Evaluating Evaluations of Medical Diagnostic Systems

System evaluation in biomedical informatics should take place as an ongoing, strategically planned process, not as a single event or a small number of episodes. Complex software systems and accepted medical practices both evolve rapidly, so evaluators and readers of evaluations face moving targets. Thus, it is crucial for readers to be able to place any individual evaluation study into proper perspective. This advice applies to the nascent technology of medical diagnostic decision support systems (MDDSS). That the editor of a prestigious medical journal judged this entire technology based on a single, well-done, but intermediate-level and partial evaluation of several systems coupled with his own anecdotal experience emphasizes our professional obligation to characterize such systems and their evaluations responsibly.

Initially, developers of broad-based medical expert consultation systems viewed their systems as free-standing “Greek oracles” that could be evaluated on their ability to solve difficult clinical problems autonomously when presented with evidence in the form of individual case findings. However, current thinking recognizes that such systems are of value only when they help users to solve users’ problems. Users, not systems, characterize and solve clinical diagnostic problems. The ultimate unit of evaluation should be whether the user plus the system is better than the unaided user with respect to a specified task or problem (usually one generated by the user). The well-written article by Elstein and colleagues in the current issue of JAMIA evaluates a broad-based, “general” (MDDSS), as opposed to narrow, “focused” MDDSS. That study appropriately compared user performance with and without system assistance, even though the tasks users performed were not generated from their own (or actual clinical) information needs. As a result, the clinical stimuli for users came from written case abstracts rather than from the ongoing care of live patients.

The nature and extent of an appropriate evaluation spectrum were characterized by Stead and his colleagues then serving on the National Library of Medicine Study Section. In attempting to apply Stead’s evaluation matrix (stage of system development versus type of evaluation methodology) to studies evaluating broad-based MDDSS including the present study, it is apparent that, under many circumstances, a third axis should be added to the matrix—that of the users’ intended form of system usage (or specific system function evaluated). The same MDDSS may serve as an electronic textbook for one user, as a diagnostic checklist generator for another user, as a consultant to determine the next useful step in a specific patient’s evaluation for a third user, and as a tool to critique or reinforce the users’ own preexisting hypotheses for a fourth user. Each system function would require a different form of evaluation whenever anticipated user benefits depend on which system function is used. Evaluations should clearly state which user objective is being studied and which of the available system functions are relevant to that objective. For example, the study by Berner et al evaluated the ability of several systems to generate first-pass differential diagnoses from a fixed set of input findings not generated by everyday clinical users. That approach was dictated by the desire to standardize system inputs and outputs for purposes of cross-system comparisons. All of the systems in that study were capable of generating questions to further refine the initial differential diagnoses, which is the intended mode of clinical use for such systems. Even though the evaluation did not examine this capability, the methods used by Berner et al. were sound. That study was not intended to produce a definitive rating or comparison of the systems themselves, so the involved systems were not placed in the hands of end users, nor were the systems used in a manner to address common end-user needs. Generating a differ-
ential diagnosis is a first step but is not useful clinically in either the purely human or computer-assisted contexts without subsequent evidence-gathering, reflection, and refinement. Berner's primary goal was to develop methods and metrics that would characterize aspects of system performance in a manner useful for rationally comparing different systems and their functions.

A checklist for placing an evaluation of an MDDSS into context should include as a minimum the following five viewpoints:

**Viewpoint 1: What is the current level of system maturity, and is the evaluation design appropriate for level of system maturity?** The discussion of Stead et al. explains this viewpoint. There are a number of relevant questions. Is the system a prototype, a fully developed commercial product, or something in between? What is the stage of development and maintenance of the program's underlying knowledge base? What evaluations have occurred during earlier stages of system development, and are those evaluations relevant to the current system configuration and to the current evaluation? Preliminary system evaluations (during early development) should explore feasibility, performance, reliability, and safety on an informal, almost anecdotal, level—"in vitro" testing outside of the environment intended for eventual usage, in a manner that does not put clinical subjects at risk (e.g., retrospective case analyses). Intermediate system evaluations should rigorously test system components and portions of overall system function. More mature systems should be evaluated more formally, initially in the context in which they were developed and eventually in the hands of external reviewers studying users who utilize the system in the manner intended for eventual widespread dissemination—"in vivo" testing. The latter process should involve rigorous study designs including, when relevant, randomized controlled trials.

**Viewpoint 2: What is the purpose of the study?** Identify the specific end-user information needs that the study investigates and which corresponding system functions or uses it evaluates. Are the questions being asked problems that clinical users generate during clinical practice, or are they artificial problems generated by the study design team? Is the case material accurately based on actual patient cases? Note that there can be no truly verifiable diagnosis when artificial, manually constructed or computer-generated cases are used. Are the evaluation subjects clinical users whose participation occurs in the clinical context of caring for the patients used as "test cases," clinical users evaluating abstracts of cases they have never seen, or, nonclinical personnel evaluating abstracted clinical cases using computer systems? Are users free to use all system components in whatever manner they choose, or is it likely that the study design will constrain users to exercise only limited components of the system?

**Viewpoint 3: What are the criteria for efficacy?** Are the criteria for "successful" system performance similar to what clinical practitioners would use or require during actual practice? Diagnosis, itself or, more properly, "diagnostic benefit" must be defined in such contexts because "near misses" raised as possibilities by a system may lead the user to order a definitive test that leads to the "correct" diagnosis. Similarly, what it means to establish a diagnosis must be carefully defined. For example, it is not adequate to accept hospital discharge diagnoses at face value as a "gold standard" because discharge diagnoses are not of uniform quality; they have been documented to be influenced by physician competency, coding errors, and economic pressures. Furthermore, some discharge diagnoses may be "active" (undiagnosed at admission and related to the patient's reason for hospitalization), while others may be relevant but inactive. Criteria for the establishment of a "gold standard" diagnosis should be stated prospectively before data collection in an evaluation study begins. This of note in the study by Elstein et al., which used discharge diagnoses as being definitive.

**Viewpoint 4: How well does the evaluation address the boundaries or limitations of the knowledge-base and available system functions?** The reader must determine the limits of the system, and ask whether they are evaluated in the study. A system may fail when presented with cases outside its knowledge base domain, but an evaluation may only use cases from within that domain—for example, Elstein and colleagues limited case material in the current study to those in the domain of the system's knowledge base; Berner did not. The limits of a system's knowledge base are a concern because patients do not accurately triage themselves to present to the most appropriate specialists. A patient with atypical appendicitis may present to an internist, and a patient with abdominal pain due to lead poisoning may first see a surgeon.

**Viewpoint 5: If the study outcome shows "lack of system effect," does the study design allow identification of the causes for this result?** The reader, as well as the authors of the evaluation study, should construct a model of all of the possible influences on the evaluation outcomes. These would include system-related factors (knowledge base inadequacies, inadequate synonyms within vocabularies, faulty algorithms, etc.), user-related factors (lack of training or experience with the
system, failure to use or understand certain system functions, lack of medical knowledge or clinical expertise, etc.), and external variables (lack of available gold standards, failure of patients or clinicians to follow-up during the study period). It is important to recognize that studies that focus on one aspect of system function may have to make compromises with respect to other system or user-related factors in order to have an interpretable result. Additionally, in any MDDSS evaluation, the user's ability to generate meaningful input into the system and the system's ability to respond to variable quality of input from different users are important concerns. Evaluations of MDDSS must each take a standard objective (which may be only one component of system function) and measure how effectively the system enhances users' performances—using a study design that incorporates the most appropriate and rigorous methodology relative to the stage of system development. The ultimate clinical end-user of a given system must determine if published evaluation studies examine the system's function in the manner that the user intends to use it. This is analogous to a practitioner determining if a given clinical trial (of an intervention) is relevant to a specific patient by matching the given patient's characteristics to the study's inclusion and exclusion criteria, population demographics, and the patient's tolerance for the proposed forms of therapy as compared with alternatives. The reporting of an individual "negative study" of system performance, should not, as it often does now, carry the implication that the system is globally suboptimal. A negative result for one system function does not mean that, for the same system, some users cannot derive significant benefits from other system functions. Similarly, complete evaluation of a system over time should examine basic components—e.g., the knowledge base, ability to generate reasonable differential diagnoses, ability to critique diagnoses, etc.—as well as clinical functionality: For example, can novice users, after standard training, successfully employ the system to solve problems that they might not otherwise solve as efficiently or completely? The field of MDDSS evaluation will become mature only when clinical system users regularly derive the same benefit from published MDDSS evaluations as they do from evaluations of more standard clinical interventions.

RANDOLPH A. MILLER, MD

References


Correspondence and reprints: Randolph A. Miller, MD, Informatics Center, Eskind Biomedical Library, Vanderbilt University Medical Center, 209 Garland Avenue, Nashville, TN 37232. e-mail: randy.miller@mcmail.vanderbilt.edu

Received for publication: 7/19/96; accepted for publication: 7/22/96.