Research Paper ■

Effects of a Decision Support System on Physicians' Diagnostic Performance

ETA S. BERNER, EDD, RICHARD S. MAISIAK, PHD, MSPH, C. GLENN COBBS, MD, O. DAVID TAUNTON, MD

Abstract **Purpose:** This study examines how the information provided by a diagnostic decision support system for clinical cases of varying diagnostic difficulty affects physicians' diagnostic performance.

Methods: A national sample of 67 internists, 35 family physicians, and 6 other physicians used the Quick Medical Reference (QMR) diagnostic decision support system to assist them in the diagnosis of written clinical cases. Three sets of eight cases, stratified by diagnostic difficulty and the potential of QMR to produce high-quality information, were used. The effects of using QMR on three measures of physicians' diagnostic performance were analyzed using analyses of variance.

Results: Physicians' diagnostic performance was significantly higher ($p < 0.01$) on the easier cases and the cases for which QMR could provide higher-quality information.

Conclusions: Physicians' diagnostic performance can be strongly influenced by the quality of information the system produces and the type of cases on which the system is used.

■ **JAMIA.** 1999;6:420–427.

Diagnostic decision support systems (DDSSs) are software programs designed to assist physicians in making clinical diagnoses.[1] Although some highly specialized programs are in routine use,[2] many of the broad-based diagnostic programs are still not widely used, possibly because it is unclear how much they can assist clinicians. Studies of system accuracy have been done,[3–24] but few studies systematically examine how the use of a DDSS affects physicians' diagnostic performance.[25–29]

Affiliations of the authors: University of Alabama at Birmingham (ESB, RSM, CGC) and private practice (ODT), Birmingham, Alabama.

This work was supported by grant LM05125 from the National Library of Medicine.

Correspondence and reprints: Eta S. Berner, EdD, University of Alabama at Birmingham, School of Health Related Professions, Department of Health Services Administration, Health Informatics Program, 1675 University Boulevard, Room 544, Birmingham, AL 35294-3361.
e-mail: ⟨eberner@uab.edu⟩.

Received for publication: 4/30/99; accepted for publication: 5/19/99.

Many of the studies that have been done have had expert users (often the system developers themselves) provide the input data and interpret the results of system performance. It is not clear how well the results of these studies generalize to use by non–system developers. One key area that has not been well explored is how well the ordinary user can interpret the system output, especially since many of the systems admittedly provide a lengthy list of suggestions, not all of which are likely to be appropriate. In an evaluation of four DDSSs, an expert panel of clinicians considered more than half of the top 20 diagnoses suggested by each DDSS to be inadequate interpretations and syntheses of the clinical case data.[23,24] Although the mode of usage of the DDSS was somewhat artificial, since it was limited to the first pass through the data rather than based on extensive interaction, such data raise concerns about the risks of using these systems or, as has been suggested, concerns that DDSSs are unlikely to be very useful to physicians.[30] It is possible that nonexpert physicians will be unable to distinguish useful from misleading information or will reject all information provided because some is irrelevant. On the other hand, it is possible that users,

especially those who have made an investment in purchasing the systems, will be too gullible and take all the output equally seriously. Evaluation studies to date do not shed light on these issues. Because physicians' diagnostic performance is known to vary across different clinical cases,[31] it is important to determine the overall effect of these systems on physicians' diagnostic performance and also to examine how the quality of the information provided by a DDSS, as well as characteristics of the test cases, influences the physicians' diagnoses.

We have engaged in a series of studies designed to explore how DDSSs influence physicians' performance, how physicians use the systems, and how they react to the information provided. The purpose of the study reported here was to determine whether physicians' diagnostic performance when using a DDSS on clinical cases was affected by the quality of the information provided by the DDSS and the difficulty of the cases. Our hypothesis was that physicians' performance when using a DDSS would vary on cases that varied in the relevance of information provided by the DDSS; that if there were a difference, performance would be better when more relevant information was provided; and that this enhancement would be most marked on more challenging cases. A secondary aim was to validate our selection of comparatively difficult and easy cases. We hypothesized that if our case selection criteria were reasonable, overall performance should be better on the easier cases than on the more difficult ones.

Methods

The DDSS chosen for this study was Quick Medical Reference (QMR).[32] The version used for the present study was designed to address more than 650 internal medicine disorders. The participant sample consisted of physicians who had purchased QMR (version 2.2.2 for DOS, 3.6.1 for Windows, or the comparable Macintosh version). All registered users of QMR in the fall of 1993 were sent a letter inviting them to participate in the study and to provide information on their clinical experience and their experience with QMR. Respondents who returned their questionnaires, which included a signed consent form, were stratified by clinical experience and QMR experience and randomly assigned to one of three sets of eight written clinical cases for which they were asked to use QMR to assist them in developing a differential diagnosis. The cases each contained a history and physical and basic laboratory data, but none included a definitive test that confirmed the diagnosis. Participants were not given any other information about the cases, such as whether the correct diagnosis was in QMR's knowledge base or whether the cases were expected to be comparatively easy or difficult.

Participants were asked to use, as a minimum, the case analysis function of QMR, but they were permitted to use any additional functions they wished. To record their interaction with QMR, participants were asked to save as a text file each screen they viewed, which permitted an analysis of their approach to the cases and QMR's output as well documenting that they actually used the program. They were asked to use QMR alone, without other resources, to develop a differential diagnosis for each case, with space allowed for up to 20 diagnoses. They were also asked to indicate how they would work up the diagnoses, rate their confidence in their differential, and rate how useful QMR was to them in solving the case. When participants completed their cases, they signed a form attesting that they had followed all procedures. If they did not follow all procedures, they were asked to describe deviations from the instructions. There was no indication that participants used other resources or that there were any serious deviations from instructions.

Initially, 120 physicians were randomly selected and stratified into four cells by combinations of high and low clinical and QMR experience. The data from the background questionnaire were used to classify respondents into categories of high and low clinical experience and high and low QMR experience as follows:

- Clinical Experience: *High experience*—above the median years out of medical school (graduated 1978 or earlier). *Low experience*—below the median years out of medical school (graduated after 1978).

- QMR Experience: *High experience*—1) Purchased QMR between six months and one year ago and uses it at least once a week *or* 2) purchased QMR more than a year ago and uses it at least once a month. *Low experience*—1) Purchased QMR less than six months ago *or* 2) purchased QMR between six months and one year ago and uses it less than once a week *or* 3) purchased QMR over a year ago and uses it less frequently than once a month.

Within a cell, physicians were stratified by specialty to represent the proportions in the total sample of respondents. Also, each subject in each of the four cells was randomly assigned to one of three case sets. To control for the possibility that the physicians might become more skillful at using QMR across the eight cases, the order of the presentation of the cases was balanced across participants. Within each cell of the study, each physician received the cases in a different

order. The sequencing of the cases was arranged so that no case type was presented more often early in the sequence than other case types. When an initially selected participant failed to complete the cases, that participant was replaced randomly by another respondent with similar specialty, clinical, and QMR experience. To ensure that all the participants used the same version of QMR, replacement stopped when a new version of QMR was distributed, which resulted in 108 participants' completing the cases. Three performance scores were calculated on the basis of the physicians' differential diagnosis lists.

Cases were selected to represent both ends of the spectrum of case difficulty and quality of the information that QMR provided. The physician sample, case selection process, and method of calculating the performance scores are described in detail below.

Physician Sample

The recruitment letter and background questionnaire were sent with a routine update of QMR to all of the approximately 2,100 QMR users. A total of 254 respondents returned the questionnaire, describing their specialty, board certifications, practice setting, clinical experience, frequency of QMR use, and comfort with using 12 major QMR functions. Six respondents did not identify a specialty. Of those who did, 65 percent were in internal medicine, 31 percent were in family medicine, and the rest were in other specialties. Thirteen respondents who were not currently involved in direct patient care or who lived outside North America were excluded.

Clinical Cases

The cases were selected from 105 cases that had been previously classified by an expert panel in terms of difficulty and on which there were performance data for QMR and three other DDSSs.[23,24] Each case was classified with respect to its difficulty and the likelihood of QMR providing high- or low-quality information.

Primary Categorization of Cases

Information Quality. In this study, information quality refers to the appropriateness of the system's diagnostic suggestions for given case data. Cases with potentially high information quality were those for which, in the previous study,[23,24] QMR had provided the correct diagnosis within the top ten diagnoses and three or more of QMR's top five diagnoses were considered relevant by the DDSS test committee. These cases were labeled *high-information-quality* cases, since QMR had the correct case diagnosis in its knowledge base and also provided highly relevant information. Cases with potentially low information quality were ones not only in which the correct case diagnosis was absent from QMR's knowledge base but for which QMR had, in the previous study, produced "irrelevant" diagnosis lists—that is, lists on which fewer than four of the top five diagnoses were considered relevant by the expert panel. These cases were labeled *low-information quality* cases, since the correct case diagnosis was not in QMR's knowledge base and, in addition, QMR tended to provide irrelevant information.

Because the participants were allowed freedom in how they used QMR, QMR's performance could differ from its performance in the previous study,[23,24] in which all case data were entered in a standard way and a single "first-pass" case analysis was the only QMR function used. Thus, the likelihood that QMR would produce better information on the high-information-quality cases than on the low-information-quality cases was high, but not 100 percent. Similarly, while the correct case diagnosis could not be suggested by QMR on the low-information-quality cases, it is possible that QMR could produce more relevant diagnoses than it did in the previous study.

Categorization of Case Difficulty. All 105 cases had presented diagnostic challenges, but the cases represented varying levels of difficulty. For the purposes of the present study, we used a combination of expert committee judgment and DDSS performance criteria to identify comparatively easy or difficult cases. *Easy* cases were those that had been previously classified by the expert panel as typical presentations. Either they were common cases or they were rare cases for which all three other DDSSs had, nevertheless, included the correct diagnosis as one of their suggestions. *Difficult* cases were those for which at least one of the other three DDSSs failed to include the correct diagnosis on its list of suggestions and that had been classified as rare, or they were atypical presentations, or they were complex cases with multiple diseases presenting simultaneously.

Our categorization method for both information quality and difficulty had dual criteria built into the classifications. To be selected as either high or low information quality or difficulty, a case had to meet both criteria. Cases that met one but not both of the criteria for either high or low information quality or difficulty were considered ambiguous as to their classification and were excluded. For example, a case with the correct diagnosis in QMR's knowledge base would meet

one of the high-information-quality criteria, but if the correct diagnosis was far down on QMR's list, it would fail to meet the second criterion for high information quality. Another example is a rare, atypical, complex case (meeting one criterion for a difficult case) for which all four DDSSs had suggested the correct diagnosis, which would exclude the case from the difficult category. Sixty of the original 105 cases that did not clearly fit into either the difficulty or information quality categories were excluded.

Confirmatory Categorization of Cases

The primary method of case categorization described above relied on specific criteria for defining difficulty and information quality. Since these criteria were unique to the particular data set, we also assessed case difficulty and quality of QMR information with the more common, but more subjective, method of expert opinion.

In the previous study,[23,24] from which the cases were taken,[23,24] QMR's performance on each case had been judged by the individual case authors in terms of the accuracy, relevance, and comprehensiveness of QMR's diagnostic suggestions. From the 45 cases that were left after excluding the 60 cases, the authors' QMR information ratings were used to select 12 high-information-quality cases with similarly high ratings, and 12 low-information-quality cases with similarly low ratings. In each group of cases, half were easy cases and half were difficult cases, according to the previously described categories.

To reconfirm the difficulty categorization, two members of the original expert panel independently judged the difficulty of the 24 cases on a five-point scale, on which a score of 1 indicated very easy and 5 very difficult. The average of their ratings was used as a further check on the homogeneity of the case difficulty classifications for the high-information-quality and low-information-quality cases.

The mean difficulty ratings assigned by the judges to the easy high-information-quality cases was 2.3, and the mean for the easy low-information-quality cases was 2.7. The mean difficulty ratings on the five-point scale for the difficult high-information-quality cases and the difficult low-information-quality cases were 3.9 and 3.7, respectively. Analysis of variance showed that the difference in mean ratings between the difficult and easy cases was significant ($F = 24.4$, $P < 0.001$). There was no significant difference in difficulty ratings between the high-information-quality and low-information-quality cases, nor was there any significant interaction between the two factors. The case difficulty ratings assigned by the two judges who confirmed the initial difficulty classifications were moderately but significantly correlated ($r = 0.54$, $P = 0.007$).

Performance Scores

Within cases, diagnoses were aggregated across participants and, for each of the 24 cases, a subset of the expert panel from the previous study[23,24] judged the appropriateness of each diagnosis that was included in at least one subject's differential diagnosis. The panel reviewed the cases in random order, and panel members were blinded as to the category into which the cases were classified and how many participants had generated a particular diagnosis for the case. As in the previous study, the correct case diagnosis and other diagnoses that were considered an appropriate interpretation and synthesis of the case data were classified as appropriate. Several performance scores were generated for each subject:

- *Accuracy*. The mean diagnostic accuracy score for each physician was computed as the proportion of cases for which the correct case diagnosis was listed on the physician's differential diagnosis list.
- *Relevance*. The diagnostic relevance score for a physician on a particular case was computed as the proportion of diagnoses on a physician's list that were considered appropriate for that case. Mean relevance scores were the means of the subject's individual case relevance scores.
- *Comprehensiveness*. The diagnostic comprehensiveness score for each case was computed as the proportion of appropriate diagnoses for a particular case that the physician included on the differential diagnosis. Appropriate diagnoses included all diagnoses from the previous study that were judged appropriate for consideration and any new diagnoses, suggested by one or more participants, that the expert panel classified as appropriate for the case. Mean comprehensiveness scores were the means of the subject's individual case comprehensiveness scores.

The mean diagnostic accuracy, relevance, and comprehensiveness scores were, thus, proportions that could range from 0 to 1.

Statistical Analysis

The primary dependent variables in the study analyses were the mean accuracy, relevance, and comprehensiveness scores for each physician. The sole units of analysis in this study were physicians and not cases. To examine the possibility of selection bias, dif-

ferences in background characteristics between selected and unselected study subjects were tested using independent groups *t*-tests. An analysis of variance approach was used to examine the influence of case difficulty and quality of QMR's suggestions on the means of the dependent variables. Preliminary analyses were conducted to ensure that the assumptions of the analyses of variance were appropriate. A three-factor analysis of variance was used. It included the two within-groups case factors—case difficulty (easy or difficult) and quality of QMR information (high information quality or low information quality)—and one nested random factor, case set (A, B, or C). All main effects and interactions were tested using the multivariate analysis of variance procedure of the SPSS statistical analysis program.[33] The two-sided criterion for significance testing was set at alpha = 0.05. SPSS software was used for all analyses.[33]

Informed Consent

Informed consent of the participants was obtained, and this study was approved by the University of Alabama at Birmingham Institutional Review Board.

Results

Response Rate and Respondent Characteristics

A total of 120 initially selected and 70 replacement participants were offered the opportunity to participate, and 108 completed the cases. All but one participant, who had one incomplete case, provided usable data on all the cases. However, it was discovered that 12 participants had used an outdated version of QMR. These participants were asked to redo their cases using the current version, and 9 of the 12 complied. Because there were almost no changes in the diagnoses of the nine participants who redid their cases using the new version, all 12 have been included.

Table 1 shows the background characteristics and self-reported QMR usage of the 133 eligible respondents who were either not invited to participate or did not complete the study, compared with the 108 subjects who completed the study. The participants who completed the study were similar in most respects to the nonparticipants. There were no significant differences in the variables used to define QMR experience and

Table 1 ■

Characteristics of Eligible Selected and Unselected Subjects

Sample Characteristics	Selected (n = 108)	Unselected (n = 133)	P Value
Stratification variables:			
Mean (SE) year of medical school completion	1976 (0.78)	1978 (0.86)	0.06
Percentage with internal medicine specialty	64	65	0.84
Percentage with family medicine specialty	33	31	0.71
Mean (SE) frequency of use of QMR in last 6 mo†	3.75 (0.15)	4.07 (0.13)	0.11
Mean (SE) length of time using QMR‡	2.30 (0.08)	2.42 (0.07)	0.25
Other demographic variables:			
Mean (SE) year of primary residency completion	1980 (0.86)	1983 (0.70)	0.02*
Percentage general specialty board certified	86	78	0.11
Percentage general board eligible	9	13	0.36
Percentage of participants who reported "comfort" using specific QMR functions:			
Exploring QMR disease profile/associated disorders	89	86	0.47
Case analysis	61	56	0.46
Asserting diagnosis	42	33	0.17
Work-up protocol	58	53	0.44
Critiquing a case	34	24	0.09
Differential diagnosis	94	86	0.02*
Comparing two diseases	45	31	0.02*
Questions for a particular diagnosis	64	50	0.04*
Rule-in/rule-out diagnoses	73	61	0.04*
Saving a case	63	41	0.001*
Saving a case to a text file	49	26	0.000*
Printing a window	61	41	0.001*
Mean (SE) number of QMR functions participants were comfortable using	7.34 (0.33)	5.88 (0.30)	0.001*

*T-test, $P < 0.05$ considered significant difference between groups.
†Ordinal scale with seven categories (1, use every day, to 7, have not used during last six months).
‡Ordinal scale with three categories (1, purchased QMR within last 6 mo; 2, purchased QMR 6 to 12 mo ago; 3, purchased QMR more than a year ago).

Table 2 ■

Physicians' Diagnostic Performance When Using the Quick Medical Reference (QMR) Decision Support System (N = 108)

Type (No.) of Cases	Accuracy Mean (SD)	Relevance Mean (SD)	Comprehensiveness Mean (SD)
Difficult (4)	0.32* (0.18)	0.56* (0.15)	0.22* (0.08)
Easy (4)	0.67 (0.18)	0.67 (0.18)	0.25 (0.08)
High information quality (4)	0.75* (0.18)	0.68* (0.15)	0.25* (0.09)
Low information quality (4)	0.24 (0.20)	0.55 (0.18)	0.22 (0.08)
Difficult, high information quality (2)	0.59 (0.33)	0.64 (0.18)	0.26** (0.12)
Difficult, low information quality (2)	0.05 (0.15)	0.48 (0.23)	0.19 (0.09)
Easy, high information quality (2)	0.92 (0.18)	0.73 (0.22)	0.24 (0.09)
Easy, low information quality (2)	0.43 (0.33)	0.62 (0.19)	0.25 (0.12)
Total (8)	0.50 (0.14)	0.61 (0.14)	0.23 (0.06)

*$P < 0.01$, analysis of variance, significant main effect of case difficulty and information quality.
**$P < 0.01$, analysis of variance, significant interaction effect.

clinical experience between those who completed the study and the nonparticipants. However, the average year of completion of residency was almost three years earlier for subjects who completed the study than for the nonparticipants. Also, there were no differences in self-reported comfort with using five of the QMR functions, but significantly higher percentages of participants than nonparticipants said they felt confident using the remaining seven functions. In addition to the data shown in Table 1, there were no statistically significant differences between subjects and nonparticipants in their estimates of average number of patients, time spent in patient care, and usual amount of time spent using QMR.

Diagnostic Performance

The mean length of the participants' differential diagnosis lists was 5.21, with a standard deviation of 2.56. The mean number of diagnoses was significantly ($P < 0.01$) higher for difficult cases compared with easy cases, but there was no significant difference in the number of diagnoses listed between the high-information-quality and low-information-quality cases.

To check for possible sources of confounding of the performance scores, we first performed preliminary tests to determine whether the case sets, the order of cases, or the background characteristics of the study physicians were associated with different levels of diagnostic performance. The correlations (r = 0.15, r = 0.12, r = 0.14, respectively) of the physicians' years since being awarded an MD degree and their accuracy, relevance, and comprehensiveness scores were low and not significantly different from zero ($P < 0.05$, Fisher test). There were no differences among the means of any of the three performance scores associated with case order, physician specialty (internal medicine versus other), or physician experience. The only significant association was case set, with the mean relevance score indicating that one of the sets produced higher relevance scores than the others.

The means and standard deviations of the accuracy, relevance, and comprehensiveness scores stratified by case difficulty and quality of QMR information are presented in Table 2. In terms of physicians' performance, the results showed that the means of all three performance scores (accuracy, relevance, and comprehensiveness) were significantly higher ($P < 0.01$) for easy cases than for difficult cases. These three performance scores were also each significantly higher ($P < 0.01$) for the high-information-quality cases than for the low-information-quality cases. The results also indicated a significant interaction effect on the mean comprehensiveness scores. Further inspection revealed that the positive effect of the high- compared with the low-information-quality cases on the mean comprehensiveness scores was greater for the difficult cases than for the easy ones.

The better performance overall, and the higher accuracy scores in particular, on the high-information-quality cases suggest that the DDSS served a prompting function by reminding physicians of the correct diagnoses on these cases. However, it is important to examine alternative explanations. For instance, despite the attempt to match the difficulty of the high- and low-information-quality cases, it is possible that the high-information-quality cases were in some unknown way easier than the low-information-quality cases.

A more precise test of the prompting effect was possible, since we could examine performance on the *same cases* when the correct diagnosis was, or was not, displayed by QMR. A review of the participants' data on their interaction with QMR (saved as text files) revealed variability in the case data selected as relevant, in the particular QMR terms selected, and in the specific QMR functions used. Thus, QMR did not always suggest the correct diagnosis even on the high-information-quality cases, for which the correct case diagnosis was in QMR's knowledge base. In terms of QMR's ability to provide a reminder of the correct diagnosis, the high-information-quality cases for which the correct diagnosis was not displayed could be considered similar to the low-information-quality cases. The difference in diagnostic performance when the correct case diagnosis was, or was not, displayed was then compared for the high-information-quality cases only. To use the within-physician analysis to control for physicians' diagnostic skill, only physicians who had experienced both types of events could be included. For 43 of the 108 physicians, QMR displayed the correct diagnosis on some (either one, two, or three) but not all of the four high-information-quality cases. For this subgroup of physicians, the mean accuracy scores (0.91 versus 0.34), the mean relevance scores (0.69 versus 0.60), and the mean comprehensiveness scores (0.30 versus 0.21) were significantly higher ($P < 0.01$, paired t-test) for cases for which the correct diagnosis was displayed by QMR than for cases for which it was not shown. There were too few eligible physicians to perform reliable tests on easy or difficult cases separately.

Another possible explanation for the better performance on the high-information-quality cases is that better diagnosticians might be able to identify the correct diagnosis prior to using QMR or might be more likely to enter appropriate case data, or both. Thus, rather than QMR's prompts leading to better physician performance, the more diagnostically astute physicians might be better able to direct QMR to the correct diagnosis. If this were the case, those physicians should also have higher accuracy scores on the low-information-quality cases than the other physician participants.

To test this hypothesis, we examined whether getting QMR to display the correct diagnosis on more of the high-information-quality cases was associated with higher diagnostic performance on low-information-quality cases. Using the physician as the unit of analysis, the correlation between the number of high-information-quality cases for which QMR displayed the correct diagnosis and each of the low-information-quality diagnostic performance scores was calculated. The respective correlation coefficients ($r = 0.12$, $r = -0.04$, and $r = 0.10$ with low-information-quality accuracy, relevance, and comprehensiveness, respectively) were each low and not significantly different from zero ($P < 0.05$, Fisher test). The results of analyses examining the effect of the display of the correct diagnosis on the QMR screens on diagnostic performance with an adjustment for accuracy on low-information-quality cases were similar to results of the above analyses in which the adjustment was excluded. The consistency of results produced from a variety of analyses strongly supported the finding that the quality of the information displayed by the DDSS influenced physicians' performance.

Discussion

The most important finding, which confirmed our main hypothesis, was that the performance of physicians using the DDSS is likely to be better on the cases where the DDSS provides better information. Even within the high-information-quality cases, the physicians were more likely to include the correct diagnosis when QMR displayed it. Difficulty differences between the high-information-quality and low-information-quality cases, ability differences between the physicians who are able, or unable, to direct QMR to produce the correct diagnosis, or the reverse effect of knowing the correct diagnosis leading to data selection so that QMR displays it, did not appear to fully explain the results. Other findings from this study—e.g., that physicians performed better on the cases judged a priori as easier—were not unexpected but did serve to confirm our case selection criteria.

The results support the idea that a DDSS can perform less than perfectly and still assist physicians. In the present study, the average length of the participants' differential diagnosis lists was approximately five diagnoses per case. The participants included the correct diagnosis on half the cases (mean accuracy score) and, on average, 61 percent of the diagnoses on their differential lists were considered appropriate (mean relevance score). The physician participants in the present study selectively used the DDSS suggestions to develop a shorter and more focused differential diagnosis than that produced by the DDSS alone in previous studies, in which the average length of the diagnosis list was 21 and less than half the suggested diagnoses were relevant.[23,24]

The results of this study also support the idea that physicians can utilize helpful DDSS suggestions even

when other irrelevant suggestions are also provided, that DDSS can prompt physicians to consider diagnoses that they might not otherwise consider, and that the use of a DDSS can improve diagnostic performance, especially in difficult clinical cases.

The authors acknowledge the contribution of Alwyn A. Shugerman, MD, who participated in the committee that reviewed the appropriateness of the diagnoses. The authors also appreciate the contribution of the participants in our study who used QMR and provided us with their results.

References ■

1. Berner ES (ed). Clinical Decision Support Systems: Theory and Practice. New York: Springer-Verlag, 1999.
2. Bleich HL. Computer evaluation of acid-base disorders. J Clin Invest. 1969;48:1689–96.
3. Barness LA, Tunnessen WW Jr, Worley WE, Simmons TL, Ringe TBK Jr. Computer-assisted diagnosis in pediatrics. Am J Dis Child. 1974;127:852–8.
4. O'Shea JS. Computer-assisted pediatric diagnosis. Am J Dis Child. 1975;129:199–202.
5. Swender PT, Tunnessen WW Jr, Oski FA. Computer-assisted diagnosis. Am J Dis Child. 1974;127:859–61.
6. Wexler JR, Swender PT, Tunnessen WW Jr, Oski FA. Impact of a system of computed-assisted diagnosis: initial evaluation of the hospitalized patient. Am J Dis Child. 1975;129: 203–5.
7. Waxman HS, Worley WE. Computer-assisted adult medical diagnosis: subject review and evaluation of a new microcomputer-based system. Medicine. 1990;69:125–36.
8. Georgakis DC, Trace DA, Naeymi-Rad F, Evens M. A statistical evaluation of the diagnostic performance of MEDAS: the medical emergency decision assistance system. Proc 14th Annu Symp Comput Appl Med Care. 1990:815–9.
9. Nelson SJ, Blois MS, Tuttle MS, et al. Evaluating RECONSIDER: a computer program for diagnostic prompting. J Med Syst. 1985;9:379–88.
10. Hammersley JR, Cooney K. Evaluating the utility of available differential diagnosis systems. Proc 12th Annu Symp Comput Appl Med Care. 1988:229–31.
11. Feldman MJ, Barnett GO. An approach to evaluating the accuracy of DXplain. Comput Methods Programs Biomed. 1991;35:261–6.
12. Heckerling PS, Elstein AS, Terzian CG, Kushner MS. The effect of incomplete knowledge on the diagnosis of a computer consultant system. Med Inform. 1991;16:363–70.
13. Lau LM, Warner HR. Performance of a diagnostic system (Iliad) as a tool for quality assurance. Comput Biomed Res. 1992;25:314–23.
14. Bouhaddou O, Lambert JG, Morgan E. Iliad and Medical HouseCall: evaluating the impact of common sense knowledge on the diagnostic accuracy of a medical expert system. Proc Annu Symp Comput Appl Med Care. 1995;742–6.
15. Bankowitz RA, Lave JR, McNeil MA. A method for assessing the impact of a computer-based decision support system on health care outcomes. Methods Inf Med. 1992;31:3–11.
16. Bankowitz RA, McNeil MA, Challinor SM, Miller RA. Effect of a computer-assisted general medicine diagnostic consultation service on housestaff diagnostic strategy. Methods Inf Med. 1989;28:352–6.
17. Berman L, Miller RA. Problem area formation as an element of computer aided diagnosis: a comparison of two strategies within quick medical reference (QMR). Methods Inf Med. 1991;30:90–5.
18. Middleton B, Shwe MA, Heckerman DE, et al. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base, part II: evaluation of diagnostic performance. Methods Inf Med. 1991;30:256–67.
19. Miller RA, Pople HE Jr, Myers J. Internist-I, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med. 1982;307:468–76.
20. Miller RA, Masarie FE Jr. The quick medical reference (QMR) relationships function: description and evaluation of a simple, efficient ''multiple diagnoses'' algorithm. Medinfo. 1992:512–8.
21. Miller R, McNeil M, Challinor S, Masarie F, Myers J. The Internist-1/Quick Medical Reference Project: status report. West J Med. 1986;145:816–22.
22. Sumner W II. A review of Iliad and Quick Medical Reference for primary care providers: two diagnostic computer programs. Arch Fam Med. 1993;2:87–95.
23. Berner ES, Webster GD, Shugerman AA, et al. Performance of four computer-based diagnostic systems. N Engl J Med. 1994;330:1792–6.
24. 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.
25. Murphy GC, Friedman CP, Elstein AS, et al. The influence of a decision support system on the differential diagnosis of medical practitioners at three levels of training. Proc AMIA Annu Fall Symp. 1996:219–23.
26. Elstein AS, Friedman CP, Wolf FM, et al. Effects of a decision support system on the diagnostic accuracy of users: a preliminary report. J Am Med Inform Assoc. 1996;3:422–8.
27. Wolf FM, Friedman CP, Elstein AS, et al. Changes in diagnostic decision making after a computerized decision support consultation based on perceptions of need and helpfulness: a preliminary report. Proc AMIA Annu Fall Symp. 1997:263–7.
28. Bacchus CM, Quinton C, O'Rourke K, Detsky AS. A randomized cross-over trial of quick medical reference (QMR) as a teaching tool for medical interns. J Gen Intern Med. 1994;9:616–21.
29. Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. JAMA. 1998;280:1339–46.
30. Kassirer JP. A report card on computer-assisted diagnosis: the grade, C. N Engl J Med. 1994;330:1824–5.
31. Berner ES. Paradigms and problem solving: a literature review. J Med Educ. 1984;59:625–33.
32. Quick Medical Reference (QMR). San Bruno, Calif.: First DataBank Corp, 1994.
33. Norusis MI. Statistical Package for Statistical Solutions Reference Guide. Chicago, Ill.: SPSS, 1990.