Published in final edited form as:
Neuroimage. 2008 March 1; 40(1): 248–255. doi:10.1016/j.neuroimage.2007.09.056.

Assessing multiple-group diagnostic problems with multi-dimensional receiver operating characteristic surfaces: Application to proton MR Spectroscopy (MRS) in HIV-related neurological injury

Constantin T. Yiannoutsos[1], **Christos T. Nakas**[2], **Bradford A. Navia**[3], and **the proton MRS Consortium**

[1]Division of Biostatistics, Indiana University School of Medicine, Indianapolis, IN

[2]Laboratory of Biometry, School of Agricultural Sciences, University of Thessaly, Magnesia, Greece

[3]Tufts Medical School Boston, MA

Abstract

We present the multi-dimensional Receiver Operating Characteristic (ROC) surface, a plot of the true classification rates of tests based on levels of biological markers, for multi-group discrimination, as an extension of the ROC curve, commonly used in two-group diagnostic testing. The volume under this surface (VUS) is a global accuracy measure of a test to classify subjects in multiple groups and useful to detect trends in marker measurements. We used three-dimensional ROC surfaces, and associated VUS, to discriminate between HIV-negative (NEG), HIV-positive neurologically asymptomatic (NAS) subjects and patients with AIDS demential complex (ADC), using brain metabolites measured by proton MRS. These were ratios of markers of inflammation, Choline (Cho) and myoinositol (MI), and brain injury, N-acetyl aspartate (NAA), divided by Creatine (Cr), measured in the basal ganglia and the frontal white matter. Statistically significant trends were observed in the three groups with respect to MI/Cr (VUS=0.43; 95% confidence interval (CI) 0.33-0.53), Cho/Cr (0.36; 0.27-0.45) in the basal ganglia and NAA/Cr in the frontal white matter (FWM) (0.29; 0.20-0.38), suggesting a continuum of injury during the neurologically asymptomatic stage of HIV infection, particularly with respect to brain inflammation. Adjusting for age increased the combined classification accuracy of age and NAA/Cr (p=0.053). Pairwise comparisons suggested that neuronal damage associated with NAA/Cr decreases was mainly observed in individuals with ADC, raising issues of synergism between HIV infection and age and possible acceleration of neurological deterioration in an aging HIV-positive population. The three-dimensional ROC surface and its associated VUS are useful for assessing marker accuracy, detecting data trends and offering insight in disease processes affecting multiple groups.

Contact information for corresponding author Constantin T. Yiannoutsos, Ph.D. Indiana University School of Medicine Division of Biostatistics 1050 Wishard Blvd, RG 4101 Indianapolis, IN 46202 Tel (317) 278-3045 Fax (317) 274-2678 cyiannou@iupui.edu.

Introduction

There are numerous instances where discrimination between two groups of individuals is based on the result of a test. For example, statistical tests are often used to distinguish between healthy subjects and individuals having a certain disease. Invariably, these tests are based on the values of a biological measurement. Frequently this entails using a threshold value above or below which the test is positive (that is, it indicates that the measurement has been obtained from a diseased individual). In the opposite case, the test is negative. All test methodologies are subject to false positive (i.e., mistaken determination that a healthy subject suffers from the disease in question) or false negative errors (i.e., mistaken determination that a diseased subject is healthy). All else being equal, the probability of these errors can be increased or decreased by altering the threshold value of the biological measurement. Assessment of the ability of a marker to discriminate between healthy and diseased subjects is routinely performed by plotting the true positive rates (also called "sensitivity" of the test) versus the false positive rates or true negative rates ("specificity") of the test generated from consideration of all such cutoff values. The resulting plot is called a Receiver Operating Characteristics (ROC) curve (Swets and R. 1982; Shapiro 1999; Zhou et al. 2002). A global index of the ability of a test to diagnose the condition of interest is the area under the ROC curve (AUC). For example, if higher marker values indicate increased likelihood of disease, the area under the ROC curve is the average chance that a randomly selected diseased individual will have higher marker measurements compared to a randomly selected healthy individual (Bamber 1975). A number of statistical approaches have been developed to estimate the AUC, as reviewed in (Zhou et al. 2002).

ROC curve methodology was originally developed for use in signal detection theory, but has since been utilized for a large number of pair-wise disease classification problems (e.g.,disease state – signal versus healthy state – noise), from psychometrics (Dorfman and Alf 1968) to radiology (Metz 1986). Recently, ROC methodology was extended to three-class diagnostic problems by constructing a three-dimensional surface. (Mossman 1999; Dreiseitl et al. 2000; Heckerling 2001; Nakas and Yiannoutsos 2004) The volume under this three-dimensional surface (VUS) is a direct extension of the AUC. The interpretation of the VUS is similar to the AUC. If, for example, the marker values in Group 1 are expected to be lower than those in Group 2 and, likewise, the marker values in Group 2 are expected be lower than those in Group 3, the VUS is the average chance that the marker measurements obtained from three randomly selected individuals of Groups 1, 2 and 3 will have the expected ordering (Nakas and Yiannoutsos 2004).

The VUS method is also powerful in that generalization to more than three groups is straightforward. In addition, appropriately defined classification criteria (see next section), result in a VUS that assesses a *trend* in the marker measurements (e.g., Group 1 < Group 2 < Group 3) as these are frequently suggested by the ordinal nature of the data. This is important for a number of situations in which both the simultaneous classification of a subject, in three or more groups, and the existence of a trend in test measurements are of scientific importance. One eminent such example is classification of disease state or injury based on measurements obtained from proton Magnetic Resonance Spectroscopy (^{1}H-MRS) into HIV-negative (NEG), HIV-positive neurologically asymptomatic (NAS) or suffering from HIV-related cognitive impairment (ADC). In addition to showing that MRS markers can distinguish between neurologically asymptomatic and symptomatic patients, another important question is whether a continuum of increasing neurological injury exists between the NEG, NAS and ADC groups. Evidence for such a continuum would be demonstrated if MRS measurements obtained from HIV-positive asymptomatic subjects tended to lie between those of HIV-negative subjects and those of HIV-positive neurologically impaired patients. The direct implication of this finding would be that HIV-associated neurological

injury is present even during the neurologically asymptomatic (as evaluated by neuropsychological assessment) stage HIV-infection. In addition, the existence of such trend would offer partial validation for the use of MRS measures as early markers of neurological injury and as a test to identify subjects at highest risk for neuro-cognitive impairment, well before overt clinical symptoms have been observed. Results would also indicate that measures such as the VUS are potentially important instruments to assess the classification accuracy of a biomedical marker and the existence of a trend in the marker values from one group to the other.

In this article, we use the volume under the ROC surface methodolocy in a three-group classification problem as a generalization of the area under the ROC curve in three instead of two-group diagnostic situations. We apply the VUS in an example of three-group classification involving HIV-negative and HIV-positive neurologically asymptomatic and neurologically impaired patients.

Patients and methods

Study subjects and imaging protocols

Data from an MRS study undertaken under the auspices of the AIDS Clinical Trials Group (ACTG; protocol 700) were analyzed for this research. The specifics of this study have been presented in more detail elsewhere (Lee et al. 2003; Chang et al. 2004). Briefly, data were collected at 11 participating imaging centers: Massachusetts General Hospital (MGH, the central site), University of California Los Angeles, Harbor-UCLA Medical Center, University of Pennsylvania, University of Washington, University of Rochester, Mt. Sinai Medical Center, University of Texas Galveston, University of Nebraska and University of California San Francisco. MRS examinations were performed on Signa 1.5T MR imagers (GE Medical Systems, Milwaukee, WI) operating with system version 5.6 or 5.7 using the standard GE quadrature head coil and a PRESS sequence with CHESS water suppression. MRS data were obtained with the automated GE pulse sequence PROBE-P, a PRESS sequence (Bottomley 1987) with CHESS water suppression (Chang et al. 1996). MRS parameters included TE=35ms, TR=3000ms, voxel size=6cm^3 (20×20×15mm^3), 128 acquisitions, spectral width=2500Hz, and 2k data points. For consistency, the sequence was compiled and distributed to participating sites by the MGH-NMR Center (Lee et al. 2003). Spectra were localized to three brain regions corresponding to different tissue types: mid-line posterior parietal cortex (gray matter), mid-frontal centrum semiovale (white matter) and basal ganglia (deep gray matter). Data and images from completed examinations, showing voxel placement and raw spectroscopic data, were transferred to MGH for processing (for voxel placement and spectroscopic data see (Lee et al. 2003), Figure 1 and 2). All raw data were analyzed in an identical manner using the commercial software package Sage IDL (GE Medical Systems, Milwaukee, WI) (Webb et al. 1994).

In this study, we focus on the ratios of N-acetyl aspartate (NAA) a marker metabolism of mature neurons (Urenjak et al. 1992), choline (Cho), a marker of brain inflammation (Tracey et al. 1996) and myoinositol (MI), a marker of glial-cell metabolism (Tourbah et al. 1999), over Creatine (Cr) measured in the white matter (NAA/Cr) and the basal ganglia (Cho/Cr and MI/Cr). Cho/Cr and MI/Cr have been reported to increase and NAA/Cr to decrease with progressive neurological injury (Tracey et al. 1996; Lopez-Villegas et al. 1997), so they are candidate markers for HIV-related damage to the central nervous system. Metabolite data were derived from subjects with a diagnosis of AIDS dementia complex (ADC). An ADC diagnosis was based both on ADC staging (ADC stages 1-3) (Price and Sidtis 1990) as well as on impaired performance on neuropsychological tests compared to age and education-matched normative values derived from HIV-negative controls participating in the Multicenter AIDS Cohort Study (MACS). There were eight

neuropsychological tests involved: Timed gait, grooved pegboard, performed with the dominant and non-dominant hand (Lafayette Instrument Company 1989), symbol digit (Smith 1973), trail making series A and B (Reitan 1958) and two reaction time tests (simple and sequential reaction time) from the California Computerized Assessment Package (CalCAP (Miller 1990)). Age and education were grouped in five categories (age categories: <30, 30-39, 40-49, 50-59 and ≥60 years; education categories: <12 years, 12 years, 12-15 years, 16 years, >16 years). Appropriately transformed test scores (Box and Cox 1964) were then averaged over these categories. Where the number of representatives from the normative sample was small, predicted mean values were imputed. These were derived from a regression model involving the remaining categories (EN Miller and CT Yiannoutsos, unpublished data). Data were also obtained from neurologically asymptomatic HIV-seropositive individuals (ADC stage 0 and no impairment on neuropsychological tests) and similarly chosen neurologically unimpaired HIV-negative subjects without neurological impairment. Informed consent was obtained from all study participants; the protocol was approved by the human subjects review board of each participating site.

Volume under the three-dimensional ROC surface

The goal of this research is to introduce the volume under the ROC surface as an instrument to evaluate the ability of proton MRS-measured brain metabolites to distinguish between HIV-infected individuals suffering from neurological complications and HIV-positive asymptomatic/HIV-negative controls. In a recent publication from our consortium (Chang et al., 2004) it was suggested that the levels of these metabolites are ordered with MRS measurements obtained from NEG and ADC subjects at the two extremes and those acquired from NAS subjects always between them. Generally, increasing trends were observed with respect to Cho/Cr and MI/Cr while a decreasing trend was evident with respect to NAA/Cr levels, particularly in the white matter. Taken together, these observations indicate the possible existence of a continuum of injury that is evident during the neurologically asymptomatic stages of HIV infection. Thus, we would like to focus on measures that assess the accuracy of classification into NEG, NAS and ADC using each metabolite, taking into account the existence of a possible trend in the measurements among the three groups.

We introduce in this context the Receiver Operating Characteristc (ROC) surface, a three-dimensional generalization of the ROC curve (Mosman 1999, (Nakas and Yiannoutsos 2004). The ROC *surface* is a visual representation of the correct classification of an experimental unit under three possible alternatives (as opposed to two in the case of the ROC *curve*). We focus here on ROC surfaces arising from continuous measures in groups that have an expected inherent ordering as reported previously (Nakas and Yiannoutsos 2004). Such is the case with the HIV-patient cohort considered here.

As depicted in Figure 1, the three-dimensional ROC surface shows pictorially how a particular metabolite differentiates among subjects belonging in these three *ordered* groups simultaneously. The three classification rules based on a continuous measurement X and two ordered cutoff points $c_1 \leq c_2$ are as follows (Figure 1):

- If $X < c_1$ → Assign into group 1
- If $c_1 \leq X \leq c_2$ → Assign into group 2
- If $X > c_2$ → Assign into group 3

Maintaining the restriction that $c_1 \leq c_2$ and sweeping c_1 and c_2 across all possible values for the measurement under consideration, one obtains the ROC surface depicted in Figure 2.

Given that there are three groups, there are three possible correct classifications and six possible incorrect classifications,

$$\begin{aligned} TC_k &= P(T_k|D_k), \quad k=1,2,3 \\ FC_{ij} &= P\left(T_i|D_j\right), \quad i,j=1,2,3 \text{ and } i \neq j \end{aligned} \quad (1)$$

where T_i is the class assigned by the test and D_i is the true group membership. Thus, for example, the true classification rate for the third group is $TC_3=P(T_3|D_3)$, the probability that an individual that belongs in the third group will be correctly classified into the third group. Alternatively, the false classification rate into the first group of an individual that belongs in the second group is $TC_{12}=P(T_1|D_2)$.

To quantify the overall classification accuracy of an MRS measure, we use the volume under the ROC surface (VUS). The VUS is a direct analog to the area under the ROC curve (AUC) and has similar properties (Dreiseitl et al. 2000; Heckerling 2001; Nakas and Yiannoutsos 2004). If all subjects are classified perfectly by the measurement in question, the ROC surface will cover the whole cube and the VUS will be equal to one. Conversely, if the measurement is useless to correctly classify any of the subjects, any correct classification will be by chance and thus the VUS will be approximately 1/6. Statistical tests can be developed (Heckerling 2001; Nakas and Yiannoutsos 2004) that provide a rigorous criterion about whether a measurement classifies subjects among three groups at significantly higher rates (in the statistical sense) compared to chance. In particular, statistical tests of trend have been introduced, which address the focused question of whether a specific trend exists in the three groups and statistical testing can be used to compare the volumes under ROC surfaces resulting from two measurements, providing in this manner a statistical criterion of the superiority of one measure in three-group classification versus another (Nakas and Yiannoutsos 2004).

As demonstrated in a previous analysis (Chang et al. 2004), the age of each individual has a significant effect on the level of the MRS-measured metabolite, in addition to the overall group effect. Thus, estimates of the VUS were adjusted for the age of each individual. This analysis is similar to regression approaches of the two-dimensional ROC curve (Beam 1995; Gatsonis 1995), thus illustrating how ancillary predictors such as age can be incorporated into the analysis of three-dimensional data. In this case, we fit a multinomial logistic regression model to describe the probability of belonging to one of the three groups (NEG, NAS or ADC) based on each MRS measure under consideration, as well as the age of each individual. Then, the three-dimensional ROC surface was constructed based on the estimated probabilities derived from this model instead of the raw MRS data, thus adjusting for subject age.

Pair-wise classification with the three-dimensional ROC surface

It is frequently of great interest to establish the pair-wise classification accuracy of certain measures. This can be accomplished by constructing the three-dimensional ROC surface and simply considering the projections of the surface on the three coordinate planes (Figure 2). Each of these projections will result in a two-dimensional ROC curve plotting the true classification (true positive) rates. This is slightly different than the usual ROC curve where the true positive rate (sensitivity) is plotted against the false positive rate. However, the interpretation is similar (i.e., larger AUC correspond to more accurate pair-wise classification). In this manner, by considering the ROC surface, we gain all the advantages of simultaneously classifying subjects in three groups without forgoing the ability to make pair-wise statements about the classification accuracy between two groups. For the current

study, ROC curves were produced by rotating the ROC surfaces to view the projections of the surfaces on the sides of the unit cube using MATLAB 6.5.

True-class predictive values

Another carry over from ROC curves is the true-class predictive value, the multidimensional equivalent of the positive or negative predicted value in the two-dimensional case. Predictive values are often of more interest to patients and clinicians than volumes under the ROC surface. Their calculation proceeds in a straightforward manner by application of the Bayes formula. For example, in a three-class problem the true-class predictive value for class 1 $P(D_1 \mid T_1)$ is given by the following formula, which is an immediate extension of the two-class case:

$$P(D_1 \mid T_1) = \frac{P(D_1)P(T_1|D_1)}{P(D_1)P(T_1|D_1)+P(D_2)P(T_1|D_2)+P(D_3)P(T_1|D_3)} = \frac{P(D_1)TC_1}{P(D_1)TC_1+P(D_2)FC_{12}+P(D_3)FC_{13}} \quad (2)$$

where the true classification rates are defined as in (1) above and $P(D_i)$ is the prevalence of class i in the population, with i=1,2,3,… classes. True-class predictive values for classes 2 and 3 are calculated in an identical manner.

Statistical analysis

Point estimates and 95% confidence intervals are presented for all parameters (i.e., for the volume under the ROC surface as well as the area under the ROC curve). The confidence intervals for the VUS are constructed following methodology presented by Nakas and Yiannoutsos. It is based on bootstrap estimates of the variability of the VUS estimator (Nakas and Yiannoutsos 2004). Point estimates of the AUC were generated via the trapezoidal rule applied to the non-parametric estimate of the ROC curve and confidence intervals were produced by the method of DeLong, DeLong & Clarke-Pearson implemented by STATA software version 9.0 (STATA Corporation, College Station, TX) (DeLong et al. 1988). Age was incorporated into the VUS and AUC estimates by carrying out a linear regression on the Box-Cox-transformed MRS ratio,

$$y^{(p)} = \begin{cases} (y^p - 1)/p & \text{if } p \neq 0 \\ \log(y) & \text{if } p=0 \end{cases} \quad (3)$$

(Box and Cox 1964) where $y^{(p)}$ is a transformation of each MRS ratio y and p is estimated from the data. The linear regression included indicators associated with the classification group (i.e., NEG, NAS or ADC) as well as each patient's age. Then, the ROC surface and the corresponding ROC curves were constructed by using the estimated values of the MRS ratio resulting from the linear-regression model. In this manner, age (as well as any other factor, continuous or discrete) can straightforwardly be added to the model resulting in a combined diagnostic index. Thus, when age is added to the estimation of the ROC surface or curve, the resulting estimate reflects the combined diagnostic capability of all predictive factors. Point estimates and confidence intervals were produced in the same manner described earlier. The same methods (i.e., boostrap for the VUS and DeLong, DeLong and Clarke-Pearson for the AUC) were used to compare the VUS and AUC between two diagnostic indices.

Results

Cho/Cr in the basal ganglia

ROC surfaces for all possible orderings of the NEG, NAS and ADC groups were constructed. The greatest VUS was produced by the ROC surface corresponding to the order NEG <NAS< ADC. The VUS = 0.36 (95% confidence interval 0.27 – 0.45). This is significantly higher than the reference level of 1/6, suggesting that Cho/Cr measured in the basal ganglia can classify three subjects, one from each group, in the correct ordering 36% of the time. A complete listing of VUS associated with Cho/Cr measured in the basal ganglia is provided in Table 1.

Considering the pair-wise projections to the sides of the unit cube of the ROC surface, we obtain the pair-wise classification between NEG and NAS subjects based on Cho/Cr levels in the basal ganglia (AUC 0.61), the classification between NEG and ADC (AUC 0.76) and between NAS and ADC (AUC 0.68). Among these, the first and third pair-wise comparisons are of particular interest. The first, NEG versus NAS, assesses the ability of the Cho/Cr to distinguish between NEG and NAS subjects (e.g., assessing injury among HIV-positive individuals during the neurologically asymptomatic stage). The third, NAS versus ADC, measures differences associated with the clinical manifestation of the neurological disease among HIV-infected individuals.

MI/Cr in the basal ganglia

Similarly, MI/Cr in the basal ganglia produces the highest VUS=0.43 (95% CI 0.33 – 0.53) corresponding to the trend NEG<NAS<ADC. This result is statistically significant, suggesting that MI/Cr is a marker that can differentiate among three subjects, one from each group, in the correct order about 43% of the time (Table 1). In addition, the VUS associated with MI/Cr is higher than the one associated with Cho/Cr measured in the same region. A statistical test can be used to formally assess the possibility that MI/Cr is a superior marker compared to Cho/Cr both measured in the basal ganglia. The result of this test was inconclusive (p=0.126).

Considering the areas under the pair-wise ROC curve, we obtain three AUC for the comparison between NEG versus NAS (AUC=0.65), NEG versus ADC (AUC=0.79) and NAS versus ADC (AUC=0.73) suggesting that MI/Cr a moderately strong classifier between HIV-positive subjects without attendant neurological disease and subjects suffering from AIDS-related cognitive impairment.

NAA/Cr in the white matter

The same analysis was conducted for NAA/Cr in the white matter; results indicate a decreasing trend between the NEG, NAS and ADC groups. The VUS of this trend is 0.29 (95% CI 0.20 – 0.38; Table 1). This is statistically significant compared to the reference level of 1/6, but the size of the VUS suggests weak classification accuracy among the three groups.

Considering the pair-wise areas under the two-dimensional ROC curves by projecting the ROC surface onto the three coordinate planes, the generated AUC for the three pairwise comparisons are AUC=0.70 for the comparison between ADC and NAS, AUC=0.51 for NAS versus NEG and AUC=0.70 for the comparison NEG versus ADC. Given that the AUC resulting from the comparison between NAS and NEG is almost equal to the reference value of 0.5, the area under the two-dimensional ROC curve, these two groups are virtually indistinguishable with respect to NAA/Cr measured in the white matter.

Effect of age on diagnostic accuracy

Inclusion of age in the model increased the classification accuracy of the resulting diagnostic index derived by considering both the MRS ratio and the age of each subject. Estimates of the VUS and the AUC, with age included in the model, are given in Tables 1 and 2. In all cases adding age increased the diagnostic accuracy of the combined index albeit in a non-statistically significant manner. A possible exception is the increase in predictive accuracy relating to the NAA/Cr in the white matter. There, the VUS increased from 0.29, in the univariate analysis, to 0.42 in the analysis that included age (p=0.053), suggesting a potentially substantial improvement in subject classification when subject age is taken into consideration.

Discussion

The three-dimensional ROC surface is a useful tool that can visually summarize the ability of a biological marker to classify individuals between more than two groups. The volume under the ROC surface (VUS) is a measure of the global classification ability of the discriminating criterion. Suitably defined (Nakas and Yiannoutsos 2004), the VUS can be a global measure of the presence of a trend in the marker values in three groups. This is particularly relevant in the analysis of brain metabolite data measured with MRS among HIV-positive patients suffering from ADC and HIV-positive/HIV-negative neurologically asymptomatic controls. The presence of a trend with respect to certain MRS-measured metabolites, where the values corresponding to HIV-positive controls are always between those of the ADC patients and those of the HIV-negative controls, lends support to the existence of a continuum of central nervous system injury associated with HIV infection. In addition, the availability of the pair-wise ROC curves, as projections of the ROC surface on the three coordinating planes, provides further insights in the brain functioning of these three groups. Significant tertiary factors, such as subject age, can be included into both the construction of the ROC surface and the resulting estimation of the VUS in a straightforward manner.

The ROC surface generated by Cho/Cr (Table 1) measured in the basal ganglia produced a VUS of 0.36 (95% CI 0.27 – 0.45), suggesting that 36% of the time, Cho/Cr levels in three randomly selected individuals, one from each of the three groups, will have the correct ordering (NEG → NAS → ADC). This result is significantly higher than 1/6, the reference level, as suggested by the lower bound of the 95% confidence interval. The VUS associated with MI/Cr is even higher 0.43 (0.33 – 0.53), suggesting that a higher proportion of MI/Cr levels among three subjects, one from each group, have the anticipated ordering. Comparison between 43% and 36% was inconclusive, but the results nevertheless suggest that the use of MI/Cr in the basal ganglia might be a potentially useful quantity by which to classify subjects in these three groups. The VUS associated with NAA/Cr measured in the white matter was 0.29 (0.20 – 0.38). This is weakly significant as suggested by the proximity of the lower bound to the reference level of 1/6, implying the existence of a weaker trend of NAA/Cr levels in the three groups. This is also evident by the corresponding ROC surface (Panel C in Figure 2) that lies close to the diagonal plane of the cube bordered by the three origin points (0,0,1), (0,1,0) and (1,0,0).

The pairwise comparisons (Table 2) provide additional information about the distinction between each pair of groups considered separately, as well as provide insight into the underlying disease process. Of particular interest are comparisons between the NEG and NAS groups, which imply that metabolic perturbations may be present in asymptomatic HIV-positive patients. The comparison between NAS and ADC is also important as it assesses metabolic changes specific to the manifestation of clinical neurological impairment. With respect to Cho/Cr, the area under the pair-wise ROC curve corresponding to the

comparison between NEG and NAS was 0.61. The same AUC corresponding to the comparison between NAS and ADC is 0.68. This suggests a trend between the three groups and is consistent with reports of persistent brain tissue inflammation at all stages of HIV infection (Glass et al. 1993; Jarvik et al. 1993; Jarvik et al. 1996; Tracey et al. 1996; Lopez-Villegas et al. 1997; Salvan et al. 1997; Chang et al. 1999). The areas under the ROC curves associated with MI/Cr levels in the basal ganglia were 0.65 corresponding to the NEG versus NAS comparison and 0.73 in the NEG versus ADC comparison, suggesting that glial-cell metabolism (gliosis) may be increasing in conjunction with observable, overt neurological impairment. With respect to NAA/Cr measured in the white matter, we noticed the near absence of an overall trend as implied by the three-dimensional VUS near the cutoff value of 1/6. Inspection of the areas under the pair-wise ROC curves provides a clue as to the reason for this result. The AUC associated with the comparison of NEG versus NAS is 0.51, making these two groups virtually indistinguishable with respect to their respective NAA/Cr levels. By contrast, the AUC generated by the ROC curve associated with the comparison between NAS and ADC is 0.70 ($p=0.001$). This strongly suggests that significant NAA/Cr decreases in the white matter are associated with overt clinical disease (ADC group), but there is no evidence of neurological damage (reflected by decreases of NAA/Cr) among HIV-positive persons who do not demonstrate clinically manifested neurological disease. This result is consistent with a number of previous reports that suggest that, while inflammation (as reflected by changes in MI/Cr and Cho/Cr) is observable throughout HIV infection, NAA/Cr decreases happen later during progression to clinically manifested neurological disease (Tracey et al. 1996; Lopez-Villegas et al. 1997).

Adding age (as well as other factors) to the derivation of a diagnostic index is straightforward and leads to the construction of both ROC curves and surfaces. In our case we added age to the diagnostic index derived from each metabolite ratio. The reason is that age is a potentially significant factor associated with brain as changes in metabolites reflecting aging are similar to those related to HIV (Chang et al. 1996; Schuff et al. 1999; Chang et al. 2004). Incorporation of aging into the diagnostic procedure, slightly increased the accuracy of the combined index although this increase did not reach statistical significance.

A limitation of the above procedure is a possible overestimation of the accuracy of the testing modality, particularly in the case of age-adjusted estimates. In the case of unadjusted VUS estimates (i.e., with no-covariates, such as the estimates presented in the unadjusted and univariate model columns of Tables 1 and 2 respectively), using cross-validation is equivalent to employing a jackknife estimate to obtain confidence intervals for the VUS or AUC estimates. We have studied this issue in a previous paper (Nakas and Yiannoutsos 2004) and have concluded that resampling (i.e., bootstrap, jackknife) and U-statistics approaches (used in both our previous article and the current paper) produce similar results in general. So the cross-validation-derived VUS estimates and their corresponding confidence intervals will not be materially affected by performing cross validation in the absence of covariates.

Cross-validation can be important when age-adjusted VUS or AUC estimates are produced. In the leave-one-out cross-validation procedure for example, age-adjusted scores will be needed to produce N-1 ROC surfaces ($N=n+m+k$, where n, m and k are the sizes of the three groups). The new estimate will then be the average of the N-1 VUS and the corresponding confidence intervals produced using the cross validation procedure. This will result in revised age-adjusted VUS estimates and confidence intervals. This procedure will be repeated three times, once for each marker. This is an alternative approach but not a more valid estimation procedure than the U-statistics procedure employed here. Other validation

procedures can be considered as well. In the case where a large sample is available, observations can be randomly assigned to a training sample and a validation sample.

Results from this investigation indicate that the three-dimensional ROC curve and the attendant measure of volume under its surface (VUS) are useful instruments to visualize and assess the utility of continuously measured markers, particularly when there is an inherent ordering among the groups under comparison. This procedure is also very useful because it lends itself to the generation of all possible pair-wise ROC curves, thereby offer further insight in the relationships between pairs of groups. Used in the setting of proton Magnetic Resonance Spectroscopy (MRS), this methodology provides a wealth of information with respect to MRS marker validation and assessment, as well as valuable insight into the underlying disease process.

Acknowledgments

Data were collected through AIDS Clinical Trials Group protocol ACTG 700. This research was also supported by grant NS 36524 to the third author from the National Institute for Neurological Diseases and Stroke.

References

Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J Math Psych. 1975; 12:387–415.

Beam CA. Random-effects models in the receiver operating characteristic curve-based assessment of the effectiveness of diagnostic imaging technology: concepts, approaches, and issues. Acad Radiol Suppl. 1995; 1:S4–13.

Bottomley PA. Spatial localization in NMR spectroscopy in vivo. Ann N Y Acad Sci. 1987; 508:333–348. [PubMed: 3326459]

Box GEP, Cox DR. An analysis of transformations. J. Roy. Stat. Soc. Ser. B. 1964; 26:211–243.

Chang L, Ernst T, et al. Cerebral metabolite abnormalities correlate with clinical severity of HIV-1 cognitive motor complex. Neurology. 1999; 52:100–8. [PubMed: 9921855]

Chang L, Ernst T, et al. In vivo proton magnetic resonance spectroscopy of the normal aging human brain. Life Sci. 1996; 58:2049–56. [PubMed: 8637436]

Chang L, Lee PL, et al. A multicenter in vivo proton-MRS study of HIV-associated dementia and its relationship to age. Neuroimage. 2004; 23:1336–47. [PubMed: 15589098]

DeLong ER, DeLong DM, et al. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988; 44:837–45. [PubMed: 3203132]

Dorfman DD, Alf E Jr. Maximum likelihood estimation of parameters of signal detection theory--a direct solution. Psychometrika. 1968; 33:117–24. [PubMed: 5239566]

Dreiseitl S, Ohno-Machado L, et al. Comparing three-class diagnostic tests by three-way ROC analysis. Med Decis Making. 2000; 20:323–31. [PubMed: 10929855]

Gatsonis CA. Random-effects models for diagnostic accuracy data. Acad Radiol. 1995; 2(Suppl 1):S14–21. discussion S57-67, S61-4 pa. [PubMed: 9419701]

Glass JD, Wesselingh SL, et al. Clinical-neuropathologic correlation in HIV-associated dementia. Neurology. 1993; 43:2230–7. [PubMed: 8232935]

Heckerling PS. Parametric three-way receiver operating characteristic surface analysis using mathematica. Med Decis Making. 2001; 21:409–17. [PubMed: 11575490]

Jarvik JG, Lenkinski RE, et al. Proton MR spectroscopy of HIV-infected patients: characterization of abnormalities with imaging and clinical correlation. Radiology. 1993; 186:739–44. [PubMed: 8430182]

Jarvik JG, Lenkinski RE, et al. Proton spectroscopy in asymptomatic HIV-infected adults: initial results in a prospective cohort study. J Acquir Immune Defic Syndr Hum Retrovirol. 1996; 13:247–53. [PubMed: 8898669]

Lafayette Instrument Company. Grooved Pegboard Instruction Manual. Lafayette Instrument Company; Lafayette, IN: 1989.

Lee PL, Yiannoutsos CT, et al. A multi-center 1H MRS study of the AIDS dementia complex: validation and preliminary analysis. J Magn Reson Imaging. 2003; 17:625–33. [PubMed: 12766890]

Lopez-Villegas D, Lenkinski RE, et al. Biochemical changes in the frontal lobe of HIV-infected individuals detected by magnetic resonance spectroscopy. Proc Natl Acad Sci U S A. 1997; 94:9854–9. [PubMed: 9275215]

Metz CE. ROC methodology in radiologic imaging. Invest Radiol. 1986; 21:720–33. [PubMed: 3095258]

Miller, EN. California Computerized Assessment Package (CalCAP). Norland Software; Los Angeles: 1990.

Mossman D. Three-way ROCs. Med Decis Making. 1999; 19:78–89. [PubMed: 9917023]

Nakas CT, Yiannoutsos CT. Ordered multiple-class ROC analysis with continuous measurements. Stat Med. 2004; 23:3437–49. [PubMed: 15505886]

Price RW, Sidtis JJ. Evaluation of the AIDS dementia complex in clinical trials. J Acquir Immune Defic Syndr. 1990; 3(Suppl 2):S51–60. [PubMed: 2231303]

Reitan R. Validity of the Trail Making Test as an indicator of organic brain damage. Percept Mot Skills. 1958; 8:271–276.

Salvan AM, Vion-Dury J, et al. Brain proton magnetic resonance spectroscopy in HIV-related encephalopathy: identification of evolving metabolic patterns in relation to dementia and therapy. AIDS Res Hum Retroviruses. 1997; 13:1055–66. [PubMed: 9264293]

Schuff N, Amend DL, et al. Age-related metabolite changes and volume loss in the hippocampus by magnetic resonance spectroscopy and imaging. Neurobiol Aging. 1999; 20:279–85. [PubMed: 10588575]

Shapiro DE. The interpretation of diagnostic tests. Stat Methods Med Res. 1999; 8:113–34. [PubMed: 10501649]

Smith, A. Symbol digit modalities test: manual. Western Psychological Services; Los Angeles: 1973.

Swets, J.; P., R. Evaluation of diagnostic systems: Methods from signal-detection theory. Academic Press; New York: 1982.

Tourbah A, Stievenart JL, et al. Localized proton magnetic resonance spectroscopy in relapsing remitting versus secondary progressive multiple sclerosis. Neurology. 1999; 53:1091–7. [PubMed: 10496272]

Tracey I, Carr CA, et al. Brain choline-containing compounds are elevated in HIV-positive patients before the onset of AIDS dementia complex: A proton magnetic resonance spectroscopic study. Neurology. 1996; 46:783–8. [PubMed: 8618683]

Urenjak J, Williams SR, et al. Specific expression of N-acetylaspartate in neurons, oligodendrocyte-type-2 astrocyte progenitors, and immature oligodendrocytes in vitro. J Neurochem. 1992; 59:55–61. [PubMed: 1613513]

Webb PG, Sailasuta N, et al. Automated single-voxel proton MRS: technical development and multisite verification. Magn Reson Med. 1994; 31:365–73. [PubMed: 8208111]

Zhou, XH.; Obuchowski, NA., et al. Statistical Methods In Diagnostic Medicine. John Wiley & Sons; New York: 2002.

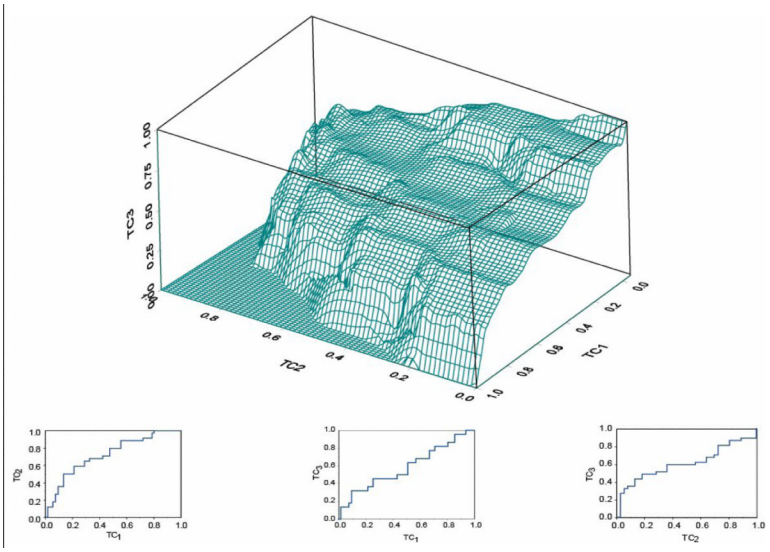

Figure 1.

Example of three-dimensional ROC surface with all three pair-wise two-dimensional ROC curve projections

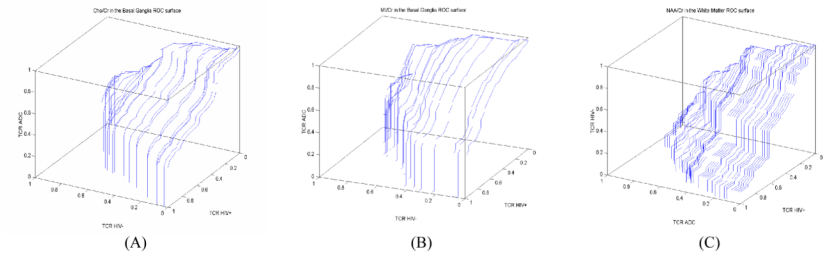

Figure 2.
ROC surfaces corresponding to Cho/Cr (A) and MI/Cr levels (B) in the basal ganglia and NAA/Cr levels (C) in the white matter.

Table 1

Volume under the three-dimensional ROC surface assessing the presence of a trend among three groups of subjects with respect to MRS metabolite levels. Unadjusted (univariate model) estimates and adjusted estimates for age derived from a multinomial logistic regression model that included age are shown. P-values reflect the comparison between the VUS estimates produced by the unadjusted dand adjusted models

	VUS (95% CI)		
Metabolite	**Unadjusted**	**With age added***	**p-value**
Cho/Cr in the basal ganglia			
NEG<NAS<ADC	0.362 (0.272, 0.452)	0.435 (0.339, 0.531)	0.324
MI/Cr in the basal ganglia			
NEG<NAS<ADC	0.430 (0.328, 0.532)	0.438 (0.335, 0.542)	0.925
NAA/Cr in the white matter			
NEG<NAS<ADC	0.287 (0.197, 0.377)	0.423 (0.329, 0.517)	0.053

Table 2

Area under the two-dimensional ROC curve assessing the overall accuracy of each marker to differentiate between individuals from the two groups under comparison

	AUC (95% CI)		
Metabolite	**Univariate model**	**With age added***	**p-value**
Cho/Cr in the basal ganglia			
NEG versus NAS	0.612 (0.482 – 0.742)	0.638 (0.507 – 0.770)	0.737
NEG versus ADC	0.767 (0.669 – 0.866)	0.804 (0.706 – 0.902)	0.519
NAS versus ADC	0.690 (0.584 – 0.795)	0.751 (0.651 – 0.852)	0.396
MI/Cr in the basal ganglia			
NEG versus NAS	0.649 (0.516 – 0.781)	0.628 (0.491 – 0.766)	0.797
NEG versus ADC	0.793 (0.688 – 0.898)	0.803 (0.698 – 0.907)	0.891
NAS versus ADC	0.726 (0.611 – 0.840)	0.764 (0.657 – 0.870)	0.600
NAA/Cr in the white matter			
NEG versus NAS	0.507 (0.374 – 0.640)	0.621 (0.489 – 0.753)	0.100
NEG versus ADC	0.714 (0.615– 0.811)	0.788 (0.688 – 0.889)	0.166
NAS versus ADC	0.714 (0.615 – 0.816)	0.751 (0.612 – 0.816)	0.596

*Box-Cox transformed