Comparison of Measures to Assess Change in Diagnostic Performance Due to a Decision Support System

Richard S. Maisiak, PhD, MSPH, and Eta S. Berner, EdD
University of Alabama at Birmingham, Birmingham, AL

Little has been done to examine the relative merit of measures used to assess the impact of diagnostic decision support systems (DDSS) on physician performance. In this study, 10 different single-measures of diagnostic performance were compared empirically. The measures were of three types: rank-order, all-or-none, and appropriateness. The responsiveness (RESP) of each measure was estimated under two repeated-measures experimental conditions. RESP is the degree to which a measure could detect differences between conditions of low and high performance. The diagnostic performance of 108 physicians was compared on medical cases of varying diagnostic difficulty and with or without a high level of assistance from a DDSS. The results showed that the RESP among the measures varied nearly tenfold. The rank-order measures tended to provide the highest RESP values (maximum = 2.14) while appropriateness measures provided the lowest RESP values (maximum = 1.41). The most responsive measures were rank-orders of the correct diagnosis within the top 5 to 10 listed diagnoses.

INTRODUCTION

Diagnostic decision support systems (DDSS) are designed to improve the diagnostic performance of physicians. A variety of measures of diagnostic performance with good reliability and validity have been reported in the literature, [1-6] but it is not clear how these measures compare with respect to other important measurement properties. The purpose of this study was to improve the measurement of medical diagnostic performance by examining the relative responsiveness, i.e. the degree to which a measure is able to detect differences among persons or constructs.[7-13] Measures with higher responsiveness increase statistical power, lower the risk of Type II error, and lower the sample size required for studies of diagnostic performance. This is especially important in studies of the impact of DDSS on physician performance, since it is often difficult to obtain large samples of physicians and/or appropriate cases on which to test the systems.

Several study protocols have involved physicians using a DDSS and then providing a rank-ordered list of reasonable diagnoses for a given medical case.[2-5] In some studies[4-5] the maximum number of diagnoses was restricted, while in others[2-3] there was a possibility of a large number of diagnoses. Three types of measures have been used to measure the impact of DDSS on physician diagnostic performance. Rank-order measures score the position or location (rank) on the list of the correct diagnosis for the case. All-or-none measures score performance on any one case as correct or not depending on whether the correct diagnosis is in the list. Diagnostic appropriateness measures are computed from scores assigned to each diagnosis in a list that correspond to the relevance[2-3] or plausibility[4-5] of the diagnosis to that case.

Another measurement strategy that has been used within each of these approaches is to limit the diagnostic list to some number of the top-ranked diagnoses.[1, 4-6] The immediate effect of list restriction is to lower the absolute performance scores, but its effect on responsiveness has not been investigated. The aims of the present study were to estimate and compare the responsiveness of the three types of measures using different list restrictions under two different experimental conditions. Previous studies by the authors used three of the measures (two based on appropriateness and one all-or-none measure) and found that physician diagnostic performance was higher on easier to diagnose cases and on cases for which physicians were assisted with high-quality information vs. low-quality information from the DDSS.[2-3] Friedman et al. used a measure that represented a combination of one rank-order and one appropriateness measure of performance before and after using a DDSS.[4-5] In this study, we used the data set from the Berner et al. studies and examined the responsiveness of all of the individual measures used in both studies as well as newly developed measures reflecting different levels of restriction of the number of diagnoses.

METHODS

The study used a repeated-measures design in which each participating physician attempted to diagnose eight medical cases that were stratified into

1067—5027/00/$5.00

categories of either low or high diagnostic difficulty and those in which a DDSS provided high-quality or low-quality information. The subjects were 108 internists or family practice physicians recruited nationally for a study on diagnostic decision support systems.[2-3] As part of the DDSS study, the difficulty level of each of 24 written cases was determined by a group of experienced physicians and then stratified into three sets of eight cases with each set containing four cases of low difficulty and four cases of high difficulty level. Each physician was randomly assigned to diagnose the cases in one eight-case set. Physicians were instructed to list up to 20 diagnoses for each case and to rank each diagnosis from best to worst. In the first analysis, the physicians' mean diagnostic performance on the four easier cases was compared to their mean diagnostic performance on the four difficult cases using a t-test for paired differences. In the second analysis, the physician's mean diagnostic performance on the four cases with a low level of assistance from a DDSS was compared to their mean diagnostic performance on the four cases with a high level of assistance from a DDDS. The physician diagnosis list for each case was scored 10 different ways. We compared the mean within-group differences on all 10 performance measures using a t-test for paired differences.

Measures of Diagnostic Performance

Rank-order Measures. There were four different rank-order measures. We used the location score from Friedman et al.[4-5] (Rank6), which restricted diagnoses only to the top six, and three other rank-order measures: the top five (Rank5), the top 10 (Rank10), or the top 20 (Rank20) listings. A physician's rank-order score was the rank of the first correct diagnosis that was listed subtracted from the total number of restricted diagnoses plus one. The score was zero if the correct diagnosis was not present in the restricted list. Thus, a higher score indicated better performance.

All-or-None Measures. For the three all-or-none measures, physician diagnoses were restricted to either the top five (Correct5), the top 10 (Correct10), or the top 20 (Correct20) on the ranked list. The top 20 score reflects the accuracy score used by Berner et al.[2-3] A physician's score on a single case was either 1, if the correct diagnosis was in the restricted list, or 0, if it was not included.

Measures of Appropriateness. The three measures in this category were intended to better assess a physician's apparent understanding of the diagnostic possibilities even when the correct diagnosis was not listed or when many diagnoses were included. The Plausibility score used by Friedman et al. was the average of values assigned to each diagnosis in the top six to indicate its plausibility given the case data.[4-5] The relevance score used by Berner et al. was the proportion of all the physician's diagnoses for any one case that were considered relevant.[2-3] The Comprehensiveness score reflected the completeness of the diagnosis list. It was based on the proportion of all possible relevant diagnoses that were listed by the physician for a case.[2-3]

Several methods have been proposed to calculate responsiveness,[7-13] sometimes termed sensitivity to change, with no one method yet shown to yield superior scores. The standardized response method proposed by Cohen[12-13] was chosen for this study since it is consistent with other formulas of effect size and because estimates of sample size are easily calculated from it. Responsiveness (effect size) was calculated as the mean difference of paired scores divided by the standard deviation of the difference scores or $ES = M_d / \sigma$.

RESULTS

Table 1 shows the Spearman intercorrelations of all 10 measures. All the rank-order measures and all-or-none measures tended to be highly correlated with one another with coefficients greater than 0.80 except for the Rank20 measure. The appropriateness measures tended to have much lower correlations with the rank-order measures or the all-or-none measures. Table 2 shows the mean difference score of the performance measures by subtracting the score on difficult cases from the score on easier cases and also shows the effect size for each measure. The results indicated that all 10 measures were able to statistically detect significant ($p<01$) differences between diagnostic performance on easier vs. difficult to diagnose cases. In general, the rank-order measures showed higher effect sizes (maximum of 2.04) than measures based strictly on being correct (maximum effect sizes of 1.69) or based strictly on appropriateness (maximum effect sizes of 1.36). Measures based only on diagnoses within the top 10 or less of the physicians' lists tended to have higher effect sizes than measures that included all diagnoses from the list. Table 3 shows the mean difference score for each of the performance measures by subtracting the score on low DDSS assistance cases from the score on high DDSS assistance cases along with the effect size for each measure. The results indicated that, again, the rank-order measures tended to provide greater effect sizes, with a maximum of 2.14. The all-or-none measures tended to have the

Table 1. Spearman Intercorrelations of 10 Measures of Diagnostic Performance Based on the Mean Across All Eight Cases

Measure	Rank5	Rank6	Rank10	Rank20	Correct5	Correct10	Correct20	Plausibility	Relevance	Comp.
Rank5	1.00	.99	.96	.43	.91	.85	.83	.60	.34	.21
Rank6	.99	1.00	.98	.42	.94	.88	.87	.60	.34	.28
Rank10	.96	.98	1.00	.38	.97	.96	.94	.56	.32	.35
Rank20	.43	.42	.38	1.00	.34	.27	.19	.56	.64	.48
Correct5	.91	.94	.97	.34	1.00	.95	.94	.55	.34	.39
Correct10	.85	.88	.96	.27	.95	1.00	.99	.45	.21	.45
Correct20	.83	.87	.94	.19	.94	.99	1.00	.45	.24	.50
Plausibility	.60	.60	.56	.56	.55	.45	.45	1.00	.88	-.02
Relevance	.34	.34	.32	.64	.34	.21	.24	.88	1.00	-.09
Comp.	.21	.28	.35	.48	.39	.45	.50	-.02	-.09	1.00

Note: Correlation coefficients greater than 0.20 were significantly different from zero at the 0.01 level (2-tailed).

Table 2. Responsiveness (Effect Size) of 10 Measures of Diagnostic Performance Based on a Paired Samples Test for Detecting Differences Between Easy and Hard to Diagnose Cases

	Mean Difference	Standard Deviation	t-value	p-value	Effect Size
Rank-order Measures					
Rank5	1.79	0.88	21.21	.000	2.04
Rank6	2.16	1.07	21.02	.000	2.02
Rank10	3.63	1.88	20.08	.000	1.93
Rank20	2.74	2.22	12.81	.000	1.23
All-or-None Measures					
Correct5	0.37	0.22	17.59	.000	1.69
Correct10	0.36	0.23	16.59	.000	1.59
Correct20	0.35	0.23	16.21	.000	1.56
Appropriateness Measures					
Plausibility Score	1.00	0.73	14.12	.000	1.36
Relevance Score	0.11	0.17	6.95	.000	0.67
Comprehensiveness Score	0.03	0.09	3.00	.003	0.29

Table 3. Responsiveness (Effect Size) of 10 Measures of Diagnostic Performance Based on a Paired Samples Test for Detecting Improvement Due to Greater DDSS Assistance

	Mean Difference	Standard Deviation	t-value	p-value	Effect Size
Rank-order Measures					
Rank5	2.24	1.06	21.96	.000	2.11
Rank6	2.73	1.28	22.16	.000	2.13
Rank10	4.74	2.22	22.23	.000	2.14
Rank20	2.12	2.14	10.27	.000	0.99
All-or-None Measures					
Correct5	0.49	0.25	20.50	.000	1.97
Correct10	0.51	0.25	20.89	.000	2.01
Correct20	0.52	0.25	21.31	.000	2.05
Appropriateness Measures					
Plausibility Score	1.14	0.81	14.67	.000	1.41
Relevance Score	0.13	0.17	8.07	.000	0.78
Comprehensiveness Score	0.03	0.11	2.92	.004	0.28

next greatest effect sizes with a maximum value of 2.05. The appropriateness measures tended to have the lowest effect sizes with a maximum of 1.41. Except for Correct20, measures that used all listed diagnoses produced relatively low effect sizes compared to measures using the first 10 or fewer diagnoses.

DISCUSSION

The main finding of this study was that the magnitude of the responsiveness of the measures of diagnostic performance differed widely among the three approaches and among the diagnostic list restrictions. These results occurred even though all 10 measures of diagnostic performance produced significant p-values under two different experimental conditions. The rank-order measures of diagnostic performance tended to provide the greatest responsiveness while appropriateness measures produced the lowest responsiveness. Limiting the physician's diagnostic list to 10 or fewer diagnoses also tended to produce measures of higher responsiveness. The measures of rank-order that were based on the first five to 10 diagnoses listed by a physician were the most responsive measures of diagnostic performance in this study.

Investigators who examine the outcomes of interventions for improving diagnostic performance will find that rank-order measures will be able to detect changes or differences with lower risk of Type II error, with greater statistical power, with fewer subjects, or with fewer cases needed than other approaches. For example, assume that the experimental conditions are similar to Table 2 but with a sample size of 10 physicians and a two-sided alpha of .001. The Type II error for the study when using Rank5 would be only 0.12, which would be less than *one-third* of either the Type II error, 0.69, using Rank20, or the Type II error, 0.42, using Correct20.[13]

The high intercorrelations among the rank-order and all-or-none measures indicated that they assessed the same underlying construct and demonstrated good concurrent validity. Since the appropriateness measures were less correlated with the other study measures, it is possible that they tap a component of diagnostic performance that is slightly different from the other measures.

The all-or-none measures may have been less responsive because they failed to take into account the degree of confidence that the physician has in the correct diagnosis. Appropriateness measures also showed lesser levels of responsiveness under the present conditions, but they could become better measures of diagnostic performance in situations where either the elicitation of only relevant diagnoses or of all possible relevant diagnoses are considered of great importance. The former case could occur when the consideration of unlikely diagnoses is judged to be wasteful. The latter case could occur when the failure to consider an unusual but serious medical diagnosis is deemed an error.

Another situation where the choice of outcome measure requires caution is the evaluation of the diagnostic output of the DDSS itself. Clearly a DDSS is less likely to serve a prompting function if it does not suggest the actual diagnosis, or a closely related one, that the patient has. However, by suggesting other relevant diagnoses, a DDSS could still provide useful prompts. Furthermore, the more prevalent diagnoses are likely to be listed ahead of the very rare diseases in the DDSS output. While this is as it should be, it is these diagnoses that the physician user is most likely to have already considered, and thus the lower-ranked diagnoses may provide unique and more helpful information to the user. For these reasons, evaluating DDSS performance by examining the rank-order of the correct case diagnosis within a restricted number of diagnoses may be highly responsive but not always appropriate to use.

One limitation of this study was the exclusion of combined measures of diagnostic performance such as the Quality score.[4-5] These types of measures can be formed by combining scores from two or more of the three approaches examined in this study. Combined measures may be highly responsive, but their meanings may be difficult to interpret. It is also unclear which of the many ways single measures could be combined is the best. Further study is needed to examine these issues.

In summary, while all of the measures examined were able to detect significant differences among the constructs, the measures varied widely in terms of responsiveness. Investigators who are evaluating the output of DDSS or its impact on physician diagnoses should choose, as a first priority, the measures that best address their specific research goals. However, if multiple measures can serve the research purpose and if investigators can enroll only a small number of subjects, as many studies evaluating decision support systems do, they should consider the responsiveness of the measures. Based on the results of this study as well as previously published studies, we now have data to make more informed decisions about the

properties of measures that assess diagnostic performance.

Acknowledgments

This study was supported in part by grant # LM05125 from the National Library of Medicine.

References

1. Berner ES, Webster GD, Shugerman AA et al. Performance of four computer-based diagnostic systems N Engl J .Med 1994;330:1792-6.

2. Berner ES, Maisiak RS, Cobbs CG, and Taunton OD. Effects of a decision support system on physicians' diagnostic performance. J Am Med Informatics Assoc 1999;6:420-7.

3. Berner ES and Maisiak RS. Influence of case and physician characteristics on perceptions of decision support systems. J Am Med Informatics Assoc 1999;6:428-43

4. Friedman CP, Elstein AS, Wolf FM, et al. Measuring the quality of diagnostic hypothesis sets for studies of decision support. Medinfo98, 9 Pt 2;1998:864-8.

5. Friedman CP, Elstein AS, Wolf FM, et al. Enchancement of Clinicians' Diagnostic Reasoning by Computer-Based Consultation. A Multisite Study of 2 Systems. JAMA 1999;282:1851-6.

6. Berner ES, Jackson JR, and Algina, J. Relationships among performance scores of four diagnostic decision support systems. J Am Med Informatics Assoc;1996:208-15.

7. Liang MH. Evaluating measurement responsiveness. J Rheumatol 1995;22:1191-2.

8. Guyatt G, Walter S, and Norman G. Measuring change over time:assessing the usefulness of evaluative instruments. J Chronic Dis 1987;40:171-8.

9. Liang MH, Larson MG, and Cullen KE. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum 1985;28:542-7.

10. Beaton D and Hogg-Johnson S. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. J Clin.Epidemiol 1997;50:79-93.

11. Wright J and Young N. A comparison of different indices of responsiveness. J Clin Epidemiol 1997;50:239-46.

12. Cohen J. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1977.

13. Borenstein M, Rothstein H, Cohen J. Power and Precision. New Jersey:Biostat, 1997.