



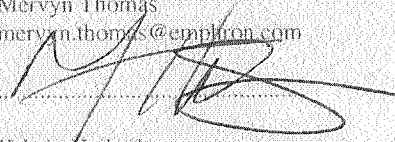
Analysis of Performance of a Novel Diagnostic for Ovarian Cancer

Version 4.0 (Final)

Emphron Informatics Pty Ltd

www.emphron.com

Report Number: 2008.001
Client: Healthlinx
Date: December 23, 2008
Author: Mervyn Thomas
Email: mervyn.thomas@emphron.com

Signature: 

Reviewer: Kristin Hatherley

Signature: 

Contents

1	Executive Summary	2
2	Introduction	2
3	Data	3
4	Methods	3
4.1	Comparisons Performed	3
4.2	Measuring and Comparing Diagnostic Performance	4
5	Diagnosis of Ovarian Cancer	6
6	Diagnosis of Early-Stage Ovarian Cancer	9
7	Power to Detect Differences in Diagnostic Performance	11
8	Summary and Conclusions	13

List of Tables

1	Composition of Training and Testing Data Sets	3
2	Illustrative Diagnostic Performance of Healthlinx and Reference. . .	6
3	Sensitivity and Specificity for Healthlinx Diagnostic	7
4	Reference Diagnostic prediction Success	8
5	Healthlinx Diagnostic prediction Success Threshold of 0.3	8
6	Healthlinx Diagnostic prediction Success Threshold of 0.5	9
7	Sensitivity and Specificity for Healthlinx Diagnostic - Early Stage Ovarian Cancer	9
8	Reference Diagnostic prediction Success	11
9	Healthlinx Diagnostic prediction Success Threshold of 0.3	11
10	Healthlinx Diagnostic prediction Success Threshold of 0.5	11

11 Healthlinx Power Calculations for Diagnostic Efficiency. Scenario 1 represents independent outcomes for reference diagnostic and Healthlinx diagnostic. Scenario 2 represents perfect correlation between tests. 12

List of Figures

1 **Bootstrap Test for Equality of AUCs - All Testing Cases** . A positive value implies superiority of the Healthlinx diagnostic. 8

2 **Bootstrap Test for Equality of AUCs - Early Testing Cases** . A positive value implies superiority of the Healthlinx diagnostic. 10

1 Executive Summary

1. This report describes an analysis of diagnostic efficacy of the Healthlinx ovarian cancer diagnostic and the current reference diagnostic;
2. The Healthlinx diagnostic clearly out-performs the reference diagnostic, with statistically significantly higher ROC AUCs;
3. The Healthlinx diagnostic is demonstrably able to detect early stage ovarian cancer.

2 Introduction

Healthlinx are developing a multi-marker technology for the detection of ovarian cancer. Their technology is based on a suite of markers, and a diagnostic rule derived using machine learning approaches. The precise nature of the machine learning algorithm is not described in this report. There is a current reference diagnostic for the detection of ovarian cancer. The Healthlinx markers are used to augment this reference marker, not to replace it. That is, the Healthlinx diagnostic rule is based on the current reference marker, plus the additional Healthlinx markers.

The objective of this report is to compare the performance of the Healthlinx diagnostic rules with the current reference diagnostic. In particular it seeks to identify whether or not the Healthlinx diagnostic panel is statistically significantly better than the reference diagnostic.

This report represents a minor modification to an existing Emphron report, following re-implementation of the machine learning algorithm. The new implementation is numerically more stable and more efficient than the previous implementation - although the fundamental structure of the algorithm remains unchanged.

The primary objective of this study was to test the hypothesis that the area under the receiver operator characteristic curve (ROC) for a multi-marker panel was significantly greater than that observed for CA125 alone.

3 Data

362 patient records were recruited randomly from a pathology practice and from the Peter MacCallum Cancer Institute Biobank. Of those 362 patients 150 had confirmed diagnoses of ovarian cancer; 212 were free of the disease. The data set was randomly partitioned into a training set and a validation set. There were 179 records in the training set and 183 records in the validation set. The composition of the data set is shown in Table 1.

Data Set	Disease	No Disease	Total
Training	82	97	179
Testing	68	115	183
Total	150	212	362

Table 1: Composition of Training and Testing Data Sets

The training data were used to derive the diagnostic rules which are the subject of this report. The validation data were used to measure the efficacy of those rules.

Each data set contained analyte information for each sample, the true sample classification (disease or no disease), and where available, the disease stage (cancer disease stage I-IV or normal). The analytes in each data set were identified as: HTX001, HTX003, HTX004, HTX007, and HTX009. Marker HTX001 is the current reference diagnostic for ovarian cancer (CA125). The other markers have not been identified beyond the codes used in this report.

4 Methods

4.1 Comparisons Performed

The performance of the Healthlinx diagnostic as compared to the existing standard was evaluated for the different scenarios:

- the diagnosis of ovarian cancer regardless of the stage of the disease;
- the diagnosis of early stage ovarian cancer only (i.e. Stage I & II);

Each of these diagnostic scenarios was analysed by fitting the model on the training data set and predicting the validation data set. The early stage ovarian cancer performance was tested using the model fitted to the entire training set - not merely those members of the training set who did not have late stage disease.

Stage data were not available for four cases in the validation set. These cases were not used in the analysis of performance of early stage disease, that is, they were treated as having late stage disease.

4.2 Measuring and Comparing Diagnostic Performance

Each classifier was developed on a training set, and tested on a validation set. Since the data sets were randomly partitioned into training and validation sets, the validation set is independent of the training set. In many machine learning applications the training set is much larger than the validation set. In this analysis, however, we use an approximately 50/50 split. This is because for our purposes, quantification of diagnostic performance is at least as important as optimal training of the classifier. This mandates a larger validation set - especially since we expect classification to be excellent.

Two approaches were used to measure and compare diagnostic performance:

ROC curves Diagnostic performance was summarised through the calculation of ROC curves, and by calculation of the area under the ROC curve [7]. The ROC curve plots the true positive rate against the true negative rate. The ROC curve for an essentially useless diagnostic (i.e. random decision making) is a straight line through the origin, with slope 1. If the ROC curve lies above this line then diagnostic performance is better than random. If the ROC curve lies below the line, then the diagnostic is 'worse than useless'. The use of ROC curves is extremely well established in the measurement of diagnostic performance [1, 8, 9, 10]. The advantages of using ROC curves over sensitivity and specificity are:

- ROC curves make more efficient use of the data. Comparisons between ROC curves are generally statistically more powerful than comparisons between sensitivity and specificity. This is especially the case when the performance of both diagnostics being compared is very high (sensitivity and specificity > 90%).
- The ROC area is less sensitive to the mix of true positives and negatives in the training data than are sensitivity and specificity. Many algorithms which do not explicitly consider prior probabilities will give varying results for sensitivity and specificity as the case-mix changes. Essentially, as the number of true negatives increases so specificity is increased at the expense of sensitivity. Conversely, as the number of true positives increases so sensitivity is increased at the expense of specificity. This behaviour may be observed in the results presented in this report. The ROC area, however, is not affected.
- In early stage diagnostic development, there is still some flexibility about setting cut-off values for negatives and positives. These cut-offs are generally settled later in the development process. Sensitivity and specificity are heavily influenced by these cut-offs which in early stage work are essentially arbitrary. ROC areas are invariant with respect to any particular cut off.

Against these advantages it must be remembered that for clinical use of the diagnostic a defined 'cut point' is required - leading to a defined sensitivity and specificity.

The ROC curve for the reference diagnostic was based simply on the marker concentration. The ROC curve for the Healthlinx diagnostic was based on the predicted posterior probability of membership of the Disease group. The area under the ROC curve (AUC) was calculated using the Wilcoxon statistic [7].

Again, the AUC for the reference diagnostic and for the Healthlinx diagnostic are not statistically independent, since they are based on the same patients. The difference in AUC between the diagnostics was tested using a bootstrap procedure [3]. The bootstrap is a general computationally-intensive tool for evaluating the statistical accuracy of an estimator, such as the mean, or in this case the area under the ROC curve. A bootstrap analysis is done by repeatedly drawing samples with replacement N times for some large N , evaluating the estimator for each sample, and then evaluating the accuracy of the estimator by looking at the distribution of the results obtained over the N replications. In the case of this report the number of bootstrap samples used was $N = 100,000$, the estimators considered were the area under the ROC curve as well as the difference between the AUCs, and the measures of accuracy were the 95% confidence intervals.

The bootstrap is an extremely well established statistical procedure. It is backed by very substantial theoretical developments [5, 6] as well as many practical applications [2]. The main advantages of using non-parametric resampling techniques such as bootstrap for the analysis of ROC data are simplicity of implementation, and the fact that minimal assumptions are made on the true distribution of the data. The bootstrap has been found to be particularly helpful when determining ROC confidence intervals from small data samples [10].

In detail, the procedure adopted was:

1. Sample the cases with replacement;
2. Generate the AUC for the reference diagnostic in the sample, using the HTX001 marker concentrations;
3. Generate the AUC for the Healthlinx diagnostic in the sample, using posterior probability of disease;
4. Generate the difference in AUCs (Healthlinx minus reference);
5. Accumulate quantiles and descriptive statistics;
6. Repeat, drawing the next bootstrap sample.
7. 95% confidence intervals for the Healthlinx diagnostic AUC, the reference AUC and the difference between Healthlinx and reference AUC are calculated. Bootstrap confidence intervals were calculated using the bias corrected and adjusted bootstrap (BC_α procedure) [4, Chapter 14]
8. A significance test for the difference in AUC between the Healthlinx diagnostic and the reference diagnostic was conducted by inverting the BC_α confidence intervals to obtain the tail area probability as α , where the $1 - \alpha$ confidence interval *just* includes 0.

Sensitivity and Specificity Sensitivity is the probability of correctly diagnosing a patient who actually has the disease. Specificity is the probability of correctly diagnosing a patient who does not have the disease. Sensitivity is estimated by the proportion of true positive patients correctly classified in the validation set. Specificity is measured by the proportion of true negative patients correctly classified in the validation set.

Estimates of the sensitivity and specificity of the Healthlinx diagnostic and the reference diagnostic are not statistically independent - they are based on the same subject. Comparison of the sensitivity and specificity must take account of this dependency. We adopt a method based on a contingency table. The success of the Healthlinx diagnostic was tabulated against the success of the reference diagnostic, as illustrated by Table 2. In that table, we define:

- a** cases correctly classified by both diagnostics;
- b** cases misclassified by both;
- c** cases correctly classified by the reference diagnostic but incorrectly classified by the Healthlinx diagnostic;
- d** cases incorrectly classified by the reference diagnostic but correctly classified by the Healthlinx diagnostic;

The comparison of the diagnostic techniques focuses on those observations which are correctly classified by one of the techniques, and misclassified by the other. It raises the question "where the diagnosis differs, are the methods equally likely to be correct?". The probability that it is the Healthlinx diagnostic which is correct for such situations is estimated by $\hat{p}_H = \frac{d}{c+d}$. The test for superiority of the Healthlinx diagnostic is then a test of the alternative hypothesis $p_H > 0.5$ against the alternative hypothesis $p_H \leq 0.5$. This is a test for marginal homogeneity in the contingency table.

	Healthlinx	
Reference	Correct	Incorrect
Correct	<i>a</i>	<i>c</i>
Incorrect	<i>d</i>	<i>b</i>

Table 2: Illustrative Diagnostic Performance of Healthlinx and Reference.

This study was not powered to detect differences in sensitivity, specificity or diagnostic efficiency using this approach. Very large sample sizes would be required, since only cases which have a different classification under the two diagnostics are informative.

5 Diagnosis of Ovarian Cancer

All tabulations presented in this section are based on the validation data set alone. That is, the classifier has been developed using the training data set, and its performance tabulated for the validation data set.

Inspection of the ROC curve for the Healthlinx diagnostic reveals that at a specificity similar to that of the reference diagnostic (approximately 90%), the diagnostic achieves a sensitivity of 98%: a greater than 5% increase over that of the reference diagnostic. Similarly, at a sensitivity equal to that of the reference diagnostic (92.6%) the Healthlinx diagnostic achieved a specificity of 94% (approximately 4.5% greater than that of the reference diagnostic).

Healthlinx Diagnostic Performance

Threshold	sensitivity	specificity
0.15	0.98	0.90
0.15	0.97	0.90
0.18	0.96	0.90
0.29	0.95	0.91
0.45	0.94	0.94
0.50	0.93	0.94
0.61	0.92	0.94
0.77	0.91	0.95
0.77	0.90	0.95

Table 3: Sensitivity and Specificity for Healthlinx Diagnostic

The area under the ROC curve was 0.960 (bootstrap confidence interval 0.913 to 0.984) for the reference diagnostic and 0.988 (bootstrap confidence interval 0.975 to 0.995) for the Healthlinx diagnostic. The bootstrap test for equality of the two AUCs was statistically significant at the 1% level¹.

Figure 1 shows the bootstrap distribution of the difference between the Healthlinx AUC and the standard diagnostic AUC. A positive value represents superiority of the Healthlinx diagnostic. The probability mass of the distribution is overwhelming for positive values of the statistic.

¹The difference is statistically significant, despite the overlapping marginal confidence intervals, because of the strong correlation between reference and Healthlinx AUCs in bootstrap realisations.

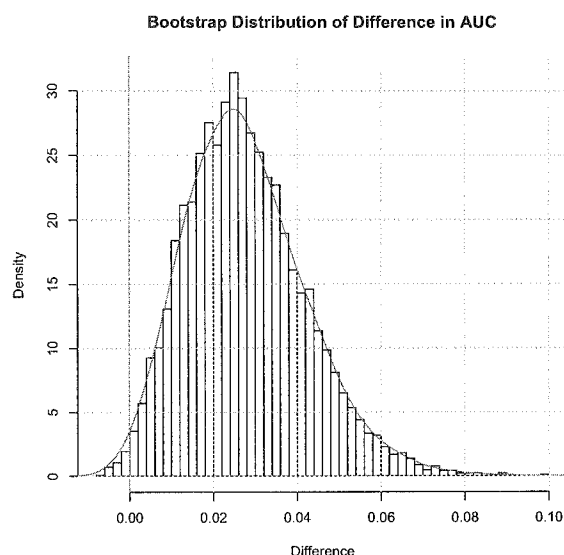


Figure 1: **Bootstrap Test for Equality of AUCs - All Testing Cases** . A positive value implies superiority of the Healthlinx diagnostic.

The sensitivity and specificity of the reference diagnostic for this data set is shown in Table 4. The sensitivity and specificity of the Healthlinx diagnostic is shown in Table 5, when the cut off for a positive diagnosis is 0.3. Similar results for a cut off of 0.5 are shown in Table 6. The lower cut off, of course, increases sensitivity at the expense of reducing specificity. A broader indication of the relationship between specificity and sensitivity is shown in Table 3. This Table shows the specificity of the Healthlinx diagnostic for a range of required sensitivities. The reference diagnostic achieved a sensitivity of 92.6% for a specificity of 89.6%.

True State	Predicted State		Success Rate
	Disease	No Disease	
Disease	63	5	Sensitivity = 92.6%
No Disease	12	103	Specificity = 89.6%

Table 4: Reference Diagnostic prediction Success

True State	Predicted State		Success Rate
	Disease	No Disease	
Disease	64	4	Sensitivity = 94.1%
No Disease	10	105	Specificity = 91.3%

Table 5: Healthlinx Diagnostic prediction Success Threshold of 0.3

Healthlinx Diagnostic Performance

True State	Predicted State		Success Rate
	Disease	No Disease	
Disease	63	5	Sensitivity = 92.6%
No Disease	7	108	Specificity = 93.9%

Table 6: Healthlinx Diagnostic prediction Success Threshold of 0.5

Although the sensitivity and specificity of the Healthlinx and Reference diagnostics are very similar, examination of the ROC curves provides strong evidence that the Healthlinx diagnostic is superior. The lack of ability to detect differences in sensitivity and specificity is not surprising. A very large sample size would be required to achieve reasonable power for this comparison.

6 Diagnosis of Early-Stage Ovarian Cancer

The validation data considered for this diagnostic scenario consisted of all samples not having the disease, together with the early stage disease samples (disease stages I and II). It comprised a total of 154 patients, 39 of whom were ovarian cancer positive. This data set excluded 22 patients with stage 3 disease, 3 patients with stage 4 disease and a further 4 patients with un-staged disease.

Inspection of the ROC curve for the Healthlinx diagnostic reveals that at a specificity similar to that of the reference diagnostic (approximately 90%), the diagnostic achieves a sensitivity of 98%: a greater than 8% increase over that of the reference diagnostic. Similarly, at a sensitivity equal to that of the reference diagnostic (89.7%) the Healthlinx diagnostic achieves a specificity of 94% - approximately 4.5% greater than that of the reference diagnostic.

Threshold	sensitivity	specificity
0.14	0.98	0.90
0.15	0.97	0.90
0.17	0.96	0.90
0.19	0.95	0.90
0.27	0.94	0.91
0.36	0.93	0.93
0.45	0.92	0.94
0.48	0.91	0.94
0.50	0.90	0.94

Table 7: Sensitivity and Specificity for Healthlinx Diagnostic - Early Stage Ovarian Cancer

The area under the ROC curve was 0.937 (bootstrap confidence interval 0.866 to 0.976) for the reference diagnostic and 0.985 (bootstrap confidence interval 0.963 to 0.994) for the Healthlinx diagnostic. The bootstrap test for equality of the two AUCs was statistically significant at the 1% level².

Figure 2 shows the bootstrap distribution of the difference between the Healthlinx AUC and the standard diagnostic AUC. A positive value represents superiority of the Healthlinx diagnostic. The probability mass of the distribution is overwhelming for positive values of the statistic.

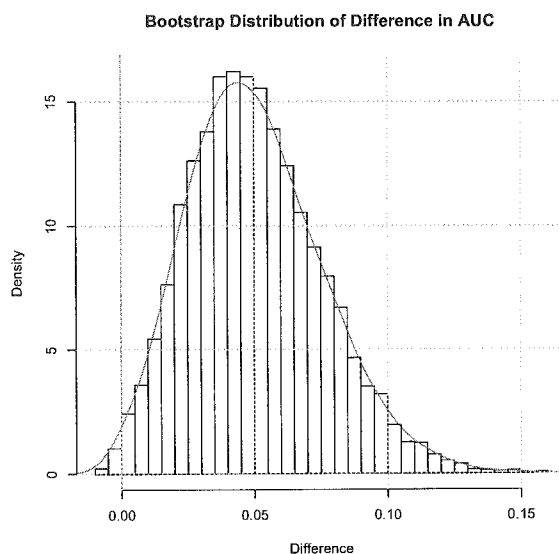


Figure 2: Bootstrap Test for Equality of AUCs - Early Testing Cases . A positive value implies superiority of the Healthlinx diagnostic.

The sensitivity and specificity of the reference diagnostic for this data set is shown in Table 8. The sensitivity and specificity of the Healthlinx diagnostic is shown in Table 9, when the cut off for a positive diagnosis is 0.3. Similar results for a cut off of 0.5 are shown in Table 10. The lower cut off, of course, increases sensitivity at the expense of reducing specificity. A broader indication of the relationship between specificity and sensitivity is shown in Table 7. This Table shows the specificity of the Healthlinx diagnostic for a range of required sensitivities.

²The difference is statistically significant, despite the overlapping marginal confidence intervals, because of the strong correlation between reference and Healthlinx AUCs in bootstrap realisations.

Healthlinx Diagnostic Performance

True State	Predicted State		Success Rate
	Disease	No Disease	
Disease	35	4	Sensitivity = 89.7%
No Disease	12	103	Specificity = 89.6%

Table 8: Reference Diagnostic prediction Success

True State	Predicted State		Success Rate
	Disease	No Disease	
Disease	36	3	Sensitivity = 92.3%
No Disease	10	105	Specificity = 91.3%

Table 9: Healthlinx Diagnostic prediction Success Threshold of 0.3

True State	Predicted State		Success Rate
	Disease	No Disease	
Disease	35	4	Sensitivity = 89.7%
No Disease	7	108	Specificity = 93.9%

Table 10: Healthlinx Diagnostic prediction Success Threshold of 0.5

Although the sensitivity and specificity of the Healthlinx and Reference diagnostics are very similar, examination of the ROC curves provides strong evidence that the Healthlinx diagnostic is superior. The lack of ability to detect differences in sensitivity and specificity is not surprising. A very large sample size would be required to achieve reasonable power for this comparison.

7 Power to Detect Differences in Diagnostic Performance

In this section we consider the power to detect differences in diagnostic performance as a function of the diagnostic performance of the reference diagnostic, the diagnostic performance of the Healthlinx diagnostic, and the sample size. The test for differences in diagnostic performance is based on only those observations which are discordant - that is, those observations in which the Healthlinx diagnostic and the reference diagnostic give different results. The *effective* sample size is therefore the number of discordant observations. This number depends on the probability of successful diagnosis under both procedures, but also on the correlation between success of the reference diagnostic and success of the Healthlinx diagnostic.

We consider two scenarios for this correlation:

Scenario 1 The independence scenario in which success on one diagnostic is statistically independent of success on the other,

Scenario 2 The perfect correlation scenario. Success on the weaker diagnostic guarantees success on the stronger. A proportion of those cases misclassified on the weaker diagnostic will be correctly classified on the stronger. Let p_1 be the probability of successful diagnosis on the weaker diagnostic. Let p_2 be the probability of successful diagnosis on the stronger diagnostic. Then the proportion of those misclassified on the weaker diagnostic which are correctly classified on the stronger is given by:

$$q = \frac{p_2 - p_1}{1 - p_1}$$

Table 11 shows the power of the test for equal diagnostic power as a function of the probability of success for the Healthlinx diagnostic, and the sample size. All power calculations in this table are based on the assumption that the probability of success for the reference diagnostic is 0.8. For the independence scenario, with a 10% improvement in the Healthlinx diagnostic (i.e. 90% diagnostic efficiency compared with 80% for the reference) a sample size of approximately 250 is required to achieve an 80% power. More than 300 samples are required to achieve 90% power. Power is much better for the perfect correlation scenario - but this is unlikely in practice. The true result should lie somewhere between the power calculated for the independence and the perfect correlation scenario. For design purposes, power should be based on the independence scenario - since this give the more conservative result. We recommend a sample size of 300 patients. This result power for this sample size is displayed in bold in Table 11.

Note that these power calculations are appropriate for the size of the validation set - not the total data set (training plus validation). They illustrate the great difficulty in demonstrating marginal improvement over a diagnostic which is already efficient. Comparisons may be performed either in terms of diagnostic efficiency (in which case the 300 patients represents the total (true negative plus true positive) patients, or for sensitivity (in which case the 300 patients represent only those patients who are disease positive).

Sample Size	Power under:			
	Scenario 1		Scenario 2	
	90%	95%	90%	95%
100	0.364	0.850	0.942	0.998
150	0.570	0.968	0.998	1.000
200	0.718	0.994	1.000	1.000
250	0.822	0.999	1.000	1.000
300	0.892	1.000	1.000	1.000

Table 11: Healthlinx Power Calculations for Diagnostic Efficiency. Scenario 1 represents independent outcomes for reference diagnostic and Healthlinx diagnostic. Scenario 2 represents perfect correlation between tests.

8 Summary and Conclusions

1. The Healthlinx diagnostic is clearly superior to the reference diagnostic, in terms of AUC. This will generate beneficial patient outcomes;
2. The Healthlinx diagnostic is able to detect early stage ovarian cancer. It is clearly superior to the reference diagnostic in this regard;

The superiority of the Healthlinx diagnostic is demonstrable through the analysis of ROC AUCs. This is an appropriate and well established statistical approach.

References

- [1] S. V. Beiden, R. F. Wagner, G. Campbell, and H.-P. Chan. Analysis of uncertainties in estimates of components of variance in multivariate roc analysis. *Academic Radiology*, 8(7):616–622, 2001.
- [2] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application*. Cambridge University Press, 1997.
- [3] B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics, 1982.
- [4] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.
- [5] P. Hall. Theoretical comparison of bootstrap confidence intervals. (with discussion.). *Annals of Statistics*, 16:927–985, 1988.
- [6] P. Hall. Asymptotic properties of the bootstrap for heavy-tailed distributions. *Annals of Probability*, 18:1342–1360., 1990.
- [7] J. A. Hanley. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.
- [8] M.-L. T. Lee and B. A. Rosner. The average area under correlated receiver operating characteristic curves: a nonparametric approach based on generalized two-sample wilcoxon statistics. *Applied Statistics*, 50(3):337–344, 2001.
- [9] C. Lloyd. The use of smoothed ROC curves to summarise and compare diagnostic systems. *Journal of the American Statistical Association*, 93:1356–1364, 1998.
- [10] D. Mossman. Resampling techniques in the analysis of non-binormal roc data. *Medical Decision Making*, 138(15):358–366, 1995.