A Scale of Methodological Quality for Clinical Studies of Radiologic Examinations

PURPOSE: To develop and evaluate a scale for assessment of the methodological quality for clinical investigations of radiologic studies.

MATERIALS AND METHODS: A scale was developed that included methodological standards compiled from established sources for assessing the methodological quality of study designs in clinical research and characteristics related to biases commonly observed in clinical radiologic research. The scale was composed of 15 standards and was tested with the results of 96 studies on imaging of liver hemangioma. Interrater reliability was measured between two observers by using percent- age agreement and $\kappa$ statistics. Interrater reliability between two observers for a composite quality index that encompassed the 15 standards was measured with the intraclass correlation coefficient.

RESULTS: Agreement between the two observers was almost perfect ($\kappa$ value, 0.8–1.0) for 11 standards and substantial ($\kappa$ value, 0.74–0.78) for four standards. Agreement between the observers with regard to the composite quality index also was high (intraclass correlation coefficient $r$, 0.91 [95% CI: 0.87, 0.94]).

CONCLUSION: The scale appears to be reliable for the assessment of methodological quality of clinical investigations of radiologic studies.

The use of reliable data to support medical and public health decisions is essential if the growing demand for health care is to be met with limited resources. It is not easy to determine the accuracy of radiologic results on the basis of clinical research data, especially if studies addressing the same radiologic examination produce conflicting results (1,2). In addition, it is almost impossible for any individual to keep up to date with the health care literature (3).

Review articles provide physicians with the potential to keep abreast of developments in their field. Traditionally, individuals often considered to be experts in a field have conducted narrative reviews of the literature associated with a particular health topic (eg, imaging of children with seizures) by using informal and subjective methods to collect and interpret information. The difficulties in verification and replication in narrative reviews have been highlighted repeatedly during the past 10 years (3–6).

Proponents of evidence-based medicine propose a more systematic evaluation of literature. The concept of evidence-based medicine provides specific criteria for the collection and evaluation of clinical trial results. One of the methods derived from evidence-based medicine is the systematic review of literature. A systematic review is one in which there is a comprehensive search for relevant reports on a specific topic, and those identified are then appraised and the results synthesized according to a predetermined and explicit method (7–9). This approach provides the reader with a particular advantage over any other type of review: the ability to replicate the review.

A meta-analysis is the statistical combination of at least two studies to produce a single estimate of the effect of the examination under consideration (10). Systematic reviews and meta-analyses can help resolve controversies between conflicting studies, guide clinical research by providing new hypotheses, and identify areas in which insufficient research has been performed (11). However, variability in the methodological quality of studies included in systematic reviews can lead to confusing results (10,12,13).

Assessment of the methodological quality of the primary studies, therefore, has been identified as one of the key components of systematic reviews and meta-analyses (8). Scales
have been developed for the assessment of methodological quality of clinical studies. Most of the scales were designed for assessment of the quality of randomized controlled trials (RCTs) only (14). Such a systematic approach is not currently used for evaluation of radiologic studies (15). Nevertheless, improvements are needed to increase the credibility that can be given to and the confidence in claims for new radiologic examinations (15). In addition, the standards for demonstrating the value of radiologic examinations have the same role as the standards for evaluating the quality of RCTs. Furthermore, biases arising in radiologic trials are relatively well known (16–19).

Therefore, the purpose of our study was to develop and evaluate a scale to assess the methodological quality that can be applied to clinical studies of radiologic examinations. Specifically, this scale was designed to facilitate evaluations of the methodological quality of original research studies (excluding reviews, case reports, and technical notes) in which the accuracy of a radiologic examination is assessed. The reliability of this scale was evaluated with a series of reports on imaging of liver hemangio-

MATERIALS AND METHODS

Methodological Quality Scale

The scale we developed was based on a list of methodological standards that can be applied to any clinical study of radiologic examination evaluation. The methodological standards evaluated in the present study were compiled from established resources for assessing the methodological quality of investigational design for studies in clinical research and from literature related to biases commonly observed in radiologic research (1,2,15–27). Fifteen standards related to study design, population characteristics, and description of imaging analyses were included. Observers were first asked to determine the type of study design used (described in the next paragraph). For the remaining 14 criteria, they could respond “Yes,” “Partially,” or “No” with regard to how well the article fulfilled the criterion. The standards were categorized as follows.

Standard 1: study design.—The study could be classified as a prospective study, retrospective study with consecutive patients, or retrospective study with nonconsecutive patients.

Standard 2: adequate definition of the purpose of the study.—To meet this standard, a single precise and specific purpose should be clearly defined and developed in the study. The standard was partially met when several purposes were defined and developed, and it was not met when the purpose of the study was not clearly indicated.

Standard 3: reference standard.—To meet this standard, we required that the article include a clear definition of the reference standard and that the reference standard be an accurate method for assessing the presence of disease. This standard was partially met when the reference standard was not clearly defined or was not optimal or when the result of the radiologic examination under investigation was actually incorporated into the evidence used as the reference standard (so-called incorporation bias). The standard was not met when the reference standard was not defined.

Standard 4: inclusion criteria.—To meet this standard, a precise plan for population inclusion must be described (ie, the method of including patients is described in sufficient detail to allow a similar group of patients to be selected if the study were to be repeated). The standard was not met if no information was provided about the method of population inclusion and was partially met if the method of inclusion was described but not in sufficient detail to allow replication.

Standard 5: indeterminate examination results.—This standard required statements regarding the existence and frequency of indeterminate examination results and the manner in which indeterminate examination results were accounted for in the estimation of accuracy. It was not met when no information was provided about indeterminate examination results and was partially met when the description of indeterminate results was mentioned but without precision about how these results were accounted for.

Standard 6: exclusion criteria.—This standard was met when both the explicit exclusion criteria and the number of excluded subjects were provided. It was not met when exclusion criteria were not defined and was partially met when excluded subjects were mentioned without a clear definition of exclusion criteria.

Standard 7: spectrum of patients.—This standard was met if (a) the age distribution, (b) the sex distribution, and (c) a summary of clinical symptoms at presentation and an indication of disease severity were provided. It was not met if no information about the patients was provided and was partially met if only one or two of points a, b, and c were detailed.

Standard 8: method of analysis.—This standard was met when details on performing the examination and interpreting the results were provided, including explicit descriptions of the techniques used and of the analysis process (eg, “images were reviewed without clinical information in a blinded retrospective manner by four cross-sectional imaging radiologists during three sessions at three times”). The standard was not met when no information was provided and was partially met when the description of the analysis method was insufficient for replication in a further study.

Standard 9: analysis criteria.—This standard was met when the analysis criteria were described in adequate detail (eg, images were analyzed for liver size, tumor location, and tumor extension to the portal vein) for both positive and negative results and when analysis criteria had been defined before the study commenced. The standard was not met when the analysis criteria were not defined and was partially met when the description of the analysis criteria was insufficient for replication in a further study.

Standard 10: avoidance of verification or work-up bias.—This standard was met either when all the study patients underwent both the radiologic examination under evaluation and the reference standard procedure or, in the case of an invasive reference standard, when a validated adjustment to correct for verification bias was used. It was not met when selection for verification was dependent on the results of the examination being evaluated and was partially met when a reference standard procedure could not be performed in all cases but the absence of reference standard did not depend on the results of the examination being evaluated.

Standard 11: avoidance of diagnostic-review bias.—To meet this standard, a statement was required that the results of the reference standard were interpreted independently of the results of radiologic examination being investigated. It was not met when the results of the examination being evaluated affected the review of data used to establish the diagnosis and was partially met when independence between the radiologic examination and the reference standard was not clearly defined.

Standard 12: avoidance of test-review bias.—To meet this standard, a statement was required indicating that the radiologic examination results were inter-
Two observers (L.A., R.R.) independently applied the eligibility criteria described in the previous paragraph to select potentially relevant studies from the references retrieved during the literature search. Disagreements between the observers regarding the inclusion of studies were resolved by consensus. Simple analysis of the title of the article allowed exclusion of 990 articles because they were not related to liver hemangioma, they were case reports or review articles, or an abstract was not available. Article abstracts were then inspected or, if necessary, study articles were reviewed for the remaining 266 references. In the second step, 168 articles were excluded because they had been written in a language other than English, French, or Spanish \( (n = 72) \), included fewer than 10 liver hemangiomas \( (n = 35) \), were restricted to scintigraphic or angiographic techniques or reported preliminary results of feasibility studies \( (n = 24) \), were review articles \( (n = 16) \) or case reports \( (n = 11) \), included no indication of a reference standard \( (n = 8) \), concerned pregnant women \( (n = 1) \), or had no abstract available \( (n = 1) \). Finally, two articles were not found in any library. After the eligibility criteria had been applied, 96 studies were retrieved and retained for our analysis.

### Inclusion of Studies

The scale was tested with clinical studies on imaging of liver hemangioma. To identify reports on the discriminative properties of radiologic examinations used in liver hemangioma imaging, we searched the MEDLINE database for articles published between 1987 and 1997. The search command included the term “liver hemangioma” associated with “CT,” “MRI,” “radiology,” “ultrasound” (ie, all terms beginning with “ultra-“), and “pathology.”

After suppression of redundancies, the search yielded 1,256 eligible articles. We excluded review articles; case reports; studies in which fewer than 10 liver hemangiomas were reported; studies with no available abstract; studies written in a language other than English, French, or Spanish; studies with no indication of a reference standard; studies concerning special populations (eg, pregnant women, children); and studies restricted to scintigraphic or angiographic techniques or that included preliminary results of feasibility studies (eg, “preliminary experience with a new double-echo half-Fourier single-shot turbo spin-echo acquisition in the characterization of liver lesions”).

Two observers (L.A., R.R.) independently applied the eligibility criteria described in the previous paragraph to select potentially relevant studies from the references retrieved during the literature search. Disagreements between the observers regarding the inclusion of studies were resolved by consensus. Simple analysis of the title of the article allowed exclusion of 990 articles because they were not related to liver hemangioma, they were case reports or review articles, or an abstract was not available. Article abstracts were then inspected or, if necessary, study articles were reviewed for the remaining 266 references. In the second step, 168 articles were excluded because they had been written in a language other than English, French, or Spanish \( (n = 72) \), included fewer than 10 liver hemangiomas \( (n = 35) \), were restricted to scintigraphic or angiographic techniques or reported preliminary results of feasibility studies \( (n = 24) \), were review articles \( (n = 16) \) or case reports \( (n = 11) \), included no indication of a reference standard \( (n = 8) \), concerned pregnant women \( (n = 1) \), or had no abstract available \( (n = 1) \). Finally, two articles were not found in any library. After the eligibility criteria had been applied, 96 studies were retrieved and retained for our analysis.

### Evaluation of Studies

The same two observers independently reviewed the 96 articles to determine which standards had been met, partially met, or not met. The observers were a 4th-year radiology resident who had worked for 6 months in our radiology department and a radiology professor with expertise in abdominal and vascular radiology. The scale had previously been developed in the radiology department to aid in performance of a systematic review of the literature on studies of skeletal radiology. Both observers were provided with text summaries of each standard, as described earlier. No instructions were given to the observers other than those outlined earlier in this article. To ensure that the standards were understood, the two observers reviewed together five sample articles related to liver metastases.

A composite quality index was then created for each standard by assigning two points if the standard was met, one point if it was partially met, and no points if it was not met. This composite quality index resulted in a 0–30-point scale for each study. Interrater reliability for individual standards was measured by using percentage agreement and the \( \kappa \) statistic \( (28) \). Degrees of agreement were categorized as follows: \( \kappa \) of 0.0–0.2, slight agreement; \( \kappa \) of 0.2–0.4, fair agreement; \( \kappa \) of 0.4–0.6, moderate agreement; \( \kappa \) of 0.6–0.8, substantial agreement; and \( \kappa \) of 0.8–1.0, almost perfect agreement. Results of the two observers for each study were correlated. Interrater reliability of the composite quality index was measured by using the intraclass correlation coefficient \( (29) \).

### RESULTS

Individual ratings for each of the 15 standards of the 96 published studies are shown in the Table, along with the percentage agreement and the measure of concordance \( (\kappa \text{ value}) \) between the two observers.

Agreement in ratings between observers for the 15 standards was high, with \( \kappa \) values of 0.9–1.0 for eight standards, 0.8–0.9 for three standards, and 0.7–0.8 for the remaining four standards. Therefore, agreement between the two observers was almost perfect for 11 of 15 methodological standards and substantial for four standards.

Individual ratings for the composite quality index of the 96 published studies are shown in the Figure, along with the measure of correlation between the two observers \( (y = 1.632 + 0.912x, r = 0.91) \). Agreement in ratings for the composite quality index also was high, with an intraclass correlation coefficient \( (r \text{ value}) \) of 0.91 (95% CI: 0.87, 0.94).

### DISCUSSION

The methodological quality of a clinical study could be defined as the likelihood that a study design will generate unbiased results and approach “the truth.” Accurate evaluation of methodological quality could be used by researchers and referees to assess study protocols, by editors and readers of journals to identify scientifically sound reports, by researchers to monitor the likelihood of bias in research reports (including their own), and by individuals involved in evidence-based medicine to perform differential analyses based on the quality of the individual primary studies \( (9,20,30–32) \).
The assessment of the methodological quality of primary studies has been identified as one of the key components of systematic reviews, meta-analyses, and evidence-based medicine. According to Jadad et al (33,34), there are at least five ways to incorporate quality assessments into systematic reviews and meta-analyses: (a) simply describe the quality assessment, (b) include or exclude studies from a review, (c) conduct sensitivity analyses allowing comparisons between the results of trials of different quality, (d) perform cumulative meta-analyses by using quality assessment as the input sequence, and (e) weight studies according to their quality.

The development of scales for systematic assessment of the methodological quality of studies is a relatively new phenomenon. The first scale of which we are aware was published in 1981 (30). By 1999, approximately 25 additional scales had been developed, and most of them were designed to assess the quality of RCTs only (14,35,36). Such a systematic approach is not currently used in evaluation of clinical studies of radiologic examinations. Nevertheless, the increasing number, cost, and sophistication of radiologic methods and their substantial influence on patient care require that new methods be subjected to rigorous assessment before they are adopted in clinical practice. In addition, the risk of bias is a serious inherent problem in studies of diagnostic efficacy.

Bias can arise from several sources, including an inadequate standard of reference, the omission of examinations that produce indeterminate results, and the absence of an explicit description of the method of analysis. Bias can also be caused by the influence of extraneous factors on the interpretation of the examination results because radiologic results are assessed subjectively (16). In addition to these potential biases, there are a number of extrapolation factors that can influence the generalization of study conclusions, including variations in examination effectiveness among patients, hospitals, and the radiologists who interpret the results (18). Many of these problems are not easily removed. The standards for demonstrating the value of radiologic examinations have the same role as those for evaluating the quality of RCTs (17). Therefore, we advocate that techniques developed from the concept of evidence-based medicine should be applied to the field of radiology. Systematic assessment of the methodological quality of radiologic studies is one of the key components of a systematic evaluation of radiologic literature (37).

The present study has many limitations. So far, our scale was tested only with clinical studies related to imaging of liver hemangioma. In addition, our search strategy was restricted to a simple MEDLINE search. It has been shown (7,33,34) that only approximately half of the available RCTs can be found by searching only the MEDLINE database. On the other hand, we did not find any data concerning the reliability of a MEDLINE search of the radiologic literature. Anyway, our purpose was not to perform a systematic review of the literature but to obtain a representative sample of clinical studies so that we could test the scale developed for the present study.

Our choice of standards was not strictly rigorous. Other standards, such as analysis of pertinent subgroups or random assignment of patients, were not used in the present study. The standard “statistical analysis” was definitely insufficient and imprecise. It should be separated, when necessary, into several parts, including appropriateness of the statistical tests used, incremental value, summary measures of diagnostic accuracy, precision of summary measures, and sufficient sample size.

Because there is little evidence on the relative importance of the individual quality criteria with regard to the control
of systematic bias, we did not try to elaborate a weighting scheme; different schemes have been described (20). Scales with complex scoring systems take more time to complete than do simple approaches, however, and they have not been shown to provide more reliable assessments of methodological quality (33).

The “not applicable” response was not available in the present scale; therefore, standards that did not apply to an article were scored as if they had not been met. We may add a “not applicable” response to our scale.

The development of the scales of methodological quality is not strictly rigorous (33). For the most part, the items chosen for use by developers of scales are based on what are called “accepted criteria,” as found in many standard clinical trial textbooks. Although these criteria may be useful, some of them are based on conviction, whereas others are based on empirical evidence. Evaluations of the methodological quality of RCTs by means of scales, however, have shown that RCTs with high scores, as measured with scales, were more likely to approach the “truth.”

Several investigators (38,39) have shown that RCTs that achieve low scores on methodological quality scales could alter estimates of the effectiveness of an intervention. In the present study, we synthesized and selected the various criteria from scales available in the medical literature that were used to evaluate the methodological quality of clinical trials and adapted these scales to the evaluation of diagnostic imaging examinations.

The problem of the validity of the standards used in our scale is, therefore, similar to the problem of the validity of standards used in the scales developed to evaluate the methodological quality of RCTs. Because our standards were adapted from those used in RCTs, however, we can expect a similar result (ie, radiologic studies that achieve a high score should be more likely to approach the truth). Actually, the objective of the present study was to test the reliability of a scale with a series of studies related to imaging of liver hemangioma. Therefore, the scale proposed in the present study could be considered a step in the development of a valid instrument of assessment of the methodological quality of radiologic studies.

Although a high level agreement was achieved between two observers with different backgrounds, reproducibility should be further tested with other studies involving multiple observers from different backgrounds and concerning different topics in diagnostic radiology. Modification of the present scale with a weighting scheme could be tested to evaluate studies characterized by a strong need to correctly evaluate a specific point. For example, if it should be necessary to obtain a correct evaluation of intra- and interobserver reliability for a specific type of examination, it would be easy to introduce a specific scheme that weighted those two items with a higher score. Finally, it is of primary interest to verify that a high score correctly identifies a high-quality study. This may be difficult, however, because it is not easy to determine a valid standard of reference for the real methodological quality of a study. Moreover, methodological quality is complex and difficult to define because it could encompass the design, conduct, analysis, external validity, and reporting of the study.

Finally, another major barrier hindering the assessment of trial quality is that, in most cases, the only way a reviewer can assess quality is by relying on the information contained in the written report (14,33,34). The problem is that a study in which the design is biased but which is well reported could be judged as being of high quality, while a well-designed but poorly reported trial could be judged as being of low quality (33). Therefore, the methodological quality of a study can be determined only to the extent that the study design and analytic methods are reported. However, because our scale allows measurement of criteria that are crucial to achieve high methodological quality, these criteria should be adequately reported in any article in which a clinical study is described. It is possible, however, to attempt to obtain additional clarifying data from the original investigators, although this may be difficult and markedly time-consuming (33).

Other specific aspects of studies that can be used to estimate study quality include the literary style of the report, the clinical relevance of the research question, and the ethical implications of the intervention evaluated (14,34). Our preference was to focus on one aspect of methodological quality; thus, we did not concern ourselves with other important issues such as clinical relevance, cost-effectiveness, or ethical implications. Finally, we developed this scale to assess the methodological quality of studies on the accuracy of radiologic examinations. Other studies in which radiologic examinations are evaluated may focus on patient outcome or societal issues, such as analysis of societal costs and benefits of a diagnostic imaging technology (eg, mammography for breast cancer screening) (21,40). The present scale could not be used to assess the methodological quality of such studies.

Although the validity of the methodological standards used in the present study was generally accepted, we wanted to check for variability in use of the criteria for rating each standard. Agreement between the two observers was almost perfect for 11 of the 15 standards used and was substantial for four standards. There was also a high correlation between the two observers for the composite quality index. Because of this high reliability, we expect that the present scale could be used by reviewers from diverse backgrounds. In addition, because no weighting scheme was used for this scale, a composite quality index should be easy to calculate.

The present scale should be modified, however, because of the different gaps we exposed. A scale of methodological quality must be flexible enough to allow some modifications as new evidence becomes available regarding the merits of maintaining or developing particular standards, discarding others, and including emerging ones. This process would undoubtedly lead to improvements in our scale to obtain a valid instrument for the assessment of the methodological quality of clinical studies of radiologic examinations, which is one of the important steps in evidence-based radiology.

Acknowledgment: We thank Pascale Dono for her kind editorial assistance.

References