JAMIA

Editorial **Comments**

Reference Standards in Evaluating System Performance

The paper in this issue by Hripcsak and Wilcox, "Reference Standards, Judges, and Comparison Subjects: Roles for Experts in Evaluating System Performance,"[1] is well written and presents a thoughtful analysis of the topic. As the authors acknowledge, however, there is more to the evaluation of clinical informatics systems than can be accomplished through comparison to experts.[2,3] Hripcsak and Wilcox focus on "how to use experts in evaluating systems when one needs them," whereas this commentary focuses on the question, "when should one use experts as part of a system's evaluation" The two perspectives are complementary rather than contradictory.

As noted previously,[4]

> System evaluation in biomedical informatics should take place as an ongoing, strategically planned process, not as a single event or small number of episodes. Complex software systems and accepted medical practices both evolve rapidly, so evaluators and readers of evaluations face moving targets. … [C]urrent 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.…

If the ultimate evaluation of a system depends on whether users of the system perform a specified task better when they use the system than when they don't, then there must be public, objective criteria (a "gold standard") made available before an evaluation begins, to determine the quality of performance of an individual on a task (independent of whether the individual uses a decision-support tool).

Hripcsak and Wilcox state that experts can be used in three evaluation settings: to generate, through introspection and expertise, the reference standard per se (e.g., by providing a list of "correct" diagnoses or of "correct" therapeutic interventions based on a reading of the problem at hand); to judge (and label) the individual behaviors of subjects in the study—on a scale ranging from "optimal" through "acceptable" to "inadequate"—without providing an absolute list of "correct answers"; and as actual subjects in the study, to make it possible to rate how well the system performs in comparison with the performance of human experts. Hripcsak and Wilcox's first two scenarios assume that no absolute, independent gold standard is available, so that the experts' opinions represent the next best metric; in their third scenario, a gold standard must exist against which both experts and study subjects are graded in performance. In each of these three settings, as in any formal, summative evaluation of a clinical informatics system, it is best to compare subjects' performances with and without the system, no matter what absolute metric of performance is used.

Hripcsak and Wilcox state,

> Experience shows that accurate reference standards rarely exist; if it were easy to obtain the correct response, a medical informatics system would be unnecessary.

However, while it is indeed difficult to find clinical settings in which absolute answers are available, it should be the goal of system evaluators to first ask the question, "Can we design an evaluation for the system that involves use of a reliable, objective, external gold standard?" For example, each clinicopathologic conference published in "Case Records of the Massachusetts General Hospital" in the *New England Journal of Medicine* involves a definitive procedure (laboratory test, biopsy, or autopsy) as a

"gold standard" to establish the patient's correct diagnosis. Some formative system evaluations of diagnostic systems in clinical informatics have used such retrospective published cases to evaluate systems because they provide an external "gold standard." By definition, such retrospective studies cannot test system performance "on the front lines" of clinical care provision. In situations in which a prospective, summative evaluation of a clinical diagnostic system is required, the evaluation should ideally be performed "on the front lines" at a time when assistance is truly required and no definitive answer is available.

Rather than using experts as a gold standard, however, it may often be possible to develop a protocol by which patients can be used as subjects when their diagnoses are unknown, but then patients are closely followed by protocol for a long time until a diagnosis is established by objective, predefined criteria.[5] If at the end of the follow-up interval, no diagnosis can be determined by the preset criteria, the case should be labeled as "unable to establish/confirm a diagnosis" and dropped from inclusion in the study. Only when no reliable, external gold standard can be identified should experts' opinions be used.

In the area of systems for therapy and prognosis, expert opinions may play a role when randomized controlled studies cannot be carried out. Each patient can only follow their own trajectory of responses to interventions, so if subjects are allowed to select from a set of potential interventions, even if "real" patient case data are used, one can only hypothesize that with a different intervention than was actually used in the case, the patient's outcome might have been different. Ideally, only randomized controlled trials matching patients in the intervention group to patients in the control group, with the objective of tracking and comparing specific outcomes, can determine whether clinicians using decision support tools provide "better" care than the same (or similar) clinicians without decision support tools. Such studies are difficult at best, requiring large numbers of clinicians and patients and long follow-up intervals. Matching physicians with like abilities in control and intervention groups is arduous; matching patients with "equivalent" degrees of "equivalent" illnesses between intervention and control groups is extremely difficult.

The use of experts can be misleading in the absence of a gold standard. Imagine a scenario in which patient case records are presented to students who are asked to provide diagnoses, and experts' opinions are sought for use as "gold standard" diagnoses. This may be appropriate if the evaluation of the students' diagnoses is aimed at probing their reasoning abilities. However, consider a different situation, in which instead of actual patient data, a computer-based diagnostic knowledge base is used to generate sample "patient cases." If a case with findings of "fever, arthralgias, skin rash, and abdominal pain" is presented, would one accept the disease template used to generate the findings, systemic lupus erythematosus, as being the "correct" diagnosis? What if an expert panel determined that with the same nonspecific findings, Lyme disease were the "best" diagnosis? In the absence of a pathognomonic weight of evidence, a "definitive" opinion by experts must be taken with at least one grain of salt, since a truly expert opinion would be that the weight of the evidence in the case could not lead to the conclusion of any specific diagnosis. Experts rarely offer such opinions when they are being consulted as experts.

In the current era of evidence-based medicine, the opinions of experts should be tempered by an attempt to measure the "weight of the evidence" that the experts interpret. Even human experts are susceptible to the "garbage in, garbage out" phenomenon.—RANDOLPH A. MILLER

References ■

1. Hripcsak G, Wilcox A. Reference standards, judges, comparison subjects: roles for experts in evaluating system performance. J Am Med Inform Assoc. 2001;9:1–15.
2. Stead WW, Haynes RB, Fuller S. Designing medical informatics research and library-resource projects to increase what is learned. J Am Med Inform Assoc. 1994;1(1):28–34.
3. Wyatt JC, Friedman CP. Evaluation Methods in Medical Informatics. New York: Springer, 1997.
4. Miller RA. Evaluating evaluations of medical diagnostic systems. J Am Med Informatics Assoc. 1996;3:429–31.
5. Bankowitz RA, McNeil MA, Challinor SM, et al. A computer-assisted medical diagnostic consultation service: implementation and prospective evaluation of a prototype. Ann Intern Med. 1989; 10:824–32.

Affiliation of the author: Informatics Center, Vanderbilt University Medical Center, Nashville, Tennessee.

Correspondence and reprints: Randolph A. Miller, MD, Associate Director of Informatics Center, Professor and Chair, Department of Biomedical Informatics, 436 Eskind Biomedical Library, 2209 Garland Avenue, Nashville TN, 37232.

Received for publication: 9/28/01; accepted for publication: 10/1/01.

Telehealth:

The Need for Evaluation Redux

Generally, we do not publish papers describing the background and methods for a research project until the results are available for inclusion. In the case of the papers by Shea and Starren in this issue, we decided to make an exception. Our decision reflects the following factors. First, the literature does not contain examples of adequate evaluation of telemedicine despite years of application of the technology and several calls to action. Second, this trial is a major effort, and the results will not be available for some time. Third, it is unlikely that multiple large-scale trials are underway in this area. Accordingly, we decided that access to the methods would inform the community of the type of research that is needed, regardless of the outcome of this specific trial. The urgent need for more evaluation argues against publication delay. The lack of competing parallel efforts limits the chance of introducing bias by early publication.—WILLIAM W. STEAD, MD

Five years ago, the journal published a set of five articles describing telehealth applications. Accompanying those papers was an editorial lamenting the lack of adequate evaluation for these studies.[1] This editorial appeared shortly after an Institute of Medicine report was published that reached the same conclusions about telemedicine in general.[2]

Unfortunately, this situation has not changed in the ensuing half-decade. Two years ago, concerned with political pressure to reimburse telemedicine services through Medicare despite an unclear picture about efficacy and cost-effectiveness, the Health Care Financing Authority, along with the Agency for Healthcare Research and Quality, awarded a contract to the Evidence-based Practice Center at Oregon Health & Science University, to prepare an evidence review on the efficacy of telemedicine interventions in terms of diagnosis, clinical outcomes, satisfaction, access to care, and cost. The original review assessed telehealth applications for the Medicare population,[3] while a supplemental study analyzed pediatric and obstetric population applications.[4]

Our conclusions from these reviews, which were exhaustive analyses of the peer-reviewed literature in telemedicine, echoed the previous observations—that while telemedicine research has led to novel and creative uses of the technology, the quality of the evaluation studies is poor. It is important to note that the major problems we found were with the methodologies of the studies. Thus, we were careful not to conclude that telemedicine technologies were not efficacious but rather that the low quality of studies assessing them precluded any conclusions about their efficacy. Examples of the problems we found included:

- Diagnostic efficacy studies in which the telemedicine and in-person assessments were performed by the same individuals
- A paucity of clinical outcome studies in clinical areas in which telemedicine is widely used
- Satisfaction studies with extremely low response rates and use of nonstandardized instruments
- Access studies that failed to use appropriate measures of access to care
- Cost studies that focused solely on the tradeoff of system cost vs. patient travel or emergency transport cost, ignoring the effects of adverse outcomes or the cost of the whole episode of care

In our reviews, we noted that medical informatics investigators have demonstrated for many years the capability to carry out well-designed studies assessing the application of information technology in the health care setting.[5–7] To the potential concern that the technology is changing too rapidly to achieve adequate research control conditions, we also stated that techniques (e.g., "tracker trials") have been developed to cope with clinical trials of changing technologies.[8] We also took journal editors to task for publishing papers that had well-written descriptions of systems and issues in the use of technologies but were marred by poor evaluation designs. Thus, we concluded that the telemedicine community still had not met the a challenge of defining the efficacious use and cost-effectiveness of their technologies.

This issue of the Journal features two papers describing a large-scale telemedicine project in New York.[9,10] The ongoing evaluation study is naturally of great interest to us. We are pleased that a large-scale evaluation has been incorporated into this project from the outset and note that its methodology is vastly superior to those of most of the studies we reviewed for our evidence report. However, we do have two concerns about the methodology that we hope the authors will address in this or future studies. The first is that the control intervention appears to consist of no intervention at all. Since the experimental intervention consists of both a telemedicine intervention and intensive nursing case management, a positive outcome of the study will not enable us to discern whether benefit accrued from the telemedicine intervention, the extensive case management, or both. Somewhat ironically, this issue has also plagued the various studies of the Diabetes Control and Complications Trial.[11] Although these experiments

were purported to demonstrate the beneficial effects of tight blood sugar control in diabetes mellitus, they may in reality have demonstrated the value of nursing case management. A better control group in the Shea et al. study would be one in which comparable nursing case management was also delivered by non-telemedicine means. This would enable us to assess the added value of telemedicine per se.

The second concern is a hope that Shea et al. will look beyond intermediate patient outcome measures of glycosylated hemoglobin and blood pressure to actual outcomes, such as development of complications, morbidity, and mortality. Although the measurement of intermediate outcomes provides more statistical power, the ultimate aim of health care interventions is to directly benefit patients, not improve their test results. A related problem with the intermediate outcome measures is their use in cost-effectiveness calculations. Although the current evaluation will allow this intervention to be compared with other interventions that affect the intermediate measures (i.e., other diabetes interventions), it will not allow comparisons that a policy maker might wish to make, such as comparison with treatments for osteoporosis. A better cost-effectiveness measure would be quality-adjusted life years.

Nonetheless, we applaud this large-scale trial and eagerly await the results. We hope the findings will not only enable us to get a better picture of the value of this form of telemedicine application for diabetic care but also serve as a means of expanding our knowledge about the contribution of control conditions for future evaluations. Such high-quality studies should also serve as a standard and raise the bar for the methodology of future evaluations in telemedicine and telehealth applications.—WILLIAM R. HERSH, MD, PATRICIA K. PATTERSON, RN, PHD, DALE F. KRAEMER, PHD

References ■

1. Masys DR. Telehealth: the need for evaluation. J Am Med Inform Assoc. 1997.4:69–70.
2. Field MJ (ed). Telemedicine: A Guide to Assessing Telecommunications for Health Care. Committee on Evaluating Clinical Applications of Telemedicine, Institute of Medicine. Washington, DC: National Academy Press, 1996.
3. Hersh WR, Wallace JA, Patterson PK, et al. Telemedicine for the Medicare Population. Rockville, Md.: Agency for Healthcare Research and Quality, July 2001. Evidence Report/Technology Assessment 24. AHRQ publication 01-E011. Available at: http://www.ahrq.gov/clinic/telemedsum.htm.
4. Hersh WR, Wallace JA, Patterson PK, et al., Telemedicine for the Medicare Population: Pediatric, Obstetric, and Clinician-indirect Home Interventions. Rockville, Md.: Agency for Healthcare Research and Quality, August 2001. Evidence Report/Technology Assessment 24 (supplement). AHRQ publication 01-E059. Available at: http://www.ahrq.gov/clinic/telmedsup.htm.
5. Tierney WM, Miller ME, Overhage JM, McDonald CJ. Physician inpatient order writing on microcomputer workstations: effects on resource utilization. JAMA. 1993;269:379–83.
6. Evans RS, Pestotnik SL, Classen DC, et al. A computer-assisted management program for antibiotics and other anti-infective agents. N Engl J Med. 1998;338:232–8.
7. Bates DW, Leape LL, Cullen DJ, et al. Effect of computerized physician order entry and a team intervention on prevention of serious medication errors. JAMA. 1998;280:1311–6.
8. Lilford RJ, Braunholtz DA, Greenhalgh R, Edwards SJ. Trials and fast-changing technologies: the case for tracker studies. BMJ. 2000;320:43–6.
9. Starren J, Hripcsak G, Sengupta S, et al. Columbia University's Informatics for Diabetes Education and Telemedicine (IDEATel) project: technical implementation. J Am Med Inform Assoc. 2002;9:25–36.
10. Shea S, Starren J, Weinstock RS, et al. Columbia University's Informatics for Diabetes Education and Telemedicine (IDEATel) project: rationale and design. J Am Med Inform Assoc. 2002;9:49–62.
11. Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. N Engl J Med. 1993;329(14):977–86.

Response from Dr. Shea:—Drs. Hersh, Patterson, and Kraemer appropriately highlight two limitations of the design for the IDEATel project. The first, boiled down, is whether the medium can be evaluated separately from the message. In designing IDEATel, we could not develop a practical, realistic way to separate the electronic medium for delivering diabetes care from the care itself. To do so would have required an artificial simulation of what electronically delivered care would be and then to have found an effective and methodologically convincing way to deliver this care non-electronically. This has never been done, it is not clear what this really means, and we did not believe we could do it successfully. For example, would each home telecare visit in the intervention group need to be matched by an in-person house call or by an in-person visit to a diabetes center in northern Manhattan or in Syracuse (nearly 800 miles for the most distant of the upstate region participants)? These two diabetes centers are where the

Affiliation of authors: Division of Medical Informatics and Outcomes Research, Oregon Health & Science University, Portland, Oregon.

Correspondence and reprints: William R. Hersh, MD, Professor and Head, Medical Informatics, Oregon Health & Science University, Mail Code BICC, 3181 SW Sam Jackson Park Road, Portland, OR 97201; e-mail: <hersh@ohsu.edu>.

Received for publication: 9/28/01; accepted for publication: 10/2/01.

intervention case managers are located. Our design choices were conditioned by what we believed we could do on a very large scale, with very short start-up time, and with very complex demands in terms of mounting the intervention technology. With these demands on the intervention side, it was desirable to keep the control side as simple as possible.

The second point is that an optimal design would focus on "actual" outcomes, such as cardiovascular events, amputations, and death, rather than intermediate outcomes such as blood pressure, glucose control, and lipid levels. This was not a viable option without a much larger sample size and longer follow-up time. Many clinicians would accept that improvement in these intermediate variables would clearly indicate benefit to patients.

We believe current and emerging technologies in medical informatics and telecommunication will alter not only the way care is delivered but what care is delivered, and that these two changes will occur in tight linkage. Telemedicine makes it possible to provide diabetes center-based case management to people who are not otherwise getting it. The most important factor driving the adoption of these technologies is that patients want to have access to health information, self-care resources, and the health care system electronically, remotely, and 24/7, which implies asynchronism. We hope that the IDEATel project will demonstrate that people now on the far side of the digital divide are there because of resources, not because of unwillingness or inability to use these technologies, and that these changes will be beneficial in terms of health outcomes and cost.—STEVEN SHEA, MD

Affiliation of author: Columbia University College of Physicians and Surgeons, New York, New York.

Correspondence and reprints: Steven Shea, MD, Division of General Medicine, 622 W. 168th Street, New York, NY 10032; e-mail: <ss35@columbia.edu>.

Received for publication: 10/1/01; accepted for publication: 10/2/01.