Dirk Stengel, MD, MSc
Kai Bauwens, MD
Grit Rademacher, MD
Sven Mutze, MD, PhD
Axel Ekkernkamp, MD, PhD

Published online before print
10.1148/radiol.2361040791
Radiology 2005; 236:102–111

Abbreviations:
CI = confidence interval
FAST = focused abdominal ultrasonography for trauma
QUADAS = Quality Assessment of Studies of Diagnostic Accuracy included in Systematic Reviews
STARD = Standards for Reporting of Diagnostic Accuracy

[1] From the Clinical Epidemiology Division, Department of Orthopedic and Trauma Surgery (D.S., K.B., A.E.), and the Institute of Radiology (G.R., S.M.), Unfallkrankenhaus Berlin Trauma Center, Warener Str 7, 12683 Berlin, Germany. Received May 1, 2004; revision requested July 6; revision received July 9; accepted August 15. **Address correspondence to** D.S. (e-mail: *dirk.stengel@ukb.de*).

Authors indicated no financial relationship to disclose.

Author contributions:
Guarantor of integrity of entire study, D.S.; study concepts and design, D.S., K.B.; literature research, D.S., K.B., G.R.; data acquisition, D.S., K.B., G.R.; data analysis/interpretation, all authors; statistical analysis, D.S.; manuscript preparation, D.S., K.B.; manuscript definition of intellectual content, all authors; manuscript editing, D.S.; manuscript revision/review and final version approval, all authors

Association between Compliance with Methodological Standards of Diagnostic Research and Reported Test Accuracy: Meta-Analysis of Focused Assessment of US for Trauma[1]

PURPOSE: To study whether compliance with methodological standards affected the reported accuracy of screening ultrasonography (US) for trauma.

MATERIALS AND METHODS: Meta-analysis was conducted of prospective investigations in which US was compared with any diagnostic reference test in patients with suspected abdominal injury. Reports were retrieved from electronic databases without language restrictions; added information was gained with manual search. Two reviewers independently assessed methodological rigor by using 27 items contained in the Standards for Reporting of Diagnostic Accuracy (STARD) checklist and the Quality Assessment of Studies of Diagnostic Accuracy included in Systematic Reviews (QUADAS) instrument. Inconsistencies were resolved by means of consensus. Summary receiver operating characteristics and random-effects meta-regression were used to model the effect of methodological standards and other study features on US accuracy.

RESULTS: A total of 62 trials, which included a total of 18 167 participants, were eligible for meta-analysis. The average proportion of men or boys was 71.7%, the mean age was 30.6 years ± 10.8 (standard deviation), and the mean injury severity score was 16.7 ± 8.3. The prevalence of abdominal trauma was 25.1% (95% confidence interval [CI]: 21.1%, 29.1%). Pooled overall sensitivity and specificity of US were 78.9% (95% CI: 74.9%, 82.9%) and 99.2% (95% CI: 99.0%, 99.4%), respectively. Varying end points (hemoperitoneum or organ damage) did not change these results. US accuracy was much lower in children (sensitivity, 57.9%; specificity, 94.3%). Strong heterogeneity was observed in sensitivity, whereas specificity remained constant across trials. There was evidence of publication bias. Initial interobserver agreement with methodological standards ranged from poor ($\kappa = 0.03$, independent verification of US findings) to perfect ($\kappa = 1.00$, sufficiently short interval between US and reference test). By consensus, studies fulfilled a median of 13 methodological criteria (range, five to 20 criteria). In investigations that lacked individual methodological standards, researchers overestimated pooled sensitivity, with predicted differences of 9%–18%. The use of a single reference test, specification of the number of excluded patients, and calculation of CIs independently contributed to predicted sensitivity in a multivariate model. In 16 investigations (1309 subjects), a single reference test was used, which provided a combined sensitivity of 66.0% (95% CI: 56.2%, 75.8%).

CONCLUSION: Bias-adjusted sensitivity of screening US for trauma is low. Adherence to methodological standards included in appraisal instruments like STARD and QUADAS is crucial to obtain valid estimates of test accuracy.

Supplemental material: *radiology.rsnajnls.org/cgi/content/full/2361040791/DC1*

Although diagnostic tests play a key role in the management of health and disease, increased awareness of methodologic norms is an issue of diagnostic research. In methodologically weak studies, the accuracy of therapeutic interventions is overestimated (1,2), and there is empirical evidence that similar trends occur in diagnostic studies, as well (3,4). Unfortunately, there is a lack of diagnostic meta-analyses, which are ultimately needed to study the scientific background of tests currently used in clinical practice.

Screening ultrasonography (US) for trauma was introduced 30 years ago, when results from a study conducted in Denmark proved it was feasible for use in the discovery of abdominal organ injury (5). After preliminary data were collected, focused abdominal US for trauma (FAST) spread throughout German emergency departments (6) and gradually replaced diagnostic peritoneal lavage.

Reports from the United States showed that US decreased the number of computed tomographic (CT) scans requested in the emergency department (7,8). Moreover, the subcommittee of the American College of Surgeons recently proposed an increased role for FAST examination in advanced trauma life support (9).

Diagnostic algorithms designed for primary trauma survey demand a high level of accuracy. On the other hand, they must keep invasiveness and radiation exposure to the necessary minimum, and US is widely regarded as a method to quickly and easily accomplish these goals. However, it is unclear if certain qualities of the source trials, mainly their methodologic soundness, influenced current estimates of US accuracy.

Several authors proposed checklists to help clinicians distinguish valuable tests from worthless tests (10,11), but there is still no accepted approach to appraise the methodologic rigor of a diagnostic study.

Recently, two promising instruments were introduced to overcome this problem. The Standards for Reporting of Diagnostic Accuracy (STARD) checklist was designed to systematize diagnostic research (12). Quality Assessment of Studies of Diagnostic Accuracy included in Systematic Reviews (QUADAS) was developed to enable researchers to evaluate methodologic issues of individual studies for meta-analyses (13). Thus, the purpose of our study was to evaluate whether compliance with methodologic standards affected the reported accuracy of screening US for trauma.

MATERIALS AND METHODS

Study Design

We conducted a systematic review of studies in which researchers evaluated the accuracy of emergency US in the detection of free intraperitoneal fluid (according to the original intent of FAST) or organ injury after abdominal trauma.

Data gained from this structured search of the scientific literature were combined in a meta-analysis. For simplification, we will use the term *FAST+* throughout this article to indicate protocols that target both free fluid and organ injury.

We included reports indexed as prospective investigations, in which patients who had experienced blunt or penetrating abdominal trauma were enrolled and in which US was compared with any diagnostic reference standard. Suitable reference tests included CT, diagnostic peritoneal lavage, laparotomy, clinical observation, outpatient follow-up, or autopsy findings. We retrieved peer-reviewed scientific material, as well as abstract presentations, book chapters, and gray literature (ie, non–peer-reviewed material, press releases, and presentations) that were available on the Internet. Animal studies, technical investigations, and case reports were excluded from this review.

Search Strategy

Three authors (D.S., K.B., G.R.) conjointly searched the MEDLINE, EMBASE, and CINAHL databases and the Cochrane Central Register of Controlled Trials, beginning with the first citation of diagnostic US and ending with January 2004. Table E1, which is available as supplemental material on the *Radiology* Web site (*radiology.rsnajnls.org/cgi/content/full/2361040791/DC1*), details our MEDLINE and EMBASE search strategy with medical subject headings, Emtree keywords, and free-text items.

An Internet search was conducted by using Google. We also used the search engine of the Radiological Society of North America (available at: *www.radiology.rsnajnls.org*), which serves *Radiology* and *RadioGraphics*. The Lippincott Williams and Wilkins Web site (available at: *www.lww.com*) allowed for comprehensive searching of the *Journal of Trauma* and *Annals of Surgery* back to 1996. Other relevant periodicals were traced with the Springer Web site (available at: *www.springerlink.com*). No restrictions applied to language.

In accordance with the Cochrane method, we collected a first set of potentially eligible articles after scrutinizing the title or abstract. In case of vague, insufficient, or conflicting information, we retrieved the full text of articles.

Two reviewers (D.S., K.B.) cross-referenced bibliographies of all original manuscripts for publications not identified with the electronic search and manually searched all journal issues containing the study of interest for other relevant investigations.

Data Collection

Two reviewers (D.S., K.B.) independently extracted information on a data abstraction sheet. We assessed publication language, year, recruitment periods, sample sizes, demographic details, and injury severity. We recorded features of the index test (eg, transducer types, video or hard-copy storage of images, interpretation of US images by surgeons or radiologists) and the diagnostic reference standard or standards used in individual trials.

The abstraction form listed items and subitems of STARD (original tool, 25 items) and QUADAS (14 items) that referred to the Materials and Methods section and the Results section of individual articles. Accounting for some items included in both instruments, two researchers (D.S. and K.B.) selected 27 methodologic standards (22 of which can be found in STARD, and 14 of which are contained in QUADAS) for evaluation of methodologic quality (Table 1).

In addition, we graded individual studies according to the Oxford Centre for Evidence-based Medicine Levels of Evidence, 2001 revision, which is available online at *www.cebm.net/levels_of_evidence.asp*.

In particular, appropriateness of the reference test and outcome definitions were investigated. According to QUADAS, a proper reference standard is likely to allow physicians to classify the target disease correctly. Discussion between the surgeons (D.S. and K.B.) and the radiologists (G.R. and S.M.) left CT as the imaging standard of choice, demanding sufficient information on conventional or helical examination techniques and application of intravenous or oral contrast agents. Accepted invasive verification procedures were diagnostic peritoneal lavage, laparotomy, laparoscopy, or autopsy.

Three authors (D.S., K.B., and A.E.) agreed on clinical observation as a suitable reference test if authors detailed the

TABLE 1
Methodologic Standards Derived from STARD and QUADAS Instruments

Instrument	Item description
QUADAS	Selection of a representative spectrum of patients
	Appropriateness of the reference test to classify disease
	Sufficiently short interval between index and reference test (<4 h)
	Blinding against US findings
	Blinding against reference test
STARD	Defined study hypothesis
	Consecutive patient enrollment
	Prospective design
	Recruitment period reported
	Specification of the number of excluded patients
	Precise definition of end points (free fluid, detection of organ injury)
	Specification of sonographer's expertise
	Blinding
	Application of methods to determine variability between readers
	Application of methods to determine test reproducibility
	Announcement of true-positive, false-positive, true-negative, and false-negative findings
	Calculation of CIs
	Announcement of adverse events with US or reference tests
QUADAS and STARD	Specification of inclusion and exclusion criteria
	Specification of the number of patients who dropped out after enrollment
	Illustration of the patient profile (ie, demographic variables, injury severity)
	Specification of the US protocol (ie, examined planes, transducer type)
	Specification of the reference test (ie, helical CT, use of contrast agents)
	Use of a single reference test
	Independent application of a reference standard (irrespective of US results)
	Announcement of the interval between US and verification
	Specification of methods to handle indeterminate results

Note.—CI = confidence interval.

number of patients hospitalized and discharged, time of surveillance, or the interval between physical examinations.

We considered detection of free intraabdominal fluid, organ injury, or both to be satisfactory definitions of CT and US results. Also, proper classification of disease was assigned for exact descriptions of diagnostic peritoneal lavage findings (eg, a red blood cell count of more than 100 000 per cubic millimeter [1×10^{12}/L] of fluid).

We considered board-certified radiologists to be experts in the performance of US and interpretation of US findings. If surgeons conducted US examinations, we assumed expertise if operators had attended formal training lessons or were supervised by an expert in US.

Statistical Analysis

After independent data abstraction and rating of studies, we calculated κ statistics to assess interobserver agreement beyond chance. After completion of independent ratings, two authors (D.S. and K.B.) resolved inconsistencies in consensus and discussed areas of concern with others (G.R., S.M., and A.E.). Secular changes in the number of fulfilled methodologic standards were evaluated with unweighted linear regression analysis.

For each study, we calculated descriptive statistics (sensitivity, specificity, and likelihood ratios) with their 95% CIs. Summary receiver operating characteristic curves were obtained by using the method of Moses et al (14).

We used Hasselblad diagnostic d (the standardized distance between the means of the healthy patients and those with disease) as a global measure to discover publication bias (15,16). We used the Egger regression method to test for funnel plot asymmetry (17).

A plot of sample size versus treatment effect should be shaped like a funnel if there is no publication bias. If the number of studies in which small effect sizes are found is lacking because these reports have less chance of being published, the plot will become skewed. Publication bias is assumed if the intercept of the regression line fitted through the data cloud of standardized effects versus precision is significantly away from zero. Statistical heterogeneity was explored by using a Galbraith diagram (18,19).

In a plot of standardized effect measures (that is, the point estimate divided by its standard error) against inverse standard errors, a homogeneous set of trials will scatter with constant variance along a regression line fitted through the data cloud. Trials situated beyond two standard errors of this line substantially contribute to statistical heterogeneity. In addition, we calculated Cochran Q as an estimate of heterogeneity with the diagnostic odds ratio (18). Q is χ^2 distributed with k minus one degree of freedom, where k is the number of studies included in the meta-analysis. As a rule of thumb, statistical heterogeneity is assumed if Q exceeds the number of studies.

We applied random-effects meta-regression with restricted maximum likelihood estimation to combine sensitivity and specificity and to examine sources of heterogeneity in the data set (20,21). Controversy exists as to the value of quality scores. We followed recommendations and tested the influence of individual quality standards on outcomes rather than sum scores (22,23).

Preferentially, we determined the effect of individual methodologic standards, patient risk profiles (eg, age, sex, and injury severity), and other study features (eg, transducer type and surgeon or radiologist operators) on test accuracy.

Variables that significantly affect test accuracy in the univariate analysis ($P < .25$) were included in a multivariate meta-regression model (24). We used a stepwise selection procedure and excluded variables with a P value of more than .1.

The final model was selected on the basis of the degree of unexplained variance, as marked by τ^2, which decreases with better model fit (18). In contrast to postregression diagnosis with traditional methods, indicators of model fit like R^2 or the Akaike Information Criterion are not available with the mixed model extension used for meta-analysis. We used Stata Release 8.0 software (Stata, College Station, Tex) for all analyses.

RESULTS

Systematic Review

The search strategy delivered 957 articles published between November 1968 and November 2003. By assessing the ti-

tles or abstracts, 580 articles were determined to be beyond the scope of this study. The remaining 377 articles seemingly provided relevant information, and they were retrieved as full-text articles for a detailed assessment. Although designed for meta-analyses of randomized controlled trials, we considered the flowchart proposed by the Quality of Reporting of Meta-analyses, or QUOROM, panel (25) to be a convenient way to depict the trial selection procedure (Fig 1).

A total of 62 studies comprising a total of 18 167 subjects form the basis of this study. Eligible articles were published between November 1982 and June 2003. In Table E2 (*radiology.rsnajnls.org/cgi/content/full/2361040791/DC1*), the profile of individual investigations is explained. In Table E3 (*radiology.rsnajnls.org/cgi/content/full/2361040791/DC1*), individual estimates of test accuracy are summarized.

Most articles were published in English (n = 56, 90.3%). On average, 71.7% of subjects in the studies were men or boys (95% CI: 68.5%, 74.9%). Mean age was 30.6 years (95% CI: 27.6%, 33.7%). Average injury severity score was 16.7 (95% CI: 13.1%, 20.3%), which indicates that many patients had multiple injuries. In 50 studies, patients with suspected blunt abdominal injury were enrolled. Ten other studies included varying but still moderate numbers of subjects with stab and gunshot wounds (average proportion, 11.7%; 95% CI: 4.6%, 18.7%), whereas two studies included only patients with penetrating trauma.

A list of studies excluded from this systematic review is available from the authors on request.

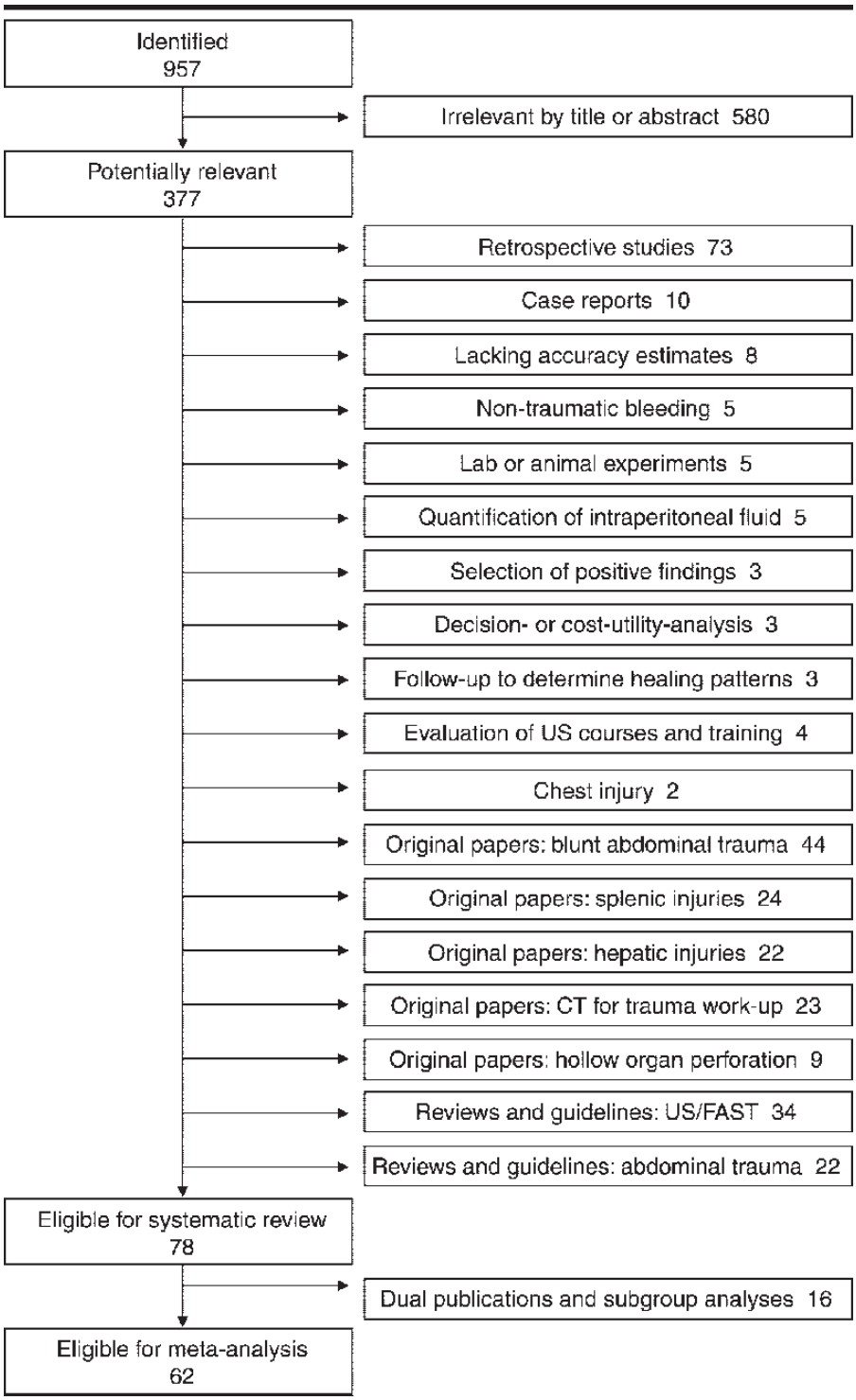

Figure 1. Quality of Reporting Meta-analyses flowchart shows the study selection procedure.

Quality Assessment

Most investigations fulfilled a grade B recommendation for diagnostic studies. This included studies meeting level 2b (ie, exploratory cohort study with good reference standards) and level 3b (ie, nonconsecutive study or study without consistently applied reference standards) evidence.

In one study, researchers explored diagnostic accuracy of US in the experimental arm of a randomized trial (therapeutic interventions: level Ib evidence, grade A recommendation). The aim of this study was to investigate the effectiveness of US-based clinical pathways to reduce the number of requested CT examinations.

In two other studies, researchers addressed similar outcomes by using a quasi-randomized format (allocating patients by date or time of admission), which is currently not explicitly covered by the available grading schemes.

During early appraisal, two authors (D.S. and K.B.) disagreed about the presence of work-up bias if clinical follow-up examinations were used as the sole confirmation procedure. This also influenced their ratings of the appropriateness of the reference standard to classify the target disease.

In several articles, it was unclear whether all subjects with negative findings at US who did not undergo further imaging studies were admitted to the hospital or deliberately followed as out-

TABLE 2
Methodologic Criteria, Early Interobserver Agreement with Individual Items, and their Influence on Predicted Sensitivity Derived from Meta-Regression Analysis

Item	No. of Patients	Studies Meeting the Individual Standard (%)	Interobserver Agreement (%)	κ Value	Predicted Difference in Sensitivity (%)	*P* Value
Representative spectrum*	61	98.4 (91.3, 100.0)	98.3	...	...	...
Defined study hypothesis	45	72.6 (59.8, 83.1)	76.7	0.38	−8.9 (−17.5, −3.1)	.042
Selection criteria specified	43	69.4 (56.3, 80.4)	98.3	0.96	0.0 (−8.5, 8.9)	.963
Consecutive enrollment	23	37.1 (25.2, 50.3)	83.1	0.62	−4.5 (−12.7, 3.7)	.284
Prospective design	51	82.3 (70.5, 90.8)	70.0	0.18	−2.9 (−13.5, 7.8)	.598
Recruitment period reported	56	90.3 (80.1, 96.4)	95.0	0.70	9.9 (−4.4, 24.3)	.173
No. of excluded patients specified	15	24.2 (14.2, 36.7)	91.7	0.76	−9.5 (−18.6, 0.5)	.038
No. of patients who dropped out specified	34	54.8 (41.7, 67.5)	58.3	0.23	−9.3 (−17.2, −1.7)	.017
Illustration of the patient profile	38	61.3 (48.1, 73.4)	78.3	0.53	−3.1 (−11.4, 5.2)	.460
US protocol detailed	55	88.7 (78.1, 95.3)	93.3	0.63	−9.5 (−21.8, 2.7)	.126
Reference test detailed	17	27.4 (16.9, 40.2)	86.4	0.67	−15.1 (−23.4, −6.8)	.000
Reference test appropriate	45	72.6 (59.8, 83.1)	82.6	0.12	−15.6 (−35.2, 4.0)	.110
Single reference test	16	25.8 (15.5, 38.5)	86.7	0.61	−16.5 (−25.2, −7.8)	.000
Independent verification	41	66.1 (53.0, 77.7)	51.7	0.03	−9.3 (−17.4, −1.3)	.023
Time interval noted	13	21.0 (11.7, 33.2)	89.5	0.34	−12.1 (−21.5, −2.8)	.016
Time interval sufficiently short	8	12.9 (5.7, 23.9)	100.0	1.00	12.4 (−17.8, 42.6)	.421
Precise definition of end points	53	85.5 (74.2, 93.1)	88.3	0.47	−3.7 (−15.2, 7.7)	.522
Handling of indeterminates noted	16	25.8 (15.5, 38.5)	81.7	0.48	1.7 (−7.6, 10.9)	.721
Sonographer expertise specified	47	75.8 (63.3, 85.8)	63.3	0.25	−5.4 (−14.7, 3.9)	.251
Blinding	17	27.4 (16.9, 40.2)	84.5	0.50	−9.8 (−18.7, −0.9)	.031
Blinding against sonogram	9	15.5 (7.3, 27.4)	93.1	0.63	−16.1 (−27.3, −4.9)	.005
Blinding against reference test	9	15.5 (7.3, 27.4)	91.4	0.57	−8.5 (−20.5, 3.5)	.167
Methods to determine variability	20	33.9 (22.3, 47.0)	78.3	0.48	−4.2 (−12.7, 4.4)	.340
Methods to assess reproducibility	21	32.3 (20.9, 45.3)	71.9	0.27	−0.1 (−8.4, 8.6)	.981
Clear depiction of TP, TN, FP, FN results	48	77.4 (65.0, 87.1)	73.3	0.48	−5.7 (−15.1, 3.7)	.238
Calculation of CI	8	12.9 (5.7, 23.9)	88.5	0.31	−18.3 (−30.2, −6.3)	.003
Adverse events reported	6	9.7 (3.6, 19.9)	92.9	0.47	3.5 (−9.8, 16.7)	.611

Note.—Negative differences in predictions indicate overestimation of sensitivity by studies lacking the particular methodologic standard. Data in parentheses are percentages. FN = false-negative, FP = false-positive, TN = true-negative; TP = true-positive.

* Because all but one study were considered to enroll a representative spectrum of patients, no κ values or differences in predictions were calculated for this item.

patients. K.B. found no exact statements regarding the follow-up policy in 12 (19.4%) reports, whereas D.S. suspected 31 (50.0%) studies did not aim at independent clinical surveying.

All authors agreed that statements such as, "all patients in this study were observed for 72 hours" (26), "patients were observed in a holding area for 4–6 hours and discharged at the discretion of the attending surgeon" (27), or "patients had follow-up as outpatients" (28) indicated independent confirmation of US findings. A reappraisal of all studies involving a third reviewer (G.R.) left 21 (33.9%) investigations without description of clinical follow-up rules in case of a negative US finding.

In addition, D.S. and K.B. dissented on the true prospective design of some investigations. After consensus was reached, methods of data collection remained ambiguous in 11 (17.7%) studies. Table 2 summarizes the proportion of methodologic standards fulfilled by the study pool and the interobserver agreement achieved during independent rating. A representative spectrum of patients was covered by all but one study, which used a quasi-case control design (level 4 evidence, grade C recommendation) (29).

Only 15 (24.2%) of all reports provided the number of potentially eligible patients screened but excluded during the recruitment period (risk of selection bias), used a single reference standard (potential verification bias), or reported methods used to handle indeterminate results.

Altogether, studies satisfied a median of 13 methodologic standards (range, five to 20 standards). Of 14 items contained in QUADAS and 22 items listed in STARD, studies fulfilled a median of seven items (range, two to 10 items) and 12 items (range, four to 17 items), respectively.

There was a rising trend in the number of studies meeting standards during the publication period (ie, from 1982 to 2003). Predicted annual increases in the number of items were 0.313 for the entire catalog ($P = .001$), 0.305 for STARD ($P < .001$), and 0.173 for QUADAS ($P = .003$).

Assessment of Publication Bias and Statistical Heterogeneity

Funnel plot analysis provided evidence of publication bias (intercept, 1.06; 95% CI: −0.27, 2.39) (slope, 2.29; 95% CI: 1.83, 2.75) ($P < .001$), with better test performance found in smaller studies (Fig 2).

Figure 3 shows diverging effects noted in the data set. *Q* statistics suggested significant heterogeneity (χ^2, 314.3; $P < .001$). Specificity was remarkably constant across studies, with only one investigation located below the margin of 2 standard errors. In contrast, sensitivity ranged from 30.8% to 100.0%, with at least 17 investigations substantially contributing to heterogeneity.

Aggregating data into a combined estimate requires a homogenous distribution of effect measures. Random-effects models aim at correcting for heterogeneity, but they still leave some degree of variance unexplained. Thus, we calculated overall estimates of sensitivity and summary receiver operating characteristic for

exploratory reasons and to illustrate changes caused by variations in study design and source populations. Estimates must be interpreted with caution.

Meta-Analysis

The prevalence of abdominal injury (organ damage, free fluid, or both) was, on average, 25.1% (95% CI: 21.1%, 29.2%). Overall sensitivity and specificity of US was estimated at 78.9% (95% CI: 74.9%, 82.9%) and 99.2% (95% CI: 99.0%, 99.4%).

There was no significant difference in test characteristics between FAST and FAST+ investigations (Fig 4). Pooled sensitivities were 77.8% (95% CI: 72.1%, 83.5%) and 80.3% (95% CI: 74.7%, 85.9%), whereas pooled specificities were 99.4% (95% CI: 99.2%, 99.6%) and 98.9% (95% CI: 98.5%, 99.5%), respectively. Few reports detailed the ability of US to depict hepatic or splenic tears, which suggests sensitivities of 66.5% and 62.6% and specificities of 97.2% and 96.6%, respectively.

We found no differences in reported test accuracy between trials that included patients with penetrating injuries and those that did not (sensitivity, 75.1% vs 79.4%; specificity, 99.1% vs 99.0%, respectively). Figure 5 shows the secular decrease of laparotomy and diagnostic peritoneal lavage rates. Utility of CT did not show a clear time-dependent trend.

The available studies did not provide enough data to allow us to evaluate differences in US accuracy between hemodynamically stable and unstable patients, trivial and nontrivial injury, or initial and repeated examinations. Of note, Boulanger et al (30) reported an overall US sensitivity of 81% (compared with diagnostic peritoneal lavage or CT), which increased to 98% for the need for acute therapeutic laparotomy. Nunes et al (31) reported a 16% increase in sensitivity for intraabdominal fluid with serial US scans.

Nine trials (involving a total of 905 patients) included children with a mean age of 9.6 years ± 3.5 (standard deviation) and show unexpectedly low accuracy of US when compared with the general population. Pooled sensitivity was only 57.9% (95% CI: 44.9%, 70.9%), whereas combined specificity was 94.3% (95% CI: 90.1%, 98.5%). This mainly reflected strict adherence to methodologic standards (eg, single well-described confirmation procedures and blinded reading of both index and reference tests). Studies of children met a median of 17 methodologic criteria (range, six to 19 criteria). Figure 6 displays the related summary receiver operating characteristics.

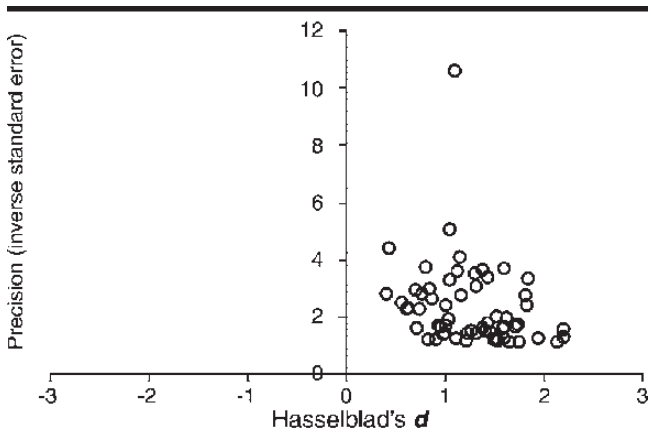

Figure 2. Graph shows assessment of publication bias with funnel plot asymmetry. Hasselblad diagnostic difference *(d)* represents a global measure of test accuracy. The shape of the plot was skewed, indicating a lack of small studies with small effect sizes.

Exploring Sources of Heterogeneity

Given the constant high specificity across studies, further analyses focused on the influence of study characteristics on the reported sensitivity of trauma US.

Table 2 summarizes associations between methodologic standards and predicted sensitivity gained from meta-regression.

Methodologic rigor had a major effect on accuracy estimates, and it confounded all further calculations. Studies missing certain design features yielded higher sensitivity than did investigations with stricter adherence to methodologic standards.

With univariate analysis, issues related to the choice of the particular reference test substantially influenced sensitivity, as did blinding and accurate reporting of patient flow (that is, specification of the number of screened patients and those who later dropped out).

Table 3 summarizes other variables that influenced sensitivity (eg, publication language, proportion of male subjects, mean age of the study population, proportion of CT scans, frequency of US probes, and area of expertise of operators). There was a slight association between predicted sensitivity and injury severity scores; however, data could be abstracted from only 23 investigations.

Figure 7 shows a linear decrease in predicted US sensitivity between 1980 and 2004, although stratified analysis revealed effect modification by increasing methodologic strength during the period studied.

The use of a single reference test, announcement of the number of screened patients, and specification of CIs independently contributed to the final multivariate model. Reports providing confidence limits met a median of 16 additional methodologic items (range, 12–17 items) compared with a median of 13 items (range, five to 20 items) fulfilled by the remaining trials. Unexplained variance of this model was 0.0112, compared with 0.0205 estimated with the constant-only model, without inclusion of covariates.

No investigation fulfilled all of these standards. Pooled sensitivity of 16 studies (that included a total of 1309 participants) that verified US findings with a single reference test was only 66.0% (95% CI: 56.2%, 75.8%), compared with 83.2% (95% CI: 79.5%, 86.9%) in case of multiple reference tests.

DISCUSSION

In 2001, we performed the first meta-analysis of prospective investigations of trauma US and showed its limited capacity to exclude intraabdominal injury (32). In retrospect, we missed some important issues. As a major drawback of our early review, we had not graded methodologic quality or explored other possible sources of heterogeneity.

In this study, we used a searching algorithm that was far more sensitive than that used in our previous study. This study also included nearly twice as many investigations. We arrive at the same conclusions; that is, a sonogram that is

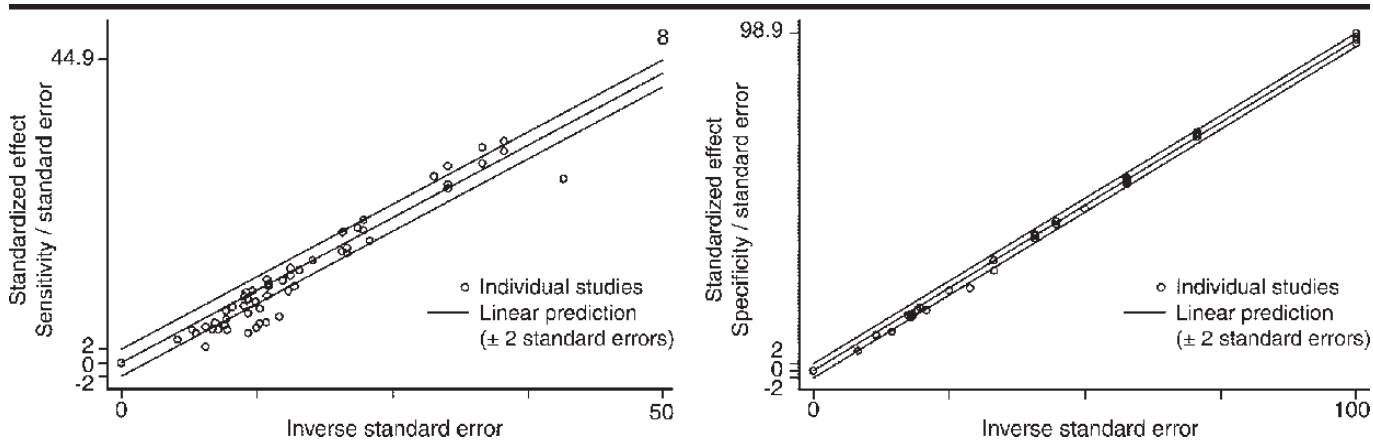

Figure 3. Assessment of heterogeneity with Galbraith diagrams. Note the scattering of sensitivity (left) compared with the homogenous distribution of specificity (right).

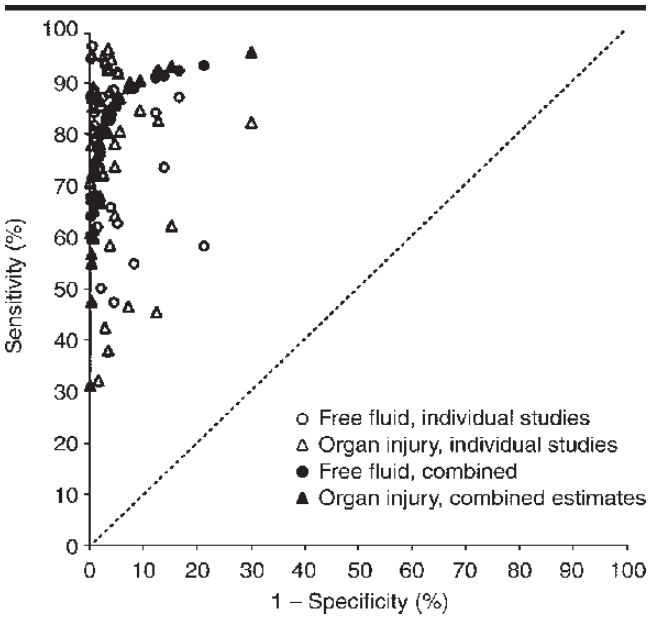

Figure 4. Summary receiver operating characteristics of FAST and FAST+ investigations. The scatterplot matrix and the summary receiver operating characteristic curves provide no evidence of a difference in test characteristics between both US protocols. The dotted line represents a test with no discriminatory ability.

positive for fluid or organ damage is decisive, whereas a negative sonogram is not. However, this study provides more detailed insight into the distribution of effect measures, and it contributes empirical evidence of the size and direction of design-related bias (33). It is open to debate whether the present results are confined to the issue of trauma US, or if they might suit diagnostic test research in general. There is no accepted hierarchy of validity criteria, and it is simply not reasonable to assign levels of universal importance to methodologic features. Standards that influence the accuracy of a particular test may not apply to another.

As a common finding of meta-analyses, methodologic quality of the original studies was, at best, average. Those who perform meta-analysis are often blamed for monotonously criticizing study quality. There is no real alternative available to caregivers; therefore, they must use the best available scientific data and take its possible limits into account.

Of note, after correcting for methodologic drawbacks, pooled sensitivity of trauma US settled close to 65% compared with a combined sensitivity of 80% projected from the entire pool of studies. Assuming a constantly high specificity of around 99%, corrections for bias shift negative likelihood ratios from 0.20 to 0.35. In other words, quality-adjusted estimates of test accuracy suggest minor changes in the pretest probability of abdominal trauma by negative US results (34).

Although perfect specificity is desirable for a screening tool to avoid unnecessary additional tests, it must provide high sensitivity so that a certain condition is not missed. One might ask whether a screening test that fails to lead to recognition of 11% of patients with abdominal injury is helpful in clinical practice.

Several points of our meta-analysis merit further discussion. First, application of the STARD checklist as a measure of methodologic rigor is beyond its intended use. The STARD steering committee admitted the limited evidence of linking particular items to potential bias (12).

Most suggestions provided by STARD match the methodologic norms included in QUADAS and show that they consistently affect accuracy estimates. Nevertheless, the expanded use of STARD may have introduced artificial relationships. For example, it is not obvious why providing measures of numerical precision independently affected sensitivity. We interpreted CIs as a surrogate of consequential planning, accomplishment, and analysis of the particular study. We cannot, however, exclude a false-positive correlation due to chance.

Clearly, bad reporting does not necessarily mean bad performance of a study (35). We did not contact authors, which might have clarified some controversies. We tried to contact researchers during

our preceding review, but we were unable to obtain unpublished information.

Second, interobserver agreement ranged from poor to almost perfect. Although reviewers consistently agreed on items that later significantly affected test sensitivity, different ideas of bias related to certain design features might influence the external validity of our results.

Assessing retest reliability or internal consistency (usually measured with Cronbach α) of STARD or QUADAS was beyond the scope of this study. The inconsistency in ratings may point to the limited utility of these appraisal tools and should prompt further investigation.

Third, we restricted our search to prospective investigations, which ignore important evidence from reviews of trauma registries. We hypothesized that data collection according to a prospectively defined protocol minimizes misclassification of exposure and disease. However, prospective patient sampling did not influence sensitivity in the meta-regression model, and it may be of minor importance if authors respect other methodologic key items.

Sirlin et al (36) recently published an excellent retrospective review of a huge trauma database. The study was extensive and provided high internal validity. Among 4000 patients who fulfilled the entry criteria for the multiple trauma outcomes study, 3641 had true-negative US findings, whereas 38 had false-negative findings.

Sirlin et al (36) infer from their data that negative findings at screening US have high clinical value. We think that this interpretation is problematic. US produced a high negative predictive value (99%), which is not surprising given the low prevalence of abdominal injury in the studied population (with a maximum of 9%, if one assumes there were no false-positive findings). The published data comply with a range in US sensitivity of 0% to 89% (if all 321 positive sonograms had been either false-positive or true-positive) and a range in US specificity of 92% to 100% (assuming opposite scenarios). These test characteristics compare well with those noted in this meta-analysis, further centering the projected point estimates.

Excluding citations after judging titles or abstracts bears the risk of omitting potentially relevant work. We adhered to accepted methods to avoid this bias, however, there is little empirical evidence on the possible effect of falsely neglected studies on the findings of a meta-analysis.

US findings must always be interpreted in the clinical context (36). The likelihood of intraabdominal disasters depends on the physical condition at arrival and the presence of index or multiple injuries. As a convenient means for scheduled bedside examinations, US may objectify progressive bleeding in case of clinical worsening. The available data did not allow for a more detailed analysis of the accuracy of serial scans, which may represent a major, if not the only, advantage of US over CT.

Regardless of US findings, hemodynamically stable patients will almost always undergo CT scanning before they are transferred to the intensive care unit. This is reasonable since a negative sono-

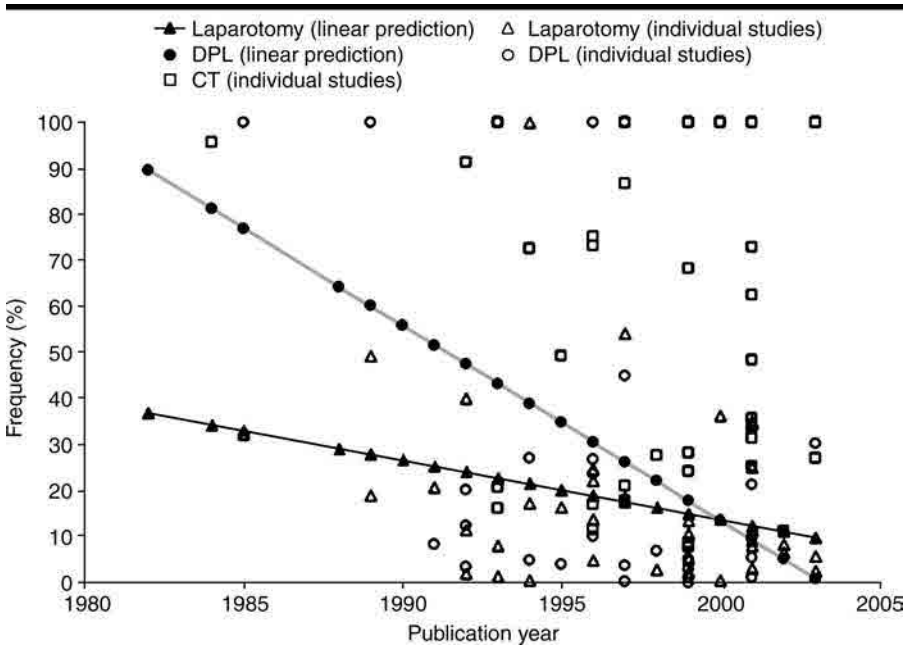

Figure 5. Graph shows significant decrease in the number of invasive procedures over time. Because of the wide distribution of CT frequencies across studies, we did not fit a regression line.

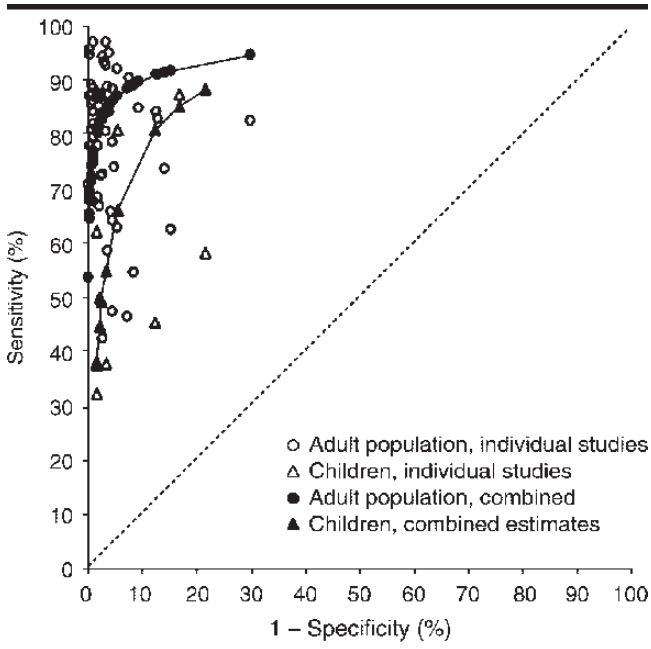

Figure 6. Graph shows differences in receiver operating characteristics of US between pediatric and adult populations. Note the poorer test performance of US in children. The dotted line represents a test with no discriminatory ability.

TABLE 3
Influence of Variables Other than Methodologic Standards on Predicted Sensitivity of Trauma US

Variable	No. of Studies Providing Sufficient Information	Predicted Difference in Sensitivity (%)	*P* Value
Publication year	62	−0.9 (−1.8, −0.1)	.025
Recruitment period	56	0.1 (−0.1, 0.2)	.368
Mean age	51	0.8 (0.5, 1.2)	.000
Proportion of male subjects	47	0.3 (0.0, 0.7)	.128
Mean injury severity score	23	0.4 (−0.4, 1.1)	.334
Proportion of CT scans	40	−0.2 (−0.4, −0.1)	.004
Proportion of diagnostic peritoneal lavage procedures	30	0.0 (−0.2, 0.1)	.621
Proportion of laparotomies	39	0.0 (−0.3, 0.3)	.977
Transducer frequency (≥3.5 MHz vs <3.5 MHz)	52	11.9 (2.5, 21.5)	.013
Language (English- vs non-English-language publications)*	62	−12.9 (−25.4, 0.5)	.041
Operator (radiologist vs surgeon)	60	8.7 (−1.9, 19.3)	.110

Note.—Negative predicted differences indicate overestimation of sensitivity in studies lacking the particular methodologic standard. For continuous data, sensitivity changes with any single unit of the variable of interest. Data in parentheses are 95% CIs.

* English- and non-English-language publications respected a median number of eight (range, five to 10) and 13 (range, five to 20) methodological standards, respectively.

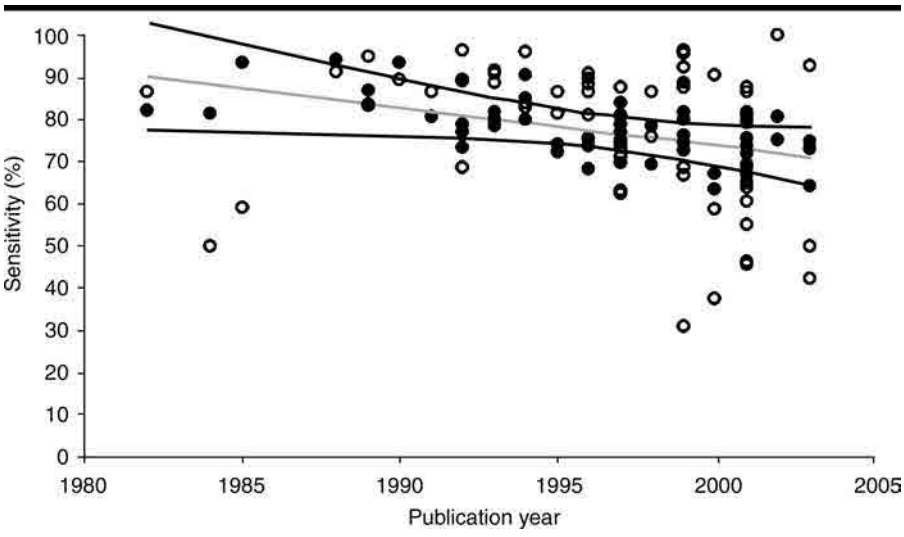

Figure 7. Graph shows the time-dependent decrease in reported US sensitivity was confounded by increasing methodologic quality. The regression line (with 95% confidence limits) was fitted between publication year and sensitivity estimates (○). ● = predictions adjusted for the number of methodologic standards.

gram is not convincing, whereas a positive sonogram is definite and warrants both organ injury scaling and exclusion of accompanying injury. US contributes little to decision making in these subjects. In hypotensive patients, because of its unsatisfactory sensitivity, a negative sonogram hardly decreases the high prior probability of intraabdominal injury and impels enforcement of the correct diagnosis by definite standards.

Several authors (37–39) emphasized the prognostic importance of increasing quantities of fluid and the ability of US scores to signal the need for therapeutic laparotomy. One might speculate that a continuous or ordinal classification of hemoperitoneum, rather than a dichotomous classification, improves the receiver operating characteristics of FAST.

On the other hand, other researchers (40,41) stressed the lack of free intraabdominal fluid as a limit of screening US. The prevalence of organ tears without accompanying hemoperitoneum occasionally exceeds 35% (42). Also, US rarely depicts hollow viscus rupture. This entity clearly remains a domain of CT scanning (43,44).

With continuing developments in CT technology, such as fast multi–detector row CT scanners that are almost at the initial point of care and skilled interdisciplinary critical care teams, there are now few subjects whose medical condition precludes definite diagnostic imaging in a reasonable time frame (45–47).

The possible benefits of emergency US are confined to patients with refractory hypotension and positive findings. Many protocols excluded potentially eligible patients who, by clinical judgment alone, required laparotomy without delay. Obviously, this would have increased rather than decreased the number of true-positive findings; however, it cannot be tested on a grand scale because the number of excluded subjects was rarely disclosed.

It is unclear how much importance doctors assign to a predictably positive sonogram in their decision to perform surgery in hypotensive patients with a high prior probability of abdominal trauma. Fryback and Thornbury (48) proposed six factors to consider in the appraisal of diagnostic test research: *(a)* technical measures, *(b)* accuracy, *(c)* influence on diagnostic thinking (that is, the difference between prior and posterior odds of disease), *(d)* influence on therapeutic decisions, *(e)* patient outcome, and *(f)* societal benefits.

The third factor is of interest in US-guided emergency laparotomy. In a decision analysis report, Brown et al (49) showed a linear decrease in the expected utility of US with rising prevalence of intraabdominal injury. This supports the thesis of a reassuring rather than an essential role of US findings in inducing further actions.

Meta-analyses are valuable instruments with which to generate theories. Because of their retrospective nature, however, they cannot be used to prove hypotheses. Variations in the individual prevalence of abdominal injury, local differences in trauma algorithms, and heterogenous source populations increase the risk of comparing apples with oranges. Obviously, multivariate random-effects modeling of published information reduced but did not fully dissolve the degree of unexplained variance. Demographic details and other important study features that might have contributed to heterogeneity were only partially available. Thus, we had limited choices in building hierarchical models on variables other than methodologic standards (eg, transducer frequency and operator expertise).

Researchers who are planning to con-

duct a diagnostic study must deal with various systematic errors and proper methods to minimize them efficiently. The same applies to the reader of a scientific article who wants to appraise the validity of the published results. A detailed assessment of methodologic strengths, as guided by the STARD checklist or the QUADAS instrument, may affect conclusions drawn from studies of diagnostic test accuracy. Referring to US performed to assess abdominal trauma, our data warrant additional research on topics like serial examinations, quantitative sensitivity, different end points, and influence of operator expertise, including proper methods to corroborate US findings.

The future reputation of US in the assessment of abdominal trauma will depend on the methodologic rigor of investigations currently under way.

Acknowledgment: We thank the Cochrane Injuries Group, London School of Hygiene and Tropical Medicine, London, England, for contributing additional information.

References

1. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ 1998; 317:1185–1190.
2. MacLehose RR, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AM. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. Health Technol Assess 2000; 4:1–154.
3. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research: getting better but still not good. JAMA 1995; 274:645–651.
4. Lijmer JG, Mol BW, Heisterkamp S. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282:1061–1066.
5. Holm HH, Kristensen JK, Rasmussen SN, Pedersen JF, Hancke S. Indications for ultrasonic scanning in abdominal diagnostics. J Clin Ultrasound 1974; 2:5–15.
6. Tiling T, Bouillon B, Schmid A, Schweins M, Steffens H. US in blunt abdominothoracic trauma. In: Borden J, Algoewer M, Ruedi T, eds. Blunt multiple trauma. New York, NY: Marcel Dekker, 1990; 415–433.
7. Healey MA, Simons RK, Winchell RJ, et al. A prospective evaluation of abdominal US in blunt trauma: is it useful? J Trauma 1996; 40: 875–883.
8. Rose JS, Levitt MA, Porter J, et al. Does the presence of US really affect computed tomography scan use? a prospective randomized trial of US in trauma. J Trauma 2001; 51:545–550.
9. American College of Surgeons. Advanced Trauma Life Support (ATLS) Student Manual. 7th ed. Chicago, Ill: American College of Surgeons, 2003.
10. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA 1994; 271:703–707.
11. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? The Evidence-Based Medicine Working Group. JAMA 1994; 271:389–391.
12. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Radiology 2003; 226:24–28.
13. Whiting P, Rutjes AW, Retsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003; 3:25.
14. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data analytic approaches and some additional considerations. Stat Med 1993; 12:1293–1316.
15. Song F, Khan KS, Dinnes J, Sutton AJ. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. Int J Epidemiol 2002; 31:88–95.
16. Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. Psychol Bull 1995; 117:167–178.
17. Egger M, Davey Smith G, Schneider M, Miner C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997; 315:629–634.
18. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Methods for meta-analysis in medical research. Chichester, England: Wiley, 2000.
19. Lijmer JG, Bossuyt PMM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. Stat Med 2002; 21:1525–1537.
20. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Stat Med 1999; 18:2693–2708.
21. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Stat Med 2002; 21:1559–1573.
22. Greenland S, O'Rourke K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. Biostatistics 2001; 2:463–471.
23. Greenland S. Quality scores are useless and potentially misleading. Am J Epidemiol 1994; 140:300–301.
24. Hosmer DW, Lemeshow S. Applied logistic regression. New York, NY: Wiley, 2000.
25. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement—quality of reporting of meta-analyses. Lancet 1999; 354:1896–1900.
26. Pak-art R, Sriussadaporn S, Sriussadaporn S, Vajrabukka T. The results of focused assessment with sonography for trauma performed by third year surgical residents: a prospective study. J Med Assoc Thai 2003; 86(suppl 2): S344–S349.
27. Brown MA, Casola G, Sirlin CB, Patel NY, Hoyt DB. Blunt abdominal trauma: screening US in 2,963 patients. Radiology 2001; 218: 352–358.
28. McKenney KL, Nuñez DB, McKenney MG, Asher J, Zelnick K, Shipshak D. Sonography as the primary screening technique for blunt abdominal trauma: experience with 899 patients. AJR Am J Roentgenol 1998; 170:979–985.
29. Krupnick AS, Teitelbaum DH, Geiger JD, et al. Use of abdominal US to assess pediatric splenic trauma: potential pitfalls in the diagnosis. Ann Surg 1997; 225:408–414.
30. Boulanger BR, McLellan BA, Brenneman FD, et al. Emergent abdominal sonography as a screening test in a new diagnostic algorithm for blunt trauma. J Trauma 1996; 40:867–874.
31. Nunes LW, Simmons S, Hallowell MJ, Kinback R, Trooskin S, Kozar R. Diagnostic performance of trauma US in identifying abdominal or pelvic free fluid and serious abdominal or pelvic injury. Acad Radiol 2001; 8:128–136.
32. Stengel D, Bauwens K, Sehouli J, et al. Systematic review and meta-analysis of emergency US for blunt abdominal trauma. Br J Surg 2001; 88:901–912.
33. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004; 140: 189–202.
34. Goodman SN. Toward evidence-based medical statistics. II. The Bayes factor. Ann Intern Med 1999; 130:1005–1013.
35. Soares HP, Daniels S, Kumar A, et al. Bad reporting does not mean bad methods for randomised trials: observational study of randomised controlled trials performed by the Radiation Therapy Oncology Group. BMJ 2004; 328:22–24.
36. Sirlin CB, Brown MA, Andrade-Barreto OA, et al. Blunt abdominal trauma: clinical value of negative screening US scans. Radiology 2004; 230:661–668.
37. Ma OJ, Kefer MP, Stevison KF, Mateer JR. Operative versus nonoperative management of blunt abdominal trauma: role of US-measured intraperitoneal fluid levels. Am J Emerg Med 2001; 19:284–286.
38. Huang MS, Liu M, Wu JK, Shih HC, Ko TJ, Lee CH. US for the evaluation of hemoperitoneum during resuscitation: a simple scoring system. J Trauma 1994; 36:173–177.
39. McKenney KL, McKenney MG, Cohn SM, et al. Hemoperitoneum score helps determine need for therapeutic laparotomy. J Trauma 2001; 50:650–654.
40. Emery KH, McAneney CM, Racadio JM, Johnson ND, Evora DK, Garcia VF. Absent peritoneal fluid on screening trauma US in children: a prospective comparison with computed tomography. J Pediatr Surg 2001; 36:565–569.
41. Taylor GA, Sivit CJ. Posttraumatic peritoneal fluid: is it a reliable indicator of intraabdominal injury in children? J Pediatr Surg 1995; 30:1644–1648.
42. Shanmuganathan K, Mirvis SE, Sherbourne CD, Chiu WC, Rodriguez A. Hemoperitoneum as the sole indicator of abdominal visceral injuries: a potential limitation of screening abdominal US for trauma. Radiology 1999; 212:423–430.
43. Allen TL, Mueller MT, Bonk RT, Harker CP, Duffy OH, Stevens MH. Computed tomographic scanning without oral contrast solution for blunt bowel and mesenteric injuries in abdominal trauma. J Trauma 2004; 56:314–322.
44. Hawkins AE, Mirvis SE. Evaluation of bowel and mesenteric injury: role of multidetector CT. Abdom Imaging 2003; 28:505–514.
45. Foley WD. Special focus session: multidetector CT: abdominal visceral imaging. RadioGraphics 2002; 22:701–719.
46. Rademacher G, Stengel D, Siegmann S, Petersein J, Mutze S. Optimization of contrast agent volume for helical CT in the diagnostic assessment of patients with severe and multiple injuries. J Comput Assist Tomogr 2002; 26:113–118.
47. Philipp MO, Kubin K, Hormann M, Metz VM. Radiological emergency room management with emphasis on multidetector row CT. Eur J Radiol 2003; 48:2–4.
48. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991; 11:88–94.
49. Brown CK, Dunn KA, Wilson K. Diagnostic evaluation of patients with blunt abdominal trauma: a decision analysis. Acad Emerg Med 2000; 7:385–396.