CompareDx: a Software Toolkit for Measuring the Performance of Programs that Generate Multiple Diagnoses

Hamish S. F. Fraser MRCP MSc[1,2], William J. Long PhD[2]
Tufts-New England Medical Center[1] and the Massachusetts Institute of Technology[2]

Introduction Evaluations of medical diagnosis programs have been carried out for several decades but for programs which produce multiple diagnoses there is a lack of suitable, well validated performance metrics. If a program reasons about only one (or a few) types of diagnosis, then the sensitivity and specificity of the program can readily be determined given a suitable standard diagnosis. However if the program is designed to reason about the possibility of dozens or hundreds of diagnoses other metrics may be 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. We have refined performance metrics for assessing diagnostic accuracy, evaluated them with data from a clinical evaluation study, and developed a Java program to implement the metrics efficiently.

Methods Two metrics, Comprehensiveness and Relevance were based on work by Berner et al [1] in assessing performance of general medical diagnosis programs. The other metrics calculated were the weighted means of Sensitivity, Specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV), similar to Moens[2]. Test data came from the evaluation of the Heart Disease Program with clinical data on 114 patients [3]. For each case there was a differential diagnosis from the HDP and from the physician. This was compared to a Final Diagnosis from chart review. The Compdx software takes into account partially matching diagnoses, and also incorporates a score of clinical importance for each diagnosis. Comprehensiveness is the proportion of Final Diagnoses suggested by the HDP or physicians for each case, Relevance is the proportion of HDP or physician diagnoses that were correct. As Comprehensiveness and Relevance scores are calculated on an individual case basis they allow statistical comparison of diagnostic performance (in this instance using the Wilcoxon Signed Rank Test).

Results The table shows the performance of the HDP and the physicians compared to the Final Diagnosis. Results are also shown for the combination of the HDP and Physicians diagnoses compared to the Final Diagnosis.

	HDP& Physician	HDP	Physician
Sensitivity	61.3	53.0	34.8
Comprehen -siveness	66.6	57.3*	39.4 *
Specificity	77.0	75.6	93.9
PPV	29.1	25.4	56.2
Relevance	28.1	28.0	55.4

Table: Performance compared to final diagnosis. * significant at P<0.001

Discussion The Compdx software allows effective comparison of diagnosis programs even when evaluation studies produce sparse data. It should be noted that Comprehensiveness is not only theoretically similar to sensitivity but gives similar results in this study (within 6%). A equivalent relationship is seen with Relevance and PPV. It is hoped that refining and validating these newer metrics and automating their calculation will encourage their wider use. This should simplify the process of evaluating diagnostic programs and performing statistical comparisons of results.

References

1. Berner, E S, Webster, G D, Shugerman, A A, Jackson, J R, et al *Performance of four computer-based diagnostic systems.* N Engl J Med, 1994. 330(25): p. 1792-6.
2. Moens, H.J.B., *Validation of AI/Rheum Knowledge Base with Data from Consecutive Rheumatological Outpatients.* Meth. Inform. Med, 1992. 31: p. 175 - 81.
3. Fraser HSF, Long W, Naimi S. *Differential diagnoses of the Heart Disease Program have better sensitivity than resident physicians.* in *Proc AMIA Annu Fall Symp.* 1998.