Diagnostic value of tests that discriminate between exudative and transudative pleural effusions. Primary Study Investigators

JE Heffner, LK Brown and CA Barbieri

Chest 1997;111;970-980

Diagnostic Value of Tests That Discriminate Between Exudative and Transudative Pleural Effusions*

John E. Heffner, MD, FCCP; Lee K. Brown, MD, FCCP; and Celia A. Barbieri, MS, for the Primary Study Investigators[†]

Study objective: **To (1) determine appropriate decision thresholds and diagnostic accuracies for pleural fluid (PF) tests that discriminate between exudative and transudative pleural effusions, and (2) evaluate the quality of the primary investigations.**
Design: **Formal meta-analysis of studies that report the diagnostic value of pleural fluid tests.**
Setting: **Data collected from international academic medical centers.**
Patients: **Hospitalized patients undergoing thoracentesis for pleural effusions.**
Interventions: **Primary investigators were requested to transmit original data from patients described in their studies.**
Measurements and results: **Eight primary studies described 1,448 patients with one or more of the following tests: protein (P)-PF, P-PF/serum ratio (R), bilirubin (BILI)-R, lactate dehydrogenase (LDH)-PF, LDH-R, cholesterol (C)-PF, C-R, and albumin gradient. We found that all eight tests had similar diagnostic accuracies when evaluated by receiver operating characteristic (ROC) analysis except for BILI-R, which was less diagnostically accurate. Decision thresholds determined by ROC analysis differed from previously reported values for LDH-PF (>0.45 upper limits of normal) and C-PF (>45 mg/dL). Paired and triplet test combinations tended to have higher diagnostic accuracies compared with individual tests, but examination of the odds ratios with 95% confidence intervals did not identify a clearly superior test combination. Limitations of the primary studies presented a high likelihood of bias affecting their results.**
Conclusions: **Several strategies exist for clinicians in utilizing PF tests to classify effusions as exudates or transudates but accurate interpretations of these test results will require better designed studies.** ***(CHEST 1997; 111:970-80)***

Key words: diagnosis; pleural disease; pleural effusion; receiver operating characteristic

Abbreviations: A=albumin; AUC=area under the curve; BILI=bilirubin; C=cholesterol; CI=confidence interval; G=gradient; LDH=lactate dehydrogenase; P=protein; PF=pleural fluid; PV=predictive value; ROC=receiver operating characteristic

Assessing the exudative or transudative nature of a pleural effusion is the initial step in determining its etiology. Early studies examined the diagnostic value of pleural fluid (PF) specific gravity[1] and PF protein (P)[2] for identifying exudative effusions. A subsequent landmark study by Light and coworkers[3] reported that a PF to serum P ratio (R) >0.5, PF to serum lactate dehydrogenase (LDH) ratio >0.6, and PF LDH concentration >200 IU/L combined in a parallel "or" rule (criteria of Light et al) resulted in fewer misclassifications. The PF LDH criterion was later modified to more than two thirds the upper limit of a laboratory's normal LDH range to account for variations in assay methods.[4] Subsequent studies have examined the diagnostic accuracy of other PF tests, some of which[5] have recently been considered superior to the criteria of Light et al.

Studies that report the diagnostic value of PF tests, however, vary considerably in case mix, study design, and the statistical analyses employed. Because of these variations and the small sample sizes reported, we have gained the impression that valid conclusions cannot be drawn from individual studies regarding the comparative diagnostic accuracies of PF tests for identifying exudative effusions. Therefore, we performed a meta-analysis using formal

*From the Department of Medicine, St. Joseph's Hospital and Medical Center, Phoenix, and the University of Arizona Health Sciences Center, Tucson, Arizona.
[†]A complete list of participants is located in the appendix.
Supported by the Mercy Health Services Research Fund.
Manuscript received August 30, 1996; revision accepted November 11.
Reprint requests: Dr. Heffner, St. Joseph's Hospital and Medical Center, 350 West Thomas Road, Phoenix, AZ 85013

techniques examining the discriminative properties of previously reported PF tests. We received original investigative data from the primary study investigators so as to create a large data set and the best estimate of the tests' relative and absolute discriminative properties. We also (1) assess the validity and reliability of the scientific evidence from the primary studies, (2) make recommendations for the clinical use of PF tests, and (3) provide recommendations for study design of future investigations examining the diagnostic accuracy of PF tests.

Materials and Methods

The study was a meta-analysis of the published English-language literature designed to examine the diagnostic accuracy of the eight biochemical tests reported to be of value in separating exudative from transudative pleural effusions. We used statistical techniques to generate receiver operating characteristic (ROC) curves and odds ratios with their 95% confidence intervals (CIs) to compare, respectively, the discriminative properties of individual tests and tests combined in parallel (test results considered simultaneously) with an "or" rule.

Study Identification

Articles that report the discriminative properties of PF tests in separating exudates from transudates published since 1976 were found by searching MEDLINE. Two investigators and a professional librarian independently developed search strategies to identify pertinent studies. Additional articles were identified by reviewing the reference lists of retrieved articles, review articles on pleural effusions, and pulmonary textbooks. The principal investigators of the retrieved articles were asked to participate in the meta-analysis by providing their original data.

Study Selection

Two authors (J.E.H. and C.A.B.) independently examined articles for study eligibility on the basis of preestablished inclusion criteria. We included studies if (1) original data were available from the investigators or if unavailable data sets including results of individual PF tests could be extracted from the published articles, (2) sufficient information was available to ascribe all PF test results to individual patients, and (3) individual patients were categorized as having exudative or transudative pleural effusions by a reference standard. A study was accepted for analysis if both examiners agreed. Disagreements were reviewed independently by a third author (L.K.B.) and the article was accepted or excluded by majority rule.

Determining Diagnostic Cutoff Points of Individual Tests

Studies were identified that examined the discriminative properties of eight PF tests: PF to serum P ratios (P-R), P-PF, LDH-PF, PF to serum LDH ratios (LDH-R), PF cholesterol (C-PF), PF to serum C ratios (C-R), PF to serum albumin gradients (A-G), and PF to serum bilirubin ratios (BILI-R). ROC curves were generated for each of the eight individual PF tests previously reported using all of the data available for each test. Because not every study patient had all eight PF test results available, the size of the data sets used to generate each of the ROC curves varied.

The entire available data sets for each test were used to determine cutoff points to the highest possible level of accuracy and precision. Cutoff points for individual tests were selected by calculating the slope of a line on the basis of the costs of false results and prevalence of exudates in evaluated patients (slope=false-positive cost/false-negative cost×[1-prevalence]/prevalence).[6] The prevalence was set at 0.5 to indicate maximum uncertainty as to the exudative or transudative nature of a patient's pleural effusion, which is the usual clinical situation. Because diagnostic thoracentesis is a screening test to identify patients with exudative pleural effusions who require additional diagnostic studies, the ratio of false-positive to false-negative cost was set at 0.5. This strategy maximizes each test's sensitivity for diagnosing exudative pleural effusions, which is the goal of screening studies.[7] The calculated slope determined by the prevalence-cost equation for generating a line was 0.5. The cutoff point was selected at the intersection of the ROC curve with the generated line.

Selection of a cutoff point for LDH-PF represented a special case in that absolute values needed to be indexed against the upper limits of normal for the reporting laboratory to control for variations in assay techniques. A cutoff point was determined, therefore, as a fraction of the upper limits of normal.

Determining the Comparative Diagnostic Accuracies of Individual Tests

Sensitivities, specificities, positive predictive values (+PV), negative predictive values (−PV), odds ratios, and areas under the ROC curves (AUC) with 95% CIs were calculated for each of the eight individual PF tests using the cutoff points derived from the meta-analysis. The AUC was the primary test function used to compare the diagnostic accuracies of individual tests; the odds ratio was also calculated to allow comparison of single test strategies with strategies that employed multiple tests in combinations. An attempt was made to include into these calculations only the test results of patients who had all of the PF tests performed on their PF because accurate comparisons of diagnostic tests require that all patients or random samples of patients are submitted to all of the compared tests.[8] Because this effort resulted in a markedly decreased sample size (n=299), most of the analyses were based on a larger subset of patients who had results available from all of the following tests: P-R, LDH-PF, LDH-R, C-PF, and C-R. Analysis of the diagnostic accuracy of P-PF was performed by evaluating a subset of patients who had the above-mentioned five tests performed in addition to P-PF. The same strategy was followed for analyzing data tests that determined A-G and BILI-R. Comparative diagnostic accuracies of all of the tests were reanalyzed, however, on the basis of the data subset (n=299) that included only patients who had all eight tests performed.

Determining the Correlation of Individual Test Results

The results of PF tests in clinical practice are commonly combined in a parallel manner with an "or" rule to increase the sensitivity for detecting exudative pleural effusions. Because a parallel strategy of combining test results works best if each of the combined tests identifies different subsets of patients who have disease, the pairwise Pearson product-moment correlations of the eight PF tests were calculated. Two tests were defined as correlated and possibly unsuitable for combining into a parallel testing strategy if the absolute value of their pairwise correlations was >0.75.

Determining the Comparative Diagnostic Accuracies of Combined Tests

Sensitivities, specificities, +PV, −PV, and odds ratios with 95% CI were calculated for all parallel testing strategies using the PF tests in paired and triplet combinations with an "or" rule. The diagnostic accuracies of individual tests and test combinations were primarily compared by considering the overlap of 95% CI limits calculated for the tests' odds ratio, which is an overall measure of a test's discriminative properties.[7] Secondary comparisons considered the overlap of the 95% CI values calculated for the tests' sensitivities and specificities. Patient data were entered into combinations consisting of pairs or triplets of P-R, LDH-PF, LDH-R, C-PF, and C-R if all patients had all of these tests performed. Smaller data sets were used for calculation of the diagnostic performance of combinations that included P-PF, A-G, and BILI-R because a smaller subset of patients were available in the primary studies who were evaluated by any one of these three tests in addition to all five of the other PF tests.

Assessment of Study Design

The reliability of the primary investigations included in the meta-analysis was reviewed independently by two of us (J.E.H. and L.K.B.) using prospectively developed criteria modified from previous reports.[8,9] Disagreements were resolved by the independent review of a third author (C.A.B.), and the majority opinion was accepted. The prospective criteria of study reliability included the following: (1) adequate description of sufficient reference standards; (2) independence of observations (study blinding); (3) uniform application of reference standards; (4) assessment of generalizability; (5) cohort assembly; and (6) a description of the biochemical testing techniques. Each component was scored as present, absent, or incomplete.

The reference standard was the gold standard method used to categorize effusions as exudative or transudative. Reference standards were considered "sufficient" if they were based on explicit, objective, and reproducible criteria beyond clinical judgment alone when such criteria were available for the disease being evaluated. Objective criteria for exudative effusions included studies such as positive pleural biopsy specimens for malignant cells or granulomas and positive PF cytologic test results or culture results, which firmly establish the exudative nature of a pleural effusion. Diagnoses that do not have objective reference standards, such as uncomplicated parapneumonic effusions, were accepted if adequate descriptions of clinical criteria for diagnosis were provided. Objective criteria for transudative effusions required the following: (1) an absence of clinical evidence of a disease commonly associated with exudative effusions; (2) tissue biopsy, laboratory result, or imaging documentation of a disorder commonly associated with transudative effusions; and (3) 3 or more months of patient observation to exclude delayed presentations of conditions associated with exudative effusions.

Independence of observations refers to the study methods used for blinding and preventing classification bias. Techniques evaluated included (1) methods used to blind the investigators (who collated the results of the PF chemical analyses) to the results of the reference standards, and (2) methods used to blind the clinicians, who applied the reference standards, to the PF test results.

Uniform application of reference standards indicates that a minimum battery of methods were applied consistently within the exudative and transudative pleural effusion populations to limit verification bias. Verification bias occurs when study patients who have an initial "negative" test result (*ie,* biochemical profile of a transudative effusion) are not sufficiently evaluated to exclude misclassification.

Generalizability was assessed by noting whether sufficient clinical information, such as age, gender, and underlying medical conditions, was provided to enable the reader to determine if the study results could be generalized to the reader's patient population.

Cohort assembly referred to the presence of an adequate spectrum of study patients and the detail by which the assembly of the cohort was described. This description should allow another investigator to assemble a similar cohort.

Description of the biochemical testing methods required that the techniques of collecting and testing the pleural fluid samples were described.

Data Analysis

The values for ROC AUC were calculated with a trapezoidal method[10] with values approaching 1.0 indicating high diagnostic accuracy.[11,12] The means and variances of AUC values were estimated by a bootstrap resampling procedure and used to determine 95% CI values.[10] Nonoverlapping 95% CI values indicated significant differences in AUC values and, therefore, differences in diagnostic accuracies between individual tests at the $p<0.05$ level. Sensitivity, specificity, PVs, likelihood ratios, and odds ratios with 95% CI values were calculated by standard techniques.[13,14] Nonoverlapping 95% CI values indicated significant differences in odds ratio values and, therefore, differences in diagnostic accuracies between test combinations at the $p<0.05$ level. The Pearson product-moment correlations were calculated to determine the strength of the linear relationships between each pair of test results.[15] Inverse correlations generated negative values. All data analyses were performed with software (JMP statistics software version 3.1; SAS Institute; Cary, NC) on a personal computer (Macintosh 8500/80).

Results

Eleven studies were identified by the literature search.[16-26] Four studies with a total of 408 patients[17,20,21,24] were excluded because only patients with treated heart failure were examined[17] or original data were no longer available from contacted investigators and test results could not be extracted from the published report.[20,21,24] The remaining seven studies[16,18,19,22,23,25,26] fulfilled the entrance criteria and were included in the meta-analysis. These studies contained one or more PF test results from 1,448 patients. Seventy-four percent of the effusions were exudates and 26% were transudates. Underlying diagnoses are shown in Table 1. The number of test results available for each of the PF tests were as follows: 1,393 for P-R,[16,18,19,22,23,25,26] 1,438 for LDH-PF,[16,18,19,22,23,25,26] 1,388 for LDH-R,[16,18,19,22,23,25,26] 1,187 for P-PF,[16,18,19,22,23,25] 1,348 for C-PF,[16,18,19,22,25,26] 1,123 for C-R,[16,19,22,25,26] 386 for A-G,[16,23] and 303 for BILI-R.[16]

Determination of Cutoff Points for Individual Tests

Cutoff points derived from the ROC and cost-prevalence analyses using all of the available test

Table 1—*Underlying Diagnoses of the Entire Study Population and the Patients Who Had All Five of the Seven Tests Performed*

Condition	Entire Study Population	Five-Test Group*
Transudates	377	252
Age, yr†	54.6±20.5	54.4±20.3
Male, No. (%)	744 (54.5)‡	579 (54.9)
Female, No. (%)	620 (45.5)	475 (45.1)
Congestive heart failure	270	174
Cirrhosis	43	23
Nephrosis	39	33
Hypoalbuminemia	25	22
Exudates	1,071	802
Malignancy	438	320
Tuberculosis	296	245
Infection	185	121
Miscellaneous	73	62
Pulmonary embolism	36	30
Collagen vascular disease	17	5
Trauma	14	14
Pancreatitis	5	4
Postpericardiotomy syndrome	4	0
Pleuropericarditis	1	0
Vasculitis	1	1
Yellow nail syndrome	1	0
Total	1,448	1,054

*The five-test group comprised patients who had all of the following test results available: LDH-PF, LDH-R, P-R, C-PF, and C-R.
†Mean±SD.
‡Gender data for entire study population based on 1,364 patients with available information.

results and the cutoff points previously recommended in the literature are shown in Table 2.

Cutoff points for the individual components of the criteria of Light et al[3] (LDH-R, LDH-PF, and P-R) were compared with those recommended in the literature. ROC analysis confirmed previous recommendations for LDH-R (>0.6) and P-R (>0.5)[3] but differed compared with some or all previous reports in the determined cutoff points for LDH-PF and C-PF. ROC analysis selected the best cutoff point to be >0.45 of the upper limits of a laboratory's normal values for LDH-PF, which differs from those recommended for the criteria of Light, which are either >200 IU/L[3] or more than two thirds the upper limits of a laboratory's normal values.[4] The test combination that incorporates LDH-R, P-R, and LDH-PF with an "or" rule in this study uses 0.45 of the upper limits of normal as the cutoff for LDH-PF and is termed "modified Light's criteria."

Table 2—*Cutoff Points Derived From the Meta-analysis ROC Analysis and Cutoff Points Commonly Recommended by Previous Reports*

PF Test	Meta-analysis ROC Cutoff Point	Previously Reported Cutoff Points
P-PF	>2.9 g/dL	>3 g/dL[2]
P-R	>0.5	>0.5[3]
LDH-PF	>0.45 of upper limits of normal	>200 IU/L[3,18] >2/3 of upper limits of normal[4]
LDH-R	>0.6	>0.6[3]
C-PF	>45 mg/dL	>45 mg/dL[18] >54 mg/dL[25] >55 mg/dL[26] >60 mg/dL[19,22]
C-R	>0.3	>0.3[16,19,22,25,26]
A-G	≤1.2 g/dL	≤1.2 g/dL[16,23]
BILI-R	>0.6	>0.6[16,20]

Diagnostic Accuracies and Correlations of Individual Tests

Table 3 shows the diagnostic accuracies of individual PF tests. The AUC data demonstrate similar discriminative properties of the examined tests except for BILI-R, which had a lower diagnostic accuracy compared with the other tests. Examination of the odds ratios and their 95% CI values confirms the lower diagnostic accuracy of BILI-R and suggests that P-R may perform marginally better than LDH-R ($p<0.05$). The pairwise correlations of the individual tests are shown in Table 4. The following pairs of tests were correlated ($r>0.75$) and would not, therefore, be expected to perform well together in parallel combination test strategies because inclusion of the additional test would not enhance the best test in the combination: P-PF and P-R, P-R and A-G, LDH-PF and LDH-R, and C-PF and C-R.

Diagnostic Accuracies of Combined Tests

Data for test combinations that include BILI-R are not shown because this individual test had a significantly lower diagnostic accuracy compared with other individual tests, data were derived from only one study, and all paired and triplet test combinations that included BILI-R had the lowest specificity (<55%) among individual tests. The remaining seven individual tests were entered into all possible pairwise (n=21) (Fig 1) and triplet (n=35) combinations (Fig 2). Combinations that included only results for P-R, LDH-PF, LDH-R, C-PF, and C-R analyzed data from 1,054 primary study patients,[16,22,25,26] which represented patients who had results for all of these five tests. The underlying diagnoses of these patients are shown in Table 1. Any combinations that included P-PF examined 801 patients, which represented patients who had results for all individual tests except for A-G.[16,22,25] Any combinations that included A-G included 310 patients, which represented the subset of patients who

Table 3—*Diagnostic Accuracy of Individual PF Tests for Identifying Exudative Pleural Effusions

PF Test	Sensitivity, % (95% CI)	Specificity, % (95% CI)	+PV, % (95% CI)	−PV, % (95% CI)	+LR	−LR	OR (95% CI)	AUC (95% CI)
P-PF (n=1,187)	91.5 (89.3 to 93.7)	83.0 (77.6 to 88.4)	94.6 (93.8 to 97.1)	75.0 (69.1 to 80.9)	5.38	0.10	52.6 (33.2 to 85.8)	94.2 (92.6 to 95.9)
P-R (n=1,393)	89.5 (87.4 to 91.6)	90.9 (87.4 to 94.5)	96.9 (95.6 to 98.1)	73.3 (68.4 to 78.1)	9.85	0.12	85.4 (53.6 to 141.8)	95.4 (94.3 to 96.7)
LDH-PF (n=1,438)	88.0 (85.8 to 90.3)	81.8 (77.1 to 86.6)	93.9 (92.2 to 95.6)	68.3 (63.1 to 73.6)	4.84	0.15	33.1 (22.7 to 49.0)	93.3 (91.8 to 94.8)
LDH-R (n=1,388)	91.4 (89.4 to 93.3)	85.0 (80.6 to 89.4)	95.1 (93.5 to 96.6)	75.7 (70.7 to 80.7)	6.08	0.10	60.0 (39.7 to 92.9)	94.7 (93.4 to 96.0)
C-PF (n=1,348)	89.0 (86.8 to 91.2)	81.4 (76.6 to 86.2)	93.8 (92.1 to 95.5)	70.1 (64.8 to 75.3)	4.79	0.13	35.5 (24.3 to 52.8)	93.3 (91.7 to 94.8)
C-R (n=1,123)	92.0 (90.1 to 93.9)	81.4 (76.6 to 86.2)	94.0 (92.3 to 95.7)	76.3 (71.2 to 81.4)	4.95	0.10	50.5 (33.9 to 76.6)	94.1 (92.5 to 95.7)
A-G (n=386)	86.8 (82.2 to 91.4)	91.8 (86.4 to 97.3)	95.8 (93.0 to 98.7)	76.3 (68.6 to 83.9)	10.6	0.14	73.9 (34.2 to 180.9)	94.0 (91.3 to 96.6)
BILI-R (n=303)	84.3 (79.3 to 89.3)	61.1 (51.2 to 70.9)	82.3 (77.1 to 87.5)	64.4 (54.6 to 74.3)	2.17	0.26	8.4 (4.9 to 14.9)	81.3 (76.3 to 86.4)

*LR=likelihood ratio; OR=odds ratio.

had results available for all seven individual tests evaluated in combinations.[16]

Paired test combinations that included pairs with high between-test correlations ($r>|0.75|$) did not perform better in most instances than the individual test with the higher diagnostic accuracy in the combination. Also, triplets that included two correlated tests did not perform better than the paired combination that remained when the test with the lower diagnostic accuracy of the correlated pair was excluded. Notably, the modified Light's criteria contained the correlated pair LDH-R and LDH-PF; removal of LDH-PF from the modified Light's criteria did not alter the combination's diagnostic performance (Figs 1 and 2).

Individual tests tended to have lower odds ratios compared with paired combinations and paired combinations of tests tended to have lower odds ratios compared with triplets. Extensive overlap of the 95% CI values for the odds ratios and other measured parameters did not allow selection of a superior combination of tests that clearly had the highest discriminative properties for exudative effusions. Several paired and triplet combinations that included A-G had the highest odds ratios but small sample sizes (n=310) produced wide CI values that overlapped other test combination results. Of note, the diagnostic accuracy of modified Light's criteria (sensitivity, 97.9% [95% CI, 96.9 to 98.9%]; specificity, 74.3% [95% CI, 68.9 to 79.7%]; odds ratio, 133.4 [95% CI, 78.5 to 240.4]) was not significantly better than two of the test combinations that included only PF tests. These latter combinations included the following: LDH-PF/C-PF (sensitivity, 97.5% [95% CI, 96.4 to 98.6%]; specificity, 71.9% [95% CI, 66.4 to 77.5%]; odds ratio, 100.1 [95% CI, 60.7 to 173.2]); P-PF/LDH-PF/C-PF (sensitivity, 98.4% [95% CI, 97.5 to 99.3%]; specificity, 70.4% [95% CI, 64.7 to 76.0%]; odds ratio, 143.9 [95% CI, 81.0 to 277.1]).

Table 4—*Pairwise Correlations of Individual Tests

Test	P-PF†	P-R	LDH-PF	LDH-R	A-G‡	C-PF	C-R	BILI-R‡
P-PF†	1.00							
P-R	0.93	1.00						
LDH-PF	0.28	0.26	1.00					
LDH-R	0.23	0.18	0.84	1.00				
A-G‡	−0.69	−0.78	−0.44	−0.39	1.00			
C-PF	0.68	0.69	0.21	0.17	−0.53	1.00		
C-R	0.72	0.74	0.26	0.18	−0.69	0.79	1.00	
BILI-R‡	0.37	0.34	0.13	0.16	−0.28	0.31	0.25	1.00

*Pairwise correlations based on results from 1,054 patients unless otherwise shown.
†Correlations with P-PF based on 801 patients.
‡Correlations with A-G and BILI-R based on 299 patients.

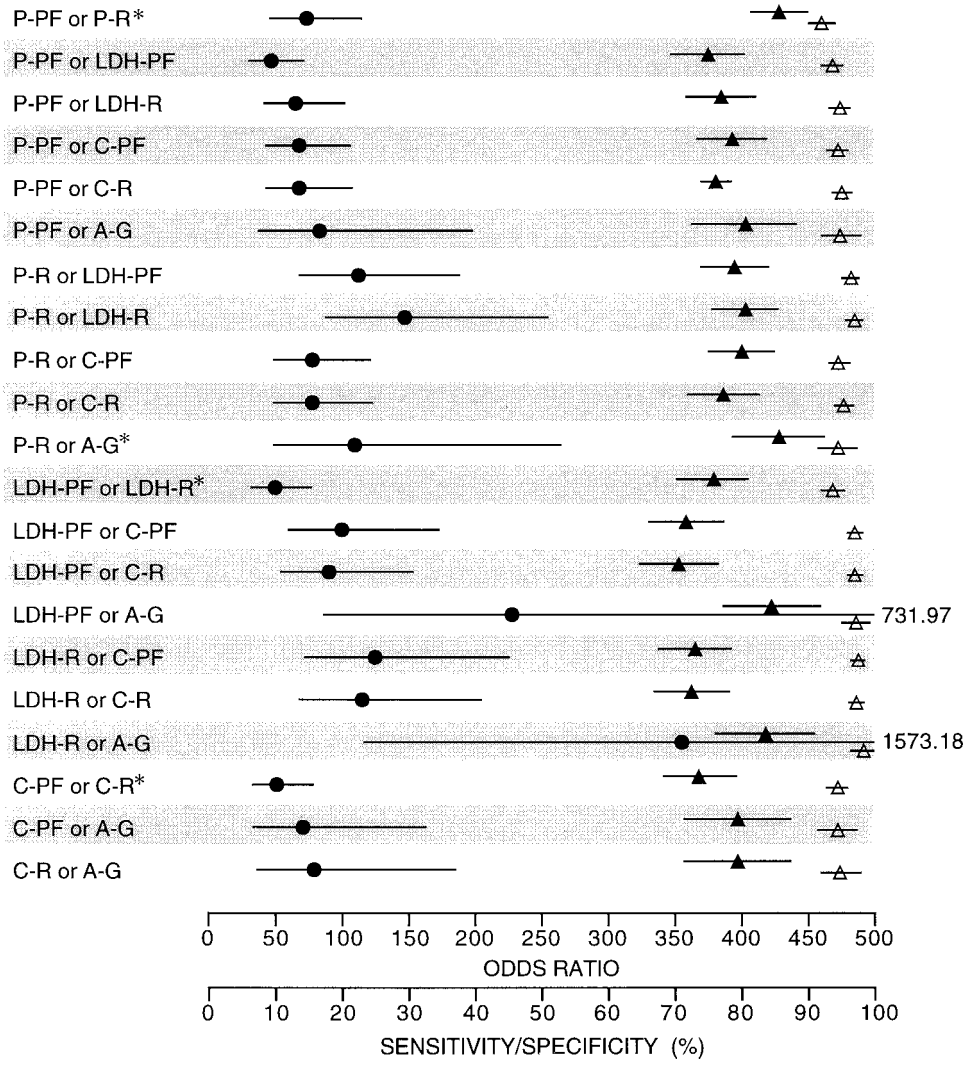

FIGURE 1. Discriminative properties of PF tests combined as pairs with an "or" rule. Open triangles represent sensitivities, closed triangles represent specificities, and closed circles represent odds ratios. Horizontal lines represent 95% CIs; values for the upper 95% confidence limit are provided if the horizontal bar is off scale. Test analyses were based on 1,054 data points except for combinations that included P-PF (n=801) and combinations that included A-G (n=310). Asterisk denotes combinations that contain correlated test pairs ($r>|0.75|$).

All of the combination tests were reanalyzed using the data set of 310 patients who had results from all seven PF tests available (data not shown). Although CIs were wider, the rank order of absolute values of odds ratios was unchanged.

Positive and negative PVs demonstrated considerable overlap in their 95% CIs between tests. For individual tests other than BILI-R, +PV ranged from 93.8 to 96.9% and −PV ranged from 68.3 to 76.3%. For paired combinations, +PV ranged from 91.0 to 95.4% and −PV ranged from 78.1 to 96.5%. For triplet combinations, +PV ranged from 89.4 to 93.5% and −PV ranged from 86.9 to 97.6%.

Assessment of Study Design

Of the seven primary studies, six provided adequate descriptions of the reference standard for the patients with exudative effusions. Conditions associated with transudative effusions were less adequately defined as reference standards: five studies completely and two studies partially excluded exudative causes of effusion; all seven studies defined the presence of a disease commonly associated with transudates; no study completely and six studies partially described explicit and clearly defined objective criteria for determining the presence of the

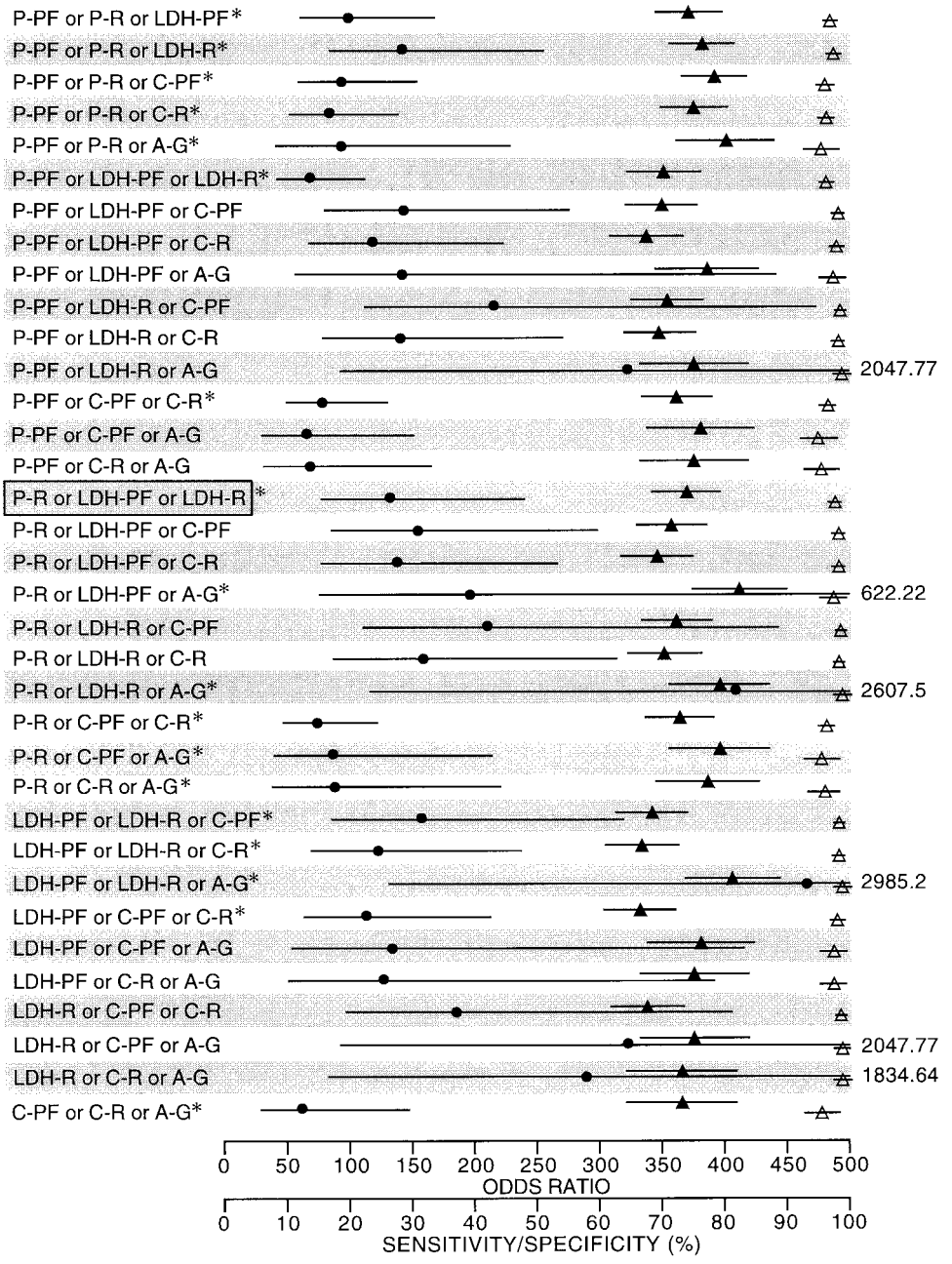

FIGURE 2. Discriminative properties of PF tests combined as triplets with an "or" rule. Open triangles represent sensitivities, closed triangles represent specificities, and closed circles represent odds ratios. Horizontal lines represent 95% CIs; values for the upper 95% confidence limit are provided if the horizontal bar is off scale. Test analyses were based on 1,054 data points except for combinations that included P-PF (n=801) and combinations that included A-G (n=310). Modified Light's criteria (P-R, LDH-PF, LDH-R) is enclosed in a box. Asterisk denotes combinations that contain correlated test pairs ($r>|0.75|$).

underlying disease. No study followed patients with transudative effusions to exclude verification bias, and no study submitted all patients to a minimum battery of diagnostic studies. None of the studies were blinded in any manner. Six of the published studies provided information on patient age and gender; these data were available from the primary investigators in one of the remaining two studies. All studies reported the patients' underlying diseases. Four of the studies provided an adequate spectrum of patients with effusions, six studies described the method of cohort assembly, and five studies described the biochemical assay techniques. PF and blood samples were obtained simultaneously in one

study, within the "same day" in one study, "usually" within the same day in one study, within 24 h in three studies, and within 8 h in one study.

Discussion

Determining the transudative or exudative nature of PF is the initial step in evaluating pleural effusions of uncertain etiology. This meta-analysis formally and critically examined the primary investigations that report the discriminative properties of various PF tests used for establishing the presence of an exudate. Our analysis resulted in two major categories of findings that pertained to (1) the diagnostic accuracies of the tests and (2) the quality of study design of the primary investigations.

First, BILI-R had the lowest diagnostic accuracy of studied individual tests and its addition to combinations of tests resulted in low discriminative properties. The remaining seven PF tests (P-PF, P-R, LDH-PF, LDH-R, C-R, C-PF, and A-G) did not differ significantly in their diagnostic accuracies when analyzed individually by ROC analysis. The meta-analysis examined the diagnostic accuracies of all permutations of test pairs and triplets rather than arbitrarily selected combinations. Combinations of tests tended to perform better and demonstrated higher sensitivities and odds ratios compared with individual tests. Wide overlap of odds ratio CIs, however, indicated that no single test or test combination could be clearly identified as being superior to others. Two test combinations that solely incorporated PF values (LDH-PF/C-PF and P-PF/LDH-PF/C-PF), however, had a similar diagnostic accuracy compared with Light's criteria, which is generally considered in the literature as the benchmark test combination.

Second, our evaluation of the primary investigations identified multiple study design limitations that would be expected to introduce bias into estimates of the tests' diagnostic accuracies. These limitations suggest opportunities for improving study design of future investigations in this area.

A strength of this meta-analysis is its use of ROC analysis and odds ratios calculated from a large set of pooled data to determine cutoff values and the comparative diagnostic accuracies of individual and combination tests. Previously reported cutoff values for some PF tests (Table 2) vary because of small sample sizes and the use of informal statistical techniques in the primary investigations. In the case of LDH-PF, the cutoff value of 200 IU/L (or more than two thirds the upper limit of a laboratory's normal LDH value) was originally selected by visual examination of dot plots in a small sample of 150 patients.[3] Subsequent investigations, which comprise the primary studies in this meta-analysis, have most often selected cutoff values and compared discriminative properties of different tests and test combinations on the basis of small study populations and the use of χ^2 or McNemar statistical analyses. These statistical tests can demonstrate statistically significant differences in study subgroups identified by various test strategies but do not formally compare the relative overall diagnostic accuracies of different PF tests or test combinations.

It is generally recognized that the AUC as determined by ROC analysis for individual tests and the odds ratios for parallel test combinations represent the most accurate techniques for comparing the discriminative properties of diagnostic tests.[7,11,12,27] ROC analysis generates a series of points along an ROC curve that represent joint sensitivities and specificities that could result from any specific cutoff value for defining an abnormal test result.[27] Comparing tests using single point estimates (single cutoff values), as reported in most previous studies, may be misleading because differences in measured discriminative properties between tests may result from the specific decision threshold chosen for analysis.

This meta-analysis selected cutoff values by first noting that screening studies benefit from having a high sensitivity to avoid a large number of false-negatives in evaluated patients.[28] In the instance of screening pleural effusions, it is more important to include a high proportion of the patients with exudates for evaluation (sensitivity) than to exclude a high proportion of patients with transudates (specificity). The cutoff values for the different tests were uniformally selected using a formal technique that maximized sensitivity without markedly decreasing specificity.[28] This technique recognizes that a fundamental characteristic of diagnostic tests is a linkage of sensitivity and specificity. This linkage produces a relatively large decrease in specificity for a small gain in sensitivity when cutoff values are selected in order to decrease false-negative results to an extreme degree. This relationship of sensitivity and specificity is graphically represented by the shape of the ROC curve, which rapidly drops in specificity but rises only marginally in sensitivity beyond an ideal cutoff value.

The primary studies and most expert reviews of pleural diagnostic testing have not emphasized this interrelationship of sensitivity and specificity. For instance, it has been stated that some test combinations, such as Light's criteria, have a higher sensitivity but a lower specificity compared with some individual tests, such as C-PF determinations.[4] This observation, however, does not result from a unique

characteristic of the reported tests but occurs as a fundamental feature of combining individual tests in parallel with an "or" rule. Combination tests in contrast to single tests in most instances have enhanced sensitivity but pay a price in lower specificity. The cutoff values reported in the present meta-analysis (Table 2) are the most accurate available in balancing joint values for sensitivity and specificity based on the pooled primary data.

With the exception of the inferior performance of BILI-R and test combinations that incorporated its values, the meta-analysis indicated that most studies had overlapping AUC and odds ratio CIs and, therefore, similar diagnostic accuracies. The statistical analyses failed to demonstrate that any particular test or test combination had superior diagnostic accuracy. The meta-analysis, however, provides insights for selecting a strategy for diagnostic testing of pleural effusions.

First, combining tests in parallel with "or" rules is useful if the discriminative properties of each of the combined tests are not highly correlated. Correlated tests used in combination will most likely identify a similar subset of true-positives compared with each of the component tests when used alone. Analysis of pairwise correlations of individual tests indicates that the pairs P-PF and P-R, P-R and A-G, LDH-PF and LDH-R, and C-PF and C-R are highly correlated. Test combinations that include these pairs would not be expected to perform better than the individual test or the paired combination that would remain after the least diagnostically accurate member of the correlated pair was removed from a test combination. The results of the meta-analysis confirmed this impression. Of note is the inclusion of the correlated pair LDH-PF and LDH-R in Light's criteria; excluding LDH-PF did not alter the diagnostic accuracy of the remaining test pair of LDH-R and P-R when these tests were combined in a parallel "or" rule in an abbreviated Light's criteria.

Second, the meta-analysis supported previous recommendations that diagnostic combinations utilizing individual tests or test combinations that only employed PF assays and did not require blood tests for calculating ratios or gradients may provide cost and convenience benefits without sacrificing diagnostic accuracy.[5,18,19] The combinations of LDH-PF/C-PF and P-PF/LDH-PF/C-PF performed as well as the modified Light's criteria combination, which incorporated blood results and calculated ratios. A strategy for excluding test combinations that incorporate tests with ratios or gradients is further supported by the meta-analysis observation that five of the seven primary studies allowed 24 h or more between PF and blood test determinations. Extensive differences in timing blood and PF collections would weaken conclusions regarding the diagnostic accuracies of the reported values for ratios and gradients and the recommendations to use blood results in diagnostic tests.

The rigorous nature of meta-analysis does not allow comparison of the tests using the entire data set. Previous studies have compared patients' test results even though all of the tests may not have been performed on all of the study patients, which violates a principle of experimental design of studies of investigative diagnostic tests.[8] In attempting to eliminate this source of bias, our meta-analysis tried to exclude patients from analysis if results from all of the tests were not available for every analyzed patient. This strategy decreased the data set from 1,448 to 299 patients. We compromised, therefore, by excluding patients from the primary analysis who did not have all of the five PF test results that represented the largest common data set: P-R, LDH-PF, LDH-R, C-PF, and C-R. Analyses of P-PF only included patients who had this test result and the previous five tests performed; analyses of A-G included only patients who had all seven test results. The smaller data sets used for P-PF and A-G indicate that comparisons of test discriminative properties that include P-PF or A-G are less precise than the other comparisons that use the larger data set. This observation explains the wider CIs for the odds ratios noted with combinations that include P-PF or A-G.

Our critical analysis of the primary studies detected considerable limitations in study design that potentially introduce bias and weaken estimates of the tests' diagnostic accuracies. These limitations are common in studies that report the performance of diagnostic studies.[8] Of fundamental importance, such studies require an explicit and independent reference standard that unequivocably establishes the presence or absence of the disease state under examination, which is the presence of a transudative or exudative effusion in this instance. Additionally, the investigators who determine the test results need to be blinded to the reference standard and, more importantly, the clinicians who determine the reference standard need to be blinded to the results of the tests under evaluation. In most of the primary studies included in the meta-analysis, rigorously explicit and objective reference standards and study blinding were absent or incompletely developed.

The other study limitations observed in the meta-analysis suggest that our understanding of the discriminative properties of PF tests can be enhanced by better study design. Reference standards for defining the presence of underlying diseases and the likely transudative or exudative nature of effusions should be explicit, described *a priori,* and based on objective test results rather than poorly reproducible

clinical impressions. The reference standard needs to be independent; the results of the PF tests under study, therefore, should not be considered by the clinicians who initiate the patients' initial diagnostic evaluation and determine the classification of the pleural effusions. This strategy requires all patients to be evaluated by a minimum battery of diagnostic studies to establish the nature of their effusions; this battery cannot include the PF tests under evaluation. Several months of follow-up for patients with transudative effusions would be required to limit verification bias: patients misclassified as having transudative effusions who present later with intrathoracic malignancies. Additionally, complete descriptions of patients' clinical features, the methods of cohort assembly, and a full spectrum of disease conditions associated with exudative and transudative effusions would enhance the generalizability of the evaluated tests. Finally, all study patients should be evaluated with each of the PF tests under consideration to allow an accurate comparison of the tests' diagnostic accuracies using formal statistical techniques, such as ROC analysis or calculated odds ratios. Recently, a multivariate technique of ROC analysis has been described to allow investigators to compare the diagnostic accuracies of test combinations.[29]

Considering the expense, practical difficulties, and entrenchment of PF tests in clinical practice, it is unlikely that a rigorously designed study of PF tests with a sufficient sample size and blinding will be performed. The wide CIs of odds ratios and similar diagnostic accuracies of the tests reported in the meta-analysis indicate that several thousand patients would be required to provide sufficient power to identify a clearly superior diagnostic test. This meta-analysis may serve, therefore, as a reasonable estimate of the diagnostic accuracies of the tests based on the available primary studies.

In considering the results of the meta-analysis, clinicians can select any one of several diagnostic strategies for identifying exudates. Light's criteria has excellent discriminative properties, but the cut-off value for LDH-PF of >0.45 of the upper limits of the laboratory's normal LDH should be used (modified Light's criteria). Furthermore, LDH-PF can be removed from Light's criteria (abbreviated Light's criteria) because it is correlated with LDH-R and does not significantly enhance the combination's discriminative properties. Alternatively, clinicians can elect to avoid blood tests without sacrificing diagnostic accuracy by utilizing test combinations that include only PF assays. The combinations of LDH-PF/C-PF and P-PF/LDH-PF/C-PF have similar diagnostic accuracies compared with Light's criteria. Future investigations with larger data sets may establish the diagnostic superiority of combinations that include A-G, as suggested by the high odds ratio in the small data sets that included this test.

Appendix

The following principal investigators participated in this study by collating and supplying their original investigative data: Lesley Burgess, University of Stellenbosch; Edgardo Cruz, de la Universidad de Santiago; Antonio Pose, Hospital de Conxo, Santiago Romero, Universidad de Alicante; Bernard J. Roth, Madigan Army Medical Center; V. Gil Suay, Hospital Universitario La Fe; and Luis Valdés, Hospital de Conxo.

References

1 Paddock FK. The relationship between the specific gravity and the protein content in human serous effusions. Am J Med Sci 1941; 201:569-74
2 Leuallen EC, Carr DT. Pleural effusion, a statistical study of 436 patients. N Engl J Med 1955; 252:79-83
3 Light RW, MacGregor I, Luchsinger PC, et al. Pleural effusion: the diagnostic separation of transudates and exudates. Ann Intern Med 1972; 77:507-13
4 Light RW. Pleural disease. 3rd ed. Baltimore: Williams & Wilkins, 1995
5 Antony VB, Holm KA. Testing the waters: differentiating transudates from exudates. Chest 1995; 108:1191-92
6 Hulley SB, Cummings SR. Designing clinical research. 1st ed. Baltimore: Williams & Wilkins, 1988
7 Jekel JF, Elmore JG, Katz DL. Epidemiology, biostatistics, and preventive medicine. 1st ed. Philadelphia: WB Saunders, 1996
8 Irwig L, Tosteson ANA, Gatsonis C, et al. Guidelines for meta-analysis evaluating diagnostic tests. Ann Intern Med 1994; 120:667-76
9 Owens DK, Holodniy M, Garber AM, et al. Polymerase chain reaction for the diagnosis of HIV infection in adults: a meta-analysis with recommendations for clinical practice and study design. Ann Intern Med 1996; 9:803-15
10 DeLeo J. Software for developing decision strategies that account for uncertainty. Washington, DC: IEEE Computer Society Press, 1993; 318-25
11 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29-36
12 McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med Decis Making 1984; 4:137-56
13 Gardner MJ, Gardner SB, Winter PD. Confidence interval analysis (CIA) microcomputer program manual. BMJ 1989; 1-30
14 Rosner B. Fundamentals of biostatistics. 4th ed. Belmont, Calif: Duxbury Press, 1995; 141-85
15 JMP manual. Cary, NC: SAS Institute Inc, 1994
16 Burgess LJ, Maritz FJ, Taljaard JJF. Comparative analysis of the biochemical parameters used to distinguish between pleural transudates and exudates. Chest 1995; 107:1604-09
17 Chakko SC, Caldwell SH, Sforza PP. Treatment of congestive heart failure: its effect on pleural fluid chemistry. Chest 1989; 95:798-802
18 Costa M, Quiroga T, Cruz E. Measurement of pleural fluid cholesterol and lactate dehydrogenase: a simple and accurate set of indicators for separating exudates from transudates. Chest 1995; 108:1260-63
19 Hamm H, Brohan U, Bohmer R, et al. Cholesterol in pleural effusions: a diagnostic aid. Chest 1987; 92:296-302
20 Meisel S, Shamiss A, Thaler M, et al. Pleural fluid to serum

bilirubin concentration ratio for the separation of transudates and exudates. Chest 1990; 98:141-44
21 Peterman TA, Speicher CE. Evaluating pleural effusions: a two-stage laboratory approach. JAMA 1984; 252:1051-53
22 Romero S, Candela A, Martín C, et al. Evaluation of different criteria for the separation of pleural transudates from exudates. Chest 1993; 104:399-404
23 Roth BJ, O'Meara TF, Cragun WH. The serum-effusion albumin gradient in the evaluation of pleural effusions. Chest 1990; 98:546-49
24 Scheurich JW, Keuer SP, Graham DY. Pleural effusion: comparison of clinical judgment and Light's criteria in determining cause. South J Med 1989; 82:1487-91
25 Suay VG, Moragón EM, Viedma EC, et al. Pleural cholesterol in differentiating transudates and exudates: a prospective study of 232 cases. Respiration 1995; 62:57-63
26 Valdés L, Pose A, Suàrez J, et al. Cholesterol: a useful parameter for distinguishing between pleural exudates and transudates. Chest 1991; 99:1097-1102
27 Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993; 39:561-77
28 Riegelman RK, Hirsch RP. Studying a study and testing a test: how to read the medical literature. 2nd ed. Boston: Little Brown, 1989
29 Shultz EK. Multivariate receiver-operating characteristic curve analysis: prostate cancer screening as an example. Clin Chem 1995; 41:1248-55