Journal of Data Science **7**(2009), 161-177

Quantifying Relative Superiority among Many Binary-valued Diagnostic Tests in the Presence of a Gold Standard

Reena Deutsch[1], Monica Rivera Mindt[2], Ronghui Xu[1],
Mariana Cherner[1], Igor Grant[1], and the HNRC Group[1]
[1] *University of California, San Diego and*
[2] *Fordham University and Mount Sinai School of Medicine*

Abstract: Comparison of more than two diagnostic or screening tests for prediction of presence vs. absence of a disease or condition can be complicated when attempting to simultaneously optimize a pair of competing criteria such as sensitivity and specificity. A technique for quantifying relative superiority of a diagnostic test when a gold standard exists in this setting is described. The proposed *superiority index* is used to quantify and rank performance of diagnostic tests and combinations of tests. Development of a validated model containing a subset of the tests may be improved by eliminating tests having a very small value for this index. To illustrate, we present an example using a large battery of neuropsychological tests for prediction of cognitive impairment. Using the proposed index, the battery is reduced with favorable results.

Key words: Binary outcomes, classification, diagnostic test, ranking, screening test, superiority.

1. Introduction

Screening patients for a specific disease or condition is commonly performed by administering one or more diagnostic tests or procedures, each one with an outcome characterized as either positive (high risk) or negative (low risk) for the disease. For instance, at the University of California, San Diego, HIV Neurobehavioral Research Center, 19 neuropsychological (NP) tests were used to classify subjects as normal vs. impaired in the cognitive domain measured by each test. These same subjects were also classified globally either as neurocognitively normal or as impaired, based on a systematic clinical rating system that relied on results from all 19 test measures and other criteria detailed in Woods *et al.* (2004) This global classification was considered to be the clinical "gold standard" with regard to characterization of neurocognitive status. A technique for quantifying the performance of each test relative to others in the prediction of this gold standard was sought. Essentially, a smaller, briefer, and relatively accurate screening battery was needed for future studies in order to efficiently classify subjects based on their global neuropsychological status.

Many criteria for the performance of a screening test exist. These criteria can be single measures, such as misclassification rate, accuracy rate, or total cost, or they can be pairs of measures, such as sensitivity versus specificity, positive versus negative predictive value (PPV vs. NPV), and likelihood ratio of a positive versus negative test (LR+ vs. LR−). If Pr means probability, "|" refers to the conditional, D reflects the presence of disease, D' reflects the absence of disease, and pred D and pred D' represent the prediction of presence or absence of disease, respectively, then sensitivity $= Pr(\text{pred}D|D)$, specificity $= Pr(\text{pred}D'|D')$, PPV $= Pr(D|\text{pred}D)$, NPV $= Pr(D'|\text{pred}D')$, LR+ $= Pr(\text{pred}D|D)/[1{-}Pr(\text{pred}D'|D')]$, and LR− $= [1{-}Pr(\text{pred}D|D)]/Pr(\text{pred}D'|D')$. There is no uniform agreement regarding which measures to apply (See discussion in Gallagher, 1998).

In the presence of a gold standard, regardless of the measure selected for assessing performance, most reported methods are designed to evaluate only two tests at a time (as in Bloch, 1997; Viana and Pereira, 2000; Leisenring, Alonzo, and Pepe, 2000) or sort through multiple screening tests to generate a model which assigns weights to the tests or specifies a sequence of tests. These strategies, such as used by logistic regression, discriminant analysis, recursive partitioning, and others (See Kraemer, 1992, pp. 165-227), may eliminate a subset of tests, but these techniques have no provision for evaluating or ranking the individual performance of a large number of tests retrospectively and simultaneously.

One complication in identifying which of many diagnostic tests better predicts a disease or condition is that there are two types of errors related to diagnosis: predicting the existence of disease when it is absent (false positive) and predicting the absence of disease when it is present (false negative). Ideally, both types of errors are taken into account when determining how well a test performs, although the two types of errors may have different weights or costs.

When a single assessment measure is used, screening tests can be ranked on their performance from best to worse simply according to the ordering of the measurement values. For instance, the overall misclassification rate is the total number of false positives and false negatives, divided by the total number of subjects. The test with the smallest misclassification rate would be ranked highest, and so forth. Misclassification rate is one criterion of relative superiority of a test compared to others. Another single measure, total cost, can be formed when one type of error is considered worse or costlier than another by weighing each type of error by its relative influence and aggregating the combined costs into a single weighted measure. Based on ascending or descending values, ranks can be assigned as appropriate. For instance, if false positives (FP) are given twice as much weight as false negatives (FN), then total weighted misclassification cost can be formulated as $[2(FP) + FN]/N$ and ranked in order of increasing value.

When comparing two diagnostic tests based on a pair of measures, such as sensitivity and specificity, complications arise if one test has better performance on one measure but the other test performs better on the second measure (e.g., one test has higher sensitivity, but the other test has higher specificity). This results in a virtual stalemate for several published techniques offering screening test comparison schemes (Leisenring, Alonzo, and Pepe, 2000; Lee, 1999; Biggerstaff, 2000). Marshall (1989) discusses this dilemma but admits to the limited clinical utility of alternate approaches in practice.

When comparisons using a pair of performance criteria are limited to only two screening tests, one simply concludes that the tests are not comparable when one test performs better for one measure and the other has superior performance based on the other measure. When there are a large number of tests, though, the pattern of pairwise noncomparability, although complex, may yield useful information on relative performance.

Some published methods for pairwise comparisons claim the reported techniques can be applied similarly when there are more than two diagnostic tests to be judged (Leisenring, Alonzo, and Pepe, 2000; Biggerstaff, 2000). However, application of these comparisons to more than two tests is not straightforward. Even if the procedure is applied repeatedly and routinely for each pair of tests, no guidance is offered for evaluating performance after making the pairwise comparisons.

Section 2 reviews pairwise comparisons of diagnostic tests when the tests have binary outcomes, the criterion for test comparison is a pair of assessment measures, and actual knowledge of whether or not the disease or condition exists for each screened subject in a sample representative of the targeted population is known, that is, when a "gold standard" is present. Section 3 introduces a new *superiority index* and its properties. This measure quantifies the superiority of a diagnostic test compared to many other tests on the basis of simultaneously optimizing a pair of screening test assessment measures. Section 4 applies the superiority index technique to data from 19 neuropsychological tests to quantify and rank the tests with respect to their relative performance, and to assist in selection of tests entered into a classification model that would result in a substantially reduced test battery. Confidence intervals are estimated for the superiority index using a bootstrap procedure. Section 5 provides a discussion of the superiority index and related issues.

2. Pairwise Comparisons

Commonly used pairs of measures to assess the performance of a screening test include: (a) *sensitivity* and *specificity*, (b) *positive predictive value (PPV)* and *negative predictive value (NPV)*, and (c) *likelihood ratio of a positive test (LR+)* and *likelihood ratio of a negative test (LR−)*. An informative discussion of these measures and how to compute them can be found in Gallagher (1998). It should be noted that large values are desirable for the first five of the six measures listed above; small values are favorable for LR−. Sensitivity and specificity will be used to illustrate the procedures which follow. The procedures apply to any of the pairs of measures listed above, although the inequalities should be reversed for LR−. The proposed methods also apply to other pairs of performance assessment measures not listed above.

Each screening test can be categorized into one, and only one, of the following four representations. When comparing Test X to Test Y, and letting $Sens(X)$ and $Spec(X)$ denote the sensitivity and specificity, respectively, of Test X, we can say:

- Test X is superior to Test Y if $Sens(X) > Sens(Y)$ and $Spec(X) > Spec(Y)$,
- Test X is inferior to Test Y if $Sens(X) < Sens(Y)$ and $Spec(X) < Spec(Y)$,
- Test X is equal to Test Y if $Sens(X) = Sens(Y)$ and $Spec(X) = Spec(Y)$, and
- Test X and Test Y are not comparable if $Sens(X) > Sens(Y)$ and $Spec(X) < Spec(Y)$ or $Sens(X) < Sens(Y)$ and $Spec(X) > Spec(Y)$.

For each *superior* or *inferior* representation above, if exactly one of the inequalities is replaced by equality, the relationship still holds.

3. Superiority index

To assess the relative performance of a diagnostic test, we will draw on the definitions of *superior, inferior, equal*, and *not comparable* from Section 2 above. The main idea is essentially that a test which is pairwise superior to a relatively large number of other tests and pairwise inferior to relatively few other tests should have a high superiority value and be ranked higher than those tests that do not perform as well. Ideally, the larger the superiority value, the more accurately a screening test is expected to predict the targeted condition compared to other screening tests, based on relatively better simultaneous performance of both assessment measures.

This approach gives more weight to a diagnostic test doing comparatively well on both measures and less emphasis on tests doing relatively poorly on both measures or even doing extremely well on one measure but performing poorly on the other measure. In its formulation, the superiority index is designed to consider the joint performance of the two assessment measures.

3.1 Formulation of Superiority Index S_i

Let a_i = the number of tests to which Test i is superior, b_i = the number of tests to which Test i is inferior, and c_i = the number of tests equal to Test i. Only non-negative values are possible for a_i and b_i, and $c_i \geq 1$, since every test is equal to itself. For convenience, from this point forward, the subscripts will be dropped when their correspondence to Test i is obvious.

Then, the superiority index S for a test can be calculated as

$$S = (a + c/2)/(b + c/2) = (2a + c)/(2b + c) \tag{3.1}$$

This superiority index is the ratio of the number of tests to which Test i is superior versus inferior after splitting the number of equal tests evenly between superiority reflected in the numerator and inferiority in the denominator. Rank 1 is assigned to the test with the largest value of S_i, rank 2 is assigned to the test with the next largest value of S_i, etc.

The number of diagnostic tests that are not comparable to Test i do not enter into the calculation of S_i. However, each diagnostic test may have different sets of tests that are pairwise comparable. Thus, the superiority index for one screening test may be based on a different set of tests than for another screening test. The importance of this feature is that quantification of overall superiority uses information from tests that are not pairwise comparable to some tests so long as they are comparable to at least one of the other tests. This property maximizes information from all tests and distinguishes the superiority index from previously reported methods that can handle only pairwise comparisons.

3.2 Properties of S_i

Holding other values constant, three properties can be observed from Equation (3.1) for test i: (a) as $a \to \infty, S_i \to \infty$; (b) as $b \to \infty, S_i \to 0$; and (c) as $c \to \infty, S_i \to 1$. These properties are all desireable. The first property represents that the more tests that exist to which Test i is superior, the higher the value of superiority, and this value is without bound. The second property shows that the more tests that exist to which Test i is inferior, the lower the value of superiority, and the index approaches its lower bound of zero in the limit. The third property indicates that the more equal tests there are, the superiority index approaches closer to one. An exact value of one reflects that a test has as many tests superior to it as inferior to it, regardless of the extent that it is equal to or not comparable to some other tests. When the superiority:inferiority ratio $(a : b)$ is greater than 1.0, S_i decreases as the number of equal tests (c) increases, and conversely, when the superiority:inferiority ratio is less than 1.0, S_i increases as c increases. This is sensible, since when given a screening test that is superior to a relatively large number of other tests compared to the number of tests to which it is inferior, additional tests that are equivalent to it will lessen the impact of the superiority. Similarly, a greater number of equivalent tests improves the overall standing of a test considered to be inferior to many other tests.

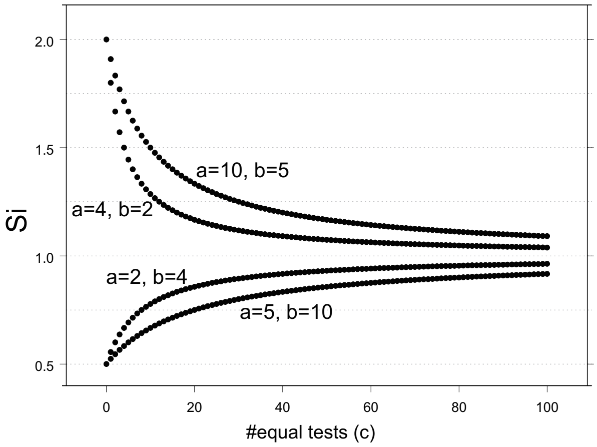

Figure 1: Changes in superiority index S_i as c varies. (a = the number of tests to which Test i is superior; b = the number of tests to which Test i is inferior; c = the number of tests equal to Test i)

Figure 1 graphically shows examples of the behavior of S_i for Test i. There are four different scenarios displayed by the four curves in Figure 1. The top two curves reflect a superiority:inferiority ratio of two, or that the given test is superior to twice as many other tests as to which it is inferior. The upper of these two curves is based on the larger number of tests that have comparable pairwise comparisons $(a = 10, b = 5$ vs. $a = 4, b = 2)$ and shows higher values of S_i than the lower of these two curves. Both curves are above 1.0, since Test i compares favorably more than unfavorably. As c increases, both curves drop closer to one from above. The bottom two curves reflect a superiority:inferiority ratio of 1/2 $(a = 2, b = 4$ vs. $a = 5, b = 10)$ and the interpretation is analogous to, but the converse of, the situation with a ratio of two. Although plotted in Figure 1, in practice, c will never equal zero, since each diagnostic test is equal to itself.

The behavior of the superiority index demonstrates that for each screening test, when the number of equal tests and the ratio of superiority:inferiority are the same, the superiority index will be further from 1.0 when more tests are comparable. This property reflects a stronger degree of superiority or inferiority when a greater number of diagnostic tests contribute to the estimate, perhaps somewhat analogous to "higher power with larger sample size" in a statistical hypothesis-testing environment.

4. Example

4.1 Data description

A large battery of neuropsychological (NP) tests was administered to participants in a longitudinal study of neurocognitive functioning in HIV-infected individuals at the University of California, San Diego, HIV Neurobehavioral Research Center. The battery consisted of nineteen separate test measures labeled as Tests #1-19 for purposes of this analysis and are identified in Appendix A. The test battery, which takes two to three hours to administer, assesses the following seven ability domains: abstraction/executive functioning, attention/working memory, learning, memory (delayed recall), motor skills, speed of information processing, and verbal skills (Woods *et al.*, 2004). Demographically corrected norms are used to convert raw scores to age-, education-, gender-, and, where possible, race/ethnicity-corrected standard scores (T-scores) using published procedures based upon large normative datasets (Benedict *et al.*, 1998; Heaton *et al.*, 1995).

Results for each individual's tests were classified as neuropsychologically impaired if the test score was more than one standard deviation below the normative mean for that test. Otherwise, an unimpaired classification was designated. An overall global rating of NP impairment vs. not impaired was assigned for each subject based on several criteria. Clinical ratings were performed by trained neuropsychologists utilizing standardized procedures which yield a global rating, as well as domain ratings, for each of the seven NP domains cited above (Woods *et al.*, 2004). This global rating was treated as the gold standard against which individual test predictions were compared. Details of the NP testing, scoring, and classifications have been published elsewhere (See Heaton *et al.*, 1994; Ellis *et al.*, 1997).

The rationale for reducing the nineteen NP tests to a smaller subset was to use a limited number of tests to screen a cohort while retaining predictive accuracy and requiring a much shorter duration of time than was necessary for the full battery, currently approximately two to three hours.

Among the 393 subjects who completed each of the 19 NP tests, 159 (40.5%) were diagnosed as NP impaired based on the gold standard. Table 1 reflects the outcomes for the full dataset on sensitivity, specificity, PPV, NPV, LR+, and LR− for each NP measure. The percentage of subjects who were classified incorrectly, the misclassification rate, is also reported. Overall, there appears to be no clear winner, or best subset of winners, among the 19 diagnostic tests according to these measures. For example, the test with the best sensitivity, NPV, and LR−, Test #8, also has the lowest specificity, PPV, and LR+.

Table 1: Performance measures for 19 neuropsychological tests on the full dataset (N=393). (Test = neuropsychological test; Sens = sensitivity; Spec = specificity; PPV = positive predictive value; NPV = negative predictive value; LR+ = likelihood ratio for a positive test; LR− = likelihood ratio for a negative test; Misclass = misclassification rate)

Test	Sens	Spec	PPV	NPV	LR+	LR−	Misclass
1	0.65	0.87	0.78	0.79	5.10	0.40	0.22
2	0.43	0.84	0.65	0.68	2.70	0.68	0.33
3	0.45	0.80	0.61	0.68	2.30	0.68	0.34
4	0.57	0.78	0.64	0.73	2.58	0.55	0.31
5	0.41	0.93	0.80	0.70	5.98	0.63	0.28
6	0.49	0.87	0.72	0.72	3.83	0.58	0.28
7	0.21	0.89	0.58	0.63	2.00	0.88	0.38
8	0.80	0.58	0.56	0.81	1.89	0.35	0.33
9	0.16	0.97	0.78	0.63	5.26	0.87	0.36
10	0.22	0.95	0.74	0.64	4.29	0.82	0.35
11	0.28	0.94	0.76	0.66	4.73	0.76	0.33
12	0.35	0.93	0.78	0.68	5.15	0.70	0.30
13	0.21	0.95	0.75	0.64	4.42	0.83	0.35
14	0.32	0.97	0.88	0.68	10.72	0.70	0.29
15	0.28	0.91	0.69	0.65	3.24	0.79	0.34
16	0.31	0.91	0.69	0.66	3.34	0.76	0.33
17	0.40	0.91	0.75	0.69	4.42	0.66	0.30
18	0.60	0.88	0.78	0.77	5.23	0.45	0.23
19	0.48	0.91	0.78	0.72	5.33	0.57	0.26

4.2 The superiority index and variability

For the purpose of building a classification model in the next subsection, the full sample was randomly split in half (Sample #1: $N = 197$, Sample #2: $N = 196$). Superiority indices were computed for Sample #1. Diagnostic test performance assessment was based on likelihood ratios of a positive and negative test, since these measures do not depend upon the neurocognitive impairment prevalence rate and were favored for the application (Gallagher, 1998). Table 2 displays superiority indices and their bootstrap bias-corrected and accelerated 95% confidence interval estimates (Efron and Tibshirani, 1993, pp. 184-188) for the Sample #1 data using the methods described in Section 3. The confidence intervals (CIs) reflect the degree of variability associated with S_i for each NP test. In identifying tests which are likely to be relatively superior, we are reasonably confident that those tests with 95% CIs having both endpoints above the value of 1.0 (NP Tests #18, 5, 14, and 16) tend to be the better performing tests, and those CIs with both endpoints below 1.0 (NP Tests #4, 7, and 9) are inferior. The others (NP Tests #6, 1, 19, 12, 17, 11, 8, 13, 2, 3, 15, and 10) all contain the value 1.0 within the interval. Thus, we cannot reject with reasonable confidence the possibility that they are both superior to and inferior to about the same number of tests.

Table 2: Sample #1 superiority indices in rank order and Bootstrap BCa 95% confidence interval estimates of mean S_i for 19 neuropsychological tests (NP = neuropsychological; S_i = superiority index for test i; BCa = bias-corrected and accelerated; CI = confidence interval).

NP Test #	Observed S_i	Bootstrap BCa 95% CI for mean S_i
18	27.0	5.7 - 35
5	21.0	2.1 - 33
14^a	13.0	1.6 - 29
16^a	13.0	1.3 - 29
6	5.0	0.5 - 27
1	3.7	0.1 - 17
19	2.6	0.2 - 23
12^b	2.2	0.1 - 23
17^b	2.2	0.2 - 25
11	1.6	0.1 - 19
8	1.0	–
13	0.6	0.03 - 13
2^c	0.3	0.03 - 3.4
3^c	0.3	0.03 - 1.8
15	0.3	0.03 - 1.3
10	0.4	0.03 - 1.0
4	0.1	0.03 - 0.33
7	0.04	0.03 - 0.24
9	0.03	0.03 - 0.33

NP test numbers with the same superscript letter (^{a,b,c}) have the same (tied) observed S_i.

4.3 Classification Models

NP Test #18 ranked highest with $S_i = 27.0$ (95% CI: 5.7, 35.0). Since a combination of screening tests may interact or provide complementary information to further reduce error, a model-building strategy was developed. As with most model development, the final model may be appropriate, useful, and effective, but it may not necessarily be unique.

Ideally, when numerous candidates are considered for model-building, two driving influences should inform selection of candidate prognostic variables: (a) avoidance of the inclusion of redundant (i.e., highly "correlated" or collinear) variables and (b) elimination of variables having little predictive value. In a classical regression setting, when a large number of candidate prognostic variables are available, a commonly used approach is to screen out variables to be included in a model by computing univariable estimates and considering only variables with a *p-value* less than, say, 0.05, 0.10, or 0.20, etc., with the criterion depending on how conservative one wishes to be. Sets of these variables containing redundant information can be pared down by eliminating all but the one with the smallest *p-value* or largest test statistic (See Schwimmer *et al.*, 2003). A method analogous to this approach for combining diagnostic tests can be used. Tests having large superiority values and thus more likely to be superior should be selected, and tests having a small superiority index value, say less than one, can be eliminated.

To examine and assess the utility of the superiority index as a screening tool for model variable-selection, Sample #1 was used as the model-training data and Sample #2 was reserved for validation of the generated model. Four different scenarios were used to determine which diagnostic tests were to be candidates for inclusion in building separate models. The first scenario included diagnostic tests having the highest superiority index values. According to Table 2, there is a clear break between S_i values for the two highest ranked NP tests, Tests #5 and #18, versus the other tests. They have superiority values of 21 and 27, respectively, in contrast to the next highest ranked tests which drop to an S_i value of 13; thus Tests #5 and 18 are indicated. The second scenario included the seven highest ranked tests. They are, in order of their ranking, NP Tests #18, 5, 14, 16 (tied with 14), 6, 1, and 19. They have superiority index values ranging from 2.6 through 27.0. The third modeling scenario dropped all of the clearly inferior NP tests as evidenced by a superiority index less than 1.0; thus, the first seven tests included in the second scenario are included, plus NP Tests #12, 17 (tied with 12), 11, and 8. The fourth modeling scenario used all 19 NP tests.

We may also assess if some tests contain highly redundant information. The Kappa statistic was used as the criteria for redundancy. The maximum Kappa value computed between every pair of NP tests was 0.46, which can be considered as only moderate agreement according to the standard of Landis and Koch (1977). Other methods are available to check for redundancy, such as the Phi coefficient, Lambda statistics, or similarity measures.

To identify the final model for classification, CART (Classification and Regression Tree) (Breiman *et al.*, 1984, pp 11-13), a recursive partitioning technique for classification, was applied using Sample #1 as training data. For each of the four scenarios, a full CART tree was grown, then the branches were pruned until the model reached a minimum misclassification error rate when applied to the validation sample. Validation is essential because of the risk of overfitting the model to the data in the training sample rather than reflecting the underlying population from which the sample is obtained.

For each of the four models based on different scenarios, the final model using CART produced the same decision rule: classify a subject as neuropsychologically normal if the subject is rated as normal on both NP Tests #5 and #18, and classify the subject as NP impaired if impaired on either of the two tests. Of note, these two NP tests were the two highest ranked tests among the nineteen with respect to the superiority index values.

Diagnostic test assessment measures for the final model for both the training and validation datasets are presented in Table 3. Most of the measures reported in Table 3 are substantially improved for both the training and validation data compared to the values for each of the two tests #5 and #18 alone.

Table 3: Final CART model results for all four variable-selection scenarios. (NP = neuropsychological, Training = Sample #1, Validation = Sample #2, Sens = sensitivity; Spec = specificity; PPV = positive predictive value; NPV = negative predictive value; LR+ = likelihood ratio for a positive test; LR− = likelihood ratio for a negative test; Misclass = misclassification rate)

NP Tests	Dataset	Sens	Spec	PPV	NPV	LR+	LR−	Misclass
#5, 18	Training	80%	86%	81%	85%	5.6	0.23	17%
	Validation	80%	80%	71%	87%	4.05	0.25	20%

To further explore what impact superiority may have on modeling and pruning, additional models were developed using the second, third, and fourth scenarios for NP test selection described above. However, this time, CART decision trees were arbitarily forced to have exactly five NP tests included in the final model. The results on several diagnostic test performance measures for these three models were compared and appear in Table 4. Note that the third and fourth scenarios produced the same final model.

Table 4: Comparison of CART results for three test-selection scenarios and forcing exactly five tests into the final model. (NP = neuropsychological, Training = Sample #1, Validation = Sample #2)

NP Test Selection Scenario	$S_i \geq 1$*; All 19 tests			7 highest ranking tests		
NP Tests in Final Model**	#18, 5, 11, 17, 6			#18, 5, 6, 1, 14		
Measure	Training	Validation	Δ***	Training	Validation	Δ***
Sensitivity	84%	69%	15%	81%	70%	11%
Specificity	88%	81%	7%	88%	84%	4%
Positive Predictive Value	84%	69%	15%	84%	72%	12%
Negative Predictive Value	88%	81%	7%	86%	82%	4%
Likelihood ratio of a + test	6.7	3.7	3.0	7.0	4.3	2.7
Likelihood ratio of a − test	0.19	0.38	0.19	0.21	0.36	0.15
Misclassification Rate	14%	23%	9%	15%	21%	6%

*S_i = superiority index, Tests in rank order: #18, 5, 14, 16, 6, 1, 19, 12, 17, 11, 8

**Tests are listed in order of appearance in each model.

***Δ =difference between results for the Training and Validation datasets, all in the expected direction.

According to Table 4 and as expected, performance on every measure is worse for the validation sample than for the training sample for all three variable-selection scenarios. The column headed by delta (Δ) shows the magnitude of worsening for each method. The key findings here are that the amount of degradation is less, and the absolute performance with the validation sample on all presented diagnostic test assessment measures is better, when only the seven highest ranking tests are presented as candidates for inclusion into the CART models, compared to when all screening tests were used, or if only tests with superiority index less than one were eliminated. This may reflect the way we apply CART here, as it may tend to overfit the data. Screening out neuropsychological tests which do not perform well relative to the other tests in the full battery appears to allow development of a better model than if more inferior tests are considered. It is expected that a model using fewer, but superior, classifiers, results in better prediction than one using more, but less informative, classifiers.

Variations on the scenarios described above, such as forcing exactly three, four, six, or more tests into the model, were applied. In every case, restricting candidate diagnostic tests to the higher values of S_i resulted in as good or better performance than no restriction (results not shown).

5. Discussion

In many fields, it is common to seek the binary-valued screening or diagnostic test which performs best among many. With our proposed methods and based simultaneously on a pair of performance measures such as sensitivity and specificity, not only is it possible to identify relative standing among a collection of diagnostic tests, but the magnitude of relative superiority or inferiority of a test can be estimated. Variability for this measure, reflected by 95% CIs, can be estimated using bootstrap.

The proposed quantification of superiority provides a computationally simple method for reducing a sometimes intractably large number of binary-valued diagnostic evaluation measures to a smaller number of more effective predictors of a disease or condition. The superiority index for a diagnostic test can be interpreted alone in a simple and meaningful way, since it is the ratio of the number of tests that it surpasses in performance to the number to which it is inferior. Pairwise noncomparability reduces the impact of the superiority strength or inferiority strength in a balanced way, and it does not cause the procedure to result in a stalemate beyond pairwise comparisons as previously reported methods do. The index reflects a relative value but does not guarantee identification of tests which perform well. The proposed methods are an option for modeling and for eliminating relatively poor screening tests from consideration. It may serve a purpose similar to pruning in CART and can be added to the collection of existing methods which are used to assign subjects to a high vs. low risk category.

The proposed diagnostic test performance assessment methods assume that various conditions exist: (a) multiple diagnostic tests with dichotomized outcomes, (b) the presence of a gold standard, and (c) desire to optimize a pair of test assessment measures. Estimates of individual screening test performance measures are easily calculated from simple script programming, commercial software, or free web-based calculators. The techniques are appropriate when any pair of measures need to be simultaneously optimized, so the criteria for assessment can be customized to fit the situation at hand. The proposed index need only be applied to tests having equal costs for the two types of misclassification. When there are different costs, the two types of errors can be combined into an overall single weighted measure of total misclassification rate or cost and evaluated and ranked accordingly. The proposed method accommodates missing data and small samples, since each individual test measure, such as sensitivity, is measured as a proportion. Nevertheless, caution should be applied when evaluating tests with a limited number of observations.

In the NP example, a two to three hour test battery of 19 tests was reduced to two tests taking only approximately ten minutes to administer, so the impact of the superiority technique was substantial. When CART was used for modeling with a specified number of tests forced into the final model, all reported validation sample performance measures were the same or better when the poorest performing tests were initially excluded based on low S_i values as compared to when CART had all tests available. The final CART model did not depend on whether or not diagnostic tests were screened out based on superiority. This seems to confirm that screening out tests based on lower Si values plays a similar role as pruning does in CART. Overfitting, a common modeling problem, is likely reduced by these techniques.

A potential concern for the superiority index is that the proportion of other screening tests from which it differs is accounted for, not the total number of tests. Although the number of equal tests in the formulation attenuates this case, this probably deserves more attention in the future. Another conceivable issue is that each pairwise comparison identifies the better test but ignores the magnitude of superiority. For example, it makes no difference if Test X has much higher or only slightly higher sensitivity and specificity than Test Y. Either way, a is incremented by one. Nevertheless, other tests will have a rougher time beating out Test X if Test X has a very high superiority value and its relative ranking will be sequenced appropriately after taking into account all pairwise comparisons.

In considering the extension of the superiority ranking to more than two performance measures, such as sensitivity, specificity, LR+ and LR−, one option would be to expand the comparisons specified in Section 2. For example, Test X is *superior* to Test Y if $Sens(X) > Sens(Y), Spec(X) > Spec(Y), LR+(X) > LR+(Y)$, and $LR-(X) < LR-(Y)$, and so on. However, the more restrictions made to indicate a test as being superior, the fewer such designations will likely be made, and the number of tests counted in computing the superiority index may be drastically reduced.

A notable strength of the superiority quantification and ranking system is that it can treat any binary covariate, including demographic or other characteristics, in the same way as it treats a diagnostic or screening test. Hence, an interesting application of the superiority index is its use in comparing competing classification models which classify subjects as high vs. low risk. Relative superiority of multiple classification models can be quantified and ranked using the proposed techniques, even if the models are not nested. Another potential utility of the superiority value is that if two screening tests are close in superiority, but one is much more expensive in either time, cost, or other resource, the more expensive test may be considered for elimination to conserve resources with the awareness that minimal superiority is lost. Thus, the proposed superiority index is adaptable in common situations and provides meaningful information with a straightforward interpretation.

Acknowledgements

This research was partially supported by the UCSD General Clinical Research Center (GCRC) grant NIH M01 RR000827 and the UCSD Biostatistics and Bioinformatics Laboratory. The authors thank the UCSD HIV Neurobehavioral Research Center (NIMH grant P50 MH45294) for providing the data and assistance associated with the example. The author is grateful to Deborah Hilton, M.S., formerly with the UCSD GCRC, for assistance in the literature search and review of a previous draft of the manuscript. An earlier version of this research was published in the *Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001* under the title "Subset selection for multiple screening tests by ranking in the presence of a gold standard."

Appendix A. Nineteen Neuropsychologic Evaluations Used in the Section 4 Example.

Test #	Test name
1	Brief Visuospatial Memory Test - Trials 1-3 Total
2	Brief Visuospatial Memory Test - Delayed Recall
3	Grooved Pegboard - Dominant
4	Grooved Pegboard - Nondominant
5	Paced Auditory Serial Addition Test
6	Story Memory Test - Learning Rate
7	Story Memory Test - Delayed Recall Trial
8	Figure Memory Test - Learning Rate
9	Figure Memory Test - Delayed Recall Trial
10	Trail Making Test - Part A
11	Trail Making Test - Part B
12	Digit Symbol subtest of the WAIS-3*
13	Letter-Number Sequencing subtest of the WAIS-3*
14	Symbol Search subtest of the WAIS-3*
15	Letter Fluency
16	Category Fluency (animals)
17	Booklet Category Test
18	Hopkins Verbal Learning Test - Trials 1-3 Total
19	Hopkins Verbal Learning Test - Delayed Recall

*Wechsler Adult Intelligence Test (3rd Edition)

References

Benedict, R. H., Schretlen, D., Groninger, L., Brandt, J. (1998). Hopkins Verbal Learning Test-Revised: Normative data and analysis of inter-form and test-retest reliability. *Clinical Neuropsychologist* **12**, 43-55.

Biggerstaff, G. J. (2000). Comparing diagnostic tests: a simple graphic using likelihood ratios. *Statistics in Medicine* **19**, 649-663.

Bloch, D. A. (1997). Comparing two diagnostic tests against the same “gold standard” in the same sample. *Biometrics* **53**, 73-85.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International Group.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

Ellis, R. J., Deutsch, R., Heaton, R. K., Marcotte, T. D., McCutchan, J. A., Nelson, J. A., Abramson, I., Thal, L. J., Atkinson, J. H., Wallace, M. R., Grant, I., and the San Diego HIV Neurobehavioral Research Center Group (1997). Neurocognitive impairment is an independent risk factor for death in HIV infection. *Archives of Neurology* **54**, 416-424.

Gallagher, E. J. (1998). Clinical utility of likelihood ratios. *Annals of Emergency Medicine* **31**, 391-397.

Heaton, R. K., Grant, I., Butters, N., White, D. A., Kirson, D., Atkinson, J. H., McCutchan, J. A., Taylor, M. J., Kelly, M. D., Ellis, R. J., Wolfson, T., Velin, R., Marcotte, T. D., Hesselink, J. R., Jernigan, T. L., Changler, J., Wallace, M., Abramson, I., the HNRC Group (1995). The HNRC 500—Neuropsychology of HIV infection at different disease stages. *Journal of the International Neuropsychological Society* **1**, 231-251.

Heaton, R. K., Kirson, D., Velin, R. A., Grant, I., the HNRC Group (1994). The utility of clinical ratings for detecting early cognitive change in HIV infection. In *Neuropsychology of HIV infection* (Edited by I. Grant and A. Martin). Oxford University Press.

Kraemer, H. C. (1992). *Evaluating Medical Tests*. Sage Publications.

Landis, R. J., Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**, 159-174.

Lee, W.-C. (1999). Selecting diagnostic tests for ruling out or ruling in disease: the use of the Kullback-Leibler distance. *International Journal of Epidemiology* **28**, 521-525.

Leisenring, W., Alonzo, T., and Pepe, M. S. (2000). Comparisons of predictive values of binary medical diagnostic tests for paired designs. *Biometrics* **56**, 345-351.

Marshall, R. J. (1989). The predictive value of simple rules for combining two diagnostic tests. *Biometrics* **45**, 1213-1222.

Schwimmer, J. B., Deutsch, R., Rauch, J. B., Behling, C., Newbury, R., Lavine, J. E. (2003). Obesity, insulin resistance, and other clinicopathological correlates of pediatric nonalcoholic fatty liver disease. *The Journal of Pediatrics* **143**, 500-505.

Viana, M. A. G., Pereira, C. A. De B. (2000). Statistical assessment of jointly observed screening tests. *Biometrical Journal* **42**, 855-864.

Woods, S. P., Rippeth, J. D., Frol, A. B., Levy, J. K., Ryan, E., Soukup, V. M., Hinkin, C. H., Lazzaretto, D., Cherner, M., Marcotte, T. D., Gelman, B. B., Morgello, S., Singer, E. J., Grant, I., Heaton, R. K. (2004). Interrater reliability of clinical ratings and neurocognitive diagnoses in HIV. *Journal of Clinical and Experimental Neuropsychology* **26**, 759-778.

Received June 18, 2007; accepted November 30, 2007.

Reena Deutsch
University of California, San Diego
9500 Gilman Drive
La Jolla, California 92093-0847 USA
rdeutsch@ucsd.edu

Monica Rivera Mindt
Fordham University & Mount Sinai School of Medicine
113 West 60th Street, LL 808-F
New York, NY 10023 USA
riveramindt@fordham.edu

Ronghui Xu
University of California, San Diego
9500 Gilman Dr, MC 0112
La Jolla, CA 92093-0112 USA
rxu@ucsd.edu

Mariana Cherner
University of California, San Diego
150 W. Washington St., 2nd floor
San Diego, CA 92103 USA
mcherner@ucsd.edu

Igor Grant
University of California, San Diego
150 W. Washington St., 2nd Floor
San Diego, CA 92103 USA
igrant@ucsd.edu