 0.05$	0.1	0.91	0.56	0.88	0.54	35%		0.87	0.54		
		0.3	0.69	0.82	0.64	0.81	40%		0.63	0.81		
	AUC = 0.85				AUC = 0.81				AUC = 0.80			
	$t = 0.10$	0.1	0.91	0.56	0.87	0.53	51%		0.87	0.53		
		0.3	0.69	0.82	0.62	0.81	58%		0.61	0.81		
	AUC = 0.85				AUC = 0.80				AUC = 0.79			
(II) $\beta_1 = 0, \beta_2 = .37$	$t = 0.05$	0.1	0.86	0.53	0.82	0.53	9%		0.80	0.52		
		0.3	0.52	0.86	0.45	0.86	20%		0.44	0.86		
	AUC = 0.79				AUC = 0.75				AUC = 0.74			
	$t = 0.10$	0.1	0.86	0.53	0.82	0.49	24%		0.81	0.49		
		0.3	0.52	0.86	0.40	0.87	40%		0.39	0.87		
	AUC = 0.79				AUC = 0.73				AUC = 0.72			

We performed 1000 simulations. On each simulation we generated 500 values $\mathbf{Z} = [z_1, z_2]^T \sim \text{truncated } MVN(\mu, \Sigma)$, where $\mu = [2, 9]^T$ and Σ is set so that z_1 and z_2 have a standard deviation of 1 and 3 respectively and a correlation of .5. The distribution is truncated using rejection sampling so that $z_1 > 0$, and $z_2 > 0$. $Y \sim \text{Bern}(\pi)$ where $\pi = 1/(1 + e^{-(\beta_0 + \beta_1 z_1 + \beta_2 z_2)})$. The empirical proportion where $Y = 1$ was 17.6% and 14.2% for case (I) and (II) respectively. Bootstrap adjusted estimates are based on methods described in Section 2.3 with 500 bootstraps.

Table 2

Simulation results with two uncorrelated diagnostic tests ($\rho = 0.0$) comparing the serial risk score classification algorithm (SRS) to logistic regression. The cost of the second diagnostic test is 10 times that of the first. Shown is the mean sensitivity (sens.), specificity (spec.) and the average cost savings (ACS) of bootstrap adjusted SRS compared to logistic regression with variable selection over 1000 simulations. The performance characteristics of SRS was validated with independent datasets in column labeled SRS validation.

	Logistic Regression				SRS Estimates				SRS Validation			
	p	Sens.	Spec.		Sens.	Spec.	ACS		Sens.	Spec.		
(I) $\beta_1 = .69, \beta_2 = .37$	$t = 0.05$	0.1	0.88	0.54	0.85	0.52	18%		0.84	0.52		
		0.3	0.60	0.84	0.53	0.83	25%		0.53	0.83		
	AUC = 0.82				AUC = 0.77				AUC = 0.76			
	$t = 0.10$	0.1	0.88	0.54	0.85	0.48	32%		0.85	0.48		
		0.3	0.60	0.84	0.50	0.84	41%		0.49	0.84		
	AUC = 0.82				AUC = 0.76				AUC = 0.75			
(II) $\beta_1 = 0, \beta_2 = .37$	$t = 0.05$	0.1	0.86	0.53	0.81	0.54	1%		0.80	0.54		
		0.3	0.51	0.87	0.44	0.86	5%		0.43	0.86		
	AUC = 0.79				AUC = 0.75				AUC = 0.74			
	$t = 0.10$	0.1	0.86	0.53	0.81	0.51	7%		0.81	0.50		
		0.3	0.51	0.87	0.39	0.87	16%		0.39	0.87		
	AUC = 0.79				AUC = 0.73				AUC = 0.72			

The empirical proportion where $Y = 1$ was 15.7% and 14.0% for case (I) and (II) respectively. See Table 1 for further details.

Table 3

Sensitivity (sens.), specificity (spec.), and average cost savings (ACS) of serial risk score classification and logistic regression applied to dataset described in Section 4. ACS refers to the average cost decrease v.s. the cost of all diagnostic tests.

<i>p</i>	Logistic Regression		Unadjusted SRS		Bootstrap Adjusted SRS		ACS
	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	
0.07	0.90	0.56	0.95	0.62	0.92	0.61	62%
0.10	0.88	0.61	0.90	0.71	0.88	0.70	82%
0.30	0.65	0.89	0.78	0.88	0.76	0.87	76%
0.50	0.65	0.89	0.73	0.91	0.71	0.90	71%
0.70	0.43	0.96	0.57	0.96	0.56	0.96	81%
AUC = 0.84		AUC = 0.91		AUC = 0.89			

Logistic regression model including all diagnostic tests: avidity, BED, CD4, and viral load. Individually each diagnostic test has an AUC of 0.85, 0.73, 0.65 and 0.56 respectively.

Table 4

Estimated regression coefficients to serial risk score rule chosen for $p = 0.30$ when $t = 0.10$. The decision rules for classification are displayed under the corresponding risk score model coefficients. Cumulative % classified refers to the percentage of observations classified ($\hat{Y} = +1$ or -1) up to and including the specified stage. Cumulative % of total cost is cost expenditure up to and including the specified stage divided $\hat{Y}$ the total cost at the end of stage 4.

	Stage			
	1	2	3	4
Reg. Coefficients				
Intercept	3.27	51.61	-72.73	-21.37
Avidity Index	-0.05	-0.54	0.71	0.21
BED-CEIA		-0.28	0.06	-0.14
Log Viral Load			0.41	0.01
CD4				2.7×10^{-3}
Decision Rule				
$\hat{Y} = -1$	$\hat{S} < 0$	$\hat{S} < 0.18$	$\hat{S} < 0.27$	$\hat{S} < 0.33$
$\hat{Y} = 0$	$0 \leq \hat{S} < 0.13$	$0.18 \leq \hat{S} < 0.51$	$0.27 \leq \hat{S} < 1.00$	—
$\hat{Y} = +1$	$\hat{S} \geq 0.13$	$\hat{S} \geq 0.51$	$\hat{S} \geq 1.00$	$\hat{S} \geq 0.33$
Cumulative % Classified	20.1	92.0	96.0	100
Cumulative % of Total Cost	10.6	67.5	81.1	100

The serial risk score algorithm presented here is based on 4 diagnostic biomarkers: avidity, BED-CEIA, CD4 and viral load. Avidity is shown measured as the avidity index (in %). BED capture immunoassay (BED-CEIA) is shown measured in normalized optical density (OD-n). The viral load assay is shown measured in copies/ml. CD4 is shown measured in cells/mm³. For each decision rule in stage i , $\hat{S}_i = 1/(1 + e^{-X_i\beta_i})$ where β_i is given in the i^{th} column of the regression coefficients.