Editorial Comments JAMIA

Editorial Comments ■

Diagnostic Decision Support Systems: How to Determine the Gold Standard?

Eta S. Berner, EdD

■ **J Am Med Inform Assoc.** 2003;10:608–610. DOI 10.1197/jamia.M1416.

In 1996 in an editorial on evaluation of decision support systems, Miller proposed that the bottom line in evaluating clinical decision support systems (CDSSs) should be "whether the user plus the system is better than the unaided user with respect to a specified task. . . ."[1] Since 1996, several studies have examined that issue, and, yet, there is still disagreement on the way to operationalize Miller's proposition. In this issue of the *Journal*, Ramnarayan et al.[2] describe a variety of metrics to evaluate the performance of a new pediatric diagnostic program, ISABEL. In a previous issue, Fraser et al.[3] also described metrics to evaluate a heart disease program, the HDP. Both Ramnarayan et al. and Fraser et al. discussed how their measures compared with the earlier measures used by Berner et al.[4] and Friedman et al.[5] to evaluate other diagnostic programs.

Why should it be so difficult to agree on a reasonable metric for evaluating these systems? Those of us who have struggled with this issue in our research have come to appreciate some of the difficulties that may not be immediately obvious in the published literature, but are important to articulate. Many of these issues are not unique to the diagnostic programs, but are a challenge in evaluating any CDSS. However, diagnostic programs are particularly challenging because, as Ramnarayan et al. indicate, diagnostic programs should influence both the diagnosis and the management plans. With that in mind, and with Miller's injunction to focus on evaluating how the system and clinician work together, I would like to discuss the problems that arise with the different "gold standards" that researchers have used and also would like to offer suggestions for researchers and developers of diagnostic CDSS.

Producing the Correct Diagnosis

Most researchers have included in their metrics the production of the "correct" diagnosis by either the CDSS or the clinicians after using the CDSS. Many have looked at the rank of the correct diagnosis on the differential, assuming that more highly ranked is better. On an intuitive basis, the CDSS's "getting the right answer" should be a good standard to use to judge the quality of a CDSS, and failure to do so has led some to dismiss the worth of these systems.[6] However, there are problems with this criterion. It could as well be argued that this is not a good criterion, since a definitive correct diagnosis is not always needed to initiate workup or treatment. Also, as Ramnarayan et al. point out, in real life the information needed to be certain of the correct diagnosis is unlikely to be known at the time that decision support is sought. In addition, if the correct diagnosis is a very rare one, it is likely that other diagnoses will be in a more prominent position on the list of both the CDSS and the clinician. This leads to a paradox; the highly ranked diagnoses are likely to already be considered by the clinician, while the lower ranked ones may not seem credible.

Quality of the Differential

In recognition of some of the problems of relying entirely on the use of the correct diagnosis as a gold standard, Ramnarayan et al. and other researchers also have included a measure of the quality of the output of the CDSS, and/or the clinicians' differential, and have relied on expert opinion to determine the "goodness" of the differential. There are several problems with this approach. If the experts use the full case data with definitive test results to judge the quality of a differential when the user and/or the CDSS did not have all of that data, there is a risk of both hindsight bias and underestimation of the quality of the performance of the CDSS. Fraser et al. discussed this possibility in their study and also noted another problem, that there are often disagreements among experts. To avoid these problems, Ramnarayan et al. had the experts develop their own

Affiliation of the author: Department of Health Services Administration, School of Health-Related Professions, University of Alabama at Birmingham, Birmingham, Alabama.

Dr. Berner's research on diagnostic decision support systems has been funded by grant number R01 LM05125 from the National Library of Medicine.

Correspondence and reprints: Eta S. Berner, EdD, Professor, Health Informatics, Dept. of Health Services Administration, School of Health-Related Professions, Susan Mott Webb Nutr. Sciences Building, 1675 University Blvd., Room 544, University of Alabama at Birmingham, Birmingham, AL 35294-3361; e-mail: <eberner@uab.edu>.

Received for publication: 07/18/03; accepted for publication: 07/23/03.

differential diagnoses and judge the appropriateness of the diagnoses with only the initial data and without knowledge of the "correct" diagnosis. However, although Ramnarayan et al. note that the final case diagnosis was always included in the list of appropriate diagnoses and was almost always ranked highly, the reliance on the collective opinion of experts is not a substitute for definitive data.

Appropriate Management Suggestions

If diagnosis is an intermediate step toward appropriate patient management, maybe a focus on how a diagnostic decision support system influences the clinician's management is a better focus than simply focusing on the correctness or quality of the CDSS diagnostic suggestions or the clinician's diagnoses. Ramnarayan et al. have included a measure of management, as well as diagnostic, quality and found a moderately positive relationship between the two measures. While examining the impact of a diagnostic system on patient management is important to do, the same clinical scenarios (whether simulated, real, consecutive cases, or particular diagnostic challenges) may not be appropriate to adequately test both kinds of suggestions. For instance, my colleagues and I examined the impact of a clinician's considering the correct diagnosis on ordering the definitive diagnostic procedure. We found that for some cases, if the correct diagnosis were not considered, the correct procedure would not be done. For others, a diagnosis of "something weird neurologic" was sufficient to lead to ordering the computed tomography (CT) scan or magnetic resonance imaging (MRI), which would ultimately provide the diagnosis.[7] The neurologic case was very difficult diagnostically, but was not a sensitive measure of management appropriateness.

User Acceptance/Satisfaction

Some researchers consider users' own judgment of helpfulness of the CDSS as the appropriate metric to use to judge its worth. This measure, too, is fraught with difficulty. Less clinically sophisticated users may be the ones most in need of decision support and, in fact, may perceive the CDSS to be quite helpful, but they also may be the least able to accurately judge their own knowledge and the appropriateness of the CDSS suggestions.

Amount of Use of the CDSS

It has been suggested that users vote with their feet (or at least with their fingers) as to the usefulness of a CDSS, and that systems that are used frequently are the most helpful. Measures such as the number of users or frequency of use are difficult to use as criteria because of the infrequent occurrence of cases in clinical practice that are perceived to be diagnostically challenging. Further, the cases for which the system would be used in a live clinical setting are likely to be those that are particularly difficult and may not be the fairest test of the CDSS.

While any of the approaches discussed above have problems in being used as the sole gold standard for evaluation of CDSS, researchers, including Ramnarayan et al., Fraser et al., and others who have conducted systematic evaluations of diagnostic CDSS, have appropriately used various combinations of these approaches to provide a multifaceted picture of CDSS performance.

Interaction of the User with CDSS

However, there is another problem in evaluating CDSS performance that still makes Miller's criterion a challenge for evaluators. Because the output of the CDSS is a combination of the adequacy of the CDSS knowledge base, its inference engine, how the user interacts with the system, and the specific data that are entered, the user can affect the performance of the CDSS. Fraser et al. noted that "giving physicians the flexibility to enter cases in their own fashion. . .can lead to cases' being entered with insufficient or inaccurate data."[3] My colleagues and I have also found that a decision support system that performed well under ideal conditions when all the case data were entered, performed less well on the same cases when clinicians were free to choose which case data to enter and what system functions to use.[8]

Also, many of the CDSS are designed to be used interactively and iteratively to provide a variety of perspectives on the patient. If the user does not utilize the system in this interactive fashion (either because of an evaluation design that standardizes the evaluation conditions, or because of lack of time or knowledge of the system capabilities in field studies), the system will perform suboptimally.

Furthermore, the influence of the CDSS on the user depends on the user's ability to interpret the CDSS output. The CDSS might suggest the "correct" diagnosis, but the clinicians may not always agree with those suggestions.[8] Tsai et al.[9] found that nonexpert users of an electrocardiogram (EKG) interpretation system also tended to be influenced by incorrect computer interpretations. These issues are not unique to medical applications. Galletta et al.[10] examined the effect of the common word processing "spell checker" and found that under certain circumstances users did worse, not better, when they used the spell-checking software. The interaction of the user and the system in data entry and output interpretation make it especially challenging to address Miller's bottom line criterion.

Suggestions for Future Development

Given the challenges in developing an appropriate gold standard for evaluation of diagnostic CDSS, it may be useful for developers of the next generation of these systems to focus more attention on the intended use of the system as well as on the information presented to the user. Diagnostically challenging cases require reflection over a period of time, examination of the case data from many different perspectives, and rethinking the case as more information becomes available. Cases such as these are memorable precisely because such challenges do not occur frequently. For such cases, a stand-alone, unintegrated CDSS may be fine. However, given that users do not always appropriately recognize diagnostic challenges, a system that can review all patient cases might be preferable, but a standalone system that is designed for extensive and iterative user interaction would be unlikely to be used routinely. Clearly, a CDSS that obtains its input data from an electronic medical record and requires minimal data entry or interaction on the part of the user will be more easily integrated into a busy

clinician's workflow. However, even with automated data entry and limited interaction with the CDSS, there is also a cognitive burden in interpreting the output of the CDSS, especially if it produces a lengthy list of diagnostic suggestions, as many of the systems do. Simply truncating the list of suggestions, for the reasons mentioned above, in the discussion of the rank of the correct diagnosis, may not be appropriate.

One approach might be to develop a diagnostic CDSS that analyzed the case data to arrive at a differential diagnosis but displayed only for the user a much smaller list of workup or management strategies. Such a system might be easier for the users to process and the researchers to evaluate. If linked to an order entry system, the CDSS might send alerts when a procedure that could rule in or rule out a highly probable diagnosis were omitted. Rather than a lengthy list of possible neurologic diagnoses, for instance, the system would suggest further workup with the neurologic imaging studies, or, if those studies were already ordered, might alert the clinician to consider ordering a vitamin B_{12} assay if pernicious anemia were also a possibility. Such a system that displayed only general categories of workup or management suggestions, rather than a list of specific diagnoses, might also be more robust in terms of being less sensitive to incomplete or inaccurate data entry. Test cases to evaluate the system would be those in which failure to consider the correct diagnosis is most likely to influence management, rather than those that are diagnostically challenging, selected by the users, or routinely seen, as is typical of most of the studies testing the diagnostic systems. The full differential could be available for the user to review if desired, and further user interaction with the CDSS might also occur. The shorter list of workup/management suggestions would be more likely to be attended to, given clinicians' limited time for interaction with the system. This approach would not negate the importance of diagnostic decision support but would target its performance where it can make the most impact on the users and, ultimately, on the patient.

References ■

1. Miller RA. Evaluating evaluations of medical diagnostic systems. J Am Med Inform Assoc. 1996;3:429–31.
2. Ramnarayan P, Kapoor RR, Coren M, et al. Measuring the impact of diagnostic decision support on the quality of clinical decision making: development of a reliable and valid composite score. J Am Med Inform Assoc. 2003;10:563–72.
3. Fraser HSF, Long WJ, Shapur N. Evaluation of a cardiac diagnostic program in a typical setting. J Am Med Inform Assoc. 2003;10:373–81.
4. Berner ES, Webster GD, Shugerman AA, et al. Performance of four computer-based diagnostic systems. N Engl J Med. 1994;330:1792–6.
5. 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.
6. Kassirer JP. A report card on computer-assisted diagnosis—the grade: C. N Engl J Med. 1994;330:1824–5.
7. Berner ES, Miller MD, Maisiak RS, Randolph V. The impact of a decision support system on physician work-up strategies. Proc AMIA Fall Symp. 2000:968.
8. Berner ES, Maisiak RS, Heudebert GR, Young KR Jr. Clinician performance and prominence of diagnoses displayed by a clinical diagnostic decision support system. Proc AMIA Fall Symp 2003; in press.
9. Tsai TL, Fridsma DB, Gatti G. Computer decision support as a source of interpretation error: the case of electrocardiograms. J Am Med Inform Assoc. 2003;10:478–83.
10. Galletta DF, Durcikova A, Everard A, Jones B. Cognitive fit and an intelligent agent for a word processor: should users take all that advice? Proceedings of the 36th Annual Hawaii International Conference on Systems Sciences (CD-ROM), Jan 6–9, 2003, Computer Society Press.