Anatomy of a meta-analysis: A critical review of "Exercise echocardiography or exercise SPECT imaging? A meta-analysis of diagnostic test performance"

Steven M. Kymes, MHA,[a] Denise E. Bruns, MBA,[c] Leslee J. Shaw, PhD,[d] Kathleen N. Gillespie, PhD,[b] and James W. Fletcher, MD[a]

Background. **Accurate diagnosis of coronary heart disease has the potential to contribute substantially to cost-effective delivery of health services. Recent work by Fleischmann et al (JAMA 1998;280:913-20) represents an effort to summarize the accuracy of exercise echocardiography and exercise single photon emission computed tomography (SPECT).**

Methods and Results. **A critique of the previous work was constructed, obtaining the 44 articles used. These articles were reviewed and summarized with established techniques for meta-analysis. The studies summarized by Fleischmann et al were found to be significantly heterogeneous (echocardiography and SPECT, both $P < .001$). In the SPECT cohort, combination of different radioisotopes and reading techniques, and inclusion of reports using experimental techniques, were sources of heterogeneity. In the echocardiography cohort, experimental techniques and an individual series were identified. When the sample was stratified for sources of heterogeneity, it was found that there was no significant difference in diagnostic accuracy between the echocardiography and SPECT techniques used in current clinical practice. Meta-regression with summary receiver operating characteristic curve techniques, after adjustment of the model for multicolinearity and outliers, revealed that there were no significant differences between SPECT as used in current clinical practice and echocardiography.**

Conclusion. **The report by Fleischmann et al contains serious flaws that limit its validity and generalizability. (J Nucl Cardiol 2000;7:599-615.)**

Key Words: Meta-analysis • summary receiver operating characteristic curve • confidence intervals • diagnostic tests • coronary artery disease

Heart disease is the leading cause of death among Americans and of premature disability in the US labor force, accounting for 19% of the disability allowance paid annually by the Social Security Administration. It is estimated that each year more than 1 million persons have a new or recurrent myocardial infarction, with a 1-year mortality rate approaching 33%. The annual cost of treating coronary heart disease in the United States includes more than $53.1 billion in direct costs (facility, drugs, and professional fees) and $46.7 billion in indirect costs (lost productivity, morbidity, and mortality). There are more than 5 million cardiac imaging procedures performed annually in an effort to diagnose and manage coronary artery disease (CAD) accurately.[1] It follows that establishing the accuracy of diagnostic modalities is an essential step in improving clinical outcomes and cost-effectiveness of care for this population.

In the common clinical algorithm for diagnosis of CAD, nuclear imaging and exercise echocardiography are used to confirm the presence or extent of CAD after presentation by a patient with angina or an abnormal electrocardiogram screening test. However, the tests differ in the risk markers provided. The most common nuclear protocol provides an assessment of regional myocardial perfusion and, more recently, ventricular function. Echocardiography provides real-time images of global and regional ventricular function. These differ-

From Saint Louis University School of Medicine[a] and School of Public Health,[b] St Louis, Mo; Molecular Biosystems, Inc,[c] San Diego, Calif; and Emory University School of Medicine, Atlanta, Ga.[d]

Ms Bruns is currently with Mallinckrodt Corporation, San Diego, Calif.

Received for publication Nov 12, 1999; final revision accepted June 1, 2000.

Reprint requests: Steven M. Kymes, MHA, Department of Nuclear Medicine (115-JC), Veterans Administration Medical Center, 915 N Grand, St. Louis, MO 63106; *kymes@slu.edu.*

1071-3581/2000/$12.00 + 0 **43/1/109027**
doi:10.1067/mnc.2000.109027

ences and the value of the information provided are areas of significant controversy in clinical cardiology.

A recent report by Fleischmann et al[2] represents one effort to compare the diagnostic accuracy of these 2 modalities with the technique of meta-analysis. This statistical technique came into wide use during the 1980s for medicine as a method to pool the results of a large number of reports, thereby creating a more precise estimate of effect. In this way it is different from primary data analysis (acquiring original data) or secondary analysis (looking at the primary data in a new way). The foundation of meta-analysis is the systematic, unbiased evaluation of published reports.[3-5] These methods emphasize the need to assemble the relevant literature and assess the quality of reports independent of data analysis[6] and have been demonstrated in national and society guideline projects with the principles of evidence-based medicine.[7]

In their review, Fleischmann et al[2] selected 44 published reports (publication dates 1990-1997) to compare the diagnostic accuracy of exercise echocardiography and nuclear imaging (see Appendix 1 for a list of the reports selected). They reported significant differences in test specificity between echocardiography and nuclear imaging. If this finding were valid, it could be concluded that, as a result of higher echocardiographic specificity, a change in health care policy should be recommended and the allocation of scarce diagnostic resources should be limited to the use of echocardiographic laboratories.[8,9] However, further review of their techniques revealed methodologic variation that warrants further analysis.[10]

The purpose of this analysis is to provide a critical review of the Fleischmann et al report and to clarify the role of meta-analysis in characterizing noninvasive test performance. This will include a reconstruction of the original meta-analysis and additional evaluation of the data through widely accepted techniques.

METHODS

Review of Literature and Analysis of Report Quality

For this critique, copies of the reports summarized by Fleischmann et al[11-54] were obtained and reviewed by 2 of the investigators, 1 from the field of nuclear imaging (S.K.) and the other from the field of echocardiography (D.B.). The reviewers were blinded to each other's findings. Because the reviewers were reconstructing previous work, blinding to the study hypothesis or the prior findings was impractical; however, it is unlikely that this had material effect on the findings. The journal name, year of publication, the study size and design, and the presence or absence of potential bias were extracted and recorded. The investigators graded the studies by adapting the definitions of Kent and Larson[55] and Fletcher et al[56] concerning proper diagnostic technology study design (see Appendix 2). Disagreements in grade were resolved by consensus. The grades were summed and a histogram was constructed with SPSS 8.0.1 (SPSS Inc, 1998). Cutpoints were identified visually and by use of frequency analysis. From this, studies were grouped into quality categories of high, medium, or low based on their grade (see Appendix 1). Note that this grading was relevant only for the literature presented and is not necessarily representative of the quality of studies in the literature at that time.

Homogeneity and Point Estimate Analysis

Midgette et al[57] described a quantitative technique for summarizing meta-analytic data. The analyst must first determine whether the true-positive and false-positive ratios are positively correlated. If this is the case, a summary receiver operating characteristic (SROC) curve can be derived from the data. The dataset is also tested for homogeneity to determine whether important differences in study design or clinical characteristics make pooling of results inappropriate. If the test shows that the data are homogeneous, point estimates of diagnostic accuracy can be made with a fixed-effects model.[3,57,58] For this critique, the Spearman nonparametric test for correlation (SPSS 8.0.1) was used to assess correlation between the true-positive and false-negative rates among the studies in the dataset. The χ^2 statistic was used to test for heterogeneity in the predictive accuracy statistic (FastPro 1.8; Academic Press, 1992).[59] The predictive accuracy statistic (true-positive + true-negative, divided by the total study participants) was used because it eliminates the contradictory results that might be obtained when sensitivity and specificity are reviewed separately. A statistically significant χ^2 statistic ($P < .05$) is indicative of statistical heterogeneity in the dataset. If the dataset is found to be heterogeneous, the reasons for heterogeneity were identified and, where appropriate, evaluation of subsets stratified by the reason for heterogeneity was made.[3,5,60]

SROC and Meta-regression Analysis

If the reports were positively correlated, SROC curves were derived with the technique described by Moses et al.[61] The dataset was first transformed into a modified ROC space, with the log-odds ratio between sensitivity and specificity on the vertical axis and the sum of the logs on the horizontal. This transformation allows linear regression to be performed identifying the "best fit" for the SROC curve among the component studies. Fleischmann et al performed this regression with a "weighted least squares" technique, with weights being a function of variability in study results and size of the study. Moses et al[61] also suggested "ordinary least squares" and "maximum likelihood" techniques as methods for deriving the regression model. The SROC curve is derived by back-transforming the line into traditional ROC space. When the dataset is in its logistic state, other variables may be tested to assess whether they modify the relation between the true-positive rate and the false-positive rate.[62] Each variable was introduced individually into the model and possible interactive effects were evaluated. All

variables with a *P* value of ≤.05 are included in the full model. For this analysis, the variables tested were the same tested by Fleischmann et al: radionuclide, scoring system, publication year, mean age of study participants, proportion of participants with prior myocardial infarction (MI), and proportion of men among study participants. The test indication given for the study (ie, whether CAD was known or suspected by the investigators) was considered as a possible variable, acting as a proxy for the degree of pretest probability of CAD.

Regression models were tested for outlying observations to identify those studies whose clinical characteristics or design were significantly different from the overall sample. Although SROC analysis does not require homogeneous data,[57] it is important that all data be selected from a similar population. If this is not the case, a single data point may distort the path of the regression line, and result in a biased estimate of effect.[63] Because the weights applied in the original regression model were strongly influenced by the size of the study and may confound the analysis, unweighted regression was used in performing outlier analysis. In this way, a single large study whose clinical characteristics are substantially different from the balance of the sample would not escape scrutiny. If outliers were identified, the reason for clinical heterogeneity was assessed and the study was deleted from the analysis if it was determined that it would result in a biased regression estimate.[63] Given the small size of the sample in this report (n = 51), the potential for multicolinearity in the model to result in a biased estimate because of outliers or small variations in the sample was very strong. Therefore the models were tested for multicolinearity with model diagnostics (SPSS 8.0.1) and by testing the model's variables for correlation.[63] If weaknesses were identified, corrective action was identified and taken. SROC curves were then plotted for each modality, with thallium and sestamibi modeled separately (no SROC curve was modeled for dual isotope imaging because the sample had only 3 studies).

RESULTS

Review of the Literature

The studies included in the meta-analysis originated from a wide variety of academic institutions or academic affiliates (ie, Veterans Administration hospitals or hospital-based physician practices at academic institutions). Twenty reports came from non-US facilities, 23 from the United States, and 2 from studies of both US and European centers. The distribution of non-US studies between the 2 modalities was consistent, with 10 of the echocardiography studies (42%) (as did both of the international studies) and 12 of single photon emission computed tomography (SPECT) studies (44%) coming from non-US facilities. Six studies involved both echocardiography and SPECT subjects (4 were from non-US facilities).

In both the echocardiographic and nuclear series, Fleischmann et al selected multiple articles from the same author and institutions (Marwick et al,[11,47-49] Roger et al,[50,51] and Van Train et al[35-37]). It is highly probable in these cases that the same patient may appear more than once in the series. The conservative approach used in the development of recent clinical guidelines is to make the assumption that latter work includes data from prior patients and select only the most recent report.[59] These series represent a large portion of participants (16% of SPECT participants, 36% of echocardiography). Although this could be a significant source of bias, analysis of the series after eliminating the potentially duplicated studies does not show a significant difference in the sensitivity or specificity estimates ($P > .05$).

Researchers in 3 of the studies[23,40,42] investigated techniques not commonly used in clinical practice. Cohen et al[40] and Dagianti et al[42] used supine ergometry to exercise echocardiography study participants, whereas Heiba et al[23] used lateral positioning of subjects for SPECT imaging. These reported sensitivity and specificity estimates significantly different from the overall series (Cohen et al, 78.4 and 86.7; Dagianti et al, 76.0 and 94.2; Heiba et al, 93.3 and 50.0; all $P < .05$). Pooling of results from experimental techniques with those of commonly used techniques may result in a biased point summary estimate.[3,64]

Analysis of Report Quality

The diagnostic accuracy of reports after stratification by grade is shown in Table 1. There were no statistically significant differences in the sensitivity estimates from study quality in the SPECT cohort. Medium-quality SPECT studies had a significantly lower specificity estimate than the high- or low-quality studies. This series includes the report by Christian et al,[19] whose primary purpose was to investigate the extent of disease, and a retrospective analysis by Chae et al.[18] Exclusion of these reports results in a new pooled sensitivity estimate that is homogeneous with the other 2 groupings (85.2%; 95% confidence interval [CI], 81.9-88.4), but the new specificity estimate is significantly higher than the other 2 groupings. Stratified analysis of specific sources of quality differences (verification bias, selection bias, multiple readers, diagnostic, and test review bias) revealed no significant differences. The source of these remaining differences may be a result of differential distribution of quantitatively scored studies and the mixing of types of radionuclides in the same dataset.

Medium-quality echocardiography reports had a significantly higher sensitivity estimate than high- or low-quality echocardiography reports. This cohort included the report by Crouse et al[41] that was identified in SROC

Table 1. Analysis of grades of study

	No. of subjects (% of total)	Value	95% CI
SPECT reports (n = 3237)			
High-quality (7 reports)	932 (28.8%)		
Sensitivity		89.6%	87.7%-91.6%
Specificity		65.7%	59.5%-72.0%
Medium-quality (10 reports)	1394 (43.1%)		
Sensitivity		86.2%	84.4%-88.1%
Specificity		57.0%*	54.6%-59.3%
Low-quality (10 reports)	910 (28.1%)		
Sensitivity		87.2%	85.0%-89.4%
Specificity		72.8%	69.9%-75.7%
Echo reports (n = 2637)			
High-quality (8 reports)	782 (29.7%)		
Sensitivity		84.3%	81.8%-86.9%
Specificity		84.1%†	81.6%-86.7%
Medium-quality (8 reports)	1088 (41.2%)		
Sensitivity		90.4%‡	88.7%-92.2%
Specificity		75.9%	73.3%-78.4%
Low-quality (8 reports)	767 (29.1%)		
Sensitivity		79.6%	76.8%-82.5%
Specificity		71.9%	68.7%-75.1%

*Specificity significantly lower than high- or low-quality SPECT reports.
†Specificity significantly higher than medium- or low-quality echocardiography reports.
‡Sensitivity significantly higher than high- or low-quality echocardiography reports.

analysis as an outlier because of its extremely high sensitivity (97.2%) and large size (n = 228). It also represents use of echocardiography as a screening test rather than to confirm the presence of disease, which may make it a source of heterogeneity. High-quality echocardiography reports had a significantly higher specificity estimate than medium- and low-quality studies. Medium-quality reports also had a higher specificity estimate than low-quality reports, although this was not statistically significant. Analysis of the echocardiography reports revealed a substantial difference in specificity where multiple readers were used and where selection bias was present. Reports for which a single reader was used had a specificity of 70.4% (95% CI, 66.3%-74.5%). Reports that had more than one reader had a specificity of 85.1% (95% CI, 81.7%-88.5%). Reports in which selection bias was present had specificity of 75.0% (95% CI, 72.9%-77.0%), and reports without selection bias had a specificity of 83.4% (95% CI, 81.0%-85.8%). All high-quality reports had more than one reader, whereas only 3 of the medium-quality and 1 of the low-quality reports had more than 1 reader. Selection bias was evenly distributed within the cohorts with 6 of the low-quality reports, 5 of the medium-quality reports, and 5 of the high-quality reports having selection bias. Therefore it is apparent that the difference in specificity estimates between quality groups may be associated with the number of readers used in interpreting images.

Homogeneity Analysis

When reports are combined with a meta-analysis, a nonsignificant heterogeneity statistic is optimal ($P > .05$) to ensure that there are no underlying differences in study design or population that might bias the estimate of effect.[3] Analysis of the echocardiography and SPECT datasets showed significant heterogeneity for both datasets (echocardiography $\chi^2 = 94.89$; SPECT $\chi^2 = 85.08$; both $P < .001$). Subgroup analysis identified specific studies that contributed to this heterogeneity. Within the echocardiography series, 2 reports by Roger et al were identified as possible sources of clinical or statistical differences: one report[50] assessed CAD extent, and another[51] was a retrospective analysis. Reports by Cohen et al[40] and Dagianti et al[42] used techniques not routinely used in clinical practice. When these 4 reports were excluded, a nonsignificant χ^2 statistic was noted (23.03, $P = .19$, n = 19), indicating homogeneity.

In subgroup analysis of the SPECT series, the report by Heiba et al[23] that used an experimental technique was excluded. Fleischmann et al also included in their analysis 3 studies of dual-isotope imaging (thallium at rest,

Table 2. Point estimates

Modality	Point estimate			
	Sensitivity	95% CI	Specificity	95% CI
SPECT fixed-effects model				
All cases (n = 27)	87.5%	86.3%-88.6%	63.7%	60.4%-67.4%
Thallium scans only (n = 9)	87.2%	85.5%-88.9%	59.0%	54.3%-63.8%
Quantitative scoring (thallium only, n = 2)	91.3%	88.9%-93.7%	69.8%	62.2%-77.4%
Visual scoring (thallium only, n = 7)	85.0%	82.9%-87.2%	53.5%	47.6%-59.4%
Sestamibi scans only (n = 15)	87.8%	85.7%-90.0%	68.8%	63.1%-74.4%
Quantitative scoring (sestamibi only, n = 5)	87.0%	83.1%-88.0%	50.8%	38.1%-63.6%
Visual scoring (sestamibi only, n = 10)	88.2%	85.6%-90.8%	74.1%	68.0%-80.2%
Dual-isotope imaging (n = 3)	88.7%	82.9%-94.5%	80.4%	69.0%-91.9%
1993 and earlier (inclusive, n = 17)	88.2%	86.8%-89.7%	63.1%	59.0%-67.2%
After 1993 (n = 10)	85.2%	82.4%-88.0%	66.4%	59.4%-73.6%
Single-reader reports (n = 12)	87.6%	85.8%-89.5%	64.4%	59.3%-69.6%
Multiple-reader reports (n = 15)	87.3%	85.5%-89.1%	63.4%	58.6%-68.3%
Homogeneous thallium (n = 5)	80.4%	75.8%-85.1%	77.8%	68.7%-86.8%
Homogeneous sestamibi (n = 8)	89.6%	86.9%-92.3%	75.4%	68.8%-82.1%
Echo fixed-effects model				
All cases (n = 24)	85.4%	83.7%-87.0%	77.3%	74.5%-80.0%
1993 and earlier inclusive (n = 10)	87.2%	85.3%-89.1%	81.5%	79.3%-83.7%
After 1993 (n = 14)	83.8%	81.9%-85.7%	74.8%	72.6%-77.0%
Single-reader reports (n = 12)	85.8%	83.7%-88.0%	70.4%	66.3%-74.5%
Multiple-reader reports (n = 12)	84.8%	82.2%-87.4%	85.1%	81.7%-88.5%
Homogeneous echo (n = 19)	87.2%	85.4%-88.9%	83.1%	80.2%-85.9%
Random-effects model				
All cases SPECT (n = 27)	86.9%	83.3%-89.8%	69.0%	60.0%-77.0%
Thallium (visual scoring only, n = 7)	81.7%	72.3%-89.0%	68.0%	47.1%-84.8%
Thallium (quantitative scoring, n = 2)	91.1%	81.9%-96.6%	70.4%	31.5%-94.9%
Sestamibi (visual scoring only, n = 10)	88.4%	82.4%-93.0%	73.2%	66.4%-79.1%
Sestamibi (quantitative scoring, n = 5)	88.0%	78.7%-94.4%	51.3%	36.2%-66.2%
Dual-isotope imaging (n = 3)	88.2%	81.4%-93.3%	78.0%	56.9%-92.2%
All cases echo (n = 24)	85.0%	81.5%-87.9%	80.5%	73.5%-86.4%

sestamibi when exercised). The pooled sensitivity and specificity estimates of these 3 studies[16,17,27] were higher than the overall pooled estimate. Although this is a technique commonly used today, these reports represented small developmental studies of the technique and possibly reflected publication and selection bias. Quantitative scoring methods were omitted from the studies Fleischmann et al selected (n = 7: 2 with thallium,[26,37] 5 with sestamibi[12,25,33,35,36]) because they represented a technique not commonly used in clinical practice. The study by Christian et al,[19] whose primary purpose was to assess the disease extent, was also excluded. The series was then stratified by radioisotope (omitting those studies of dual isotopes). Visual analysis of confidence intervals of the predictive accuracy statistics identified the report by Chae et al[18] (a retrospective study) as a potential source of heterogeneity for the thallium series, as was the report by Marwick et al[13] for sestamibi (another report examining CAD extent, although the companion echocardiography report was not a source of heterogeneity). Exclusion of these resulted in homogeneous cohorts for thallium imaging (n = 5, $\chi^2 = 2.17$, $P = .705$) and sestamibi imaging (n = 8, $\chi^2 = 11.81$, $P = .11$). To facilitate comparison with the results of Fleischmann et al, the homogeneous cohorts were analyzed as subgroups (Table 2).

Point Estimate Analysis

The pooled estimate of sensitivity of SPECT in this series of reports was 87.5% (95% CI, 86.3%-88.6%). The

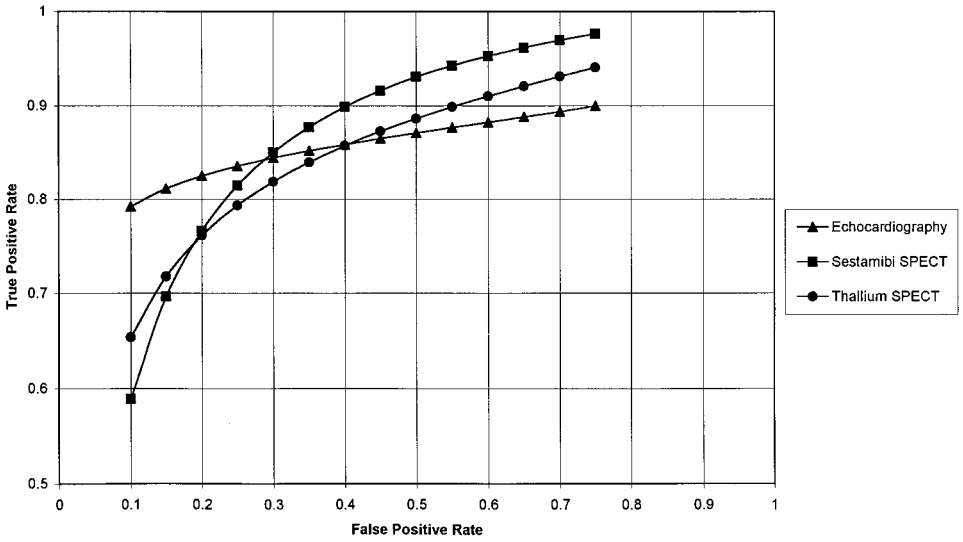

Figure 1. SROC analysis of diagnostic accuracy.

reports were stratified by the identified sources of heterogeneity: radioisotope (thallium, sestamibi, dual isotope), analysis method (visual or quantitative), year of publication (< or >1993), and multiple- versus single-reader studies. The point estimate of the homogeneous groups is also included for comparison. The pooled estimate of specificity was 63.7% (95% CI, 60.4%-67.4%). Stratifying sestamibi imaging by scoring technique results in a significantly different estimate of specificity (50.8% for sestamibi with quantitative scoring vs 74.1% for visual; $P < .05$). Stratification by scoring technique for the thallium series yields the opposite result (quantitative thallium specificity, 69.8%; visually scored studies, 53.5%; $P < .05$). Given the small size of the quantitatively scored thallium sample (n = 2) and the heterogeneity of the visual sample ($\chi^2 = 17.15$, $P = .009$), the reliability of the thallium comparison is questionable. Date of publication and the number of readers included in the study design were not significant covariates for SPECT. The SPECT homogeneous datasets had significantly improved specificity versus the original dataset ($P < .05$); it is likely that this is from the exclusion of the quantitatively scored series.

The pooled estimate of sensitivity for exercise echocardiography was 85.4% (95% CI, 83.7%-87.0%). These reports were stratified by year of publication, number of readers in the study design, and homogeneity. Stratification by these variables did not result in a significant modification of the estimate of sensitivity. The estimate of specificity was 77.3% (95% CI, 74.5%-80.0%). Publication year was found to be statistically significant in modifying the estimate of specificity (81.5% for those 1993 or earlier vs 74.8% for later reports; $P < .05$). Having more than one reader in the study significantly increased the specificity of the study (70.4% for single-reader reports and 85.1% for those with $P < .05$). The specificity of the homogeneous series was also significantly different than the overall sample (83.1% vs 77.3%, $P < .05$).

SROC and Meta-regression Analysis

Spearman analysis demonstrated a positive correlation within the echocardiography and SPECT cohorts. Meta-regression modeling of the combined series (combining the echocardiography and SPECT studies) identified the same confounding variables as were identified by Fleischmann et al: mean age of the study population, year of publication, and test indication (the latter within the SPECT series only). However, when thallium and sestamibi were modeled separately (without echocardiography), test indication was not found to be a significant covariate. The mean age of study participants and the year of publication were found to be potential sources of multicolinearity within the model. However, this was likely the result of the low variation in these variables rather than correlation between these variables and others in the model. This was confirmed by the correlation studies, which found no correlation between these 2 variables and others in the model. However, strong correlation was identified between the SPECT setting variable and the variable identifying echocardiography studies ($P < .001$). Given that 74% of the SPECT studies were coded as having “CAD known or suspected” in the setting variable, this is not surprising.

Omitting test indication from the combined model resulted in a substantial change in the coefficient comparing echocardiography accuracy to SPECT (35%) and for mean age (33%). This demonstrates that the prior

Table 3. Comparison of SROC models to Fleischmann et al[2] SROC results

Variable	Original model*		Model without test indication	
	Coefficient	95% CI	Coefficient	95% CI
Echo vs SPECT	1.20	0.71 to 1.68	0.78	0.36 to 1.19
Mean age	–0.12	–0.18 to –0.07	–0.16	–0.22 to –0.10
Publication year	–0.20	–0.30 to –0.11	–0.19	–0.29 to –0.09
SPECT mixed setting (test indication)	0.76	0.23 to 1.27	NA	NA
Adjusted R^2	0.64		0.58	

All coefficients are significant ($P < .05$).
*Fleischmann KE, Hunink MG, Kuntz K, Douglas P. Exercise echocardiography or exercise SPECT imaging? A meta-analysis of diagnostic test performance. JAMA 1998;280:913-20.

Table 4. Meta-regression by radioisotope with adjustment for outliers

Model	Adjusted R^2	Variable					
		Mean age		Publication year		Comparison of echocardiography to SPECT†	
		Coefficient	95% CI	Coefficient	95% CI	Coefficient	95% CI
Echo vs sestamibi							
All studies	.521	–.13*	–.19 to –.06	–.24*	–.57 to –.11	.57*	.002 to 1.15
Without Crouse et al[41]	.527	–.14*	–.21 to –.07	–.19*	–.04 to –.34	.49	–.08 to 1.06
Without Sylven et al[33] and Crouse et al[41]	.560	–.13*	–.20 to –.07	–.20*	–.06 to –.34	.35	–.21 to 1.06
Echo vs thallium							
All studies	.797	–.17*	–.10 to –.23	–.26*	–.14 to –.37	1.11*	.67 to 1.54
Without Crouse et al[41]	.821	–.18*	–.12 to –.24	–.20*	–.09 to –.32	.89*	.45 to 1.34
Without Nguyen et al[28] and Crouse et al[41]	.824	–.18*	–.12 to –.24	–.20*	–.09 to –.31	.90*	.45 to 1.34

*$P < .05$.
†A positive coefficient indicates superior diagnostic accuracy of echocardiography when compared with SPECT.

conclusion that the accuracy of SPECT is sensitive to test indication may be a result of statistical artifact and represents a weakness in the SROC model developed for the Fleischmann et al report.

The echocardiography report by Crouse et al[41] was identified as an outlier, as was the SPECT report by Nguyen et al.[28] The report by Nguyen et al[28] (with thallium) had moderate sensitivity (76%) but an extremely high specificity (100%), with an extremely small number of negative results (n = 5). The report by Sylven et al[33] was found to be an outlying observation for the sestamibi series. This report had extremely low diagnostic accuracy (sensitivity of 61% and specificity of 50%) as well as a postivity criterion different from any other SPECT report (persistent or transient perfusion defect).

The results of the meta-regression are shown in Tables 3 and 4. Table 3 shows the results originally reported by Fleischmann et al as well as the impact of the removal of the test indication from the model. Table 4 shows the results of the meta-regression analysis comparing echocardiography to thallium and sestamibi SPECT (separately), excluding the variable for test indication. In this analysis, the impact of the exclusion of the identified outliers was also assessed. Before exclusion of outliers, the variables for mean age of study participants, year of publication, and comparison of echocardiography to SPECT are significant. When the reports by Crouse et al and Nguyen et al are excluded from the model comparing thallium SPECT with echocardiography, all coefficients con-

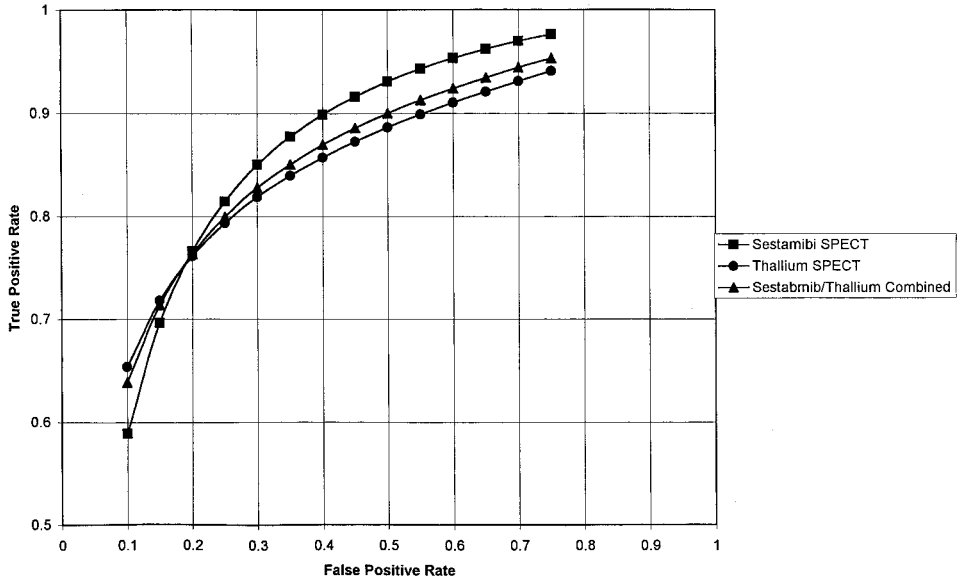

Figure 2. Effect of combining sestamibi and thallium SPECT studies.

tinue to be significant, indicating that echocardiography's advantage over thallium SPECT is not sensitive to studies selected for the meta-analysis. When the reports by Crouse et al and Sylven et al are excluded from the model comparing sestamibi SPECT to echocardiography, the coefficient comparing the 2 modalities is no longer significant, indicating that the reported advantage of echocardiography over SPECT is highly dependent on the studies selected for combination. Such a finding limits the generalizability of results.

Figure 1 shows the results of SROC analysis comparing echocardiography to sestamibi SPECT and thallium SPECT. In constructing these curves, the studies identified as outliers were omitted. Respecting the cautions of previous authors, these curves are not adjusted for covariates.[61] The curves as drawn show that at higher levels of specificity echocardiography has superior diagnostic accuracy when compared with both thallium and sestamibi SPECT. As specificity is reduced, the curves merge with sestamibi SPECT having higher sensitivity at specificity lower than 70% and thallium at specificity lower than 60%. This suggests that where higher levels of specificity are valued (ie, settings where screening for CAD is being performed), echocardiography may be the preferred diagnostic modality. In settings where higher sensitivity is valued (ie, where diagnosis of CAD is being confirmed) SPECT may be preferred. However, as noted later, without an understanding of the relative cost and morbidity of misdiagnosis this is difficult to evaluate based on the results of point estimate or SROC analysis alone.

DISCUSSION

This report presents a detailed meta-analysis of the diagnostic accuracy of exercise echocardiography and SPECT imaging examining a series reported by Fleischmann et al.[2] The findings of this analysis show significant variation from those previously reported. Contrary to those findings, exercise SPECT reports relevant to current standard of practice (ie, sestamibi SPECT) were found to have similar diagnostic accuracy when compared with echocardiography.

The purpose of a meta-analysis is to create a precise and definitive answer when the results of individual studies disagree or are inconclusive.[5] When possible, the data for the meta-analysis are taken from randomized controlled trials (RCT). Because RCTs are considered to provide unbiased estimates of effect, it follows that a meta-analysis consisting of RCTs would provide an unbiased estimate of effect. However, this report summarizes the results of a series of observational studies. Such studies may yield estimates of effect that substantially differ from the true effect within the population of interest because of uncontrolled confounding effects, bias in study design, or residual confounding (ie, confounding that remains after statistical adjustment).[5] Some have gone so far as to suggest that reporting summary estimates of a meta-analysis based on observational studies is inappropriate.[6,65] However, most authors support the position of recognizing and interpreting important clinical differences and heterogeneity between studies while cautioning against extensive statistical adjustment.[3,5,60]

One example of the problem posed by statistical adjustment is the use of the random effects model. A random effects model considers unaccounted variability in the comparison, thus yielding wider confidence estimates in the results (a more conservative estimate).[58,59] By comparison, a fixed-effects model (used in this and the previous report[2]) begins with the assumption that all of the outcome results for the general population may be explained by the population selected without confounding.[3] This is not the case in analyzing most clinical questions, particularly where clinical covariates are not considered. The use of the random-effects model, however, only resolves statistical heterogeneity, not clinical heterogeneity. In the report by Fleischmann et al, the use of the fixed-effects model provides a misleading result (ie, a point estimate summary showing echocardiography specificity superior to SPECT). The use of the random-effects model provides a result consistent with the findings presented here (Table 2).[10] However, such an estimate remains misleading because the clinical heterogeneity unaccounted for in pooling different isotopes and scoring systems results in an estimate that does not represent results obtained by any of the technologies being summarized.

Sources of Clinical Variation

Heterogeneity has been defined as a factor that causes a change in the effect size from one patient series to another so that the resulting meta-estimate (when the series are combined) is meaningless.[6] The results of the homogeneity analysis in this series revealed important differences that influence diagnostic accuracy. Most meta-analysts would discourage synthesizing the results when such variability exists. Thompson[60] has distinguished 2 types of heterogeneity: statistical and clinical. Clinical heterogeneity is defined as those clinically important differences that exist between the trials that constitute the meta-analysis. Sources of clinical heterogeneity would include the process of patient selection, baseline disease severity, differences in study design, or clinical procedures. Statistical heterogeneity exists when quantitative results in a meta-analysis are incompatible. This is often the result of clinical heterogeneity or methodologic differences in study design (as in the case of combining an RCT with a retrospective study).

The significant modifying factors Fleischmann et al identified were average age of participants in the component study, publication year of the study, and the clinical indication for the test (ie, known or suspected CAD, used by Fleischmann et al as a proxy for pretest probability of disease). However, this analysis found that other sources of clinical heterogeneity existed (ie, SPECT radionuclide and scoring system) that result in a biased report of diagnostic accuracy.

Point estimate analysis clearly shows that the radionuclide used in imaging is a source of clinical heterogeneity, and thus the results of sestamibi, thallium, and dual-isotope imaging should be separately reported. However, a more important question is whether it is appropriate to pool the results of thallium, sestamibi, and dual-isotope imaging regardless of statistical results. The purpose of performing a meta-analysis is to provide clinicians and policy makers with information that is useful in evaluating the effectiveness of the technology. A pooled estimate of sestamibi, thallium, and dual-isotope imaging does not provide useful information for these decision makers (one might just as well pool the results of echocardiography and SPECT). Decision makers are concerned with the performance of a technology as it is used. As such, in this analysis the results of sestamibi, thallium, and dual-isotope imaging should be separately reported, regardless of the statistical significance of the difference in estimate.

Figure 2 demonstrates this with SROC analysis. The curve representing sestamibi is clearly the dominant one in this set, with the curve representing thallium having considerably lower diagnostic accuracy (except at the highest levels of specificity). A third curve lies between the two, representing the combined diagnostic accuracy of the modalities. This curve clearly represents a weighted average of the diagnostic accuracy of sestamibi and thallium, yet its results are misleading. No point on this curve (except the intersection point at the far left) represents a level of diagnostic accuracy obtained by either technology. Every point on this curve (in the relevant range) is higher than that obtained by thallium imaging, and lower than that obtained by sestamibi. Thus the information provided by the curve is not useful to clinicians and policy makers.

It is not clear whether the differences seen in the diagnostic accuracy of quantitatively versus visually scored SPECT are the result of a change in the disease classification threshold or technology. If the former is the case, it is appropriate to model quantitative SPECT on the same SROC curve as visual SPECT. If it is the latter, a separate SROC curve would be needed (as seen in comparing sestamibi and thallium SPECT). It can be argued that clinicians have the option to choose quantitative versus visual scoring (as they have the option to choose thallium versus sestamibi) and as such a second curve would be justified. However, in SROC and outlier analysis quantitatively scored studies were not identified as being significantly different than visual studies. This would indicate that quantitative studies fall on the same curve, with the difference being that quantitative scoring techniques sacrifice specificity to gain sensitivity. In any case, the most conservative approach to SROC modeling would be to pool quantitative and visual studies for comparison to echocardiography, as has been done here.

The importance of the difference in point estimates of specificity in the echocardiography series between studies with single readers and those with multiple readers is also not clear. The presence of multiple readers was a significant factor when modeling point estimates, but not in SROC analysis. Therefore the differences may represent a shift in the threshold used to classify patients as diseased, with specificity being sacrificed to gain sensitivity (as was done with quantitative SPECT studies). Why this would be seen in the echocardiography series and not SPECT is not clear, but it might be an indication of lower interreader reliability with echocardiography.

SROC Analysis and Cost-effectiveness

The weakness of a point estimate of accuracy is that it fails to recognize the interrelation between sensitivity and specificity. Every test incorrectly classifies some patients who are nondiseased as diseased (specificity) or diseased as nondiseased (sensitivity). Additional sensitivity or specificity is gained by altering the threshold point at which the diagnostician determines the patient to be diseased. Without a change in technology (eg, changing the radionuclide used or use of a contrast agent) as a given test is made more specific (it correctly classifies more nondiseased patients), it will become less sensitive (there will be more diseased patients missed). The opposite of course is true when making a test more sensitive.

When this disease threshold point is varied a number of times across a continuum of confidence, sensitivity and specificity can be calculated for each threshold. The resulting pairs can be plotted in a graph with sensitivity (or true-positive rate) on the y-axis and "1 – specificity" (or false-positive rate) on the x-axis, with the result being an ROC curve. When these are constructed for a number of readers, or in summarizing a meta-analysis, the result is an SROC curve. The greater the space under the curve, the greater the diagnostic accuracy of the modality.[66] If the 2 ROC curves do not intersect, the one with the greatest space underneath it will always have superior cost-effectiveness (if the change for both tests is the same). However, if the 2 curves intersect anywhere but their most extreme points (Figure 1), further analysis is necessary to establish the most cost-effective modality. In such a case, the comparative effectiveness of the diagnostic tests will be determined in part by the pretest probability of disease in the setting in which it is used. For example, in Figure 1, the marginally better sensitivity of echocardiography at higher levels of specificity might be an indication that echocardiography is a superior choice for screening for CAD (an environment with low pretest probability of disease). However, the higher sensitivity of sestamibi SPECT at lower levels of specificity might indicate that SPECT is preferred where there is a higher pretest probability of disease (ie, confirming a diagnosis of CAD).

Taking this analysis a step further, it must be determined whether all inaccurate results for the competing tests should be evenly weighed. Failure to diagnose CAD on echocardiography or SPECT may result in an MI, an event that consumes substantial resources in the health system and, more importantly, results in substantial risks of morbidity or death for the patient. Low specificity results in unnecessary testing or treatment (in this case, the patient undergoing an unnecessary catheterization). Although substantial, it is not as high as the cost of low sensitivity (on a per-case basis). This might be considered as a reason more sensitive tests are more highly valued at this stage in the clinical algorithm. This is not always the case. Routine screening tests such as electrocardiography are often read with a low sensitivity in favor of high specificity. In such cases the total cost of false-negatives (even with such low sensitivity) in the screened population is viewed as being less than the total cost of false-positives.

The economic cost of a test is a function of several factors: size of the population of interest, the prevalence of the disease in the setting the test is given, the expense associated with diagnosis or treatment of an inaccurate result, and the morbidity or mortality rate associated with the inaccurate result. The cost of treatment or follow-up testing of accurately classified patients is not a relevant cost when comparing diagnostic tests. These costs will be the same for both modalities and are therefore not considered to be marginal costs. The charge for the individual tests is also not a relevant cost. However, the marginal charge of the more expensive test (ie, the difference in the charges between the tests being performed) should be treated as a cost assigned to all patients being tested. Therefore what should be considered is not the charge for an individual test, but what charge is necessary to offset the advantage gained by the test with the lowest cost of misdiagnosis and morbidity. This is an important consideration given the wide variation in charges for tests nationwide or even in a single metropolitan area (Note that the above is relevant only when comparing 2 tests and does not apply when considering the social or economic justification of the test itself. In such a case, all marginal economic and social costs, including follow-up treatment and testing, should be considered.)

An example for a single episode of care can be developed from the data presented here. Assume diagnostic testing for a population of 6000 persons with a disease prevalence of 73% (the size and prevalence of disease in the data reported by Fleischmann et al). The sensitivity and specificity of the competing tests, echocardiography and sestamibi SPECT, are also taken

from the data (Table 2). The cost of follow-up of a false-positive is the cost of further diagnosis (coronary angiography); for a false-negative, the relevant cost is that of treating an ensuing acute MI. For the purpose of this analysis, these costs would be $10,880 for cardiac catheterization[1] and $92,600 for 5 years of care for a patient with an acute MI.[67]

The resulting cost of misdiagnosis for the modalities was $63,267,840 for echocardiography and $52,443,800 for SPECT (see Appendix 3 for the details of the analysis). These estimates represent the incremental costs incurred as a result of the inaccuracy of each modality. The maximum allowable difference in charge that can exist between the 2 modalities (and still allow the dominant modality a cost advantage) is calculated by taking the cost advantage of the dominant modality (SPECT in this case), and dividing it by the number of persons tested. The result for this dataset is $1804, indicating that SPECT would still have a cost advantage if the difference in charge for the test is up to $1804 greater than the charge for echocardiography.

Note that this finding is highly dependent on the original assumptions made. In a population with a lower prevalence of disease than that described by Fleischmann et al, there would be a larger proportion of nondiseased individuals, reducing the cost advantage of SPECT because of its relatively superior sensitivity. Similarly, if the relative costs of the false-positive and false-negative were different, a different result would have been obtained. For example, simple algebra shows us that in this case if the cost of a false-negative had been $4600, the incremental costs of the 2 modalities would be equal. Given that the literature reflects a great deal of variability in the cost of an acute MI, this is possible.[67]

There has been no attempt made to quantify the costs of the morbidity and mortality associated with the false-negative and false-positive finding. However, it is generally assumed that in the case of this disease the probability of substantial morbidity associated with a false-negative (death or disability from an MI) is greater than that of a false-positive (death or disability resulting from complications of coronary angiography). This being the case, the findings would not change if this were included in the analysis. However, if the differential cost of the modality caused the less sensitive modality to be considered dominant, morbidity or mortality would become a factor in the analysis.

This has been a necessarily simplistic analysis of the economic impact of these 2 modalities. Proper economic analysis would factor in patient preferences and take a societal point of view.[3] However, this approach has been taken to illustrate the danger of simplistically relying on diagnostic accuracy statistics in making policy decisions.

This analysis also illustrates a weakness in current economic analysis techniques in assessing cost-effectiveness of diagnostic technology. Most cost-effectiveness analyses of diagnostic technologies rely on a single point of the SROC curve to estimate cost-effectiveness.[8,68,69] These point estimates represent the essential contribution of the diagnostic technology to the clinical algorithm. However, they fail to consider that individual users of the diagnostic technology might operate at an infinite number of points on the ROC curve. The point at which the clinician (or technology) will operate depends on the nature of the patient being diagnosed, the characteristics of the reader interpreting the study, and the nature of the equipment and facility. When one technique is clearly dominant to the other, with every point on its ROC (or SROC) curve lying above that of the second modality's curve, this difference is inconsequential (assuming similar changes for the technology). However, when the 2 curves intersect, as seen in this analysis, the use of a single point to estimate cost-effectiveness for an entire population of patients and clinicians will result in a misleading result. Some investigators who have recognized this have suggested that it might be addressed by use of multiple points to estimate cost-effectiveness of a technology (in essence performing a sensitivity analysis).[64] Such a technique would be superior to the current practice of basing cost-effectiveness on a single, idealized point, but it is still limited by its lack of scope because it fails to take into consideration the frequency at which clinicians would operate at different points on the curve.

Limitations

This analysis has 3 major limitations. First, the method used in obtaining the literature for the meta-analysis was not critiqued. The proper methods to select studies for a systematic review of the literature are described by others.[58] Second, if an accurate estimate of diagnostic accuracy is to be developed, state of the art imaging should be considered, including gated SPECT,[70] dual-isotope imaging,[17] and intravenous contrast agents to improve echocardiography.[71] The inclusion of reports in 1997 or before limits this analysis. Third, others have criticized Fleischmann et al's work based on the high prevalence of disease among the studies reported.[72] Prevalence higher (or lower) than that seen in traditional settings would limit the applicability of results and would have significant impact on our cost-effectiveness analysis.

CONCLUSION

In this critique, the work of previous meta-analysts in summarizing the diagnostic accuracy of exercise SPECT and exercise echocardiography was recon-

structed and examined. Although Fleischmann et al maintained that exercise echocardiography was a superior diagnostic modality to SPECT, our findings, using the same dataset, found no significant difference in the diagnostic accuracy of the 2 methods when comparing those techniques in common clinical practice. The major reason for this difference in findings was a failure to stratify the SPECT dataset for sources of clinical heterogeneity. In constructing SROC curves, statistical weaknesses were identified that limited the generalizability of the statistical model. Finally, given recent changes in the technology available to clinicians using these diagnostic tests, the applicability of the findings to today's practice settings is questionable.

References

1. American Heart Association. Cardiovascular disease statistics. Available at: http://www.amhrt.org/catalog/Scientific_catpage70.html. Accessed 1999.
2. Fleischmann KE, Hunink MG, Kuntz K, Douglas P. Exercise echocardiography or exercise SPECT imaging? A meta-analysis of diagnostic test performance. JAMA 1998;280:913-20.
3. Petitti DB. Meta-analysis, decision analysis, and cost-effectiveness analysis; New York: Oxford Press; 1994.
4. Irwig L, Tosteson A, Gatsonis C, Lau J. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994;120:667-76.
5. Egger M, Schneider M, Smith GD. Meta-analysis: spurious precision? Meta-analysis of observational studies. BMJ 1998;316:140-4.
6. Bailar JC III. Passive smoking coronary heart disease, and meta-analysis [editorial]. N Engl J Med 1999;340:958-9.
7. Peterson ED, Shaw LJ, Califf RM. Clincal guideline: part II; risk stratification after myocardial infarction. Ann Intern Med 1997;126:561-82.
8. Kuntz KM, Fleischmann KE, Hunink MGM, Douglas PS. Cost-effectiveness of diagnostic strategies for patients with chest pain. Ann Intern Med 1999;130:709-18.
9. Hunink MGM, Kuntz KM, Fleischmann KE, Brady TJ. noninvasive imaging for diagnosis of coronary artery disease: focusing the development of new diagnostic technology. Ann Intern Med 1999;131:673-80.
10. Kymes SM, Shaw LJ, Fletcher JF. Exercise cardiography or exercise SPECT imaging? A meta-analysis of diagnostic test performance [letter]. JAMA 1999;282:1622.
11. Hecht HS, DeBord L, Shaw R. Supine bicycle stress echocardiography versus tomographic thallium-201 exercise imaging for the detection of coronary artery disease. J Am Soc Echocardiogr 1993;6:177-185.
12. Hoffmann R, Lethen H, Kleinhans E, Weiss M, Flachskampf FA, Hanrath P. Comparative evaluation of bicycle and dobutamine stress echocardiography with perfusion scintigraphy and bicycle electrocardiogram for identification of coronary artery disease. Am J Cardiol 1993;72:555-9.
13. Marwick TH, D'Hondt AM, Mairesse GH. Comparative ability of dobutamine and exercise stress in inducing myocardial ischemia in active patients. Br Heart J 1994;72:31-8.
14. Pozzoli MMA, Fioretti PM, Salustri A, Reijs AE, Roelandt JR. Exercise echocardiography and technetium-99m MIBI single-photon emission computed tomography in the detection of coronary artery disease. Am J Cardiol 1991;67:350-5.
15. Quinones MA, Verani MS, Haichin RM, Mahmarian JJ, Suarez J, Zoghbi WA. Exercise echocardiography versus 201TI single-photon emission computed tomography in evaluation of coronary artery disease: analysis of 292 patients. Circulation 1992;85:1026-31.
16. Salustri A, Pozzoli MM, Hermans W, et al. Relationship between exercise echocardiography and perfusion SPECT in patients with single-vessel coronary artery disease. Am Heart J 1992;124:75-83.
17. Berman DS, Kiat H, Friedman JD. Separate acquisition rest thallium-201/stress technetium-99m sestamibi dual-isotope myocardial perfusion single-photon emission computer tomography: a clinical validation study. J Am Coll Cardiol 1993;22:1455-64.
18. Chae SC, Heo J, Iskandrian AS, Wasserleben V, Cave V. Identification of extensive coronary artery disease in women by exercise single-photon emission computed tomography (SPECT) thallium imaging. J Am Coll Cardiol 1993;21:1305-11.
19. Christian TF, Miller TD, Bailey KR, Gibbons RJ. Noninvasive identification of severe coronary artery disease using exercise tomographic thallium-201 imaging. Am J Cardiol 1992;70:14-20.
20. Fleming RM, Kirkeeide RL, Taegtmeyer H. Comparison of technetium-99m teboroxime tomography with automated quantitative coronary arteriography and thallium-201 tomographic imaging. J Am Coll Cardiol 1991;17:1297-302.
21. Gupta NC, Esterbrooks DJ, Hilleman DE, Mohiuddin SM. Comparison of adenosine and exercise thallium-201 single-photon emission computed tomography (SPECT) myocardial perfusion imaging. J Am Coll Cardiol 1992;19:248-57.
22. Hambye AS, Vervaet A, Lieber S, Ranquin R. Diagnostic value and incremental contribution of bicycle exercise, first-pass radionuclide angiography, and 99mTc-labeled sestamibi single-photon emission computed tomography in the identification of coronary artery disease in patients without infarction. J Nucl Cardiol 1996;3:464-74.
23. Heiba SI, Hayat NJ, Salman HS. Technetium-99-MIBI myocardial SPECT: supine versus right lateral imaging and comparison with coronary arteriography. J Nucl Med 1997;38:1510-4.
24. Ho YL, Wu CC, Huang PJ. Dobutamine stress echocardiography compared with exercise thallium-201 single-photon emission computed tomography in detecting coronary artery disease-effect of exercise on level of accuracy. Cardiology 1997;88:379-85.
25. Kiat H, Van Train KF, Maddahi J. Development and prospective application of quantitative 2-day stress-rest Tc-99m methoxy isobutyl isonitrile SPECT for the diagnosis of coronary artery disease. Am Heart J 1990;120:1255-66.
26. Mahmarian JJ, Boyce TM, Goldberg RK, Cocanougher MK, Roberts R, Verani MS. Quantitative exercise thallium-201 single photon emission computed tomography for the enhanced diagnosis of ischemic heart disease. J Am Coll Cardiol 1990;15:318-29.
27. Minoves M, Garcia A, Magrina J, Pavia J, Herranz R, Setoain J. Evaluation of myocardial perfusion defects by means of "bulls eye" images. Clin Cardiol 1993;16:16-22.
28. Nguyen T, Heo J, Ogilby JD, Iskandrian AS. Single photon emission computed tomography with thallium-201 during adenosine-induced coronary hyperemia: correlation with coronary arteriography, exercise thallium imaging and two-dimensional echocardiography. J Am Coll Cardiol 1990;16:1375-83.
29. Oguzhan A, Kisacik HL, Ozdemir K. Comparison of exercise stress testing with dobutamine stress echocardiography and exercise technetium-99m isonitrile single photon emission computerized tomography for diagnosis of coronary artery disease. Jpn Heart J 1997;38:333-44.
30. Palmas W, Friedman JD, Diamond GA, Silber H, Kiat H, Berman DS. Incremental value of simultaneous assessment of myocardial function and perfusion with technetium-99m sestamibi for prediction of extent of coronary artery disease. J Am Coll Cardiol 1995;25:1024-31.
31. Rubello D, Zanco P, Candelpergher G. Usefulness of 99mTc-MIBI stress myocardial SPECT bull's-eye quantification in coronary artery disease. Q J Nucl Med 1995;39:111-5.

32. Solot G, Hermans J, Merlo P. Correlation of 99Tcm-sestamibi SPECT with coronary angiography in general hospital practice. Nucl Med Comm 1993;14:23-9.
33. Sylven C, Hagerman I, Ylen M, Nyquist O, Nowak J. Variance ECG detection of coronary artery disease: a comparison with exercise stress test and myocardial scintigraphy. Clin Cardiol 1994;17:132-40.
34. Taillefer R, DePuey EG, Udelson JE. Comparative diagnostic accuracy of Tl-201 and Tc-99m sestamibi SPECT imaging (perfusion and ECG-gated SPECT) in detecting coronary artery disease in women. J Am Coll Cardiol 1997;29:69-77.
35. Van Train KF, Garcia EV, Maddahi J. Multicenter trial validation for quantitative analysis of same-day rest-stress technetium-99m-sestamibi myocardial tomograms. J Nucl Med 1994;35:609-18.
36. Van Train KF, Areeda J, Garcia EV. Quantitative same-day rest-stress technetium-99m-sestamibi SPECT: definition and validation of stress normal limits and criteria for abnormality. J Nucl Med 1993;34: 1494-02.
37. Van Train KF, Maddahi J, Berman DS. Quantitative analysis of tomographic stress thallium-201 myocardial scintigrams: a multicenter trial. J Nucl Med 1990;31:1168-79.
38. Beleslin BD, Ostojic M, Stepanovic J. Stress echocardiography in the detection of myocardial ischemia. Circulation 1994;90:1168-76.
39. Bjornstad K, Aakhus S, Hatle L. Comparison of digital dipyridamole stress echocardiography and upright bicycle stress echocardiography for identification of coronary artery stenosis. Cardiology 1995;86: 514-20.
40. Cohen JL, Ottenweller JE, George AK, Duvvuri S. Comparison of dobutamine and exercise echocardiography for detecting coronary artery disease. Am J Cardiol 1993;72:1226-31.
41. Crouse LJ, Harbrecht JJ, Vacek JL, Rosamond TL, Kramer PH. Exercise echocardiography as a screening test for coronary artery disease and correlation with coronary arteriography. Am J Cardiol 1991;67:1213-8.
42. Dagianti A, Penco M, Agati L. Stress echocardiography: comparison of exercise, dipyridamole and dobutamine in detecting and predicting the extent of coronary artery disease. J Am Coll Cardiol 1995;26:18-25.
43. Galanti G, Sciagra R, Comeglio M. Diagnostic accuracy of peak exercise echocardiography in coronary artery disease: comparison with thallium-201 myocardial scintigraphy. Am Heart J 1991;122:1609-16.
44. Jun T, Guizhen A, Xiaoming W, Jijun C, Jing X. Exercise echocardiography: feasibility and value for detection of coronary artery disease. Chin Med J 1996;109:381-4.
45. Luohtolati M, Saraste M, Hartiala J. Exercise echocardiography in the diagnosis of coronary artery disease. Ann Med 1996;28:73-7.
46. Marangelli V, Iliceto S, Piccinni G, De Martino G, Sorgente L, Rizzon P. Detection of coronary artery disease by digital stress echocardiography: comparison of exercise, transesophageal atrial pacing and dipyridamole echocardiography. J Am Coll Cardiol 1994;24:117-24.
47. Marwick TH, Nemec JJ, Pashkow FJ, Stewart WJ, Salcedo EE. Accuracy and limitations of exercise echocardiography in a routine clinical setting. J Am Coll Cardiol 1992;19:74-81.
48. Marwick TH, Anderson T, Williams MJ. Exercise echocardiography is an accurate and cost-efficient technique for detection of coronary artery disease in women. J Am Coll Cardiol 1995;26:335-41.
49. Marwick TH, Torelli J, Harjai K. Influence of left ventricular hypertrophy on detection of coronary artery disease using exercise echocardiography. J Am Coll Cardiol 1995;26:1180-6.
50. Roger VL, Pellikka PA, Oh JK, Bailey KR, Tajik AJ. Identification of multivessel coronary artery disease by exercise echocardiography. J Am Coll Cardiol 1994;24:109-14.
51. Roger VL, Pellikka PA, Bell MR. Sex and test verification bias: impact on the diagnostic value of exercise echocardiography. Circulation 1997;95:405-10.
52. Ryan T, Segar DS, Sawada SG. Detection of coronary artery disease with upright bicycle exercise echocardiography. J Am Soc Echocardiogr 1993;6:186-97.
53. Tawa CB, Baker WB, Kleiman NS. Comparison of adenosine echocardiography, with and without isometric handgrip, to exercise echocardiography in the detection of ischemia in patients with coronary artery disease. J Am Soc Echocardiogr 1996;9:33-43.
54. Williams MJ, Marwick TH, O'Gorman D, Foale RA. Comparison of exercise echocardiography with an exercise score to diagnose coronary artery disease in women. Am J Cardiol 1994;74:435-8.
55. Kent DL, Larson FB. Disease, level of impact, and quality of research methods: three dimensions of clinical efficacy assessment applied to magnetic resonance imaging. Invest Radiol 1992;27:245-54.
56. Fletcher J, Adams E, Flynn K, Papatheofanis F. Design of prospective clincal trials for assessment of PET. Clinical Positron Imaging 1998;1:201-11.
57. Midgette AS, Stukel TA, Littenberg B. A meta-analytic method for summarizing diagnostic test performance. Med Decis Making, 1993;13:253-7.
58. Lau J, Ioannidis JPA, Schmid CH. Quantitative Synthesis in Systematic Reviews. Annals of Internal Medicine November 1, 1997;127(9): 820-6.
59. Shaw LJ, Peterson ED, Kesler K, Hasselblad, Califf RM. A metaanalysis of predischarge risk stratification after acute myocardial infarction with stress electrocardiographic, myocardial perfusion, and ventricular function imaging. Am J Cardiol 1996;78:1327-37.
60. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. BMJ 1994;309:1351-5.
61. Moses L, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary roc curve: data-analystic approaches and some additional considerations. Stat Med 1993; 2:1293-1316.
62. de Vries S, Hunink M, Polak J. Summary receiver operating characteristic curves as a technique for meta-analysis of the diagnostic performance of duplex ultrasonography in peripheral arterial disease. Acad Radiol 1996;3:361-9.
63. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W, eds. Applied linear regression models. 3rd ed. Chicago: Irwin; 1996.
64. Littenberg B, Moses L. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making 1993;12:313-21.
65. Shapiro S. Meta-analysis/shmeta-analysis. Am J Epidemiol 1994;140: 771-8.
66. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143: 29-36.
67. Russell MW, Huse DM, Drowns S, Hamel EC, Hartz SC. Derect medical costs of coronary artery disease in the United States. Am J Cardiol 1998;81:1110-5.
68. Patterson R, Eisner R, Horowitz S. Comparison of cost-effectiveness and utility of exercise ecg, single proton emission computed tomography, positron emission tomography, and coronary angiography for diagnosis of coronary artery disease. Circulation 1995;91:54-65.
69. Schelbert H, Maddahi J. Clinical cardiac PET: quo vadis? J Nucl Cardiol 1994;1:576-9.
70. DePuey EG, Rozanski A. Using gated technetium-99-m-sestamibi SPECT to characterize fixed myocardial defects as infarct or artifact. J Nucl Med 1995;36:952-5.
71. Shaw LJ, Gillam L, Feinstein S. Use of an Intravenous contrast agent (optisontm) to enhance echocardiography: efficacy and cost implications. Am J Managed Care 1998;4:SP169-SP176.
72. Jacobs M. Review: exercise echo is a more specific and discriminitory test than exercise SPECT for coronary artery disease. ACP Journal Club 1999;130:45.

Appendix 1. Studies included in analysis

Reference No.	Institution	Year	Radionuclide	Single/multisite	Study type	Size	Selection bias	Multiple readers
SPECT Reports								
17	Cedars-Sinai	1993	Dual-isotope	Single	Prospective	63	Yes	Yes
18	PHI	1993	Thallium	Single	Retrospective	243	Yes	No
19	Mayo Clinic	1992	Thallium	Single	Prospective	688	Yes	Yes
20	U Texas	1991	Sestamibi	Single	Prospective	30	Yes	Yes
21	Creighton	1992	Thallium	Multi	Prospective	93	Yes	Yes
22	Middelheim	1996	Sestamibi	Single	Prospective	128	No	Yes
23	Kuwait U	1997	Sestamibi	Single	Prospective	34	No	Yes
24	Taiwan U	1997	Thallium	Single	Prospective	51	Yes	Yes
25	Cedars-Sinai	1990	Sestamibi	Multi	Prospective	53	Yes	No
26	Baylor	1990	Thallium	Single	Prospective	360	Yes	No
27	CETIR de Baelan	1993	Dual-isotope	Single	Prospective	54	Yes	Yes
28	PHI	1990	Thallium	Single	Prospective	30	Yes	Yes
29	Ankara U	1997	Sestamibi	Single	Prospective	70	Yes	Yes
30	Cedars-Sinai	1995	Sestamibi	Single	Prospective	70	Yes	No
31	Castelfrako	1995	Sestamibi	Single	Prospective	120	Yes	No
32	Jolomot	1993	Sestamibi	Single	Prospective	128	No	Yes
33	Karolinska Inst	1994	Sestamibi	Single	Prospective	67	Yes	No
34	Montreal	1997	Sestamibi	Single	Prospective	48	Yes	Yes
37	Cedars-Sinai	1990	Thallium	Multi	Prospective	371	No	No
36	Cedars-Sinai	1993	Sestamibi	Single	Prospective	38	Yes	No
35	Cedars-Sinai	1994	Sestamibi	Multi	Prospective	124	Yes	No
Echo reports								
38	Serbia U	1994	–	Single	Prospective	136	Yes	Yes
39	Trondheim U	1995	–	Single	Prospective	37	Yes	No
40	NJMS	1993	–	Single	Prospective	52	Yes	Yes
41	MAHI	1991	–	Single	Prospective	228	Yes	No
42	La Sapienza U	1995	–	Single	Prospective	60	Yes	Yes
43	U Florence	1991	–	Single	Prospective	53	Yes	No
44	Beijing	1996	–	Single	Prospective	47	No	No
45	Turku U	1996	–	Single	Prospective	118	No	Yes
46	U Barcelona	1994	–	Single	Prospective	80	Yes	No
47	Cleveland Clinic	1992	–	Single	Prospective	150	No	Yes
48	Cleveland Clinic	1995	–	Multi	Prospective	161	Yes	Yes
49	Cleveland Clinic	1995	–	Single	Prospective	147	Yes	Yes
50	Mayo Clinic	1994	–	Single	Prospective	150	Yes	No
51	Mayo Clinic	1997	–	Single	Retrospective	244	Yes	No
51	Mayo Clinic	1997	–	Single	Retrospective	96	Yes	No
52	U Indiana	1993	–	Single	Prospective	309	No	No
53	Baylor	1996	–	Single	Prospective	45	Yes	No
54	Cleveland Clinic	1994	–	Single	Prospective	70	Yes	Yes
Echo/SPECT comparison studies								
11*	SFHI	1993	Thallium	Single	Prospective	71	No	No
11†	SFHI	1993	–	Single	Prospective	71	No	Yes
12*	Cliniken der RWTH	1993	Sestamibi	Single	Prospective	66	No	No
12†	Cliniken der RWTH	1993	–	Single	Prospective	55	No	Yes
13*	U Louvain	1994	Sestamibi	Single	Prospective	86	Yes	Yes
13†	U Louvain	1994	–	Single	Prospective	86	Yes	Yes
14*	Erasmus	1991	Sestamibi	Single	Prospective	75	No	Yes
14†	Erasmus	1991	–	Single	Prospective	75	No	No
15*	Baylor	1991	Thallium	Single	Prospective	112	No	Yes
15†	Baylor	1991	–	Single	Prospective	112	No	No
16*	Erasmus	1992	Dual-isotope	Single	Prospective	44	Yes	No
16†	Erasmus	1992	–	Single	Prospective	44	Yes	Yes

TP, True-positive; *FN,* false-negative; *TN,* true-negative; *FP,* false-positive; *PHI,* Philadelphia Heart Institute; *CETIR,* Centro Tecnico de Isotopes Radiactivos; *U,* University; *MAHI,* Mid-America Heart Institute; *SFHI,* San Francisco Heart Institute; *RWTH,* Rheinisch-Westfalische Technische Hochschule; *NJMS,* New Jersey Medical School.
*SPECT; †echo.

Diagnostic bias	Verification bias	Test review bias	Miscellaneous	TP	FN	TN	FP	Sensitivity	Specificity
No	Yes	No		50	5	6	2	90.91%	75.00%
No	Yes	Yes		116	47	52	28	71.17%	65.00%
Yes	No	No		527	51	30	80	91.18%	27.27%
Yes	Yes	No		15	1	6	2	93.75%	75.00%
No	No	Yes		62	14	14	3	81.58%	82.35%
No	No	Yes		75	16	28	9	82.42%	75.68%
Yes	Yes	Yes		28	2	2	2	93.33%	50.00%
Yes	No	No		29	9	10	3	76.32%	76.92%
No	No	Yes	Quantitative	45	3	4	1	93.75%	80.00%
Yes	Yes	Yes	Quantitative	192	29	65	10	86.88%	86.67%
No	Yes	No		27	3	22	2	90.00%	91.67%
No	Yes	No		19	6	5	0	76.00%	100.00%
Yes	No	No		47	2	15	6	95.92%	71.43%
Yes	No	No		60	6	3	1	90.91%	75.00%
Yes	No	No		100	7	8	5	93.46%	61.54%
No	No	No		87	3	27	11	96.67%	71.05%
No	Yes	No	Quantitative	41	16	5	5	71.93%	50.00%
No	Yes	No		23	9	13	3	71.88%	81.25%
Yes	No	No	Quantitative	290	17	32	32	94.46%	50.00%
Yes	Yes	No	Quantitative	28	1	4	5	96.55%	44.44%
Yes	Yes	No	Quantitative	91	11	8	14	89.22%	36.36%
No	No	No		105	14	14	3	88.24%	82.35%
No	Yes	No		26	5	4	2	83.87%	66.67%
No	No	No		29	8	13	2	78.38%	86.67%
No	Yes	No		170	5	34	19	97.14%	64.15%
No	No	No		19	6	33	2	76.00%	94.29%
No	Yes	No		25	2	25	1	92.59%	96.15%
No	Yes	No		28	4	14	1	87.50%	93.33%
No	No	Yes		101	7	7	3	93.52%	70.00%
No	Yes	No		42	5	30	3	89.36%	90.91%
No	Yes	No		96	18	31	5	84.21%	86.11%
No	Yes	Yes		47	12	83	19	79.66%	81.37%
No	Yes	Yes		44	18	77	8	70.97%	90.59%
No	Yes	Yes		50	10	56	34	83.33%	62.22%
No	Yes	No		151	43	22	28	77.84%	44.00%
No	Yes	No		46	12	14	24	79.31%	36.84%
No	Yes	No		193	18	76	22	91.47%	77.55%
Yes	Yes	Yes		31	2	10	2	93.94%	83.33%
No	Yes	No		29	4	31	6	87.88%	83.78%
No	No	No		447	4	13	7	92.16%	65.00%
No	No	No		46	5	16	4	90.20%	80.00%
No	No	No	Quantitative	37	5	9	4	88.10%	69.23%
No	No	No		40	10	14	2	80.00%	87.50%
No	No	No		41	15	21	9	73.21%	70.00%
No	No	No		49	7	24	6	87.50%	80.00%
Yes	Yes	No		41	8	23	3	83.67%	88.46%
Yes	Yes	No		35	14	25	1	71.43%	96.15%
Yes	Yes	No		65	21	21	5	75.58%	80.77%
Yes	Yes	No		64	22	23	3	74.42%	88.46%
No	Yes	No		25	5	9	5	83.33%	64.29%
No	Yes	No		26	4	12	2	86.67%	85.71%

Appendix 2. Grading criteria

Grading criteria	Description	Possible grades
Single or multisite	Multisite studies are considered to have greater generalizability than single sites.	Multisite = 5; single site = 1
Study type	RCTs are considered to provide the best scientific evidence, prospective studies have more potential uncontrolled variables, and retrospective studies are considered inferior to both.	RCT = 5; prospective = 3; retrospective = 1
Study size	Larger study sizes are considered to be superior to smaller ones because the larger size increases statistical power and may also increase generalizability of results. The grading criteria for this is based on that suggested by Kent.[55]	>35 diseased and nondiseased subjects = 5; 35 diseased but <35 nondiseased = 3; <35 diseased and nondiseased = 1
Selection bias	Inclusion/exclusion criteria could result in a tested cohort that is different in important respects from those patients normally referred for testing at that facility (ie, inclusion criteria that exclude patients with prior MI).	Absent = 5; present = 1
Multiple readers	Test designs that result in more than 1 reader making the determination of diseased/nondiseased status on index test; it is considered to have better generalizability than one in which a single reader makes the determination.	Multiple = 5; single = 1
Diagnostic review bias	Present when results of the gold standard test affect the review of the test in question. For this report, diagnostic review bias is considered to be present unless the authors specifically stated that those interpreting the index test were blinded to catheter results or if the catheterization occurred after the index test in all cases.	Absent = 5; present = 1
Test review bias	Present when the results of the index test affect the results of the gold standard test. For this report this test review bias was considered to be present unless the authors specifically stated that the readers of the catheterization results were blinded to results of the index test.	Absent = 5; present = 1
Verification bias	Present when inclusion of the patient in the testing cohort depends on the patient undergoing the gold standard test (ie, the inclusion criterion for the study requires that the patient undergo coronary catheterization). When the gold standard is highly invasive, expensive, or relatively risky, those referred for the test are only those with a relatively high likelihood of disease.	Absent = 5; present = 1

Appendix 3. Calculation of economic costs

Test result	Have disease	No disease	Row total
Echocardiography results*			
Positive	3740	368	4992
Negative	640	1252	1008
Column total	4380	1620	6000
Sestamibi SPECT results†			
Positive	3863	420	4992
Negative	517	1200	1008
Column total	4380	1620	6000

*Sensitivity = 85.4%; specificity = 77.3%.
†Sensitivity = 88.2%; specificity = 74.1%.

Cost Calculations

Echo false-positives	368 × $10,880 =	$4,003,840
Echo false-negatives	640 × $92,600 =	$59,264,000
Total		$63,267,840
SPECT false-positives	420 × $10,880 =	$4,569,600
SPECT false-negatives	517 × $92,600 =	$47,874,200
Total		$52,443,800

Marginal Cost of Test Required to Offset SPECT Advantage

Cost of Echo inaccuracy	$63,267,840
Cost of SPECT inaccuracy	–$52,443,800
Total	$10,824,040
Total number of tests	6000
Marginal cost calculation	$10,824,040/6000 = $1804