

[¹⁸F]Fluorodeoxyglucose Uptake by Positron Emission Tomography for Diagnosis of Suspected Lung Cancer

Impact of Verification Bias

Michael S. Lauer, MD; Sudish C. Murthy, MD, PhD; Eugene H. Blackstone, MD; Ikenna C. Okereke, MD; Thomas W. Rice, MD

Background: Verification bias occurs when test findings influence the decision to perform a gold standard test. It adversely influences diagnostic test accuracy by inflating sensitivity and deflating specificity. We studied the impact of verification bias on the estimated accuracy of a test commonly used in suspected lung cancer.

Methods: We studied 534 consecutive patients referred for [¹⁸F]fluorodeoxyglucose uptake by positron emission tomography (PET). Primary outcomes were tissue diagnoses of cancer and of mediastinal lymph node metastases. A secondary outcome was 3-year mortality. We accounted for verification bias using 2 validated methods.

Results: The gold standard test, namely tissue acquisition, was performed in 419 patients (78%); mediastinal lymph node sampling occurred in 301 (56%). While the 410 patients with PET-diagnosed stage I cancer or higher

were more likely than patients with negative PET scan findings to undergo tissue diagnosis testing (92% vs 34%) ($P < .001$), there was no association between PET findings and performance of mediastinal sampling. Without accounting for verification bias, the sensitivity and specificity of PET for diagnosis of cancer were 0.95 (95% confidence interval [CI], 0.92-0.97) and 0.31 (95% CI, 0.21-0.42), respectively. After adjustment, sensitivity fell to 0.85 (95% CI, 0.81-0.89), while specificity increased to 0.51 (95% CI, 0.40-0.60). For diagnosis of mediastinal disease, verification bias had slight effects on test accuracy. There were 224 deaths, with a strong gradient between PET stage and death ($P < .001$).

Conclusion: The diagnostic accuracy of PET for assessment of suspected lung cancer is substantially affected by verification bias.

Arch Intern Med. 2007;167:161-165

VERIFICATION BIAS IS AN INCREASINGLY appreciated problem that occurs when investigators fail to account for incomplete measurement of a gold standard test among patients referred for a diagnostic test.^{1,2} For example, sensitivity and specificity of prostate-specific antigen screening for prostate cancer in men younger than 60 years are claimed to be 57% and 60%, respectively.³ However, after adjustment for verification bias, the sensitivity of prostate-specific antigen screening decreased to 18%, and specificity increased to 98%. Verification bias has also been shown to lead to substantial overestimation of sensitivity of exercise electrocardiography⁴ and noninvasive imaging⁵ for diagnosis of coronary artery disease.

Imaging of [¹⁸F]fluorodeoxyglucose uptake by positron emission tomography (PET) is a widely used tool for diagnosing and managing suspected lung can-

cer.⁶ Recent meta-analyses have documented high sensitivity and moderate specificity for lung cancer staging, particularly when combined with computed tomography (CT),^{7,8} but these studies did not adjust for verification bias or limit analyses to studies in which all patients undergoing scanning were required to undergo gold standard testing. Because results of this test may guide treatment algorithms that affect patient outcome, it is essential to examine the validity of these established accuracies. We suspect that sensitivity and specificity may be incorrectly estimated because of failure to account for verification bias.⁹

For a noninvasive test for suspected cancer such as PET, verification bias is possible if not all patients undergoing the test subsequently undergo a procedure to obtain tissue for diagnosis. We suspected that at our institution, this would be true for the diagnosis of cancer but not necessarily true for the diagnosis of metastases to

Author Affiliations:

Departments of Cardiovascular Medicine (Dr Lauer), Thoracic and Cardiovascular Surgery (Drs Murthy, Blackstone, Okereke, and Rice), and Quantitative Health Sciences (Drs Lauer and Blackstone), The Cleveland Clinic, and Department of Epidemiology and Biostatistics, Case Western Reserve University (Dr Lauer), Cleveland, Ohio.

mediastinal lymph nodes. This is because our practice is to perform mediastinoscopy in all patients referred for consideration of resection of suspected lung cancer regardless of the PET scan result. These differences provided us a unique opportunity to systematically investigate the occurrence and impact of verification bias in a large cohort of patients referred for PET to evaluate suspected lung cancer.

METHODS

PATIENTS AND CLINICAL DATA

We systematically abstracted the records of all 534 consecutive patients who were under the care of a Cleveland Clinic Foundation physician and were referred for a PET scan to evaluate suspected lung cancer between January 2000 and July 2002. No patient had tissue diagnosis at the time of the PET scan. Patients were included whether or not further diagnostic testing for cancer was performed. In addition to PET findings, patient age, sex, smoking status, and Eastern Cooperative Oncology Group (ECOG)¹⁰ performance status were abstracted, as were CT findings (staged by TNM classifications¹¹) and results of subsequent procedures to obtain tissue diagnosis (including mediastinoscopy). Institutional review board approval for this systematic medical chart review was obtained; the requirement for written informed consent was waived.

PET SCANNING

The PET scans were performed according to standard protocols.⁶ All patients received 10 Ci (37×10^{10} Bq) of [¹⁸F]fluorodeoxyglucose 45 minutes before imaging. Axial views were reconstructed into sagittal and coronal views from the top of the neck to the pelvic floor. Scans were obtained for 10 minutes in each bed position using a germanium Ge 68 pin source. All images were reviewed by a nuclear radiologist who was blinded to tissue diagnoses, clinical outcomes, and the hypothesis of this study. Scans were interpreted for abnormalities regarding the suspected primary tumor, hilar lymph nodes, mediastinal lymph nodes, and distant metastatic sites. The TNM classifications were applied to derive clinical stage of cancer based on PET interpretation.¹¹

OUTCOMES

The 2 primary outcomes were (1) tissue diagnosis of lung cancer at any site (as a gold standard) and (2) tissue diagnosis of metastases to mediastinal lymph nodes (as a similar gold standard). Because not all patients underwent an attempted tissue diagnosis, we noted whether a tissue or mediastinal sample was obtained. In a supplementary analysis, we analyzed all-cause mortality based on review of the Social Security Death Index.¹² Our research group and others have shown the high accuracy for this method of assessing mortality.^{12,13}

STATISTICAL ANALYSIS

In unadjusted analyses, sensitivity and specificity of PET were calculated according to standard definitions.¹⁴ The positive likelihood ratio, which relates posttest odds of disease to pretest odds, was calculated as sensitivity/(1 – specificity), and the negative likelihood ratio was calculated as (1 – sensitivity)/specificity.

We adjusted for verification bias, that is, failure to obtain tissue diagnosis in all subjects, using the simple formula detailed by Miller et al,¹⁴ Diamond,¹⁵ and Diamond et al¹⁶ (hereinafter, “the Diamond method”) and the more complex formula outlined by Miller et al¹⁴ and Begg and Greenes¹⁷ (hereinafter “the Begg and Greenes method”). The Diamond method calculates adjusted sensitivity ($P[\text{PET}^+|\text{CA}^+]$) as

$$(P[\text{CA}^+|\text{PET}^+, \text{ biopsy}] \times P[\text{PET}^+]) / (P[\text{CA}^+|\text{PET}^+, \text{ biopsy}] \times P[\text{PET}^+] + P[\text{CA}^+|\text{PET}^-, \text{ biopsy}] \times P[\text{PET}^-]),$$

where P means probability; |, given; CA, cancer; and biopsy, that the patient underwent a tissue diagnosis test. Analogously, the Diamond method calculates the adjusted specificity ($1 - P[\text{PET}^+|\text{CA}^-]$) as

$$1 - (P[\text{CA}^-|\text{PET}^+, \text{ biopsy}] \times P[\text{PET}^+]) / (P[\text{CA}^-|\text{PET}^+, \text{ biopsy}] \times P[\text{PET}^+] + P[\text{CA}^-|\text{PET}^-, \text{ biopsy}] \times P[\text{PET}^-]).$$

The Begg and Greenes method, as described by Miller et al,¹⁴ requires calculating an estimated probability of disease among subjects *not* having a tissue diagnosis. This probability ($P[\text{cancer}]$) is generated from a nonparsimonious logistic regression model relating biopsy findings in patients who *did* have tissue diagnosis to the prespecified variables of age, sex, smoking status, ECOG performance status, CT stage, and PET stage. The adjusted sensitivity is then calculated as

$$(N[\text{PET}^+ \text{ and } \text{CA}^+] + \sum_{\text{PET}^+, \text{ no biopsy}} P[\text{cancer}]) / (N[\text{PET}^+] + \sum_{\text{no biopsy}} P[\text{cancer}]),$$

where $N[\text{PET}^+ \text{ and } \text{CA}^+]$ is the number of patients with positive PET scan findings and a tissue diagnosis of cancer, and $N[\text{PET}^+]$ is the total number of patients with positive PET scan findings. Analogously, the adjusted specificity is calculated as

$$(N[\text{PET}^- \text{ and } \text{CA}^-] + \sum_{\text{PET}^-, \text{ no biopsy}} 1 - P[\text{cancer}]) / (N[\text{PET}^-] + \sum_{\text{no biopsy}} 1 - P[\text{cancer}]),$$

where $N[\text{PET}^- \text{ and } \text{CA}^-]$ is the number of patients with negative PET scan findings and a negative tissue biopsy, and $N[\text{PET}^-]$ is the total number of patients with negative PET scan findings.

The Diamond method has the advantage of relative simplicity and provides researchers the ability to calculate adjusted sensitivity and specificity without knowing about any patient characteristics other than PET findings, whether or not a biopsy was obtained, and biopsy findings if available. The Begg and Greenes method is arguably more intuitive in that it requires an estimate of the likely findings of a biopsy among those patients who did not undergo gold standard testing.

To obtain robust and relatively unbiased estimates of test accuracy before and after adjusting for verification bias, we used bootstrap bagging to create 1000 data sets.¹⁴ These data sets were used to generate logistic models needed for the Begg and Greenes method¹⁷ and to calculate test accuracy parameters along with 95% confidence intervals.

Survival curves were generated according to the Kaplan-Meier method. Differences in survival according to PET-estimated cancer stage were compared using the log-rank χ^2 test.

All analyses were performed using SAS statistical software, version 9.1 (SAS Institute Inc, Cary, NC). Macros in SAS for generating test accuracy measures and logistic regression models were written by 1 of us (E.H.B.) and are available on request.

RESULTS

PATIENT CHARACTERISTICS

There were 534 patients eligible for analysis. The median age was 68 years (interquartile range, 59-75

Table 1. Estimated Accuracy of PET Scanning for Detection of Cancer at Any Site Before and After Accounting for Verification Bias*

Method†	Sensitivity	Specificity	Positive Likelihood Ratio	Negative Likelihood Ratio
Unadjusted	0.95 (0.92-0.97)	0.31 (0.21-0.42)	1.37 (1.19-1.61)	0.17 (0.09-0.30)
Diamond	0.87 (0.82-0.91)	0.55 (0.43-0.64)	1.92 (1.47-2.47)	0.24 (0.15-0.38)
Begg and Greenes ¹⁷	0.85 (0.81-0.89)	0.51 (0.40-0.60)	1.72 (1.40-2.15)	0.30 (0.21-0.41)

Abbreviation: PET, positron emission tomography.

*Data are reported as point estimates (95% confidence intervals) and are based on 1000 bootstrap resamples.

†For a detailed description of the analytic methods, see the "Statistical Analysis" subsection in the "Methods" section.

Table 2. Estimated Accuracy of PET Scanning for Detection of Mediastinal Cancer Before and After Accounting for Verification Bias*

Method†	Sensitivity	Specificity	Positive Likelihood Ratio	Negative Likelihood Ratio
Unadjusted	0.50 (0.38-0.62)	0.90 (0.84-0.94)	4.75 (2.95-8.77)	0.56 (0.42-0.71)
Diamond	0.49 (0.39-0.59)	0.90 (0.86-0.94)	4.83 (2.99-8.87)	0.57 (0.45-0.70)
Begg and Greenes ¹⁷	0.44 (0.34-0.53)	0.88 (0.84-0.92)	3.56 (2.32-5.86)	0.64 (0.53-0.76)

Abbreviation: PET, positron emission tomography.

*Data are reported as point estimates (95% confidence intervals) and are based on 1000 bootstrap resamples.

†For a detailed description of the analytic methods, see the "Statistical Analysis" subsection in the "Methods" section.

years; range, 30-89 years). There were 310 men (58%), 135 current (25%) and 318 past smokers (60%). Only 83 patients (16%) had an ECOG status of 2 or higher.

PET INTERPRETATION

Among the 534 patients, PET interpretation resulted in the following clinical stage groupings: stage 0 (no cancer identified), 124 patients (23%); stage I, 189 patients (35%); stage II, 45 patients (8%); stage III, 110 patients (21%); and stage IV, 66 patients (13%). The PET results suggested mediastinal lymph node metastases in 117 patients (22%).

TISSUE PROCUREMENT

Procedures to obtain tissue for diagnosis (gold standard) were performed in 419 patients (78%). Diagnostic procedures included mediastinoscopy in 211 (40%), thoracotomy in 159 (30%), transbronchial biopsy in 87 (16%), percutaneous fine-needle aspiration in 67 (13%), thoracentesis in 13 (2%), and other techniques in 32 (6%).

PET FINDINGS AND PRACTICE

Among patients with assigned clinical stage grouping 0 by PET, only 42 (34%) of 124 underwent an attempt at tissue diagnosis. In contrast, among patients with assigned clinical stage grouping I, II, III, or IV by PET, 172 (91%), 39 (87%), 105 (95%), and 61 (92%), respectively, had a tissue diagnosis procedure ($P < .001$).

Patients who were considered candidates for surgical resection automatically had mediastinal lymph node sampling. In contrast to the strong association of obtaining tissue diagnosis and PET scan interpretation of disease, there was no evidence of association of PET scan results and performing mediastinal lymph node sampling ($P > .90$). Spe-

cifically, among the 417 patients without evidence of mediastinal lymph node metastases by PET, 181 (43%) had mediastinal lymph node sampling; among the 117 patients with evidence of mediastinal lymph node metastases by PET, 52 (44%) underwent mediastinal lymph node sampling. These data confirmed practice of our management philosophy to perform mediastinoscopy (gold standard testing) in all patients referred for possible lung cancer resection, regardless of PET scan findings.

PET ACCURACY VALUES

Without accounting for verification bias, PET had a sensitivity of 95% and a specificity of 31% for any cancer (**Table 1**). After accounting for verification bias by the Diamond and Begg and Greenes methods, sensitivity decreased to 85%, and specificity increased to 51%. All positive likelihood ratio values were well below 10, and negative likelihood ratios were above 0.10, especially after accounting for verification bias.

For diagnosis of mediastinal lymph node metastases, unadjusted sensitivity was only 50%, and specificity was 90% (**Table 2**). After adjustment for verification bias, these values changed little. Positive likelihood ratios were higher than for diagnosis of any cancer, but negative likelihood ratios were well above 0.10, whether or not adjusted for verification bias.

SURVIVAL

During a mean follow-up of 2.3 ± 1.2 years (3.1 ± 0.6 years among survivors), there were 224 deaths among 523 patients with Social Security numbers. Survival decreased monotonically with increasing PET-assigned clinical stage (**Figure**). Nonetheless, even among patients with PET stage 0, there were 24 deaths (2-year Kaplan-Meier death rate of 14%/y); 8 (33%) of these 24 deaths were in patients eventually diagnosed as having lung cancer.

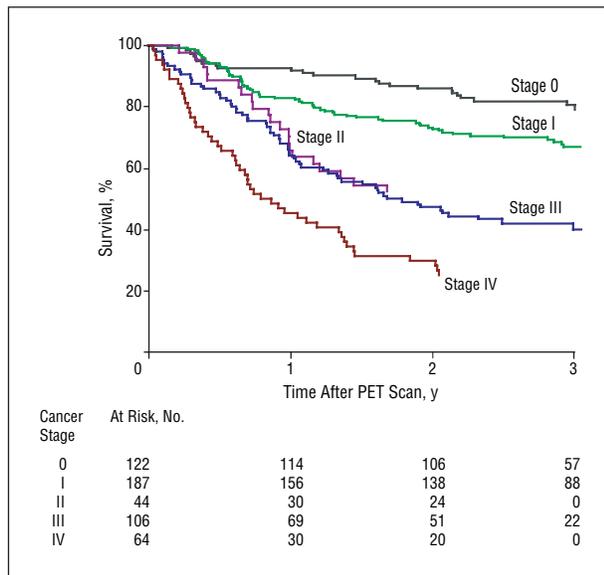


Figure. Survival according to cancer stage as determined by positron emission tomography (PET). Log-rank $\chi^2 = 97$ ($P < .001$).

COMMENT

In a large consecutive sample of 534 patients who underwent PET scanning for evaluation of suspected lung cancer, a subsequent gold standard tissue biopsy specimen was obtained in 419 patients (78%), while mediastinal lymph node sampling was done in 233 (44%). When only those patients who underwent PET scanning as well as a subsequent attempt at tissue diagnosis are considered, the estimated sensitivity of the PET scan for the detection of any cancer was 95%, which is consistent with the literature.⁸ However, after taking into account that a tissue diagnosis was not attempted in all patients, we found that the apparent sensitivity for PET scanning fell to approximately 85% to 87%, while the specificity increased.

The situation was different for the detection of mediastinal cancer: there appeared to be little association between findings on the PET scan and the clinical decision to obtain or not obtain mediastinal tissue. This is consistent with our practice to perform mediastinoscopy in all patients referred for possible resection of suspected lung cancer, regardless of PET scan findings. Thus, the apparent sensitivity of PET scanning for diagnosing mediastinal cancer was low, at only 50%, while the specificity was high, at 90%. The fact that only a minority of patients underwent mediastinal tissue sampling after PET scanning did not materially change these results.

Our observations are consistent with a natural experiment of the phenomenon of verification bias. As has been documented before, verification bias occurs when the results of a diagnostic test affect a decision to obtain a gold standard evaluation.^{9,15,18} It can result in a severe overestimation of sensitivity and underestimation of specificity.¹ In our study sample of patients who underwent PET scanning for suspected lung cancer, there was a marked association between PET scan results and the decision to obtain any subsequent tissue sample for the purpose of diagnosing any-stage cancer at any site. The de-

cision to perform mediastinal lymph node sampling, however, was not associated with PET scan results.

While our observations support the increasing concern about the importance of verification bias when assessing the value of diagnostic tests, there are some important limitations to consider. First, the ideal way to study a diagnostic test would be to perform both the diagnostic test and the gold standard test regardless of the results of the diagnostic tests. This is in practice difficult to do, although it has been done, for example, in the case of exercise testing.⁴

Second, it could be argued that our results may reflect suboptimal PET scanning techniques within our institution. We doubt this to be the case, given that our measured sensitivity without accounting for verification bias was entirely consistent with previously published observations.⁸ Furthermore, when we assessed mortality as a function of PET scan findings, we found a gradient whereby mortality increased as PET scan stage increased. Since all-cause mortality is an entirely unbiased and objective end point,¹⁹ this gradient suggests that the PET scan findings appropriately reflected pathologic stage.

Finally, the noninvasive diagnostic evaluation of lung cancer has advanced such that now PET scanning is frequently simultaneously combined with CT imaging.⁶ Since our patients did not undergo combined PET scanning and CT imaging, but rather our cohort was defined entirely by those who underwent PET scanning, we were not able to evaluate this newer technology. Nonetheless, it is noteworthy that recent reports of combined PET and CT have not sought to account for verification bias.⁶

Despite these limitations, there are important clinical implications of our findings. Previously reported values for the test accuracy of PET scanning, as well as other diagnostic tests commonly used in cancer care, have largely ignored the possibility of verification bias. In the case of lung cancer, our findings suggest that PET scanning is a poor tool for mediastinal staging and therefore should not be considered as a viable substitute for mediastinoscopy. Before other imaging technologies (such as combined PET and CT scanning) are accepted for this purpose, they should be tested in such a way as to minimize or eliminate verification bias. The best way to do this would be to prospectively obtain tissue samples in all patients undergoing imaging, regardless of the imaging findings.

Our results, along with previously published observations, argue strongly that future evaluations of diagnostic tests must take into account verification bias. Failure to take into account verification bias can lead to a misleading impression of the true diagnostic accuracy of a test. Furthermore, clinicians should be suspicious of reports of test accuracies that fail to consider verification bias; in cases of PET scanning for suspected lung cancer, clinicians might be well advised to lower their threshold for proceeding to definitive tissue diagnosis in the setting of negative PET scan findings.

One might think that the standard literature, which presents only data for patients undergoing biopsy, accurately represents the real-world setting, but in fact this is not the case. Patients who do not undergo biopsy continue to require follow-up evaluation and care and continue to be at

risk for poor outcomes if their otherwise detected disease goes undiagnosed. Although we cannot know for sure, it is possible that a more aggressive clinical approach to patients with negative findings on PET scans may have affected their poor prognosis (Figure); one third of the deaths occurred among patients eventually diagnosed with lung cancer. While ideally, prospective studies would insist that all patients undergo gold standard testing, an alternative method might be to account for all patients undergoing diagnostic testing, whether or not subsequent gold standard testing is obtained.

Accepted for Publication: October 6, 2006.

Correspondence: Michael S. Lauer, MD, Desk JJ40, Cleveland Clinic, 9500 Euclid Ave, Cleveland, OH 44122 (lauer@ccf.org).

Author Contributions: *Study concept and design:* Lauer, Murthy, Blackstone, Okereke, and Rice. *Acquisition of data:* Okereke. *Analysis and interpretation of data:* Lauer, Murthy, Blackstone, Okereke, and Rice. *Drafting of the manuscript:* Lauer, Blackstone, and Rice. *Critical revision of the manuscript for important intellectual content:* Lauer, Murthy, Blackstone, Okereke, and Rice. *Statistical analysis:* Lauer and Blackstone. *Administrative, technical, and material support:* Okereke. *Study supervision:* Lauer.

Financial Disclosure: None reported.

Funding/Support: This study was supported by grants R01 HL-66004-2, R01 HL-072771-02, and P50 HL-77107-1 from the National Institutes of Health (Drs Lauer and Blackstone).

REFERENCES

1. Choi BC. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *J Clin Epidemiol.* 1992;45:581-586.
2. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med.* 1978;299:926-930.
3. Punglia RS, D'Amico AV, Catalona WJ, Roehl KA, Kuntz KM. Effect of verification bias on screening for prostate cancer by measurement of prostate-specific antigen. *N Engl J Med.* 2003;349:335-342.
4. Froelicher VF, Lehmann KG, Thomas R, et al. The electrocardiographic exercise test in a population with reduced workup bias: diagnostic performance, computerized interpretation, and multivariable prediction. *Ann Intern Med.* 1998;128:965-974.
5. Roger VL, Pellikka PA, Bell MR, Chow CW, Bailey KR, Seward JB. Sex and test verification bias: impact on the diagnostic value of exercise echocardiography. *Circulation.* 1997;95:405-410.
6. Lardinois D, Weder W, Hany TF, et al. Staging of non-small-cell lung cancer with integrated positron-emission tomography and computed tomography. *N Engl J Med.* 2003;348:2500-2507.
7. Gould MK, Kuschner WG, Rydzak CE, et al. Test performance of positron emission tomography and computed tomography for mediastinal staging in patients with non-small-cell lung cancer: a meta-analysis. *Ann Intern Med.* 2003;139:879-892.
8. Gould MK, Maclean CC, Kuschner WG, Rydzak CE, Owens DK. Accuracy of positron emission tomography for diagnosis of pulmonary nodules and mass lesions: a meta-analysis. *JAMA.* 2001;285:914-924.
9. Blackstone EH, Lauer MS. Caveat emptor: the treachery of work-up bias. *J Thorac Cardiovasc Surg.* 2004;128:341-344.
10. Oken MM, Creech RH, Tormey DC, et al. Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am J Clin Oncol.* 1982;5:649-655.
11. Mountain CF. Revisions in the International System for Staging Lung Cancer. *Chest.* 1997;111:1710-1717.
12. Newman TB, Brown AN. Use of commercial record linkage software and vital statistics to identify patient deaths. *J Am Med Inform Assoc.* 1997;4:233-237.
13. Nishime EO, Cole CR, Blackstone EH, Pashkow FJ, Lauer MS. Heart rate recovery and treadmill exercise score as predictors of mortality in patients referred for exercise ECG. *JAMA.* 2000;284:1392-1398.
14. Miller TD, Hodge DO, Christian TF, Milavetz JJ, Bailey KR, Gibbons RJ. Effects of adjustment for referral bias on the sensitivity and specificity of single photon emission computed tomography for the diagnosis of coronary artery disease. *Am J Med.* 2002;112:290-297.
15. Diamond GA. Work-up bias. *J Clin Epidemiol.* 1993;46:207-208.
16. Diamond GA, Rozanski A, Forrester JS, et al. A model for assessing the sensitivity and specificity of tests subject to selection bias. *J Chronic Dis.* 1986;39:343-355.
17. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics.* 1983;39:207-215.
18. Cecil MP, Kosinski AS, Jones MT, et al. The importance of work-up (verification) bias correction in assessing the accuracy of SPECT thallium-201 testing for the diagnosis of coronary artery disease. *J Clin Epidemiol.* 1996;49:735-742.
19. Lauer MS, Blackstone EH, Young JB, Topol EJ. Cause of death in clinical research: time for a reassessment? *J Am Coll Cardiol.* 1999;34:618-620.