Paper 144-2008

In Search for a Golden Algorithm.

Tamara Slipchenko, DM Consulting, San Diego, CA
Candice Bowman, VA San Diego Healthcare System, San Diego, CA
Catherine Sugar, School of Public Health, UCLA, Los Angeles, CA
Allen L. Gifford, VA New England Healthcare System and Boston University, Boston, MA

ABSTRACT

This paper describes an application of data mining methods for development of an HIV- casefinding algorithm with SAS Enterprise Miner (EM) 5.2. Access to HIV care depends on accurate identification of all infected persons. The Veterans Health Administration (VHA) provides care to ~20,000 HIV-infected veterans. The current algorithm for patient identification into the Registry is based only on HIV-specific diagnostic codes. We built logistic regression (LR), decision tree (DT), and neural network (NN) models to predict a binary outcome variable HIV status. We applied these models to the VHA population to identify patients with high predicted probability of disease. In addition to the diagnostic codes we were using demographic, geographic, laboratory, pharmacy and service utilization variables. False Negative (FN) rates and Area Under the Curve (AUC) indices were used for model comparisons. Our best models outperformed the reference model (RM) both in terms of lower FN rate and higher AUC index. The lowest FN rate (0.010% vs. 0.016% for the RM) was demonstrated by the NN model, while the highest AUC index was observed for the LR model (0.995 [0.994, 0.996] vs. 0.974 [0.971, 0.977] for the RM). Non-HIV-specific variables selected by our models included age, race/ethnicity, marital status, service-connected disability, number of days hospitalized, number of primary care and social work visits, number of total and lipids lab tests, blood pressure and liver co-morbidities. Apart from those already on the registry, new algorithms have identified additional 5% new cases. Using EM, our approach can be applied to other disease registries where electronic clinical data are available.

BACKGROUND

The Veterans Health Administration (VHA) provides care to approximately 5 million veterans, of whom ~20,000 are HIV-infected. Access to HIV care depends on the ability of the organization to accurately and completely identify and register all infected persons. The Immunology Case Registry (ICR) is a large clinical database that currently includes about 20,000 HIV-infected VHA patients. In addition to positive HIV test results, the current algorithm for entry onto the ICR is based on a small number of selected HIV- and AIDS-specific ICD-9 diagnostic codes. Use of diagnostic codes can lead to error from both systematic and random sources. Completeness and accuracy of the ICR has long been questioned (Rabeneck et al., 2001)

PROBLEM AND STUDY OBJECTIVES

We sought to develop, test and evaluate an “improved” algorithm for more complete and accurate identification of HIV cases onto the VA’s registry, as compared to the current algorithm.

The objectives of the study were to:
(1) Develop and evaluate new HIV casefinding algorithms using supervised data mining methods;
(2) Compare their performance to a reference model (RM) based on current registry entry criteria;
(3) Apply the new algorithms to the VA population in order to identify new patients with high predicted probability of having HIV disease.

APPROACH

We used a supervised data mining (DM) approach to solve this problem. DM is a process of finding meaningful patterns in large data sets that explain past events (e.g., known HIV status) in such a way that these patterns can be applied to new data to predict future events (e.g., unknown HIV status) (Berry, Linoff, 1997). To train and validate predictive models, a source of “gold standard” cases with known status is needed.

DATA MINING PROJECT SUMMARY

Our data mining project involved seven steps, which can be grouped into Data Preparation and Data Analysis categories:

Data Preparation:

1. Build the analysis dataset
2. Prepare the data
3. Identify pre-classified cases
4. Create the modeling sample

Data Analysis:

5. Develop and evaluate casefinding models
6. Compare model performance to a reference model
7. Apply new algorithms to the population (score data).

Data Preparation steps are shown in Figure 1, while Data Analysis steps are summarized in Figure 2. All steps are described in further detail below.

Data Preparation:

STEP 1. BUILD THE ANALYSIS DATASET

An Analysis Dataset (AD) is usually assembled from multiple sources of data. The main characteristic of the AD is the presence of a single record for each patient in the study population. Our AD was constructed from several archived VA databases: the National Patient Care Database (NPCD), the Decision Support System (DSS), and the Pharmacy Benefits Management (PBM) package. The final AD included n=4,963,796 unique patients seen in VA outpatient care between June 1, 2004 and May 31, 2005 and consisted of 123 Demographic, Geographic, Diagnostic, Treatment, Pharmacy, Laboratory, Provider, and Utilization variables.

STEP 2. PREPARE THE DATA

This step included the following activities: cleaning and reducing the dimensionality of the data, imputing missing values, deriving new variables, and summarizing and transforming data to keep one record per patient. This step involved frequent interactions with clinical domain experts, who knew from experience which variables were likely to be important.

We used several sets of input variables for model development:

- **"Expert"** sets were selected by our clinical advisors according to *a priori* knowledge of clinical importance (different sets included 32, 54, 55, 65 variables).
- **A "Most Popular"** set of 24 variables was selected from the full set of 103 variables by any of the three logistic regression models: forward, backward, or stepwise variable selection with p=.05 variable entry and exit criteria.
- **"The Full Monty"** sets included all the variables. Based on different levels of aggregation of the diagnostic variables this led to data sets with 103, 115, or 123 predictors.

STEP 3. IDENTIFY PRE-CLASSIFIED CASES

Because only four percent of the study population had ever been tested for HIV Disease, we made the working assumption that most patients who had never been tested for HIV were negative. We used available historical HIV laboratory test results to create the target variable, HIV Status.

Based on laboratory evidence of HIV disease between October 1999 and May 2005, all patients were classified into one of the two following categories:

- **True Positive** (TP): Cases which had clear evidence of disease (n = 15,475)
- **True Negative** (TN): Cases which had either clear evidence of no disease or no available evidence of disease (n = 4,948,321)

TN cases were comprised of patients who had a recorded negative HIV test result not followed by a positive result, as well as patients who had no HIV test result available and thus were presumed to be negative. Since this classification was largely based on presumed status and not on a clear standard, we recognize that the TN designation is not technically a gold standard but is used here strictly for the sake of convenience. TP cases were comprised of patients who had either a documented positive ELISA or Western Blot test or a detectable Viral Load.

STEP 4. CREATE THE MODELING SAMPLE

Once the data had been pre-classified, we created a sample of 30,950 patients consisting of all TPs and an equal number of randomly sampled TNs. The sample was partitioned into the three subsets: Training (40%), Validatation (30%), Test (30%), which were used for model development, validation and comparison. Training data were used for preliminary model fitting; Validation data were used to assess the adequacy of the models and section of tuning parameters; and Test data were used for unbiased model performance comparisons. We intentionally oversampled TPs to create a balanced sample. Usually models, which are trained on a balanced sample, can better detect a rare class (e.g., TP) of a target variable (Predictive Modeling Using Enterprise Miner™ Software, 2003).

We developed models on a balanced sample (50% TP-50% TN); however, population priors for classes of target variable are different (99.7% TN-0.3% TP). To adjust predictions made by the models, which are fit on a balanced sample, we used priors and decision weights during analytical steps. Priors are used to adjust for the difference in balance between the sample and the population while decision weights indicate the relative importance of different outcomes (true positive, true negative, false positive, false negative). A standard choice is to use a combination of the population priors and weights that lead to the same fraction of predicted "cases" that occur in the population (a so-called conforming model). However, in our case, the extreme importance of avoiding false negatives also led us to consider models which were more aggressive in identifying patients as HIV positive.

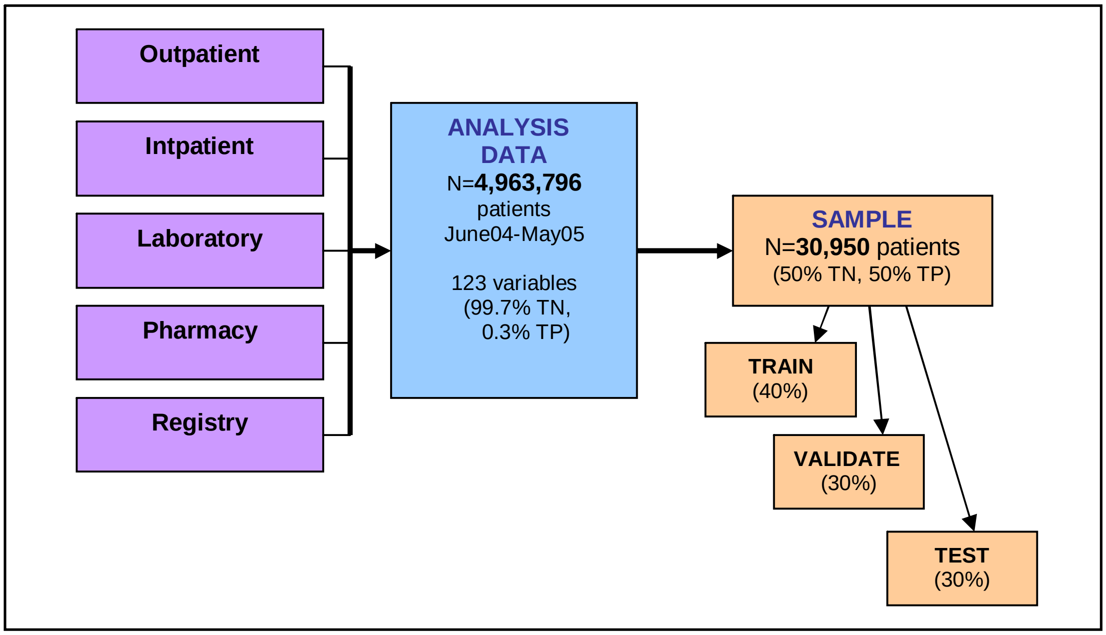


Figure 1. Summary of Data Preparation Steps: Building Analysis Dataset and Modeling Sample.

Data Analysis:

We conducted this retrospective analysis using SAS Enterprise Miner 5.2, SAS 9.3.1, and R software. The sample of pre-classified TP and TN cases was used for model development, validation and comparisons. False Negative (FN) error rates and Area Under the Curve (AUC) indices were used for model comparisons. We focused on reducing the FN rate, since we did not want our models to miss any patients with HIV, which would leave them untreated.

The process in Figure 2 shows implementation of the analytical steps taken in SAS EM 5.2. The diagram includes the following nodes:

- ***Sample Data*** node that brought the Modeling Sample from the SAS project library to the Workspace. Priors and weighs were specified in a decision processing tab of the node.
- ***Data Partitioning*** node that partitioned the sample into the three subsets.
- ***Modeling*** nodes that fit the predictive models (Regression, Decision Tree and Neural Network) to a Sample dataset. The Training dataset was used for preliminary model fitting, while the Validation dataset was used for choosing the best model based on certain selection criteria: We used a Validation Misclassification rate and Profit/Loss function for model selection. To use a Profit/Loss criterion, we had to specify decision weights for different outcomes as discussed above.

- ***Model Comparison*** node that compared models and predictions from modeling nodes using various criteria: We used two primary criteria for model comparison: (1) Area Under the Curve (AUC) indices, which were computed for the Test dataset and (2) False Negative misclassification rates, which were obtained from the Validation data.
- ***Score Data*** node that brought the Scoring data from the SAS project library to the Workspace
- ***Score*** node applied models (SAS code from modeling nodes) to the scoring data and predicted HIV status.

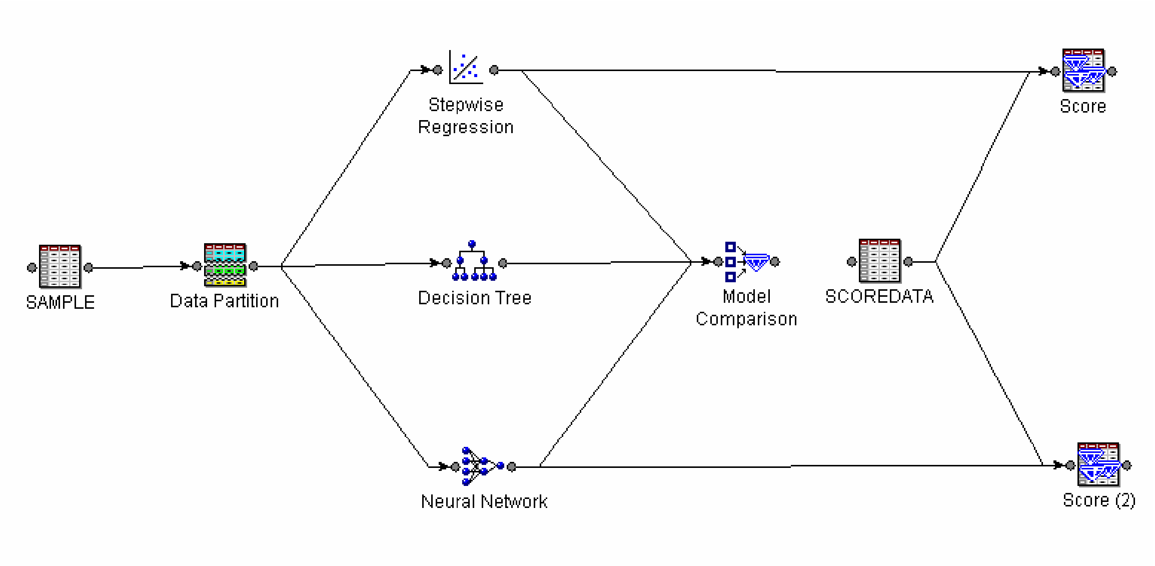


Figure 2. Simplified Process Flow Diagram (Analytical Steps 5, 6, 7).

STEP 5. DEVELOP AND EVALUATE MODELS

We built over 100 models to predict the binary target variable HIV status using the following data mining methods:

- Logistic Regression
- Decision Trees, including the newer methods of Boosting and Bagging
- Artificial Neural Networks.

For each method we used several sets of input variables (expert/popular/full Monty), different model specifications and tuning parameters, which are listed in Appendix. All models were cross-validated using Training and Validation datasets. The "best" model with the lowest False Negative (FN) error rate, and the highest Area Under the Curve, were selected for each method.

Logistic Regression (LR) predicts the probability that a binary target variable will have the event of interest as a function of one or more independent inputs. We used our expert selected, most popular and full sets of variables to build a full model plus backward, forward and mixed stepwise LR models for comparison. The best model was the mixed stepwise LR model which selected 19 input variables.

Decision Tree (DT) models are recursive procedures that iteratively partition data into segments. Partitioning is done according to splitting rules that maximize a homogeneity or purity of the target variable in each segment. This is done until further partitioning does not significantly improve classification. A DT produces a set of rules that can be applied to new data to generate predictions. To reduce the classification mistakes made by a single DT, we also implemented Boosting and Bagging techniques, which combine results of more then one tree analysis. Boosting and Bagging are examples of Majority Vote classifiers (Freund, 1996; Brieman, 1996). These recently popularized techniques work by training a large number of simple classifiers and then letting the classifiers vote on the class membership of each patient. In statistical terms, the averaging of many classifiers reduces the variance of final predictions. Our best DT model was a boosted tree model with 500 tree iterations. It identified 24 variables as having high importance for classification.

Artificial **Neural Networks** (NN) are a class of flexible non-linear models, which allow all possible interactions between input variables. They are based on computer applications that model interactions between neurons in the human brain. The most commonly used Multilayer Perceptron (MLP) architecture has three layers: input, hidden, and output. Neurons on a given layer do not link to each other, but are connected to subsequent layers by activation functions. The hidden layer makes the network more powerful by enabling it to recognize more patterns. The number of neurons in the hidden layer is determined by trial-and-error during cross-validation (Giudici, 2003). Training an NN includes a process of setting the best weights for each of the inputs. Our best model was an MLP network with 54 input variables, three neurons in the hidden layer, and an output layer with one neuron.

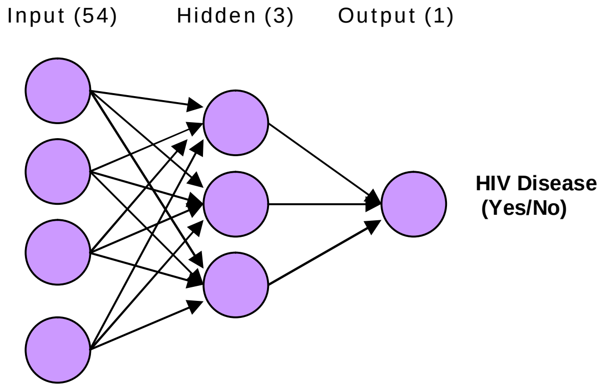


Figure 3. Artificial Neural Network architecture with 3 layers, consisting of 54 input variables, 3 hidden neurons and 1 output neuron, which are connected by network weights.

The **Reference Model (RM)** approximates the current algorithm for patient entry onto the Registry. RM is an LR model with one input variable, which is the indicator for HIV-specific ICD-9 codes (V08, 042, 043, 044x, 079.53, 795.71, 0795.8). The actual algorithm also used positive HIV laboratory test results. Since lab test results were used as a source for the initial patient classification, we limited the RM to just the diagnostic codes.

STEP 6. COMPARE MODEL PERFORMANCE

During previous steps, we selected the best models with the lowest FN error rate for each method. This section compares our three best models to the reference model:

- Stepwise LR with 19 input variables
- Boosted DT model with 24 variable
- NN with 54 input variables and 3 hidden neurons
- Reference model.

We used two criteria for the model comparisons (see Table 1):

(1) FN misclassification rate, which was computed on the Validation set
(2) AUC index, which was computed on the Test set.

First, we compared models in terms of FN misclassification rates, which were obtained from the confusion matrix. The confusion matrix compares Actual and Predicted values of the binary target variable and classifies them into four possible outcomes:

- TP predicted as negative, classified as false negative FN
- TN predicted as positive, classified as false positive FP
- TN predicted as negative, classified as true negative TN
- TP predicted as positive, classified as true positive TP.

Other useful measures derived from the confusion matrix were sensitivity (TP / [TP + FN]) and specificity (TN / [FP + TN]).

Decision Threshold. Classification of patients into positive and negative categories depends on the selection of a particular decision threshold. When a model is fit for a binary target variable (i.e., HIV Status), for each patient a predicted probability of event (i.e., TP), a number between 0 and 1, is calculated. However, the final goal is to classify each patient as either positive (TP = event) or negative (TN = non-event). To make this decision, a cut off value, or decision threshold, is selected. For example, if a predicted probability of having HIV is equal to or greater than the decision threshold, then the model classifies the patient as TP.

For model comparisons, we used classification results obtained from the *conforming* models, which used the probability of HIV in the population as the decision threshold, which in our case was 0.003. To achieve conforming models and obtain final classification results, we used a combination of priors and weights. Since the overall chance of someone having HIV in our population is 0.3%, the patients predicted by the conforming models were as likely to have HIV as someone taken at random from the VA population.

Secondly, we computed AUC indices based on Receiver Operating Characteristics (ROC) curves, which measure the predictive accuracy of the model. ROC curves are obtained by plotting (1-specificity) on the horizontal axis and the sensitivity on a vertical axis, for a range of cut-off values. Each point on the ROC corresponds to a particular cut-off. In terms of model comparison, the ideal curve coincides with the upper end of the left-hand axis.

The AUC index assesses overall model performance for a range of cut-off values. It can be interpreted as the probability that when we randomly pick one positive and one negative patient, the model will assign a higher score to the positive example than to the negative one. To test the differences in AUC indices between our models and the RM, we used the nonparametric Mann-Whitney two-sample statistic. Even though the difference in AUC indices between our models and RM seem to be very small, AUCs for all our models were significantly higher (p<0.001) then the AUC for the RM.

Table 1. Classification results and Area Under the Curve (AUC) indices for the best conforming models.

Model	Counts N=9,285				Adjusted Percent				Area Under Curve
	FN	FP	TN	TP	%FN	%FP	%TN	%TP	AUC (95% CI)
LR	180	25	4,617	4,463	0.012	0.537	99.163	0.288	**0.9952** (0.9940-0.9964)
DT	164	45	4,597	4,479	0.011	0.967	98.733	0.289	0.9945 (0.9932-0.9959)
NN	152	56	4,586	4,491	**0.010**	1.203	98.497	**0.290**	0.9947 (0.9933-0.9961)
RM	253	4	4,638	4,390	0.016	**0.086**	**99.614**	0.284	0.9737 (0.9705-0.9769)

Based on model comparison criteria, we were not able to select one model which had both the lowest FN rate and the highest AUC index. The simple LR model demonstrated better overall predictive accuracy (highest AUC index) and lowest total misclassification rate (FN%+FP%=0.549%), while more complex NN model achieved the lowest FN rate.

All of our models outperformed the RM in terms of lower FN rates and higher AUC indices. The lowest adjusted FN rate (0.010% vs. 0.016% for the RM) was demonstrated by the NN model with 54 inputs and 3 neurons, while the highest AUC index was observed for the LR model (0.995 [0.994, 0.996] vs. 0.974 [0.971, 0.977] for the RM).

Let's keep in mind that we are modeling a very rare event with 0.3% prevalence and there is not much room for improvement. A simple model, which predicts everybody as negative, will correctly classify 99,7% of the population. Therefore, even a small difference in performance criteria over the RM is a real achievement for our modeling effort.

Examples of non-HIV-specific variables selected by our models included age, race/ethnicity, marital status, service-connected disability, number of days hospitalized, number of primary care and social work visits, number of total and lipids lab tests, blood pressure and liver co-morbidities. Three best candidate models were applied to VA data to predict gaps in access to quality HIV care.

STEP 7. APPLY NEW ALGORITHMS TO THE VA POPULATION

We applied our best models to the VA data to (1) predict HIV status for all patients in a study population, (2) compare classification results for our models to the RM in terms of misclassification rates and (3) identify cohorts of new patients with high predicted probability of having HIV disease.

The scored data included all pre-classified TP and TN from the study population (N=4,963,796). Classification results obtained from applying final conforming models to the VA population are shown in Table 2.

Were our conforming models doing a better classification job then the current RM algorithm? We compared model performance in terms of (1) FN and TP rates, which measure positive predictive accuracy of the models and (2) number of "new" cases, which are patients predicted positive by our models (TP+FP) and not on the Registry. Once again, all our models outperformed the RM in terms of lower FN rate and higher TP rate. The lowest FN rate (0.010% vs. 0.016% for the RM) and the highest TP rate (0.302% vs. 0.296% for the RM) were observed for the NN model. The NN model correctly classified 305 additional positive patients compared to the RM (14,996-14,681=305). However, the improved positive predictive accuracy of the NN model was achieved at the expense of a higher FP rate (1.468% for NN vs. 0.100% for the RM). Even such a relatively small difference in FP rate (1.468%-0.100%=1,368%) translated into roughly additional 68,000 misclassified patients.

Table 2. Classification results from applying final conforming models to VA population.

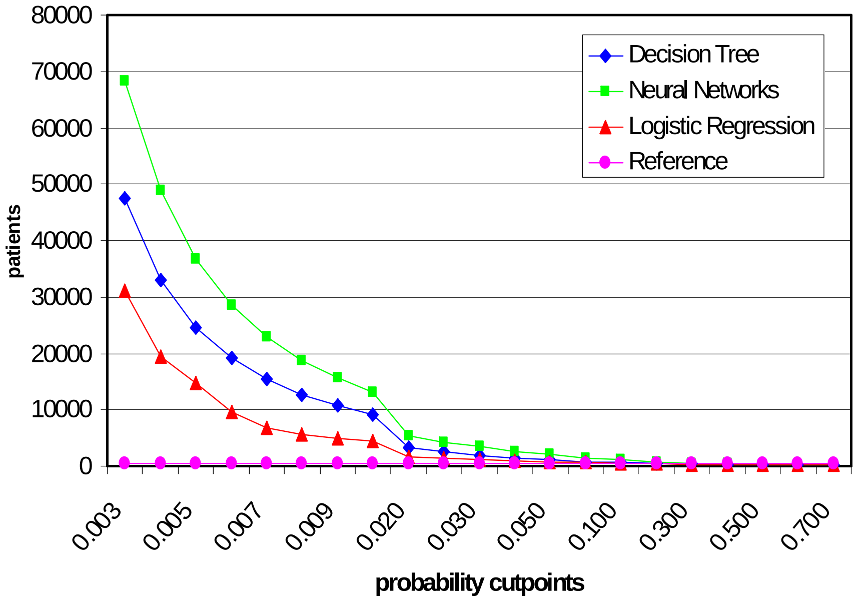
Model	Counts (N=4,963,796)				Percent				New Cases*
	FN	FP	TN	TP	%FN	%FP	%TN	%TP	N
LR	572	35,575	4,912,746	14,903	0.012	0.717	98.972	0.300	30,964
DT	526	52,125	4,896,196	14,949	0.011	1.05	98.638	0.301	47,478
NN	479	72,881	4,875,440	**14,996**	**0.010**	1.468	98.22	**0.302**	68,188
RM	794	4,966	4,943,355	14,681	0.016	**0.100**	**99.588**	0.296	519

^ New cases = predicted positive by our models and not on the Registry

We also identified "new" cases who are in theory, patients who have been missed by the system. They have a high predicted probability of HIV disease but are not on the Registry. Number of "new" cases varied from 30,964 for the LR model to 68,188 for the NN model.

It should not be forgotten that these results were obtained from the *conforming* models, which used the population probability of HIV as the decision threshold. We obtained different classification results, including number of newly identified cases, for different decision thresholds, as illustrated on Figure 4. The sharp drop around the 0.020 cutpoint for all models shown in the plot suggests a notable reduction in the FP rate. We used this change as the basis for our decision to select a conservative threshold that determined the modeling cohorts for our further analytic efforts in evaluating access to HIV care.

Figure 4: Number of newly identified patients for a range of decision thresholds**.**



The choice of the 0.05 decision threshold depended on four factors: (1) purpose of the analysis (2) FN misclassification rate for the RM (3) the trade-off between FP and FN rates and sensitivity and specificity, and (4) the level of statistical uncertainty we were willing to accept.

The purpose of analysis was to identify new patients with high predicted probability of having HIV disease. Since we wanted to improve the accuracy of the RM, we wanted to make sure that the FN rate for our models did not exceed the FN rate for the RM. Therefore, we considered a range of thresholds between 0.003 (the conforming model) and 0.075 (where the FN rate for our models exceeded the FN for the RM). We were willing to accept a trade-off between the FP and FN rates in order to lower the high FP rate for the conforming models. Additionally, we wanted to have a high level of certainty, as well as be somewhat conservative in our estimate.

Each threshold represents a trade-off between FN and FP misclassification rate, as well as between sensitivity and specificity. The lower the threshold selected, the lower the FN rate and specificity and the higher the FP rate and sensitivity will be. If the purpose of the analysis is to identify more new cases, then a lower decision threshold should be selected. However, since we wanted greater confidence in our predictions, we preferred the higher 0.05 threshold, indicating a 5% chance of having HIV. Misclassification rates, total number of predicted positives, including those already on the registry, and newly identified cases for 0.05 decision threshold are shown in Table 3.

Table 3. Misclassification rates, number of predicted positives, and number of newly identified cases using a 0.05 decision threshold.

Model	FN %	FP %	No. Predicted Positives	No. on the Registry	No. (%) New cases*
LR	0.016	0.105	21,574	20,763	811 (4%)
DT	0.016	0.111	21,910	20,805	1,105 (5%)
NN	0.015	0.131	22,908	20,798	2,110 (10%)
RM	0.016	0.100	21,348	20,817	531 (3%)

^ New cases = predicted positive by our models and not on the Registry

By using more conservative decision threshold, we significantly reduced the number of newly identified cases. Apart from those already on the registry, our new algorithms have identified additional 4% (n=811 for LR) to 10% (n=2,110 for NN) patients with high predicted probability of HIV. These newly identified patients have been used for further analysis, which aimed to identify individual patient and system characteristics that have predicted entry onto the Registry.

LESSONS LEARNED

SAS Enterprise Miner 5.2 provided a nice graphical programming interface, an adequate computing power and perfect compatibility with existing VA SAS databases. Unfortunately, some important features we were looking for in a data mining software package have not been fully implemented in SAS Enterprise Miner 5.2, so we used free R statistical software and custom programming to do this ourselves. These features included (1) the newest classification techniques, such as Boosting and Bagging; (2) full flexibility for adjusting the tuning parameters in the modeling nodes, such as weight decay in the Neural node; and (3) confidence intervals around odds ratio estimates in the SAS output from the Regression node, which is a standard feature in PROC LOGISTIC.

As health care is increasingly organized into large, integrated systems, patient data from these systems will be merged to form integrated electronic record systems such as the one we have used in the U.S. Veterans Health Administration. In turn, health policy decisions will require large comprehensive disease registries. We have learned that data mining techniques can help with the potentially laborious tasks of identifying cases for inclusion in a national registry, and with assuring that the registry is appropriately accurate and comprehensive.

STUDY LIMITATIONS

Drawing clinical conclusions from data that are in part collected for administrative and utilization-monitoring purposes always should be done with caution. Nevertheless, VA administrative and clinical databases have been used widely to support numerous published research efforts in the past, and meet a comparatively high quality standard. Also, the TN classification was based on presumed status and not on a clear standard since the large majority of patients in any healthcare system would not have been tested for HIV. Finally, the newly identified cases still need to be clinically validated to confirm their positivity.

CONCLUSIONS

1) SAS Enterprise Miner 5.2, in conjunction with R software and custom programming, provided modeling tools to achieve the study objectives;

2) We were able to improve the current HIV case finding algorithm by using supervised data mining techniques and additional variables;

3) An improved HIV casefinding algorithm, in terms of accuracy and clinical policy priorities, may serve as a practical tool for improving a disease registry for HIV care management;

4) This novel application of standard data mining methods may also be applied to development of other disease registries where electronic clinical data are available.

APPENDIX: Description of screened models

Logistic Regression (24 models screened):

- *Number of input variables:* 24, 32, 55, 65, 103, 115, 123
- *Model selection method:*
  - None (Full Regression)
  - Stepwise
  - Backward
  - Forward
- *Model selected:* Stepwise LR with 19 input variables

Artificial Neural Networks (53 models screened):

- *Architecture:* Multilayer Perceptron (MLP) with one hidden layer
- *Number of input variables and neurons:*
  - 24, 32, 54, 55, and 65 inputs with 2-10 and 15 neurons
  - 123 inputs with 3-5 neurons
- *Training techniques:* Quasi-Newton, Back Propagation, and Conjugate Gradient
- *Model selected:* MLP with 54 input variables and 3 hidden neurons, Conjugate Gradient training, linear combination function and logistic activation function

Decision Trees (115 models screened):

- *Number of input variables:* 24, 32, 54, 64 & 123
- *Approaches:*
  - Simple Tree using Gini splitting rules
  - Bagged trees with variable numbers of iterations (10 to 250 by 10 increment)
  - Boosted trees with variable numbers of iterations (50 to 1000 by 50 increment) and shrinkage factors
- *Model selected:* Boosted tree with 500 trees and shrinkage factor 0.7

REFERENCES:

(1) Rabeneck L, Menke T, Simberkoff MS, et al. *Using the national registry of HIV-infected veterans in research: lessons for the development of disease registries.* J Clin Epidemiol 2001; 54:1195-203.

(2) Michael J. A. Berry, Gordon Linoff. 1997. *Data Mining Techniques: For Marketing, Sales, and Customer Support.* New York: John Willey & Sons.

(3) Predictive Modeling Using Enterprise Miner™ Software Course Notes. Copyright 2003. SAS Institute Inc.

(4) Giudici, Paolo. 2003. *Applied Data Mining: Statistical methods for business and industry*. Chichester: John Willey & Sons.

(5) Brieman L. *Bagging Predictors*. Machine Learning. Vol. 26 #2, 1996:123-140

(6) Freund Y, Schapire R. *Experiments with a new boosting algorithm*. International Conference on Machine Learning, 1996:148-156.

CONTACT INFORMATION:

Your comments and questions are valued and encouraged. Contact the author at:

Tamara Slipchenko, PhD
DM Consulting
8875 Costa Verde blvd., #1613
San Diego, CA 92122
Phone: 858.458.0305
e-mail: tomicella@yahoo.com