Research Paper ■

Measuring the Impact of Diagnostic Decision Support on the Quality of Clinical Decision Making: Development of a Reliable and Valid Composite Score

Padmanabhan Ramnarayan, MRCP(UK), MRCPCH, Ritika R. Kapoor, MRCPCH, Michael Coren, MRCP (UK), Vasantha Nanduri, MRCP (UK), MD, Amanda L. Tomlinson, RN (Child), Dip HND, Paul M. Taylor, PhD, Jeremy C. Wyatt, MRCP (UK), DM, Joseph F. Britto, MD

Abstract **Objective:** Few previous studies evaluating the benefits of diagnostic decision support systems have simultaneously measured changes in diagnostic quality and clinical management prompted by use of the system. This report describes a reliable and valid scoring technique to measure the quality of clinical decision plans in an acute medical setting, where diagnostic decision support tools might prove most useful.

Design: Sets of differential diagnoses and clinical management plans generated by 71 clinicians for six simulated cases, before and after decision support from a Web-based pediatric differential diagnostic tool (ISABEL), were used.

Measurements: A composite quality score was calculated separately for each diagnostic and management plan by considering the appropriateness value of each component diagnostic or management suggestion, a weighted sum of individual suggestion ratings, relevance of the entire plan, and its comprehensiveness. The reliability and validity (face, concurrent, construct, and content) of these two final scores were examined.

Results: Two hundred fifty-two diagnostic and 350 management suggestions were included in the interrater reliability analysis. There was good agreement between raters (intraclass correlation coefficient, 0.79 for diagnoses, and 0.72 for management). No counterintuitive scores were demonstrated on visual inspection of the sets. Content validity was verified by a consultation process with pediatricians. Both scores discriminated adequately between the plans of consultants and medical students and correlated well with clinicians' subjective opinions of overall plan quality (Spearman ρ 0.65, $p < 0.01$). The diagnostic and management scores for each episode showed moderate correlation ($r = 0.51$).

Conclusion: The scores described can be used as key outcome measures in a larger study to fully assess the value of diagnostic decision aids, such as the ISABEL system.

■ **J Am Med Inform Assoc.** 2003;10:563–572. DOI 10.1197/jamia.M1338.

Affiliations of the authors: Department of Paediatrics, St. Mary's Hospital, London, England (PR, MC, JFB); Department of Paediatrics, Princess Alexandra Hospital, Essex, England (RRK); Department of Paediatrics, Watford General Hospital, Watford, England (VN); ISABEL Medical Charity, London, England (ALT); Centre for Health Informatics and Multiprofessional Education, London, England (PMT); Klinische Informatiekunde (KIK), Academic Medical Centre, Amsterdam, The Netherlands (JCW).

The authors thank Jason Maude of the ISABEL Medical Charity, Dr. C. Edwards, and Dr. T. Sajjanhar for their ideas and support in the development of the scoring system. This study was supported by an evaluation grant from the National Health Service Research and Development Department (NHS R&D), London. An abstract of the scoring procedure was presented at the Paediatric Education Section of the Royal College of Paediatrics and Child Health Annual Meeting, York, UK, April 2003. Dr. Joseph Britto is a Trustee and Medical Adviser of the ISABEL Medical Charity (nonremunerative post). Amanda Tomlinson works for the ISABEL Medical Charity full time as research nurse.

Correspondence and reprints: Joseph F. Britto, MD, Department of Paediatric Intensive Care, 7th Floor, St. Mary's Hospital, South Wharf Road, London W2 1NY, England; e-mail: <j.britto@ic.ac.uk>.

Received for publication: 01/29/03; accepted for publication: 05/15/03.

Many computerized systems have been developed to assist physicians during diagnostic decision making (DDSS).[1–6] Although the benefits of providing diagnostic decision support in clinical practice have been closely examined,[7,8] few studies have been able to convincingly show changes in physician behavior or improved patient outcomes resulting from the use of DDSS.[9–11] This may have occurred for two reasons: first, the precise manner and clinical setting in which DDSS might help a physician remain unclear—as an "oracle" in the uncommon clinical scenario of a diagnostic dilemma or as a simple diagnostic reminder system in routine clinical practice.[12] Second, as a consequence of this lack of clarity, a number of heterogeneous outcome measures have been used to quantify the clinical benefits of DDSS.[13–18]

Early studies expected the DDSS to be able to predict the "correct" diagnosis in a diagnostic dilemma.[19] This was the "Greek oracle" model in which the user remained a passive recipient of DDSS advice.[20] These studies examined the "diagnostic accuracy" of the system functioning in isolation. A binary metric was commonly used—the system was accurate if it displayed the "correct" diagnosis and inaccurate if it did not.[14,21] More sophisticated measures of system

performance proposed by Berner et al.[14,22] also studied the ranking of diagnostic hypotheses in a system's list and other discrete indicators of diagnostic quality, such as relevance and comprehensiveness, generated by comparing the DDSS diagnostic hypothesis set to a "gold standard" set generated by expert clinicians. Subsequent evaluations of DDSS de-emphasised the value of testing the system only, and focused on examining the *impact* of a DDSS on the *user's* diagnostic plans.[23] This reflected the belief that the clinician would serve as an active cognitive filter of DDSS advice rather than remain a passive user during system consultation in real life.[12] In this setting, it was not essential that the system possessed a high degree of diagnostic accuracy, so long as its suggestions positively influenced users' diagnostic reasoning; the clinical impact of a DDSS was assessed by measuring changes in the diagnostic quality of the clinician to whom decision support was provided. Friedman et al.[24]described a composite score for this purpose. However, in general terms, most scoring schemes aimed to objectively measure the same concept—the quality of a diagnostic hypotheses plan—irrespective of whose efforts it represented (system or user). In Berner's study, numerous discrete indicators of quality were used; in Friedman's study, a single composite score was used.

As part of the assessment of impact of a free, Web-based differential diagnostic aid on clinical reasoning in an acute pediatric setting (ISABEL, <www.isabel.org.uk>, ISABEL Medical Charity, UK),[25–28] we sought an instrument to measure the quality of initial clinical assessment, consisting of diagnostic and management plans. The impact assessment was planned in two stages—a simulated study followed by a real life clinical trial; methods validated during the simulation could be used successfully in the clinical trial. ISABEL utilizes unformatted electronic, natural language text descriptions of diseases derived from standard textbooks as the underlying knowledge base; 3,500 disease descriptions are represented in its database. Commercial textual pattern-recognition software (Autonomy, <www.autonomy.com>) searches the underlying knowledge base in response to clinical features input in free text and displays diseases with matching textual patterns arranged by body system rather than in order of clinical probability. Thus, ISABEL functions primarily as a reminder tool to suggest 8 to 10 diagnostic hypotheses to clinicians, rather than acting as an "oracle." During initial system performance evaluation, the correct diagnosis formed part of the ISABEL reminder list on greater than 90% of occasions,[29] and *all* diagnoses judged to be appropriate for each case by an expert panel were displayed in 73% of cases (data awaiting publication). Berner's comprehensiveness score (proportion of appropriate diagnoses, as judged by the panel, included in the diagnostic plan under examination) applied to ISABEL in this study was 0.82.

For the purposes of our simulated impact evaluation, many scoring systems were considered as candidate outcome measures. We needed a composite score that took into account all pertinent factors that contributed to the quality of a diagnostic and management plan. Berner's comprehensiveness score as well as relevance score (proportion of suggestions in the diagnostic plan that the panel found reasonable to consider, including retrospectively) were not considered suitable: suggestions were not weighted based on how reasonable or appropriate they were (one highly appropriate suggestion and another less appropriate suggestion contributed the same value to the score). Friedman's composite score conceptualized diagnostic quality as having two primary components: a plausibility component derived from ratings of each individual diagnosis in a set (whether "correct" or "incorrect") and a location component derived from the location of the "correct" diagnosis if contained in the set. The composite score could not be computed without knowledge of a single "correct" diagnosis, which is usually not available in the acute medical setting for which ISABEL was designed. In this setting, an initial diagnostic plan is often generated with a dataset that includes only clinical history, an initial examination, and sometimes results from a set of "first-pass" investigations. It would be difficult to assign, or even expect, a single "correct" diagnosis at this stage, the emphasis being on considering the most appropriate set of diagnoses ("high frequency, highly plausible, high impact"). The location component of Friedman's score limited the maximum number of diagnostic suggestions to 6, making it difficult to assess comprehensiveness in this setting (>6 "appropriate" gold standard diagnoses may be appropriate for some cases depending on the level of uncertainty at initial assessment). In addition, in the Friedman score, although each case might have numerous highly plausible suggestions, a list containing only the "correct" diagnosis was assigned the highest score — comprehensiveness was not rewarded. The plausibility component, where appropriate suggestions are weighted on a 0 to 7 scale by the expert panel, could not be used without modification; the appropriateness of a diagnosis may be based on its plausibility, its likelihood, as well as implications for further test ordering.

Both scores also did not have a mechanism to measure the quality of management plans; we felt that the full clinical impact of any DDSS would be manifested in changes engendered in physician's diagnostic plans *as well as* in real changes made to the patient's treatment. The measurement of a DDSS' clinical impact had to be undertaken from two separate but related views: (1) changes in diagnostic plan *and* (2) changes in management plan. Since these changes were generated purely as a result of the provision of diagnostic support, this concept is different from measuring the clinical impact of systems primarily intended to provide decision support for test ordering,[30] antibiotic prescription,[31] and critiquing patient management[32,33] or integrated hospital information systems that offer advice on all of these functions.[34] Such systems also offered appropriate advice on relatively narrow areas of decision making such as antibiotic choice in infections or the management of hypertension, in which outcome selection was simpler. This study describes the development of a scoring metric for diagnostic and management plan quality and an examination of some of its measurement properties.

Aims

The aim of this study was to develop and validate a single diagnostic quality score (DQS) and a management quality score (MQS), to sensitively measure the quality of overall clinical assessment (differential diagnosis, investigations, and treatment) at a single discrete point during decision making.

Methods

The main focus of this study was measurement of the quality of the diagnostic and management plan provided by subjects, pre- and post-DDSS consultation, rather than the diagnostic suggestions presented by the DDSS itself. A brief description of our simulated impact evaluation is provided in this context. This study involved 76 subjects, representing each grade of hospital-based pediatrician in the United Kingdom —19 senior houseofficers (junior resident/intern level), 24 registrars (senior resident level), and 18 consultants (attending physicians)—as well as 15 final-year medical students. Twenty-four textual cases based on real-life acute pediatric presentations, representing 12 different subspecialities of pediatrics, and three levels of difficulty (1, unusual; 2, usual; and 3, common; identical allocation of level performed by the author and a consultant working independently) were used for the study. Initial presenting features were summarized in 100 to 200 words, avoiding the inclusion of only positive features and textual cues.

Although it had been possible to assign a final diagnosis for all children by the end of their hospital stay, this was not always feasible at the time of initial assessment with the amount of data available. Since the cases used in this study provided details of initial clinical assessment only, a single "correct" diagnosis was not assigned to any case. Subjects assessed each case on a special ISABEL trial website <http://trial.isabel.org.uk>, first without decision support from ISABEL, to create a differential diagnosis and management plan. Immediately, without reading the case again, minimizing the time that might contribute to a "second-look" bias, case-specific diagnostic suggestions from ISABEL were displayed for the clinician's consideration. The subject then was allowed to modify his or her initial diagnostic and management plan by adding or subtracting items. DQS and MQS were calculated for each subject's pre-ISABEL and post-ISABEL diagnostic and management plan, rather than for the ISABEL diagnostic suggestion list.

Factors Contributing to Plan Quality

These were derived from two sources: informal but focused discussions with clinicians at the registrar and consultant level using fictitious case examples and previously described discrete indicators of diagnostic quality. These concepts were combined to help create a single composite quality score. In essence, the overall quality of a diagnostic or management plan was hypothesized to depend on these factors:

1. *Appropriateness of each component suggestion*: How appropriate and relevant is each individual suggestion to the case?
2. *Contribution of each component suggestion to the plan:* How much should each suggestion contribute to overall plan quality? The most appropriate diagnosis (or the "correct" diagnosis, if there was one) should contribute heavily to a diagnostic plan, and less appropriate diagnoses should contribute less.
3. *Relevance:* How focused is the plan, does it contain large numbers of inappropriate and irrelevant items?
4. *Comprehensiveness:* How inclusive is the plan, does it contain *all* items judged to be appropriate and relevant?

Thus, the quality of a differential diagnosis and a management plan is seen to depend on two principal components—the quality of individual items constituting it and other factors applicable to the entire plan considered *in toto*. The appropriateness of each individual diagnostic suggestion during the initial clinical presentation of a patient depends not only on how well it explains the clinical findings (plausibility), but also on whether it is the most likely within the setting and its potential impact on further management (how treatable, how dangerous, prognosis, genetic implications). For example, the diagnosis of bacterial meningitis may not adequately explain the presenting features of fever, irritability, and red tympanic membranes in an infant, but its inclusion in the workup would be considered highly appropriate due to its life-threatening nature. In analogous fashion, the quality of each individual test or management step was hypothesized to depend on its appropriateness in the clinical scenario (how much value it adds to reaching a conclusive diagnosis), its impact (potential to cause clinical harm), and its cost–benefit ratio.

Scoring Procedure

The scoring was performed in the order in which the main factors of quality have been listed.

Step 1: Judging Appropriateness of Component Suggestions

This involved assigning each individual diagnostic and clinical management suggestion an appropriateness value (Figs. 1 and 2). This was performed by an expert panel of two general pediatricians, both of whom had at least three years' experience at consultant level (>10 years' experience in total).

Each panel member was provided all 24 cases (rather than just a summary of clinical features). Working independently, each suggested a list of "appropriate" diagnostic and management items for the cases. Neither panel member was privy to his or her colleague's suggestions. Although this list could have been used as a "gold standard" against which subjects' suggestions were scored, it was not. From previous studies, it

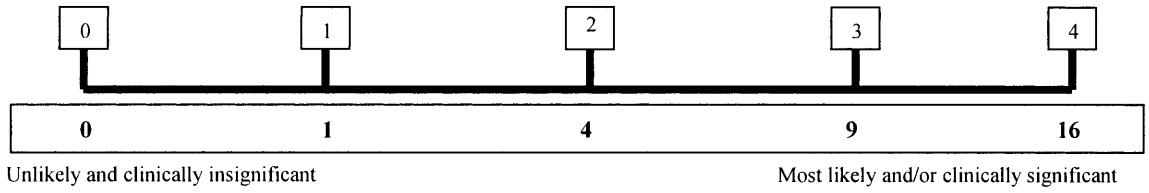

Figure 1. Scale for diagnostic suggestion ratings and proposed weighting scheme.

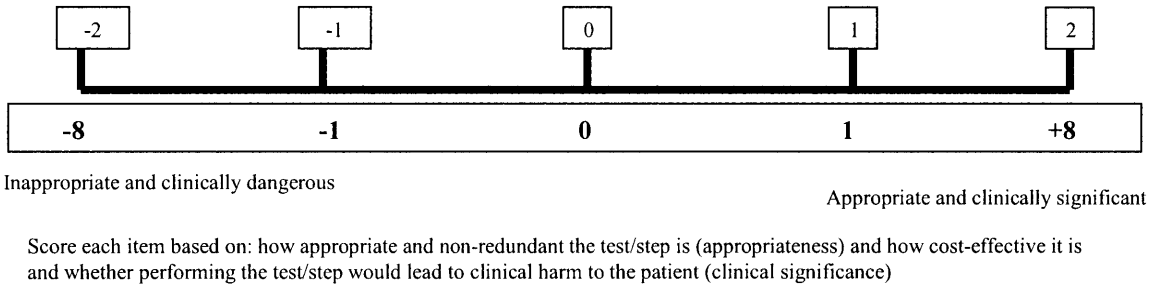

Figure 2. Scale for management suggestion ratings and proposed weighting scheme.

was obvious that some diagnoses suggested either by a DDSS or by subjects, but not present in the "gold standard" list, were also considered relevant by the panel on retrospective review.[14] For this reason, an aggregate list of unique diagnoses and management items was created. This was derived from two sources: suggestions from both panel members *and* suggestions from all subjects working their cases, provided either before or after decision support (Fig. 3). Suggestions provided exclusively by ISABEL, but not suggested by either the panel or the subjects, were not included in this aggregate list. This procedure served two purposes: first, since there was a large degree of duplication in the suggestions provided by subjects and panel members, the aggregate list was much shorter to examine; second, the aggregate list did not provide any information about the origin of each item, preventing bias in scoring.

Two weeks later, the panel worked independently and scored each item on the aggregate list for each case using the visual analog scale shown in Figures 1 and 2. Diagnoses considered to be the most likely explanation for the clinical features in the setting were assigned the highest rating on the scale (rather than a diagnosis that matched all clinical features). This rating could also be assigned to a less likely diagnosis, if it was so clinically significant that failure to consider it would imply clinical negligence. In some cases, diagnoses would satisfy both above criteria. Therefore, it was possible that more than one suggestion scored highest on the scale. Items thought to be irrelevant (but not dangerous) were assigned a score of 0 by the scorer. Other items were scored using these scores as anchors. Panel members' initial suggestions, provided in the first step prior to the creation of the aggregate list, would be expected to score highly on the scale, by their contributor as well as by the other panel member; this was examined to indicate intrarater reliability. In summary, each rater grouped all suggestions into five levels of appropriateness. The conclusive diagnosis (as established at the end of hospital stay for each case) was present in the aggregate list for all cases; the panel assigned them the highest score in 22 of 24 cases (92%), indicating a high degree of criterion validity for the panel itself.

This two-step procedure succeeded in assigning an appropriateness value to each suggestion. Following this, a list of diagnoses and management steps that scored greater than 0 on each scale was created for each case. This consisted only of appropriate suggestions, although some were less appropriate than others. This list was used as our "gold standard." Discrepancies between the panel members' scores were discussed at a single consensus meeting at the end of the entire scoring process to allow true differences in opinion to be separated from differences due to incomplete understanding of the scoring technique.

Step 2: Weighting the Contribution of Individual Suggestions to Overall Plan Quality

The contribution of highly appropriate suggestions to the overall plan must outweigh the contribution of less appropriate items; failure to consider an appropriate diagnosis should highly disadvantage the subject's diagnostic plan score. Therefore, the contribution of individual item scores

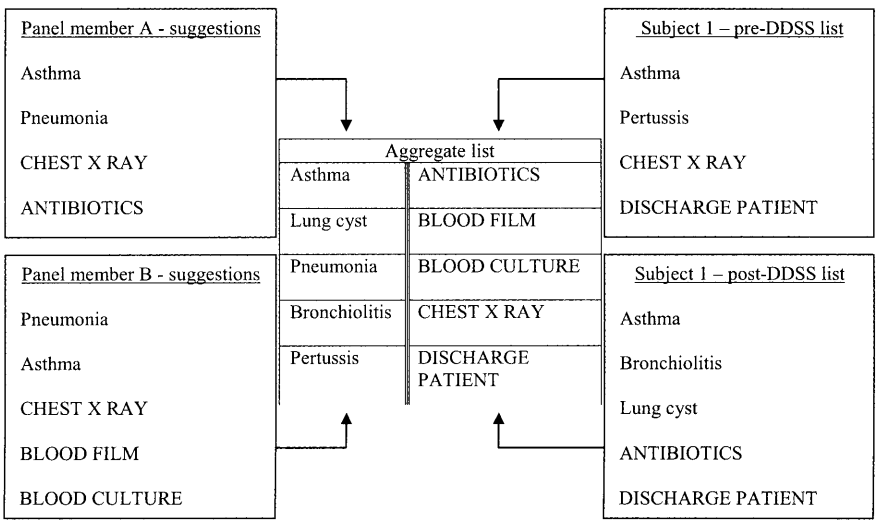

Figure 3. Creation of a fictitious aggregate list from panel members' *and* subjects' suggestions. Diagnostic suggestions are in Title case, management suggestions are in UPPER case.

Suggestions in plan (score assigned by panel)	Weighted score
Asthma (4)	16
Bronchiolitis (3)	9
Lung cyst (0)	0
Total weighted score (TWS)	25
Irrelevant suggestions scoring 0 (N_0)	1
DQS: (TWS-n_0) ÷ Gold standard score*	**0.50**

*Assuming a gold standard score of 48 (weighted sum of gold standard item ratings)

Suggestions in plan (score assigned by panel)	Weighted score
ANTIBIOTICS (+2)	+8
DISCHARGE PATIENT (-1)	-1
Total weighted score (TWS)	+7
Irrelevant suggestions scoring 0 (N_0)	0
MQS: (TWS-n_0) ÷ Gold standard score*	**0.25**

*Assuming a gold standard score of 28 (weighted sum of gold standard item ratings)

Figure 4. Example of scoring procedure for one subject's post-DDSS diagnostic and management plan.

to the overall score was weighted following an empirical weighting scheme (bottom panel, Figs. 1 and 2). The DQS and MQS were computed as a weighted sum of the ratings of the individual suggestions rather than as a simple arithmetic sum (total weighted score, TWS).

Step 3: Relevance

With the steps taken so far, two sets of diagnostic or management plans (one containing only appropriate suggestions and another containing similar appropriate suggestions as well as numerous other inappropriate or irrelevant suggestions) would score the same. To ensure that plans remained focused, one point was subtracted from the TWS for each irrelevant suggestion (n_0, items that scored 0) contained within the plan (TWS-n_0).

Step 4: Comprehensiveness

The diagnostic or management plan under examination was compared with the "gold standard" list of appropriate diagnostic and management suggestions. It was considered comprehensive if it included all suggestions contained in the gold standard list. Therefore, the final DQS and MQS were calculated by expressing TWS-n_0 as a proportion of the gold standard score (weighted sum of the gold standard suggestions' ratings).

Figure 4 illustrates how the DQS and MQS were computed for one subject's diagnostic and management plan. Figure 5 shows a summary of the entire scoring system, clearly delineating the role of subjects, DDSS, and the panel members.

Reliability and Validity Analysis of the Scoring System

Scores generated from six cases used by subjects in the main trial were included in the reliability and validity testing. Reliability is a measure of the precision of the index. Reliability was determined via an index of interrater agreement (the intraclass correlation coefficient; two-way random effects model using individual ratings; SPSS Inc., Chicago, IL) as well as intrarater consistency derived from each of the panel member's scores for diagnosis and management items.

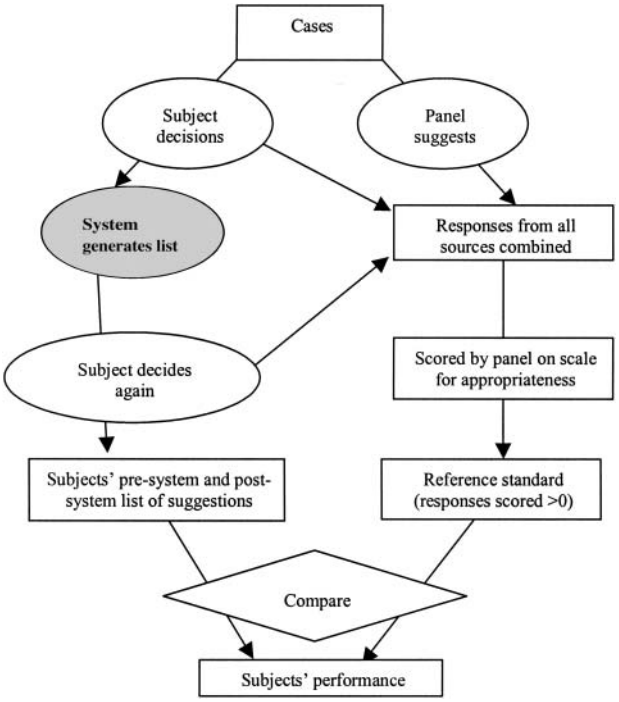

Figure 5. Graphical depiction of entire scoring scheme. Rectangles represent tasks and measurements, ovals represent actions, and the diamond represents a simple task such as tallying. The gray shaded oval indicates a step that did not directly contribute to the procedure. Adapted with permission from: Hripcsak G, Wilcox A. Reference standards, judges, and comparison subjects: roles for experts in evaluating system performance. J Am Med Inform Assoc. 2002;9:1–15.

Validity is the extent to which the score measures what it intends to measure. Validity of the score was assessed by examining the face, content, concurrent, and construct validities of the scores.

In a preliminary step, face validity was examined by one investigator to ensure that there were no obvious discrepancies between visual inspection and the scores. Sets of diagnoses that seemed "good" should have scored highly, and sets that appeared "bad" should have scored poorly. Content validity of the index was checked by consultation with ten pediatricians, of different grades and levels of experience, who were not involved in the development of the score. They reviewed the factors contributing to diagnostic and management plan quality; they also assessed each factor individually and indicated whether other factors were involved. Sets of diagnostic and management plans were used as examples during this process. Sets of diagnoses and management plans constructed without ISABEL support were considered for construct validity. This was assessed by the extent to which the scores discriminated medical students (expected to perform badly) and consultants (expected to perform well), as well as by the extent to which the scores discriminated between subjects' plans for common cases (higher scores) and unusual cases (lower scores). We compared the mean scores of the medical student group and the consultant group and computed the difference between their mean scores in standard deviation units (similarly for mean scores of common and unusual cases). In the absence of an

Table 1 ■ Main Characteristics of Proposed Scoring System

1. Complete composite score can be calculated when the "correct" diagnosis is unknown.
2. Contribution of any number of diagnoses/management steps can be considered.
3. Assesses the quality of clinical "actions" prompted by diagnostic decision support.
4. Assesses quality as a function not just of plausibility; quality is defined as "appropriateness" of decision making.

already established score for the quality of decision plans, concurrent validity of the DQS and MQS was established by comparing the score with a subjective impression of quality elicited from a panel of general pediatricians. Thirty sets of differential diagnoses and management plans were selected at random from the study data, ensuring that high-scoring, mid-scoring, and low-scoring sets were included in equal proportions. These were sent by electronic mail to ten general pediatric consultants chosen at random from the ISABEL user database. Overall quality was scored on a scale from 0 to 5. The extent to which the DQS and MQS correlated with these scores was tested by using the Spearman rank correlation test. The extent to which the two scores (DQS and MQS) correlated with each other within the same episode was tested by computing the Pearson correlation between the user scores across all cases. The study protocol was submitted to the local ethics committee; formal approval of the study was not felt to be necessary.

Results

Reliability and validity were assessed using 190 differential diagnosis and management plan sets produced by 71 subjects working on six different cases (level 1, 2 cases; level 2, 1 case; and level 3, 3 cases). Each set was composed of a pair of lists, one before, and one after, consultation with the ISABEL differential diagnosis tool.

Reliability

Aggregate lists of diagnostic and management suggestions for six cases were used to test for interrater reliability. This comprised a total of 252 diagnostic and 350 clinical management suggestions. The median number of unique diagnoses suggested per case was 37 (range, 29 to 64); unique management suggestions per case 58 (range, 38 to 80). Intrarater reliability, examined by cross-checking the scores assigned by each panel member to their own gold standard decisions submitted *a priori*, was acceptable (rater 1, 39 of 41; rater 2, 45 of 54; decisions scored highest on the scale). Intraclass correlation coefficient for the panel's scoring for all diagnoses was 0.79 (95% CI, 0.74 to 0.83; $p < 0.01$); for all management item scores it was 0.72 (95% CI, 0.67 to 0.77; $p < 0.01$), both suggesting good interrater agreement. When diagnostic suggestions were classified dichotomously using a cutoff score for "appropriateness" ("appropriate" ≥ 3; "inappropriate" < 3), there was excellent agreement (kappa statistic 0.84, 95% CI, 0.79 to 0.89) between raters. Similarly, when management suggestions were classified (appropriate ≥ 1; inappropriate < 1), agreement between panel members was good (kappa statistic 0.58, 95% CI, 0.51 to 0.64). The interrater reliability of the DQS and MQS depended only on establishing the reliability of the panel scoring procedure, since only mathematical calculations were used in subsequent steps to derive the DQS and MQS from panel scores.

Validity

Initial exploration of face validity found no instances in which the scores seemed counterintuitive. Consultation with ten pediatricians of differing grades and levels of experience established that the four main concepts that made up the scoring system were comprehensive and relevant to measuring the quality of a diagnostic and management plan (content validity). Examination of construct validity showed that the mean DQS and MQS for medical students were much lower compared with the consultants (30.15 vs. 41.59; 28.9 vs. 38.87, respectively: a difference of 0.83 SD units for both). Mean DQS for unusual cases was much lower compared with mean DQS for common cases (all users, 31.62 vs. 46.34: difference of 0.86 SD units). Concurrent validity, tested by measuring the correlation between the subjective impression of overall quality and corresponding DQS and MQS derived from the equations used this study, showed a Spearman correlation ρ 0.64 (95% CI, 0.36 to 0.81; $p < 0.001$). When only the DQS was tested, correlation improved to ρ 0.75 (95% CI, 0.24 to 0.94; $p < 0.01$), when only MQS was tested, ρ was 0.62 (95% CI, 0.24 to 0.83; $p < 0.001$; Fig. 6). The DQS and MQS showed a moderate positive correlation: Pearson's correlation coefficient, computed using all 190 sets of diagnostic and management plans, was 0.51 (95% CI, 0.39 to 0.60; $p < 0.01$; Fig. 7).

To sequentially identify whether each of the steps used in the scoring system was necessary, concurrent validity was assessed separately for three of six cases used in the validation sample. Scores were first computed using an arithmetic sum of the individual ratings, as opposed to the weighted sum. For DQS and MQS together, Spearman ρ dropped to 0.34 (95% CI, -0.15 to 0.70), for DQS alone it was 0.49 (95% CI, -0.53 to 0.93). When the step that ensured relevance of the plan (involving n_0) was eliminated, correlation dropped to ρ 0.54 (95% CI, 0.09 to 0.80). These results are summarized in Figure 8.

Discussion

The results of this study have generated a reliable and valid composite scoring metric for the measurement of quality of a clinical assessment plan in an acute medical setting. It can be used as an outcome measure in medical informatics studies that attempt to quantify the benefits of diagnostic decision support through changes in diagnostic quality as well as changes in clinical management. It may also prove useful in medical education exercises utilizing case simulations to assess examinees. The necessity to assess changes in the quality of "actions" performed as a result of DDSS use has been highlighted previously.[13] Although a study involving QMR reported changes in test ordering prompted by diagnostic decision support, the quality of these changes was not assessed objectively.[35]

Our scoring system attempts to measure diagnostic quality (as opposed to diagnostic accuracy) by examining appropriateness at a fixed point in time, rather than predicting a gold standard "final diagnosis." Our approach may prove useful in the prospective evaluation of DDSS (or clinicians) "on the front lines" when a single correct diagnosis is unavailable, as in our planned clinical trial.[36] The unavailability of a conclu-

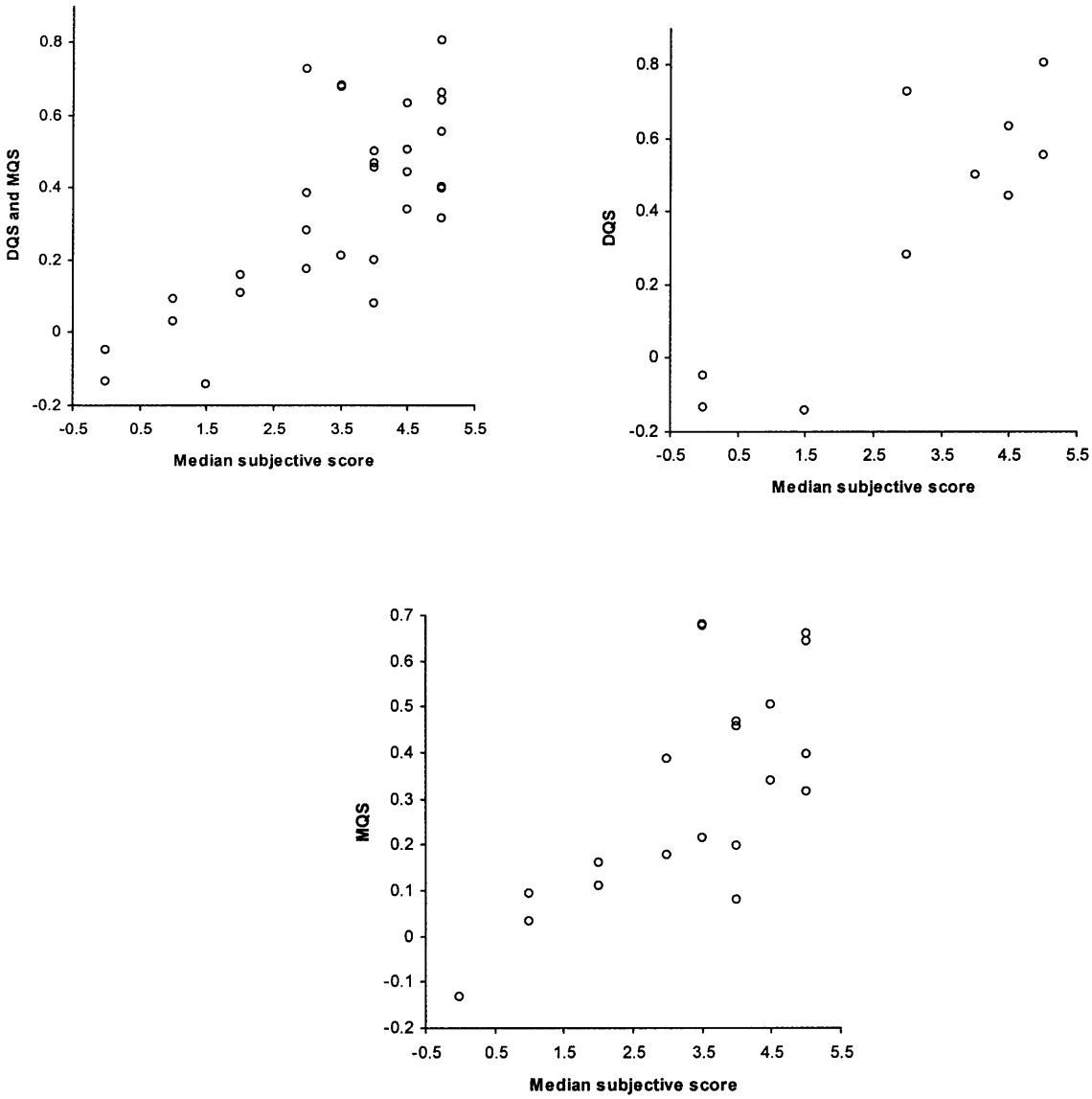

Figure 6. Concurrent validity: correlation between DQS/MQS and overall subjective score.

sive diagnosis is common early on during the clinical presentation of a patient. DDSS that are used primarily in these settings, such as ISABEL, might prove useful to clinicians by reminding them to consider potentially important diagnoses (and order relevant tests to help reach a conclusive diagnosis). Depending on how the presenting clinical features evolve and change during the course of a diagnostic assessment, one or none of the diagnostic suggestions could turn out to be the "final diagnosis." Thus, using diagnostic accuracy as the sole outcome measure may falsely undervalue the clinical utility of such tools during prospective real life testing. Previous studies have also shown that the diagnostic accuracy of other DDSS in a true diagnostic dilemma (a clinical dead-end) may not exceed that of experienced clinicians,[15] suggesting perhaps that these systems might be more useful to inexperienced clinicians early on during a patient consultation, even in cases that do not appear to be diagnostic dilemmas to senior clinicians.

Unfortunately, this approach to measuring DDSS benefit does not provide an objective gold standard against which the quality of the diagnostic (and management) plans can be judged, and forces the use of an expert panel.[37] We first outlined the factors that might contribute to this quality, by combining clinician opinion and previously used measures, and chose to use subjective assessments of appropriateness for individual suggestions, utilizing the process of panel review. To add objectivity to the panel's choices, we have shown that the real final diagnoses for most of these cases were scored highly on the scale. This provides additional criterion validity to the panel assessment procedure. The panel was not provided this final diagnosis at any time during scoring, avoiding an "outcome bias" or a "hindsight bias."[38] By creating a gold standard of appropriate decisions by merging the panel's and subjects' combined suggestions, we avoided the problem of disregarding a subject's appropriate suggestion if it was not present within the panel's list of decisions. However, ISABEL-generated suggestions that did not match any subject's or panel member's suggestions were not included in the aggregate list. This step was avoided for logistical reasons, but it may have resulted in the absence

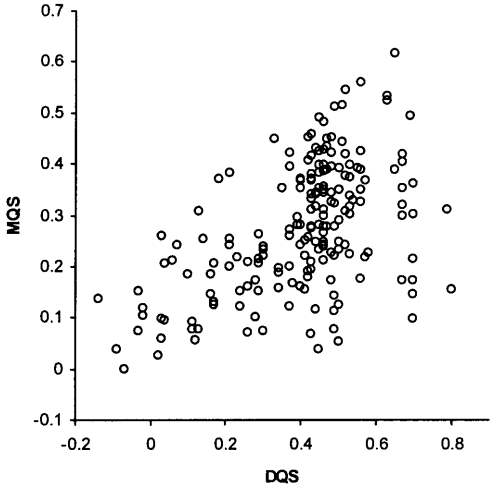

Figure 7. Correlation between DQS and MQS.

of some truly "relevant" diagnoses within the aggregate list. Creation of the aggregate list was suggested previously by Friedman et al. in their scoring system.[24] This procedure provided an opportunity to test a panel member's consistency (intrarater reliability) by checking that their own decisions scored highly within the aggregate list. The scale for diagnostic and management items treated diagnoses as "judgments" and management as "actions." Only an action that might cause clinical harm to a patient was penalized (a harmful and inappropriate test scored negatively, while an irrelevant diagnosis scored 0). Using a weighted sum of the individual ratings as a measure of overall plan quality enabled differential contribution of appropriate versus less appropriate decisions to the entire plan. The Unites States National Board of Medical Examiners (NBME) has previously used weighted scoring schemes to assess performance in their case simulations but did not consider comprehensiveness and relevance as factors. Steps taken to keep the plan focused and comprehensive were based on concepts outlined in other studies.

The DQS and MQS have been shown to be reliable for measurement purposes. Agreement between the two raters was better for diagnoses than for management items. This is consistent with the reported level of agreement for judging the appropriateness of test ordering during a peer review process.[39] As can also be gauged from the less-than-perfect agreement between raters for "appropriate" suggestions, each panel member was not forced to assume that the other's suggestions were ideal. The DQS and MQS also appear to be valid measures of quality. The scores showed construct validity by satisfactorily differentiating consultants' plans from those of medical students', and plans for easy cases from scores for difficult cases. The quality score derived from our four-step procedure correlated well with subjective opinion of diagnostic and management plan quality used to test concurrent validity; in the absence of a true gold standard, this was the best available measure of quality assessment. In addition, the two scores seemed to measure separate but related aspects of clinical performance, implied by the moderately positive correlation between the two scores. Intuitively, a greater correlation between the two scores might be expected: a subject scoring a high DQS for a case would also score a high MQS. However, when decision making is measured at one fixed point in time in the acute setting, management plans between subjects may not vary greatly with the complexity of the case presentation. Indeed, a simple management plan consisting of basic tests and supportive treatment could easily be the ideal choice for most clinical scenarios in the immediate term. Despite this, it is conceivable that the high DQS might favorably influence future actions by helping the clinician reliably interpret the results of initial tests or treatments ordered. Measuring quality at a fixed, single point in time in the acute setting was, therefore, an important limitation of this scoring system. It also led to problems during scoring: most discrepancies between panel members were related to confusion about deciding whether a particular test or step would be appropriate at that point.

Ranking of diagnostic hypotheses was not considered in our scoring system; ranking is often used as a way of implying the choice of further management steps. Since the quality of clinical "actions" was also measured in this study, ranking of diagnostic suggestions may not have provided any additional information. In addition, the rank value for the correct diagnosis (location component) did not have much relevance to our study, in which a set of appropriate suggestions was

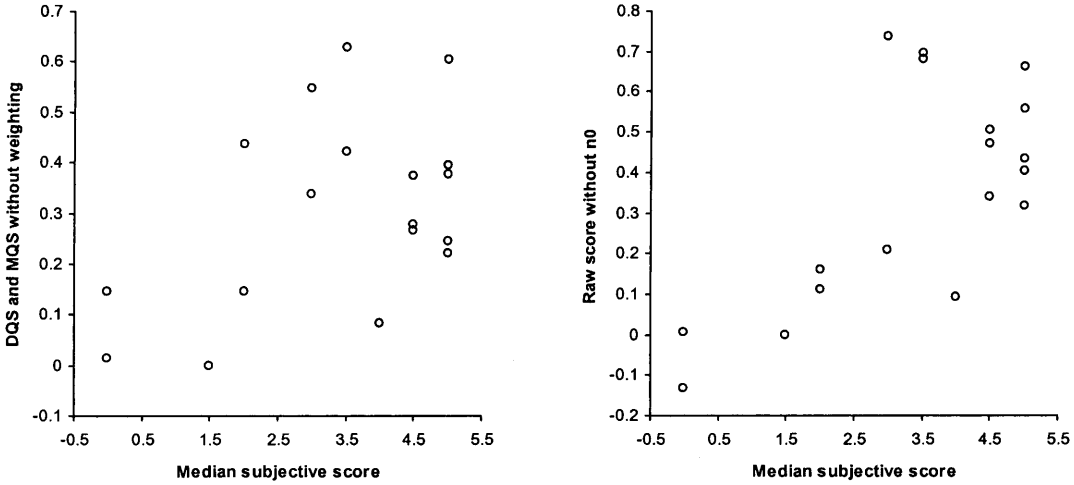

Figure 8. Concurrent validity assessed after the omission of weighting and relevance steps.

being considered. One of the other significant limitations of this scoring system stemmed from the aggregate list of diagnoses and management items for each case. Although the number of new items diminishes as a function of the number of new subjects who have considered the case, it is possible that a new item will be added, even one that is considered by the panel to be significant. The median number of 37 diagnostic suggestions in our study indicates the logistical problem associated with this: the panel had to examine between 20 and 65 diagnoses and a similar number of management items for each case. Friedman et al.[40] reported a similar phenomenon with their scoring and imply that around 28 subjects had to work a case to exhaust all plausible diagnostic suggestions. In addition, data analysis was not possible until all the subjects had completed their cases. Finally, the index described is complex and involves many steps. However, the quality of clinical reasoning (in a wide domain such as internal or pediatric medicine) is a complex and abstract concept without a clear gold standard to determine appropriateness. Most steps described were necessary to satisfy content validity.

Conclusions

This study describes the development and the assessment of reliability and validity of a new scoring index for clinical assessment quality in an acute medical setting. It is intended that this score will serve as a key outcome measure in our evaluation studies evaluating diagnostic decision support (simulated and real life). This scoring metric, as well as the methods described in this report, can be generalized to other settings and studies with similar methodology and objectives.

References ■

1. De Dombal FT, Leaper DJ, Staniland JR, McCann AP, Horrocks JC. Computer-aided diagnosis of acute abdominal pain. Br Med J. 1972;2:9–13.
2. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain: an evolving diagnostic decision-support system. JAMA. 1987;258: 67–74.
3. Miller R, Masarie FE, Myers JD. Quick medical reference (QMR) for diagnostic assistance. MD Comput. 1986;3(5):34–48.
4. Rubeck RF. ILIAD: a medical diagnostic expert system. Teach Learn Med. 1989;1:221–2.
5. Warner HR, Haug P, Bouhaddou O, et al. ILIAD as an expert consultant to teach differential diagnosis. In: Proc Annu Symp Comput Appl Med Care. 1988:371–6.
6. Nelson SJ, Blois MS, Tuttle MS, et al. Evaluating Reconsider: a computer program for diagnostic prompting. J Med Syst. 1985;9:379–88.
7. Bankowitz RA, McNeil MA, Challinor SM, Parker RC, Kapoor WN, Miller RA. A computer-assisted medical diagnostic consultation service. Implementation and prospective evaluation of a prototype. Ann Intern Med. 1989;110:824–32.
8. 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.
9. Wexler JR, Swender PT, Tunnussen WW, Oski FA. Impact of a system of computer-assisted diagnosis: initial evaluation of the hospitalized patient. Am J Dis Child. 1975;129:203–5.
10. Wellwood J, Johannessen S, Spiegelhalter DJ. How does computer-aided diagnosis improve the management of acute abdominal pain? Ann R Coll Surg Engl. 1992;74:40–6.
11. Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based clinical decision-support systems on physician performance and patient outcomes. JAMA. 1998;280:1339–46.
12. Miller RA. Medical diagnostic decision support systems—past, present and future: a threaded bibliography and brief commentary. J Am Med Inform Assoc. 1994;1:8–27.
13. Lemaire JB, Schaefer JP, Martin LA, Faris P, Ainslie MD, Hull RD. Effectiveness of the Quick Medical Reference as a diagnostic tool. Can Med Assoc J. 1999;161:725–8.
14. Berner ES, Webster GD, Shugerman AA, et al. Performance of four computer-based diagnostic systems. N Engl J Med. 1994; 330:1792–6.
15. Friedman CP, Elstein AS, Wolf FM, et al. Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems. JAMA. 1999;282:1851–6.
16. Adams ID, Chan M, Clifford PC, et al. Computer-aided diagnosis of acute abdominal pain: a multicentre study. Br Med J (Clin Res Educ). 1986;293:800–4.
17. Pozen MW, D'Agostino RB, Selker HP, Sytkowski PA, Hood WB. A predictive instrument to improve coronary-care-unit admission practices in acute ischemic heart disease: a prospective multicenter clinical trial. N Engl J Med. 1984;310:1273–8.
18. 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(1):3–10.
19. Warner HR, Toronto AF, Veasey LG, Stephenson R. A mathematical approach to medical diagnosis: application to congenital heart disease. JAMA. 1961;177:75–81.
20. Miller RA, Masarie FE. The demise of the "Greek Oracle" model for medical diagnostic systems. Methods Inf Med. 1990; 29:1–2.
21. Miller RA, Pople HE Jr, Myers JD. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med. 1982;307:468–76.
22. Berner ES, Maisiak RS, Cobbs CG, Taunton OD. Effects of a decision support system on physicians' diagnostic performance. J Am Med Inform Assoc. 1999;6:420–7.
23. 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.
24. Friedman C, Elstein A, Wolf F, et al. Measuring the quality of diagnostic hypothesis sets for studies of decision support. Medinfo. 1998;9 pt 2:864–8.
25. Greenough A. Help from ISABEL for paediatric diagnoses. Lancet. 2002;360:1259.
26. Thomas NJ. ISABEL. Crit Care. 2002;7(1):99–100.
27. McKenna C. New online diagnostic tool launched to help doctors. BMJ 2002;324:1478.
28. Ramnarayan P, Britto J. Paediatric clinical decision support systems. Arch Dis Child. 2002;87:361–2.
29. Ramnarayan P, Tomlinson A, Rao A, Coren M, Winrow A, Britto J. ISABEL: a web-based differential diagnostic aid for paediatrics: results from an initial performance evaluation. Arch Dis Child. 2003;88:408–13.
30. Kent DL, Shortliffe EH, Carlson RW, Bischoff MB, Jacobs CD. Improvements in data collection through physician use of a computer-based chemotherapy treatment consultant. J Clin Oncol. 1985;3:1409–17.
31. Wraith SM, Aikins JS, Buchanan BG, et al. Computerized consultation system for selection of antimicrobial therapy. Am J Hosp Pharm. 1976;33:1304–8.
32. van der Lei J, Musen MA, van der Does E, Man in 't Veld AJ, van Bemmel JH. Comparison of computer-aided and human review of general practitioners' management of hypertension. Lancet. 1991;338:1504–8.
33. Miller PL. Extending computer-based critiquing to a new domain: ATTENDING, ESSENTIAL-ATTENDING, and VQ-ATTENDING. Int J Clin Monit Comput. 1986;2:135–42.

34. Haug PJ, Gardner RM, Tate KE, et al. Decision support in medicine: examples from the HELP system. Comput Biomed Res. 1994;27:396–418.
35. 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.
36. Miller RA. Reference standards in evaluating system performance. J Am Med Inform Assoc. 2002;9:87–8.
37. Hripcsak G, Wilcox A. Reference standards, judges, and comparison subjects: roles for experts in evaluating system performance. J Am Med Inform Assoc. 2002;9:1–15.
38. Dawson NV, Arkes HR, Siciliano C, Blinkhorn R, Lakshmanan M, Petrelli M. Hindsight bias: an impediment to accurate probability estimation in clinicopathologic conferences. Med Decis Making. 1988;8:259–64.
39. Bindels R, Hasman A, van Wersch JW, Pop P, Winkens RA. The reliability of assessing the appropriateness of requested diagnostic tests. Med Decis Making. 2003;100: 254–5.
40. Friedman CP, Gatti GG, Murphy GC, et al. Exploring the boundaries of plausibility: empirical study of a key problem in the design of computer-based clinical simulations. Proc AMIA Symp. 2002:275–9.