

The Practice of Informatics

JAMIA

Review Paper ■

Reference Standards, Judges, and Comparison Subjects:

Roles for Experts in Evaluating
System Performance

GEORGE HRIPCSAK, MD, MS, ADAM WILCOX, PhD

Abstract Medical informatics systems are often designed to perform at the level of human experts. Evaluation of the performance of these systems is often constrained by lack of reference standards, either because the appropriate response is not known or because no simple appropriate response exists. Even when performance can be assessed, it is not always clear whether the performance is sufficient or reasonable. These challenges can be addressed if an evaluator enlists the help of clinical domain experts. 1) The experts can carry out the same tasks as the system, and then their responses can be combined to generate a reference standard. 2) The experts can judge the appropriateness of system output directly. 3) The experts can serve as comparison subjects with which the system can be compared. These are separate roles that have different implications for study design, metrics, and issues of reliability and validity. Diagrams help delineate the roles of experts in complex study designs.

■ *J Am Med Inform Assoc.* 2002;9:1–15.

Medical informatics systems are often designed to carry out complex tasks and to perform at the level of human experts. For example, diagnostic systems use clinical evidence, such as admission history, clinical

signs, and diagnostic results, to produce probabilities of disease or lists of diagnoses. Therapeutic systems suggest interventions tailored to patients. Information retrieval systems produce lists of documents that are relevant to some topic. Image processing systems detect features in a digital image.

Evaluating the function of these systems can be difficult.¹ Determining the appropriate responses that a system should have produced, deciding whether the system output matches an appropriate response, and even deciding whether a given level of performance is good enough are all challenges. Clinical domain experts have frequently been enlisted to address these challenges. In this paper, we review the many designs that have incorporated human experts into

Affiliation of the authors: Columbia University, New York, New York.

This work was supported by grants R01 LM06910 and R01 LM06274 from the National Library of Medicine and grant NY01-002153889 from Pfizer, Inc.

Correspondence and reprint requests: George Hripcsak, MD, MS, Department of Medical Informatics, Columbia University, 622 West 168th Street, VC5, New York, NY 10032; e-mail: <hripcsak@columbia.edu>.

Received for publication: 5/24/01; accepted for publication: 10/9/01.

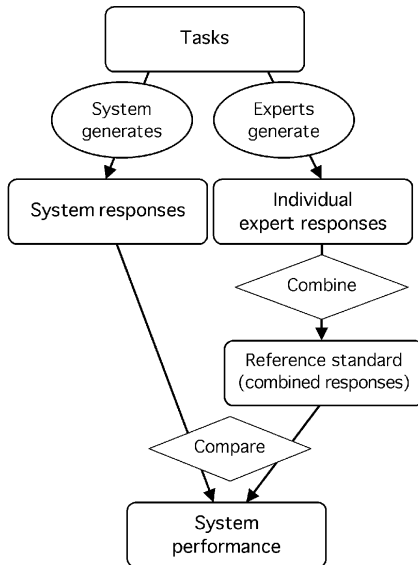


Figure 1 *Experts generate a reference standard.* Experts generate responses, which are combined by the evaluator to form a reference standard. The system also generates responses, which are compared by the evaluator with the reference standard to derive performance. It is assumed that tasks are simple enough for responses to be combined and compared unambiguously. (NOTE: Rounded rectangles indicate tasks, observations, or measurements; ovals indicate actions by the system or experts; and diamonds indicate actions that require no domain expertise, such as simple tallying.)

system evaluation, enumerate the roles that experts may play in evaluation, and provide a framework for describing designs. We draw largely on examples from clinical informatics and from information retrieval, but the framework and issues are more broadly applicable across medical informatics.

Roles for Experts

Experts may serve in one or more of several roles:

- *Reference standard.* To measure system performance (in terms of performance metrics) on tasks for which the set of correct responses is not known, appropriate responses may be generated by experts.
- *Judges.* To measure system performance on tasks for which the generation of a reference standard is impractical or impossible, experts may judge the appropriateness of system responses directly.
- *Comparison subjects.* To interpret the measured performance of a system, its performance may be compared with the performance of experts on the same tasks.

In the first two roles, the goal is to quantify performance. For example, it may be found that 75 percent of

the system responses are appropriate (either by comparing the system responses to a reference standard, as in role 1, or by having experts judge the system responses directly, as in role 2). In the third role, the goal is to interpret that quantity. For example, it may be found that experts achieve only 70 percent accuracy, so the 75 percent accuracy of a system appears reasonable. In the sections that follow, we discuss each of these roles and issues of reliability and validity. The distinction between the first and third role is further explored under “Experts as Comparison Subjects.”

We define a *task* as one unit of work, such as a diagnosis for one patient or an assessment of one document. Most evaluations comprise a set of tasks. The granularity of a task is sometimes ambiguous. A single query for relevant documents could be seen as one task with many items (documents) or as many individual tasks (deciding whether a given document is relevant). For each task, experts and systems generate observations, interpretations, advice, suggestions, measurements, or system output, which we generally refer to as *responses*.

Generating the Reference Standard

Generating and Combining Responses

Measuring the performance of a medical informatics system generally requires comparison of its responses (recommendations, plans, diagnoses, etc.) with the correct or appropriate responses, known as the “reference standard” or “gold standard.” Experience shows that accurate reference standards rarely exist; if it were easy to obtain the correct responses, a medical informatics system would be unnecessary.

When an obvious reference standard does not exist, human experts can often generate a reference standard (Figure 1).² The experts perform the same tasks as the system, and their responses are combined in some way. Experts’ responses may differ because of such qualities as subjective judgment, variation in practice, or tendencies to judge harshly or leniently.³ Their aggregated responses, while not perfect, are more reliable than any single expert’s response and may serve as a reasonable reference with which the system may be compared. An expert-generated reference standard implies two important assumptions—that the experts are performing the same task as the system and that the experts’ aggregate response is more accurate than the system response. If either assumption is not met, the system performance will usually appear lower than it really is.

Three important issues come up in the design of an expert-generated reference standard—what form the experts' responses should take, how different experts' responses to a single task should be combined, and what should be measured when the system response is compared with the reference standard. The decisions will depend on the nature of the tasks, the goal of the evaluation, and the format of the system output.

The task may be to answer a specific question, such as determining the existence, quality, or other attribute of specific entities. In this case, the experts can rate the attribute on a dichotomous (yes/no), ordinal (e.g., *definite, probable, possible, and not mentioned*),⁴ interval (e.g., Likert), or nominal (unordered list of items) scale. (When we use the term *interval data*, we include data that meet stricter criteria, such as ratio data.) The presence of pneumonia and the advisability of treating with the antibiotic ampicillin are examples of specific questions. A dichotomous scale (yes/no) may appear faster or more obvious to the expert, but the other scales will collect more specific information about a task.

The experts' responses can be combined by majority vote with random assignment for ties (dichotomous or nominal), average (dichotomous, interval, or ordinal where categories are assigned scores), median,⁵ or other functions, such as minimum or maximum for specialized applications. If the goal is to measure performance in terms of accuracy, sensitivity, specificity, or predictive value, then a reference standard with dichotomous responses is necessary. Non-dichotomous responses may be made dichotomous by using a threshold or by grouping categories.

Some tasks require a system to select items from a large set. For example, the system might select several diagnoses from a long list of diseases, or it might select several relevant documents from a corpus. In such a case, the response is a list of items that may be ranked or that may each have a score assigned (indicating, for example, probability of disease or relevance).

The system output can be seen as a set of responses to a large number of specific tasks (yes or no for every possible disease or document), but it is usually not practical for experts to view it in this way. Instead, each expert generates a list of diseases or relevant documents. Their responses may be combined with a simple union (a disease mentioned by any expert) or a threshold (diseases that at least two or a majority of experts chose). If experts assign scores to items on the list, the scores can be averaged. If the list is ranked,

then a combining algorithm—e.g., “include any disease that is on three or more experts' lists and any disease that is within the top two diagnoses on any expert's list”—can be defined.

In these approaches, the experts work independently, and the evaluator combines their responses. This allows experts to perform tasks in a single sitting, and it permits the estimation of the reliability of the reference standard (see “Reliability”). An alternative is for the experts to come to a consensus. This approach was used, for example, to generate a reference standard of diagnoses that were appropriate for an evaluation of diagnostic systems.⁶

A formal approach to generating consensus is the Delphi method and its modifications.^{1,5,7} In the Delphi method, experts' individual responses are collated into a single document by a moderator and sent back to each expert for review. An expert may comment on the responses or may change his or her own response on the basis of the others' opinions. The modified responses are collated and returned to the experts for further modification. The process is repeated until a consensus is achieved or there are no further changes. One goal is to avoid trivial errors.⁵ A second goal is to spur the experts to think more deeply about the problem and consider issues they may have missed—which is accomplished by showing them each other's responses—thus producing a more accurate standard than a mere average or majority opinion. In favor of this is the observation that experts do change their minds when faced with a different response; in one study, experts changed their minds after seeing the system output.⁸

Another approach is not to combine experts' responses at all. The reference standard consists of a vector of responses for each task. Although this approach may be less satisfying because there is no single preferred response for each task, it preserves all the information in the experts' responses.

A related approach is to use a patient's actual diagnosis or course of therapy, decided by the treating clinician, as the reference standard.^{8,9} This strategy assumes that the actual diagnosis or treatment is adequate. The authors of one such study, in which the system displayed substandard performance, concluded that the system was not necessarily in error and that the treating physicians would have benefited from the advice.⁸

In some cases, when the expertise is unique, it makes sense to use a single highly accurate expert. For example, user variability in abstracting and entering case

histories into *Quick Medical Reference* used the primary developer of the system as the reference standard.¹⁰ Here the expertise was not clinical diagnosis but use of the system and knowledge of its vocabulary.

Performance Metrics: Comparing the System with the Reference Standard

Given a reference standard, the system must then be compared with that standard and some measure of performance must be calculated. For dichotomous or nominal data, the most basic performance metric is simple accuracy, which is the proportion of correct responses.⁹ This is easy to calculate and interpret, but it has several disadvantages. It does not account for agreement due to chance, for different utilities among different types of errors, or for the effect of prevalence.

Sensitivity and specificity improve on accuracy by distinguishing two types of errors. If the reference standard and system responses are dichotomous, then the problem can be formulated as a 2×2 contingency table, and measures like sensitivity, specificity, and positive and negative predictive values can be calculated.^{5,11-14} Separating accuracy into sensitivity and specificity allows the evaluator to isolate the poor performance of a system to cases with or cases without the disease or property of interest, to assign separate costs to false-positive and false-negative results, and to correct for changes in prevalence. (In practice, because of measurement error and heterogeneous populations, the separation of prevalence from sensitivity and specificity is not perfect.^{15,16})

If the system produces an ordinal output (such as low, medium, and high probability of disease) or an interval output (such as a probability estimate), then a series of sensitivities and specificities can be calculated. From them, a receiver operating characteristic (ROC) curve¹⁷⁻²⁰ can be derived, and the area under the curve (or its equivalent, the *C* statistic²¹) can be interpreted as the classification ability of the system, ranging generally from 0.5 (chance) to 1 (perfect).

Recall and precision have been used frequently in information retrieval and knowledge-based tasks.²²⁻³¹ Recall is the proportion of responses that are considered positive in the reference standard and are marked positive by the system. Precision is the proportion of positive system responses that are positive in the reference standard. They are analogous to sensitivity and positive predictive value when an actual contingency table can be defined and responses are independent of each other, but they have been applied in more gener-

al circumstances. For example, they are useful when the number of true negative cases is unknown or ill defined.^{32,33} Heuristic combinations of recall and precision, such as the *F* measure,^{25,34} provide a single number for comparing two systems.

Some systems produce lists—of diagnoses, relevant documents, or other items—as responses, and the list may be ranked⁶ and the items on it each assigned a score. A list can be treated as a large set of dichotomous responses for each possible disease, document, or other item, and a metric for dichotomous data, such as accuracy or sensitivity, can be applied.

This approach is not ideal, because it misses the correlation among items, it does not handle ranking well, and it misses the concept of the differential diagnosis or query result as a unit. A number of heuristic approaches have been taken. The system may be given credit for having the appropriate response anywhere in its list, near the top of its list, or at the top of its list, or credit may be weighted by position on the list.

Tasks that naturally warrant interval, ordinal, or nominal responses are sometimes converted to dichotomous data by use of a threshold or by the grouping of categories. This approach allows the use of familiar metrics, but it loses useful information by compressing all responses into two categories.^{9,35} For interval data or ordinal data with a score assigned to each category, accuracy can be defined as the average linear difference between the system response and the reference standard. This is most appropriate when the responses have a physical interpretation (e.g., time).^{5,36} Alternatively, correlation coefficients can be used to express the concordance between the system responses and the reference standard.¹ Whereas accuracy rewards exact correspondence between the responses, correlation coefficients reward any positive correlation between the system and the reference standard, automatically adjusting the responses for differences in scale, and they can be corrected for attenuation.¹

When the response is a probability of a disease,³⁷ performance can be separated into two components. Predictive power indicates whether the system can distinguish positive from negative cases using a measure like area under the ROC curve. Calibration³⁸ indicates the degree to which a probability estimate is accurate, and it may be expressed graphically or as a single parameter estimate.³⁷

For nominal data, various agreement metrics may be defined. Fleiss³⁹ provides an excellent discussion of

several agreement metrics, including simple agreement (accuracy), specific agreement, and chance-corrected agreement (kappa).^{40,41} Agreement was used to compare various information retrieval methods to a reference standard,⁴² in several clinical studies,^{8,14,43} and in a study of methods for knowledge base construction.⁴⁴

If a reference standard consists of the experts' uncombined responses, then several metrics are possible. Dichotomous, interval, and ordinal responses can use a distance metric. For example, rater distance, defined as the average number of diseases per patient for which two raters disagreed, was used to quantify the performance of natural language processing.^{12,36} In another study, Euclidean, city-block, and Chebyshev distance were compared.⁴⁵

The search for the single best metric to measure performance is elusive. The goal is to pick a metric that is sound (the metric and its confidence intervals or hypothesis tests are statistically sound), intuitive (the metric is easy for the reader to grasp), and familiar (the metric has been used previously, permitting comparisons). It may take a combination of metrics to meet all these goals.

Experts as Judges

For some tasks, a single reference standard does not exist. For example, a therapeutic plan may be more difficult to assess than a diagnosis.⁸ If the task is to produce a therapeutic plan, then many responses may be possible and may have various degrees of correctness. For such a task, it is difficult to aggregate separate expert responses into one reference standard, and

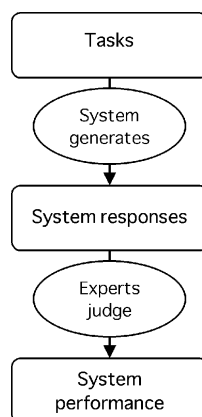


Figure 2 Experts judge system responses. Experts judge the appropriateness of responses generated by the system, and performance is calculated. (See note to Figure 1.)

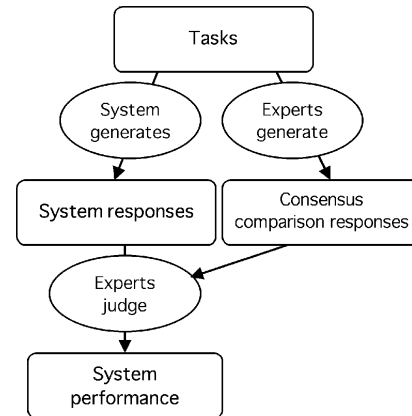


Figure 3 Experts judge system responses using comparison responses. Experts judge the correctness of system responses using comparison responses that they generate by a consensus process. The response-generating experts and the judging experts may be the same or different. This scenario differs from that represented by Figure 1 because the tasks are assumed to be more complex and therefore require expert judgment to determine appropriateness. The comparison responses do not constitute a reference standard in the sense of a single preferred response per task. Instead, they serve as a reference that can be overridden by the judgment of the experts. (See note to Figure 1.)

comparison of the plan proposed by the system with that reference standard does not capture the appropriateness of the system response. For example, if the response does not match the reference standard exactly, it remains unclear whether it is absolutely wrong, partially appropriate, or equivalent.

The experts can, however, serve as judges, reviewing the appropriateness of the system response directly (Figure 2). A similar problem arises for the evaluation of a hypothesis structure as opposed to a simple diagnostic statement⁴⁶; hypothesis summarization was followed by assignment by experts into five categories—*correct*, *possible*, *partly correct*, *wrong*, and *seriously wrong*. This approach has been used in many studies of diagnostic and therapy planning systems.^{1,14,47–49}

In the process of judging the appropriateness of a system responses, the experts may refer to one or more examples of correct (or reasonable) responses for each task. Such a set of comparison responses might be obtained, for example, by having the judges first generate their own responses and come to a consensus for each task (Figure 3). These comparison responses do not constitute a reference standard in the sense discussed previously (under “Generating a Reference Standard”), where it was assumed that a single appropriate response could be generated for each task and

that the system responses could be compared with the reference standard relatively mechanically (without the need for expert judgment to carry out the comparison). For more complex reasoning tasks, experts are needed to judge the appropriateness of system responses. A response may be judged appropriate even if it matches none of the comparison responses (e.g., a reasonable medication alternative).

There are several ways to quantify appropriateness. The easiest to interpret is a dichotomous score with a majority vote. Each expert judges the response to be satisfactory or unsatisfactory (or whatever criteria are relevant to the task), and the overall response of experts is the majority vote with random assignment for ties. Then the proportion of tasks that have been voted satisfactory represents the performance of the system, which can be loosely interpreted as how often the system will be correct. This is equivalent to accuracy as described in the previous section.

This proportion is easy to understand, but it loses some information. Tasks for which opinion was split are likely to have some intermediate appropriateness between unanimous approval or disapproval. A degree of appropriateness can be defined as the proportion of experts who deemed a single task satisfactory. More generally, the experts can review each response and assign it a score on a Likert scale or ordinal scale. Ordinal scales for therapeutic plans have included the following: *not acceptable*, *acceptable alternative to the expert's own opinion*, and *equivalent to the expert's own opinion*⁴⁸; and *ideal*, *acceptable*, *suboptimal*, and *unacceptable*.⁴⁷ For etiology, eight linguistic labels were offered (e.g., *very possible*) and assigned a probability interpretation (0.8486).⁴⁵ In a study of AI/RHEUM,⁵⁰ both Likert and ordinal scales were employed; percentages were mapped to the ordinal scale, and a qualitative scoring matrix mapped ordinal probabilities to ordinal agreements.

Experts' judgments can be combined by averaging the Likert scores or by averaging an assigned score for each ordinal category. This appropriateness score, when averaged over many tasks, represents the performance of the system.

A dual approach is possible. The proportion of tasks that a majority of experts judge to be at least "nearly fully effective" can be reported and readily understood. An appropriateness score can also be calculated. Its finer granularity allows detailed comparisons to such response questions as which kinds of tasks the system does best. It is also likely to have greater statistical power for the comparison of competing systems.

Sometimes a reference standard is difficult to create because too many tasks or items must be completed. In an adverse drug event study, experts cannot be expected to review every patient's chart, manually looking for events, and in an information retrieval study, experts cannot be expected to manually review every document in a large corpus. If several systems (or subjects) are being compared, their responses can be pooled and reviewed by expert judges blinded to the source of the responses. In this way, a reference standard is created. The performance metrics are similar to those for a reference standard generated directly by the experts. This approach was used in a study of adverse drug events.⁵¹

Similarly, in a comparison of several MEDLINE-based information retrieval systems,⁵² judges were used to decide which of all cited articles were relevant. In that study, judges assigned a relevance score from 1 (*definitely not relevant*) to 7 (*directly relevant*) to each article retrieved, and responses were made dichotomous by grouping responses 1 to 4 as not relevant and responses 5 to 7 as relevant. This approach has also been used in terminology studies.^{53,54}

The approach saves work for the experts (in fact, it may make an impossible job feasible), but it assumes that any task or item marked as negative by all the systems or subjects is truly negative. Thus, because some relevant cases will most likely be missed by all the systems, the approach is likely to overestimate sensitivity. Nevertheless, the method is useful for comparing systems if not for gauging absolute performance.

In some studies, the expert does not judge the system responses directly but instead uses the responses for some other task, which is in turn evaluated. For example, experts extracted information from two representations—one of which was generated by the system—to see whether there was a difference in the output.⁵⁵

Experts as Comparison Subjects

Gauging Performance with Comparison Subjects

Presented with the performance of a knowledge-based system, readers often wonder, "How good is good enough?" Ideally, the system would be placed into a real health care environment and its effect on patient outcomes or the health care process measured, but that is beyond the assessment of system function. In simple cases, one can estimate the cost of false-positive and false-negative system responses and the benefit of appropriate system responses; system performance then translates directly into net

benefit. More often, however, as long as system performance is neither terrible nor perfect, it remains unclear whether it is sufficient for the intended purpose. In an assessment of MYCIN, for example, it was unclear how to interpret a 75 percent approval rating by expert judges.⁵⁶

A common alternative is to ask whether the system is performing in an expert-like manner.^{4,12-14,24,48,57-59} If it performs as well as experts, then it may be able to assist non-experts or supplement experts. Therefore, the goal of some evaluations is to measure how much like experts a system is.

The experts perform the same tasks as the system. Both the system and the experts are compared with a reference standard (or their responses are judged), and their estimated performances are compared (Figure 4). Alternatively, their responses can be compared directly without actual calculation of performance (Figure 5).

It is easy to confuse two roles—generating a reference standard and serving as comparison subjects. The goal in generating a reference standard is to get as accurate a response as possible using whatever information is available and combining experts' responses into a

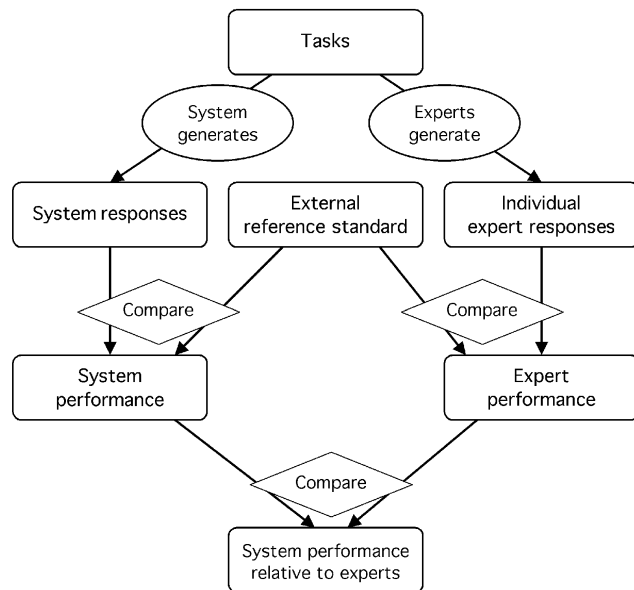


Figure 4 Experts serve as comparison subjects for interpreting performance. Experts serve as comparison subjects for setting an external reference standard. The responses of both the system and the experts are compared with an external reference standard, and performance is calculated for each. The performance of the system is then compared with that of the experts to determine whether the system performance is adequate or reasonable. (See note to Figure 1.)

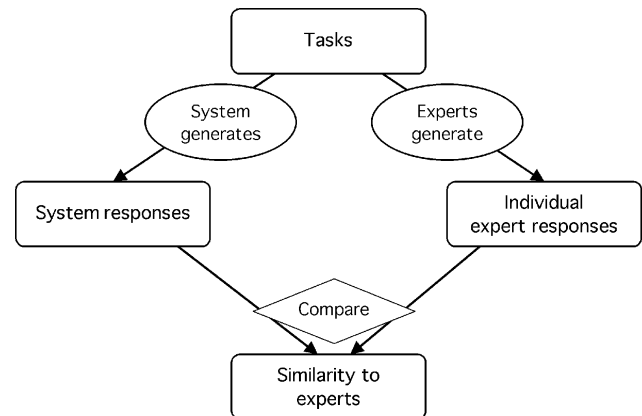


Figure 5 Experts serve as comparison subjects without a reference standard. Lacking a reference standard, the system responses are compared directly with the responses of the experts, resulting in a measure of similarity rather than of performance. This design differs from the design shown in Figure 1, because the experts' responses are not combined and no reference standard is claimed. (See note to Figure 1.)

more reliable standard. Expert comparison subjects, on the other hand, are usually given only the information they would have had in the normal course of practice, and their responses are not combined.

A patient's actual diagnosis or course of therapy has sometimes been used as a comparison subject.^{11,47,48,60,61} It must be decided whether the actual course should be considered expert-like. Similarly, non-experts (lay persons, medical students) are sometimes added as comparison subjects to show that the system has better performance.^{12,13,25,48,57}

Statistical Techniques

Perhaps the clearest way to quantify "expert-like" behavior would be to report the percentile of system performance (using one of the metrics mentioned in the previous sections) among the universe of potential experts. For example, if a system performed at the 20th percentile, then it outperformed a fifth of the experts. Even this figure requires interpretation, and it requires a clear definition of what constitutes an expert. Estimating this figure is not so easy. It requires a sufficient number of experts to estimate the empiric distribution of their performance, and it requires a sufficiently powerful test to place the system accurately within that distribution.

In the absence of these requirements, a grosser estimate can be obtained. For example, non-experts (other systems or other persons) can be included as subjects, and whether the confidence interval for the

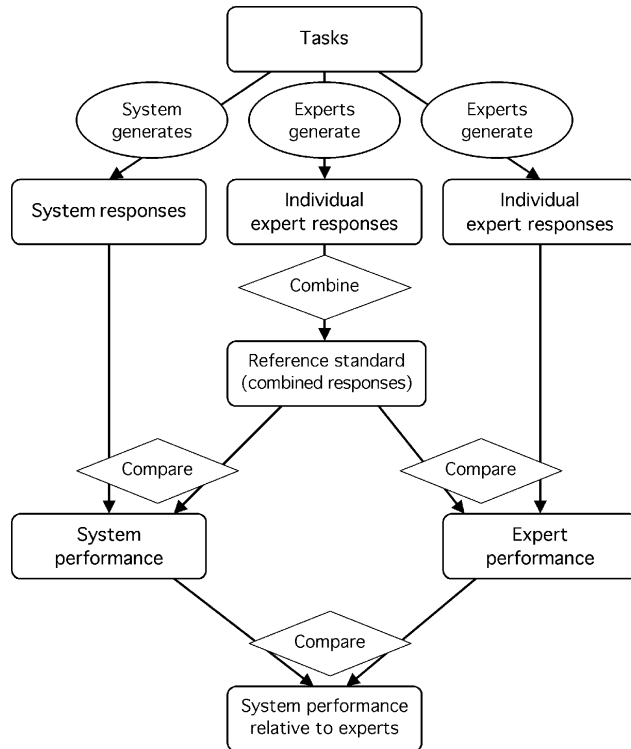


Figure 6 Experts generate a reference standard and serve as comparison subjects. Experts generate responses, which are combined into a reference standard (middle column). The system responses (left column) and the uncombined responses of the experts (right column) are compared with the reference standard, resulting in estimates of system and expert performance. The performance of the system is compared with that of the experts to determine whether the system performance is adequate or reasonable. The two ovals labeled “Experts generate” may represent two groups of experts or the same experts. In the latter case, the experts generate one set of responses, but to avoid bias, the responses of a given expert are compared only with the combined responses of the other experts, as described in the text. (See note to Figure 1.)

system performance includes only the experts (expert-like behavior), only the non-experts (inferior behavior), or both (insufficiently powerful test) can be reported. Equivalent methods have been used.^{12,13,57} Confidence intervals on differences in performance between the system and the experts are useful when acceptable limits on the differences can be set by intuition or experience.⁵

For pair-wise comparisons between the system and individual experts, the McNemar test^{27,62} is useful. This test indicates whether two subjects, the system and one expert, differ significantly in classifying cases into two categories (e.g., condition present or absent). Visual techniques have also been employed. Multidimensional scaling analysis⁶³ was applied to a

study in which experts, lay persons, a natural language processor, and keyword search methods interpreted chest radiographic reports.³⁶ Clustering has been used to visualize the similarity between a system and experts who disagree, using distance^{45,64} and weighted kappa.⁶⁵

A typical example of the use of experts as comparison subjects is found in the electrocardiogram-interpretation literature.⁴ Experts generated diagnoses, assigning a certainty on an ordinal scale. Responses by experts and systems were made dichotomous and compared with a reference standard (which was, in this case, externally generated). Performance was quantified using measures like accuracy and sensitivity extended to classification matrixes. Pair-wise performance was compared using the McNemar test, and relative performance was visualized on ROC curve axes.

Experts Serving in Multiple Roles

An evaluation may use experts in two or more of the roles defined above. In a blinded mutual audit,^{1,47,48} experts serve as comparison subjects, performing the same tasks as the system, and experts serve as judges, each assessing the other experts’ responses, the system responses, and often the actual care of the patient while blinded to the source of the response.

When different experts serve in the two roles^{48,51} (that is, there is no overlap between the two groups of experts), the evaluation is straightforward. It may be necessary to rewrite experts’ responses to use the same syntax and vocabulary as the system to ensure a blinded evaluation. Otherwise, judges may treat the system differently (e.g., more harshly) than they treat expert comparison subjects.⁵⁶ Similarly, experts may generate a reference standard while other experts serve as comparison subjects.^{23,57}

The same group of experts may serve in multiple roles—as both generators of the reference standard and comparison subjects (Figure 6)^{12,13} or as both judges and comparison subjects (Figures 7 and 8).^{14,49} Ideally, the system and the experts would be compared with an accurate external reference standard,⁵ such as external pathognomonic criteria. In the absence of an external standard, however, individual experts can be compared with their own reference standard, because the standard should be more reliable than any single expert.

One challenge is the avoidance of bias due to comparison of experts to themselves. If an expert’s responses contributed to a reference standard, then the expert’s

performance will be overestimated if that reference standard is used to make the estimate. This bias can be avoided by using an approach similar to leave-one-out cross-validation. The system is compared with a reference standard generated by all the experts. When serving as a subject, each expert is compared with a reference standard generated only by the other experts. Then the performance of the experts can be compared with that of the system. This approach cannot, however, factor out systematic problems of validity (which are discussed below) that are consistent across experts. In such cases, experts will appear more accurate than they really are.

Care must be taken in combining experts' responses into a reference standard: bias can be introduced because the number of experts generating the reference standard varies. For example, suppose there are six experts and the reference standard is positive when "at least half the experts" agree that a condition is present. The reference standard used for the system (generated by six experts with the possibility of a

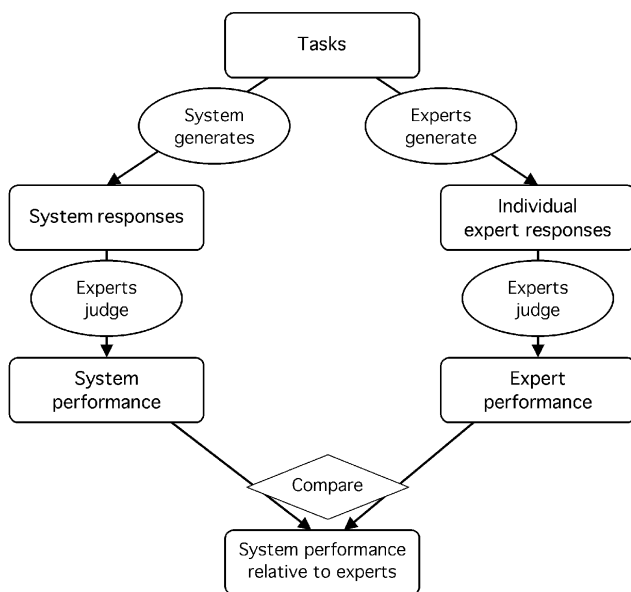


Figure 7 Experts serve as judges and as comparison subjects. The system and the experts generate responses, which are then judged by experts. The performance metrics for the system and for the response-generating experts are calculated and then compared, to determine whether system performance is adequate or reasonable. The two ovals labeled "Experts judge" indicate the same experts, and the experts are blinded to which responses (i.e., those of the system or those of the expert) they are judging. The experts who generate responses may be the same as the judging experts if bias is avoided (i.e., if judgments on their own responses are not included in the performance estimates). (See note to Figure 1.)

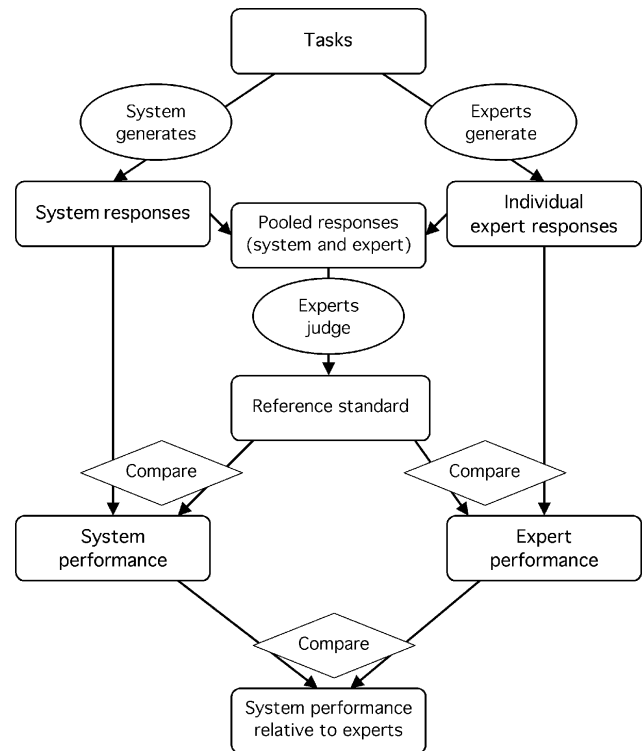


Figure 8 Generating a reference standard from the pooled responses of the system and the experts. The system (left column) and experts (right column) generate responses, which are pooled and judged by experts (middle column). The responses that are judged to be appropriate constitute the reference standard, which is then used to estimate the performance of the system and of the response-generating experts. The performance of the system is compared with that of the experts to determine whether the system performance is adequate or reasonable. Again, the same experts may generate responses and judge responses if bias is avoided. (See note to Figure 1.)

3-to-3 tie) will have slightly more positive cases on average than the reference standard used for the experts (generated by the five other experts with no possibility of a tie). The proper approach in such a case would be to redefine the standard as positive when there are more than half, negative when there are fewer than half, and either positive or negative by a fair coin flip when there is a tie.

Reliability

Estimating Reliability

Up to this point, the discussion has centered on demonstration studies—that is, studies meant to answer a question. Measurement studies provide the foundation for demonstration studies, assessing the accuracy of an evaluation method and offering feedback to improve it. In this context, a measurement

study focuses on the accuracy of the reference standard or expert judges, answering questions such as how many experts are needed, whether they are the right experts, and how many tasks are needed.

The accuracy of a reference standard can be divided into two components—reliability and validity.¹ Reliability is a measure of how precisely something is being measured. For example, if a reference standard were recreated by the same experts, how close would it be to the first version? Validity indicates how close the thing that is actually measured matches what was intended to be measured. That is, are you measuring what you wanted to measure?

Reliability has sometimes been expressed informally as the frequency with which experts agree with each other.⁹ A better approach is to quantify reliability with a reliability coefficient—that is, the proportion of variance attributable to the thing being measured, as opposed to differences in judgment or errors. A re-analysis of Hypercritic data¹⁴ was used to illustrate reliability.¹ Reliability was used in a study of ONCOCIN,⁴⁷ in which pair-wise reliability ranged from 0.11 to 0.79. It has also been used in assessing judges of adverse drug events⁵¹ and in a study of natural language processing.⁶⁶

Reliability is an essential measure for a reference standard. If it is too low, then performance measures based on it will be inaccurate. A reliability coefficient of 0.7 is generally considered sufficient for estimating overall performance.^{1,66,67} A value of 0.95 may be needed to support case-by-case assessment.^{66,67} For example, detailed discussions about why a system fails^{13,58} require assurance that the failures are indeed failures.

The reliability of a reference standard can be improved by increasing the number of experts. The expected improvement can be calculated via the Spearman-Brown prophecy formula.¹ Poor reliability may be due to a few unusual cases or experts, varying tolerance among experts, or clustering of experts.⁵⁶ In rare cases, an unusual expert may be removed, but this may result in a loss of validity. Avoiding inappropriate participants who are different from the others (clinic director, developer) and avoiding unwilling judges can improve reliability.¹

Criteria should be well defined. Formal training may improve reliability but may reduce generalizability.¹ For example, an algorithm may be used to decide whether a pneumonia is present on the basis of the paper chart; if the algorithm is purely mechanical, leaving no room for exceptions, then there is no real

expertise. That would be fine if the algorithm were accurate, but decisions involving real patients are not always so simple.

More complex analyses such as generalizability studies are possible, but they are beyond the scope of this paper. Excellent tutorials already exist (e.g., those of Friedman and Wyatt,¹ Shavelson et al.,⁶⁸ Cronbach et al.,⁶⁹ and Dunn⁷⁰ and that of Hripcsak et al.,⁶⁶ which contains a detailed example for natural language processing.)

If a technique like the Delphi method is used to combine experts' responses, then assessment of reliability is no longer straightforward, because observations are no longer independent. Non-standard approaches to the assessment of reliability have been used, however.⁵

Reliability Results in the Literature

Reporting on the reliability of the reference standard also builds the evaluation literature. For example, it allows other evaluators to estimate the number of experts needed to create a reliable reference standard for their studies. A number of pertinent measurement studies have been carried out in medical informatics.

A study of natural language processing¹³ assessed inter-rater disagreement (22 percent) and intra-rater disagreement (8 percent) and tabulated the reasons for expert disagreement—interpretation of findings (42 percent), judgment of relative likelihood (24 percent), judgment of degree of disease (21 percent), and errors in coding (13 percent). The magnitude of disagreement was somewhat lower than that obtained in the classic studies of interpreting radiographic images—30 percent inter-rater and 22 percent intra-rater disagreement.⁷¹

In a paper presented at the Fifth Message Understanding Conference, a reference standard was created by one coder who had available the responses of a second coder.⁷² A measurement study was carried out to determine the reliability of the human coders. They found that, although the primary coder relied mostly on his or her own opinion, the second coder's opinion was also incorporated. They found no significant difference among experienced and new coders.

Other studies have found that experts frequently cluster, with some achieving high consistency and others being very discordant.^{47,73} Even the degree of intra-rater variability may vary.⁷³ Intra-rater reproducibility was also assessed in an electrocardiogram study.⁴

A reliability study of experts interpreting chest radiographic reports⁶⁶ found an average per-rater reliability of 0.80, but reliability for detecting individual clinical conditions ranged from 0.67 (pleural effusion) to 0.97 (central line presence). On average, six experts were necessary to create a reference standard with a reliability of 0.95 (sufficient for case-by-case analysis). A single expert was sufficiently reliable to pass the 0.70 criterion for estimating overall performance. In practice, it would first be confirmed that an expert was not unusual by comparing that expert's responses to those of experts already known to be reliable.

Validity

Issues of validity—that is, whether you are measuring what you think you are measuring—affect every evaluation. Good summaries of validity and bias exist in several areas—medical informatics,¹ information retrieval,²² epidemiology,⁷⁴ and clinical prediction rules.⁷⁵ In this paper, we discuss those issues specific to experts and their roles.

Experts must be chosen carefully. Experts differ in their specific area of expertise, years of experience, current institution, training, and overall philosophy. The broader the sample they represent, the more generalizable the results (e.g., choosing experts from outside institutions⁴⁸) but also the more likely they are to disagree. Problems of subjectivity arise even with the consensus of recognized experts.⁷⁶ In general, developers should not serve as experts,⁵ although they can help focus a study or interpret the results. In rare circumstances, a developer may in fact be the most appropriate expert, as was the case in one study of the use rather than the performance of a system.¹⁰

Tasks must also be chosen carefully. If only those tasks with an obvious response are chosen, performance will be overestimated.⁵ An unbiased, representative set of tasks should be chosen. For example, in an evaluation of MYCIN, cases were chosen to include the major diagnostic categories of infectious meningitis.⁴⁸ Some studies have included only diseases that are present in the knowledge base,⁵⁸ but for best generalizability it is better to include cases with diagnoses that are missing from system. Similarly, unparsed cases have been dropped from performance statistics⁷⁷; this must be done carefully and clearly.

Adequate sample size is important not only to ensure statistical significance for real differences and confidence in negative results, but also to ensure a representative sample of interesting cases that challenge the system.⁵ On the other hand, carrying out tasks can be

laborious. If the load is too large, experts may not complete the tasks or may not perform the tasks with adequate care.⁵⁶ It may be necessary to enrich the test set with interesting cases to achieve an appropriate mix.

Enlisting more experts can increase the sample size without increasing the workload per expert. In a natural language processing study,¹² 200 cases were distributed among 12 experts so that every expert saw only 100 cases, and the average time expected of an expert was kept to 2 hours.

Another form of bias, reported in both the clinical prediction rule⁷⁵ and medical informatics⁷⁸ literature, occurs when the data used by experts to generate the reference standard and the data used by the system overlap. For clinical prediction rule evaluation, a goal is to estimate the rule's accuracy in assigning a diagnosis. When the actual diagnosis cannot be known, the evaluator will frequently use a surrogate, such as a simple algorithm based on available clinical data. Clearly, if the prediction rule and the algorithm used for the reference standard overlap, then the measured performance of the rule will be spuriously high. (This is an example where the reference standard is inaccurate but, due to correlation, performance is overestimated.)

A similar effect occurred in a knowledge-based system study in which the estimated acceptability of the system was 100 percent.⁷⁹ As Hilden and Habbema⁷⁸ point out, however, this analogy between clinical prediction rules and knowledge-based systems does not always hold. It makes no sense, for example, to withhold critical clinical information from either the system or the experts. (If the task is to interpret liver enzyme tests, for example, then neither the system nor the experts can be denied access to the enzyme results.) Bias is avoided by being clear about what is being compared. It can be reported that the system is highly correlated with what the experts would have done with the same data or with a particular algorithm selected by the evaluator. It is up to the evaluator to select an algorithm that is or will be acceptable to the community. The accuracy of the system with respect to the true diagnosis may not be knowable in this context.

In some cases, it makes sense to withhold information from the system. For example, a prospective diagnostic system will not have follow up and autopsy results available. They can be hidden from the system but given to the experts to generate the reference standard. In some cases, pathognomonic findings have been hidden from the system on purpose to

make a task more difficult.⁸⁰ In an analogous situation, experts were given noise-free electrocardiogram tracings while the system was given tracings with noise.⁵ Assuming that the reference standard is significantly more accurate than the system, the estimated accuracy should be close to the true accuracy.

This example highlights the importance of separating two different roles for experts—as generators of a reference standard and as comparison subjects. As generators of a reference standard, experts should have all the information possible to improve the reference standard. As comparison subjects, experts should generally have the same information as the system or the same information they would have had in a real clinical situation. Whether the experts get additional information depends on the goal of the study. For example, it is sometimes desirable to test the ability of the system to parse reports whether those reports are complete or correct or not, so the reference standard should be derived from the reports, not from the patient.⁸¹

Discussion

Expert opinion and judgment are invaluable to assessing the function of medical informatics systems. By enlisting experts to assist in evaluation, researchers can carry out studies that would not otherwise be feasible or ethical. Although they are easier than most studies of clinical impact, studies that rely on expert opinion and judgment are still time consuming. Such studies will be facilitated by future work in three areas—research into better evaluation methods, measurement studies that quantify reliability in common domains, and, potentially, the reuse of existing reference standards.

Research on new study designs and new ways to create reference standards will be helpful. For example, in an electrocardiogram interpretation task, one group found that the median response among established automated systems was a reasonable reference standard to test new systems⁵; that is, the experts were other automated systems rather than human experts.

Measurement studies serve not just the goals of the researcher who performs them but also the goals of the research community. Measurements made by one researcher can be used by other researchers to estimate the number of experts or tasks required, to select reasonable metrics, and to choose an appropriate study design. For example, a formal generalizability study established the reliability of experts in

interpreting chest radiograph reports.⁶⁶ These results were later exploited by another group.⁵⁷

A number of examples of measurement studies have been cited in this paper, but more work is clearly warranted.⁸² As the medical informatics field matures, the body of measurement literature should grow and support future researchers. Sharing need not be limited to methods and measurement. Actual reference standards may be shared and reused.^{5,24,66,83}

Published work may be mined for interesting measurement results. Studies that exploit an external (presumably accurate) reference standard but that employ experts as comparison subjects can be seen as case studies of how well expert-generated reference standards might work. In a study of electrocardiogram interpretation systems, the responses of expert comparison subjects were combined and compared with the accurate reference standard.⁴ The combined response was indeed more accurate than that of seven of the eight experts, but it was not impressively high (0.79 accuracy) and was not much higher than that of some of the systems (e.g., 0.77). Had the experts been used to generate the reference standard, the performance of the systems would generally have been overestimated.* This emphasizes the danger of relying on purely theoretic results and the need for empirical measurement studies.

The designs discussed in this paper are applicable beyond the study of information systems. For example, they can be applied when the object under study is a method rather than a system⁴⁴ or when the action of the experts themselves is the focus of a study.⁸⁴

Diagrams like those shown in Figures 1 to 8 are useful for understanding study designs, especially when the experts serve in multiple roles. Similar diagrams should be employed in evaluation publications to improve readers' understanding of the design of the studies.

The goal of this paper is to separate the three roles that experts may serve and to provide a framework for describing studies. Some systems are more complex than those described in the examples presented here; advice may be given iteratively with user interaction. The issues of separating the experts' roles and of reliability and validity remain similar, however. We have not discussed at length how experts generate a response or how they judge the appropriateness of system responses. This depends somewhat on the

*See Willems et al.,⁴ Figure 3.

application area. For patient management problems, for example, the literature on clinician licensing and credentialing can be helpful.^{85–87} More broadly, cognitive science and psychology are fruitful areas.

Conclusion

Clinical domain experts can serve in several roles in the evaluation of medical informatics systems—as sources of reference standards, as judges, and as comparison subjects. Many designs and combinations of designs are possible, although issues of reliability and validity must be handled properly.

References ■

- Friedman CP, Wyatt JC. *Evaluation Methods in Medical Informatics*. New York: Springer, 1997.
- Willems JL, Abreu-Lima C, Arnaud P, et al. Evaluation of ECG interpretation results obtained by computer and cardiologists. *Methods Inf Med*. 1990;29:308–16.
- Miller PL. Issues in the evaluation of artificial intelligence systems in medicine. *Proc Annu Symp Comput Appl Med Care*. 1985:281–6.
- Willems JL, Abreu-Lima C, Arnaud P, et al. The diagnostic performance of computer programs for the interpretation of electrocardiograms. *N Engl J Med*. 1991;325:1767–73.
- Michaelis J, Welk S, Willems JL. Reference standards for software evaluation. *Methods Inf Med*. 1990;29:289–97.
- Berner ES, Webster GD, Shugerman AA, et al. Performance of four computer-based diagnostic systems. *N Engl J Med*. 1994;330:1792–6.
- Kors JA, Sittig A, van Bommel JH. The Delphi method used to validate diagnostic knowledge in a computerised ECG interpreter. *Methods Inf Med*. 1990;29:44–50.
- Reggia JA, Tabb DR, Price TR, Banko M, Hebel R. Computer-aided assessment of transient ischemic attacks. *Arch Neurol*. 1984;41:1248–54.
- Kingsland LC. The evaluation of medical expert systems: experience with the AI/RHEUM knowledge-based consultant system in rheumatology. *Proc Annu Symp Comput Appl Med Care*. 1985:292–5.
- Bankowitz RA, Blumenfeld BH, Giuse Bettinsoli N, Parker RC, McNeil M, Challinor S, et al. User variability in abstracting and entering printed case histories with Quick Medical Reference (QMR). *Proc Annu Symp Comput Appl Med Care*. 1987:68–73.
- Goldman L, Cook EF, Brand DA, et al. A computer protocol to predict myocardial infarction in emergency department patients with chest pain. *N Engl J Med*. 1988;318:797–803.
- Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med*. 1995;122:681–8.
- Hripcsak G, Kuperman GJ, Friedman C. Extracting findings from narrative reports: software transferability and sources of physician disagreement. *Methods Inf Med*. 1998;37:1–7.
- van der Lei J, Musen M, van der Does E, Man in't Veld AJ, van Bommel JH. Comparison of computer-aided and human review of general practitioners' management of hypertension. *Lancet*. 1991;338:1504–8.
- Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat Med*. 1997;16:981–91.
- Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978;299:926–30.
- Hanley JA, McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology*. 1982;143:29–36.
- Swets JA. Measuring the accuracy of diagnostic systems. *Science*. 1988;240:1285–93.
- Hilden J. The area under the ROC curve and its competitors. *Med Decis Making*. 1991;11:95–101.
- Metz CE. Basic principles of ROC analysis. *Semin Nucl Med*. 1978;8:283–98.
- Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA*. 1982;247:2543–6.
- Hersh WR. *Information Retrieval: A Health Care Prospective*. New York: Springer, 1995:45–50.
- Haynes R, McKibbin K, Walker C, Ryan N, Fitzgerald D, Ramsden M. Online access to MEDLINE in clinical settings. *Ann Intern Med* 1990;112:78–84.
- Hersh WR, Hickam DH, Haynes RB, McKibbin KA. A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *J Am Med Inform Assoc*. 1994;1:51–60.
- Chapman WW, Haug PJ. Bayesian modeling for linking causally related observations in chest X-ray reports. *Proc AMIA Annu Symp*. 1998:587–91.
- Fizman M, Haug PJ, Frederick PR. Automatic extraction of PIOPED interpretations from ventilation/perfusion lung scan reports. *Proc AMIA Annu Symp*. 1998:860–4.
- Chapman WW, Haug PJ. Comparing expert systems for identifying chest X-ray reports that support pneumonia. *Proc AMIA Annu Symp*. 1999:216–20.
- Sager N, Lyman M, Tick LJ, Nhan NT, Bucknall CE. Natural language processing of asthma discharge summaries for the monitoring of patient care. *Proc Annu Symp Comput Appl Med Care*. 1993:265–8.
- Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. *J Am Med Inform Assoc*. 1994;1:142–60.
- Vries JK, Marshalek B, D'Abarno JC, Yount RJ, Dunner LL. An automated indexing system utilizing semantic net expansion. *Comput Biomed Res*. 1992;25:153–67.
- Chinchor N, Hirschman L, Lewis DD. Evaluating message understanding systems: an analysis of the Third Message Understanding Conference (MUC-3). *Comput Linguist*. 1993;19:409–49.
- Lin R, Lenert L, Middleton B, Shiffman S. A free-text processing system to capture physician findings: canonical phrase identification system (CAPIS). *Proc Annu Symp Comput Appl Med Care*. 1991:843–7.
- Haug PJ, Ranum DL, Frederick PR. Computerized extraction of coded findings from free-text radiologic reports. *Radiology*. 1990;174:543–8.
- Chinchor N. MUC-4 evaluation metrics. In: *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, McLean, Va.; Jun 1992. San Mateo, Calif.: Morgan Kaufmann, 1992:22–9.
- Reggia JA. Evaluation of medical expert systems: a case study in performance assessment. *Proc Annu Symp Comput Appl Med Care*. 1985:287–91.
- DuMouchel W, Friedman C, Hripcsak G, Johnson SB, Clayton PD. Two applications of statistical modelling to natural lan-

- guage processing. In: Fisher D, Lenz H-J (eds). *Learning from Data: AI and Statistics V*. Lecture Notes in Statistics, vol. 112. New York: Springer-Verlag, 1996:413–21.
37. Habbema JDF, Hilden J, Bjerregaard B. The measurement of performance in probabilistic diagnosis, part V: general recommendations. *Methods Inf Med*. 1981;20:97–100.
 38. Knaus W, Wagner D, Lynn J. Short-term mortality predictions for critically ill hospitalised patients: science and ethics. *Science*. 1991;254:389–94.
 39. Fleiss JL. *Statistical Methods for Rates and Proportions*. 2nd ed. New York: Wiley, 1981.
 40. Fleiss JL. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*. 1975;31:651–9.
 41. Carletta J. Assessing agreement on classification tasks: the kappa statistic. *Comput Linguist*. 1996;22:249–54.
 42. Brown PJB, Sonksen P. Evaluation of the quality of information retrieval of clinical findings from a computerized patient database using a semantic terminological model. *J Am Med Inform Assoc*. 2000;7:392–403.
 43. Lenert LA, Tovar M. Automated linkage of free-text descriptions of patients with a practice guideline. *Proc Annu Symp Comput Appl Med Care*. 1993:274–8.
 44. Giuse NB, Giuse DA, Miller RA, et al. Evaluating consensus among physicians in medical knowledge base construction. *Methods Inf Med*. 1993;32:137–45.
 45. Verdaguer A, Patak A, Sancho JJ, Sierra C, Sanz F. Validation of the medical expert system PNEUMON-IA. *Comput Biomed Res* 1992;25:511–26.
 46. Long WJ, Naimi S, Criscitiello MG. Evaluation of a new method for cardiovascular reasoning. *J Am Med Inform Assoc*. 1994;1:127–41.
 47. Hickam DH, Shortliffe EH, Bischoff MB, Carlisle AS, Jacobs CD. The treatment advice of a computer-based cancer chemotherapy protocol advisor. *Ann Intern Med*. 1985;103:928–36.
 48. Yu VL, Fagan LM, Wraith SM, et al. Antimicrobial selection by a computer: a blinded evaluation by infectious diseases experts. *JAMA*. 1979;242:1279–82.
 49. Quaglini S, Stefanelli M, Barosi G, Berzuini A. A performance evaluation of the expert system ANEMIA. *Comput Biomed Res*. 1987;21:307–23.
 50. Bernelot Moens HJ. Validation of the AI/RHEUM knowledge base with data from consecutive rheumatological outpatients. *Methods Inf Med*. 1992;31:175–81.
 51. Jha AK, Kuperman GJ, Teich JM, et al. Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. *J Am Med Inform Assoc*. 1998;5:305–14.
 52. Haynes RB, Walker CJ, McKibbin KA, Johnston ME, Willan AR. Performances of 27 MEDLINE systems tested by searches with clinical questions. *J Am Med Inform Assoc*. 1994;1:285–95.
 53. Cooper GF, Miller RA. An experiment comparing lexical and statistical methods for extracting MeSH terms from clinical free text. *J Am Med Inform Assoc*. 1998;5:62–75.
 54. Campbell JR, Carpenter P, Sneiderman C, Cohn C, Chute CG, Warren J. Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, and clarity. *J Am Med Inform Assoc*. 1997;4:238–51.
 55. Rocha RA, Huff SM, Haug PJ, Evans DA, Bray BE. Evaluation of a semantic data model for chest radiology: application of a new methodology. *Methods Inf Med*. 1998;37:477–90.
 56. Buchanan BG, Shortliffe EH. *Rule-based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, Mass.: Addison-Wesley, 1984.
 57. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest x-ray reports. *J Am Med Inform Assoc*. 2000;7:593–604.
 58. Miller RA, Pople HE, Myers JD. INTERNIST-1, an experimental computer-based diagnostic consultant for general practice internal medicine. *N Engl J Med*. 1982;307:468–76.
 59. Zingmond D, Lenert LA. Monitoring free-text data using medical language processing. *Comput Biomed Res*. 1993;26:467–81.
 60. de Dombal FT, Leaper DJ, Horrocks JC, Staniland JR, McCann AP. Human and computer-aided diagnosis of abdominal pain: further report with emphasis on performance of clinicians. *BMJ*. 1974;1(904):376–80.
 61. Baxt WG. Use of an artificial neural network for the diagnosis of myocardial infarction. *Ann Intern Med*. 1991;115:843–8.
 62. Daniel WW. *Applied Nonparametric Statistics*. 2nd ed. Boston: PWS-Kent, 1990:163–8.
 63. Dillon W, Goldstein M. *Multivariate Analysis*. New York: Wiley, 1984:587.
 64. Hernandez C, Sancho JJ, Belmonte MA, Sierra C, Sanz F. Validation of the medical expert system RENOIR. *Comput Biomed Res*. 1994;27:456–71.
 65. Martín-Baranera M, Sancho JJ, Sanz F. Controlling for chance agreement in the validation of medical expert systems with no gold standard: PNEUMON-IA and RENOIR revisited. *Comput Biomed Res*. 2000;33:380–97.
 66. Hripcsak G, Kuperman GJ, Friedman C, Heitjan DF. A reliability study for evaluating information extraction from radiology reports. *J Am Med Inform Assoc*. 1999;6:143–50.
 67. StataCorp. *Stata Statistical Software*. Release 4.0, vol. 2. College Station, Tex.: StataCorp, 1995:163.
 68. Shavelson RJ, Webb NM, Rowley GL. Generalizability theory. *Am Psychol*. 1989;44:922–32.
 69. Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. *The dependability of behavioral measurements: theory of generalizability of scores and profiles*. New York: Wiley, 1972.
 70. Dunn G. *Design and Analysis of Reliability Studies*. New York: Oxford University Press, 1989.
 71. Yerushalmy J. The statistical assessment of the variability in observer perception and description of roentgenographic pulmonary shadows. *Radiol Clin North Am*. 1969;7:381–92.
 72. Will CA. Comparing human and machine performance for natural language information extraction: results for English microelectronics from the MUC-5 evaluation. *Proceedings of the Fifth Message Understanding Conference (MUC-5)*; Baltimore, Maryland; August 1993. San Mateo, Calif.: Morgan Kaufmann, 1993:53–67.
 73. Gilpin EA, Olshen RA, Chatterjee K, et al. Predicting one-year outcome following acute myocardial infarction: physicians versus computers. *Comput Biomed Res*. 1990;23:46–63.
 74. Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic Research*. New York: Van Nostrand Reinhold, 1982.
 75. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules: applications and methodological standards. *N Engl J Med*. 1985;313:793–9.
 76. Aliferis CF, Miller RA. On the heuristic nature of medical decision-support systems. *Methods Inf Med*. 1995;34:5–14.
 77. Gundersen ML, Haug PJ, Pryor TA, et al. Development and evaluation of a computerized admission diagnosis encoding system. *Comput Biomed Res*. 1996;29:351–72.
 78. Hilden J, Habbema JD. Evaluation of clinical decision aids—more to think about. *Med Inform (Lond)*. 1990;15:275–84.
 79. Weiss SM, Kulikowski CA, Galen RS. *Representing expertise*

- in a computer-program: the serum-protein diagnostic program. *J Clin Lab Automation*. 1983;3:383-7.
80. 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.
81. Hersh WR, Leen TK, Rehfuss PS, Malveau S. Automatic prediction of trauma registry procedure codes from emergency room dictations. *MedInfo*. 1998:665-9.
82. Friedman CP. Toward a measured approach to medical informatics. *J Am Med Inform Assoc*. 1999;6:176-7.
83. Srinivasan P. Optimal document-indexing vocabulary for MEDLINE. *Inform Proc Manage*. 1996;32:503-14.
84. Giuse NB, Huber JT, Giuse DA, Brown CW, Bankowitz RA, Hunt S. Information needs of health care professionals in an AIDS outpatient clinic as determined by chart review. *J Am Med Inform Assoc*. 1994;1:395-403.
85. Nayer M. An overview of the objective structured clinical examination. *Physiother Can*. 1993;45:171-8.
86. Carraccio C, Englander R. The objective structured clinical examination: a step in the direction of competency-based evaluation. *Arch Pediatr Adolesc Med*. 2000;154:736-41.
87. Chambers DW, Loos L. Analyzing the sources of unreliability in fixed prosthodontics mock board examinations. *J Dent Educ*. 1997;61:346-53.



Reference Standards, Judges, and Comparison Subjects : Roles for Experts in Evaluating System Performance

George Hripcsak and Adam Wilcox

J Am Med Inform Assoc 2002 9: 1-15
doi: 10.1136/jamia.2002.0090001

Updated information and services can be found at:
<http://jamia.bmj.com/content/9/1/1.full.html>

These include:

References

This article cites 66 articles, 27 of which can be accessed free at:
<http://jamia.bmj.com/content/9/1/1.full.html#ref-list-1>

Article cited in:
<http://jamia.bmj.com/content/9/1/1.full.html#related-urls>

Email alerting service

Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

To request permissions go to:
<http://group.bmj.com/group/rights-licensing/permissions>

To order reprints go to:
<http://journals.bmj.com/cgi/reprintform>

To subscribe to BMJ go to:
<http://group.bmj.com/subscribe/>