New Approaches to Measuring the Performance of Programs that Generate Differential Diagnoses using ROC Curves and Other Metrics

Hamish S. F. Fraser MB MSc[1,2], Shapur Naimi MD[3], William J. Long PhD[2]
Children's Hospital Informatics Program[1], Boston, MA, the Massachusetts Institute of Technology[2,] Cambridge, MA and Tufts-New England Medical Center[3], Boston, MA

Abstract

Introduction: *Evaluation of computer programs which generate multiple diagnoses can be hampered by a lack of effective, well recognized performance metrics. We have developed a method to calculate mean sensitivity and specificity for multiple diagnoses and generate ROC curves.*
Methods: *Data came from a clinical evaluation of the Heart Disease Program (HDP). Sensitivity, specificity, positive and negative predictive value (PPV, NPV) were calculated for each diagnosis type in the study. A weighted mean of overall sensitivity and specificity was derived and used to create an ROC curve. Alternative metrics Comprehensiveness and Relevance were calculated for each case and compared to the other measures.*
Results: *Weighted mean sensitivity closely matched Comprehensiveness and mean PPV matched Relevance. Plotting the Physician's sensitivity and specificity on the ROC curve showed that their discrimination was similar to the HDP but sensitivity was significantly lower.*
Conclusions: *These metrics give a clear picture of a program's diagnostic performance and allow straightforward comparison between different programs and different studies.*

Introduction

Evaluations of medical diagnosis programs have been carried out for several decades but there are still significant problems in accurately measuring performance. For programs which produce more than one diagnosis there is a particular challenge in finding suitable performance metrics. If a system reasons about only one diagnosis, then the sensitivity and specificity of the program can readily be determined given a suitable standard diagnosis. This approach is also suitable if the program produces a small number of diagnoses. However if the program is designed to reason about the possibility of dozens or hundreds of diagnoses other metrics are required. Evaluating such programs usually requires a considerable amount of data per case and it is therefore difficult to collect more than 100 to 200 cases. This results in sparse data with many diagnoses appearing only once or twice in the evaluation (and many diagnoses not appearing at all). Calculating sensitivity and specificity for each diagnosis is therefore impractical, and only common diagnoses can be effectively evaluated. The approach taken by Moens[1] was to calculate a weighted mean of the sensitivity and specificity for each of the diagnoses generated by the program AI/Rheum. However they did not perform statistical comparisons with physicians or other programs.

An alternative approach was taken by Berner et al[2, 3] evaluating general medical diagnosis programs. They studied QMR, Iliad, Meditel and Dxplain using 105 detailed cases. These programs may contain several thousand diagnoses in their knowledge base and hence only a small proportion will appear at all in such a study. The researchers developed several new metrics to measure the aggregate performance on each case. However the use of new metrics that have not previously been evaluated in measurement studies makes it difficult to compare that study with others[4].

In addition, all the above metrics require a threshold to be set in their calculation. A low threshold tends to increase sensitivity and decrease specificity and vice versa. This makes comparing programs and/or physicians problematic. The Receiver Operator Characteristic (ROC) Curve[5] is calculated with a range of thresholds and the area under the curve provides a robust measure of the discrimination of a diagnostic model. We describe here a method for creating ROC curves from the output of programs which generate broad differential diagnoses. The ROC curves are compared to

other metrics using data from a clinical trial of the Heart Disease Program[6].

Background

The Heart Disease Program (HDP) is a large model based diagnosis program designed to reason about most aspects of cardiac disease[7]. The program provides a broad differential diagnosis with 186 possible diagnoses currently in the knowledge base. It generates a mean of 10.7 diagnoses per case grouped into 1 or more hypotheses that fully describe the input data. The program was recently evaluated with 114 cases directly entered by working physicians[6](111 were used here). The program's performance was compared with that of the physicians entering the cases, standard diagnoses for comparison were derived from follow-up data (Final Diagnoses). The diagnoses from the three groups: HDP, Physicians and Final Diagnosis, were coded in an identical fashion.

Methods

The basis of calculating the metrics used here is determining the number of matches between two lists of diagnoses. For example if the HDP generates 10 diagnoses for a case and the Final Diagnosis has 5, and 3 diagnoses are present in both lists, then there are 3 matches. The following metrics can then be derived.

Comprehensiveness and Relevance. These two metrics were developed by Berner et al[2] and are shown in Figure 1. *Comprehensiveness* is the number of matching diagnoses divided by the number of Final Diagnoses. *Relevance* is the number of matching diagnoses divided by the number of Test Diagnoses (HDP or Physician's). These metrics are calculated on a case by case basis and expressed as a mean for all cases.

Calculation of sensitivity, specificity and positive predictive value (PPV). Refer to Figure 2. For a diagnosis D, sensitivity is Matches (TP) divided by the frequency of D in the Final Diagnosis. Mean sensitivity is weighted by the frequency of D in the Final Diagnosis (Figure 3). PPV is Matches (TP) divided by frequency of D in the Test Diagnosis (Figure 2). Specificity is True Negatives (TN) divided by all cases without D in the Final diagnosis (True Negatives are all 1ll cases minus cases with D in the Final or Test Diagnoses). For mean specificity and PPV the weighting was by the frequency of D in the Test Diagnosis. Results of division by zero are set to zero, and excluded from analysis by the weighted mean. Figure 3 shows example results.

	Disease -	Disease +	Row total
Test -	TN	FN	
Test +	FP	TP *(Matches)*	*Test Diagnosis*
Column Total		*Final Diagnosis*	

Figure 2: TN & FN are true and false negatives, TP & FP, true and false positives.

Creating ROC Curves from the HDP output

Production of an ROC curve requires the ability to generate pairs of sensitivity and specificity at different thresholds. The HDP can generate a file with each diagnosis labeled with a pseudo-probability score normalized to the 0-1 range. The diagnoses are filtered according to that score to generate multiple diagnosis files (Figure 3) at several thresholds between 0 and 1. Finally, mean sensitivity and mean specificity are calculated for each file. The resulting pairs of sensitivity and 1 – specificity are plotted to create the ROC curve, and the area under the curve is calculated. With sufficient data, ROC curves can be created for individual diagnoses.

Testing Statistical Significance

In comparative studies[4], the performance of a diagnostic program is compared with another program, or the physicians who normally manage such patients. In these circumstances it is important to assess whether a statistically significant difference exists between the scores for the program and the physician.

Case ID#	Matches	Test (HDP) diagnoses	Final Diagnoses	Comprehensive-ness	Relevance
5403	3.0	18	3	3/3 = 100%	3/18 = 17%
5404	1.0	5	3	1/3 = 33%	1/5 = 20%

Figure 1: Part of the output file showing data for two performance metrics developed by Berner et al[2]. (*Test diagnosis* is Physicians or HDP, *Final Diagnosis* is from follow-up and investigations)

Diagnosis	HDP	FinalDx	TP	FP	FN	TN	**Sens.**	**PPV**	**Spec.**
congestive-failure	22.0	35.0	17.0	5.0	18.0	74.0	**49%**	**77%**	**94%**
constrictive-pericarditis	3.0	1.0	1.0	2.0	0.0	111.0	**100%**	**33%**	**98%**
COPD	39.0	9.0	6.0	33.0	3.0	72.0	**67%**	**15%**	**69%**
cor-pulmonale	1.0	0.0	0.0	1.0	0.0	113.0	**0%**	**0%**	**99%**

Figure 3: Part of the output file for the HDP compared to the cardiologists by individual diagnosis. Note the sparse data for some diagnoses. Sens. = Sensitivity, Spec. = Specificity, FP and FN = False Positive and Negative, TN = True Negative, TP=True Positives (equivalent to "Matches" in Figure 1).

This is straightforward to do with a single diagnosis, for example the presence or absence of myocardial infarction, using a two by two table and the Chi square test. For more complex diagnoses it is generally necessary to have a score for each different case. This works well for Comprehensiveness and Relevance, which are calculated for each case individually. When the sensitivity and specificity are calculated for each different type of diagnosis and combined using a weighted mean (Figure 3) it is not possible to obtain a score for each case, and therefore difficult to test comparisons for statistical significance. This is perhaps the most important reason to use the alternative metrics.

The simplest way to perform statistical comparisons is to use the output file containing data on Comprehensiveness and Relevance for a particular comparison, such as HDP compared to the Final Diagnoses. A second file of this type is generated for an alternative comparison such as Physicians compared to the Final Diagnoses. The first set of scores can then be subtracted from the second giving a Comprehensiveness (or Relevance) score difference for each case. The mean of these score differences is then compared to zero, testing the null hypothesis that there is no difference between scores. If the data are normally distributed the T test is used, in the examples described here a nonparametric method, the Wilcoxon Signed Rank test, was employed. The statistical package used was JMP (SAS Institute, Carey, NC).

If an ROC curve is generated from the pairs of weighted mean sensitivity and mean specificity then the discrimination of the program for the presence or absence of the correct diagnosis is expressed by the area under the curve. Using the method of Hanley and McNeil[8], the areas under two curves may be compared to determine if the difference is statistically significant.

Results: Evaluation of Metrics

The 111 cases from the Heart Disease Program study were used to test the performance of the various metrics implemented here. Figure 4 shows the comparison of traditional measures, sensitivity, specificity and PPV with Comprehensiveness and Relevance. It will be noted that there is a close relationship between the values of Comprehensiveness and Sensitivity and also a similar relationship between Relevance and PPV. The reason for the small differences in the scores between the different types of metrics may relate to the sparse data (figure 3). The presence of rare occurrences of a particular diagnosis will generate large swings in Sensitivity and PPV scores (0, 50, 100% with two cases) and this is likely to increase the random variation in the weighted mean. The Comprehensiveness and Relevance scores typically compare several different diagnoses per case, producing smoother measures.

	Sensitivity	Comprehen-siveness	Specificity	PPV	Relevance
HDP& Physician	61.3	66.6	77.0	29.1	28.1
HDP	54.5	57.3*	79.5	29.6	31.0*
Physician	34.8	39.4 *	93.9	56.2	55.4*

Figure 4: Performance (%) compared to the Final Diagnosis (* p<0.0001 by Wilcoxon Signed Rank)

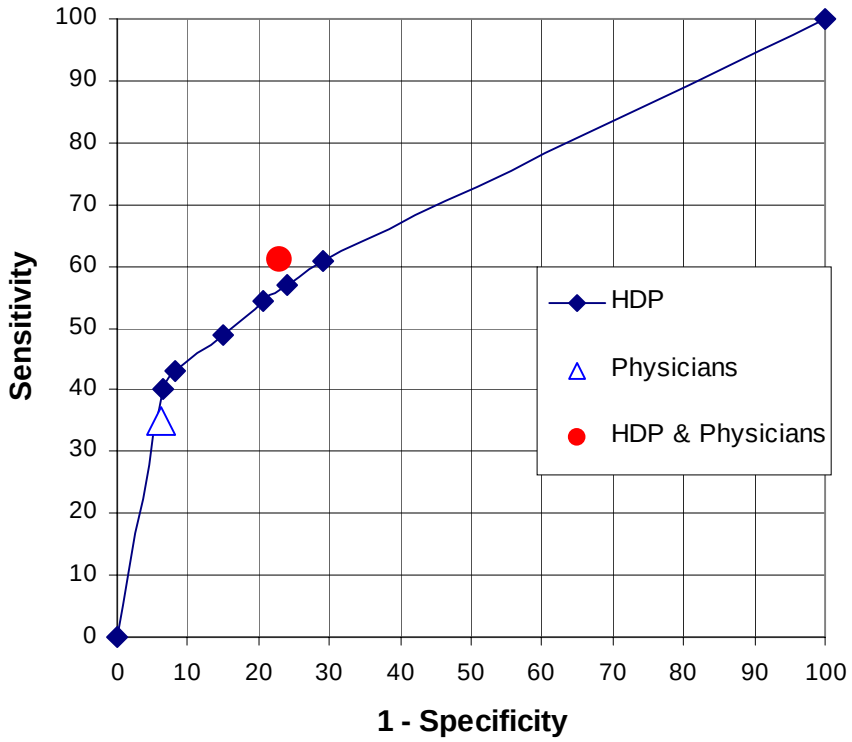

Figure 5, ROC Curve of the HDP compared to the Final Diagnoses. The performance of the Physicians and the combination of Physicians and HDP is also shown.

Figure 4 also shows the effect of combining the diagnoses of the HDP and the Physicians, creating the union of the two groups (HDP & Physician). This allows us to explore the potential effects that the program might have if the physician adopted all the diagnoses suggested into their own differential. While there is a significant increase in Comprehensiveness and sensitivity this is offset by a decrease in Relevance, PPV and Specificity.

Figure 5 shows the ROC curve of the HDP compared to the Final Diagnoses. The area under the curve is 70.1%. It will be noted that only the central area of the curve is demarcated. This is due to the design of the HDP which generates a maximum of around 25 diagnoses and generally outputs a minimum of around 4 diagnoses per case. More traditional Bayesian systems, like Dxplain or Iliad, output ranked lists of 40 or more diagnoses and would be expected to create a wider range of points.

The sensitivity and specificity of the physicians entering the case is also plotted on Figure 5. It is clear that their discrimination is very similar to the HDP, but their sensitivity is lower than the usual output of the HDP of 54.5% (Figure 4). Also plotted is the performance of the combination of the HDP and Physicians indicating that the combination gives slightly better performance than the program alone.

Discussion

We believe that this may be the first measurement study comparing these performance metrics using real clinical data of differential diagnoses. By highlighting the similarities between the measures used in Berner's study and more traditional approaches, it is hoped that the newer measures will gain wider acceptance. This may be particularly important as a means to test the statistical significance of comparisons.

The ROC curve is widely used in assessing the diagnostic performance of tests and simple diagnostic systems. Its use in differential diagnoses should allow easier comparison between different systems (in a study such as Berner's) and also between different studies.

A particular problem with programs that generate a large differential diagnosis is deciding how many of the output diagnoses to assess. An arbitrary cut-off such as 20 diagnoses has been used in the past, and forms part of the definition of metrics such as Comprehensiveness and Relevance in a previous study[2, 3]. Using ROC curve analysis avoids this problem, and in fact is an ideal way of determining how many diagnoses to consider in a clinical situation. For example a low threshold and hence larger differential diagnosis list is appropriate in initial assessment and screening.

In Figure 5, the Physicians' performance falls on the same curve as the HDP, suggesting that their discrimination is equivalent. However, the low sensitivity of the Physicians relative to the usual output of the HDP may be a disadvantage. Friedman et al [9] have shown that if the correct diagnosis for a case is present in a program's output, it is significantly more likely to be included in the Physician's differential diagnosis. The higher sensitivity of the HDP should be beneficial to the physicians as suggested by the performance of the combined diagnosis in figure 5. In addition, this analysis does not take into account the detailed explanations of the diagnoses provided by the HDP. By allowing the physician to exclude less appropriate diagnoses, explanations should increase the effective specificity and PPV of the HDP, moving the ROC curve to the left. The magnitude of this effect is difficult to measure.

Automating the process of diagnostic comparison is particularly helpful in monitoring the performance of a diagnosis program after changes have been made. It is a simple matter to add the new diagnostic output from the program to the database and rerun the comparisons. Any changes in performance can be rapidly assessed and checks made of the types of diagnoses that are affected.

We plan to test these analysis methods on data from other differential diagnosis programs. The Java software written for these analyses will be made available to interested researchers once development is complete.

Acknowledgements

This study was supported by the National Heart Lung and Blood Institute grant R01-HL33041. The authors thank Laura Smeaton for statistical expertise.

References

1. Moens HJB. Validation of AI/Rheum Knowledge Base with Data from Consecutive Rheumatological Outpatients. *Meth. Inform. Med*. 1992;31:175 - 81.
2. Berner ES, Webster GD, Shugerman AA, et al. Performance of four computer-based diagnostic systems. *N Engl J Med*. 1994;330:1792-6.
3. Berner ES, Jackson JR, Algina J. Relationships among performance scores of four diagnostic decision support systems. *J Am Med Inform Assoc*. 1996;3:208-15.
4. Friedman C, Wyatt J. Evaluation Methods in Medical Informatics. : Springer-Verlag; 1996.
5. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143:29-36.
6. Fraser H, Long W, Naimi S. Differential diagnoses of the Heart Disease Program have better sensitivity than resident physicians. In: Chute CG, ed. *Proc AMIA Annu Fall Symp*; 1998:622 - 626.
7. Long WJ, Naimi S, Criscitiello MG. Development of a knowledge base for diagnostic reasoning in cardiology. *Comput Biomed Res*. 1992;25:292-311.
8. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. 1983;148:839-43.
9. Friedman CP, Elstein AS, Wolf FM, et al. Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems. *Jama*. 1999;282:1851-6.