EDITORIAL

Caveat emptor: The treachery of work-up bias

Eugene H. Blackstone, MD[a,b]
Michael S. Lauer, MD[c]

See related article on page 396.

Nomori and colleagues[1] demonstrate the relationship between contrast ratio derived from F-18 fluorodeoxyglucose positron emission tomography (FDG-PET) in cT1 N0 M0 lung adenocarcinoma and pathologic TNM classification, carcinoembryonic antigen levels, lymphatic and vascular invasion, pleural involvement, and tumor differentiation. These observations constitute the scientific merit of the study. Quite properly, the authors go on to ask what the findings mean and, in particular, what clinical inferences they suggest. Based on what appears to be 100% sensitivity of the imaging test, they conclude that if the contrast ratio is less than 0.5, "limited lung resection could be indicated, lymph node dissection or mediastinoscopy could be reduced, or both."

At the heart of these seemingly logically derived clinical inferences lies treachery. We must be quick to state that these same or similar inferences would be drawn by more than 90% of the readership, not just in this context but also in the general context of interpreting the accuracy of any diagnostic test; the authors are well within the mainstream. It is the rare reader who knows that the lid has been blown off many diagnostic tests, particularly the ones cardiologists and cardiac surgeons have come to rely on in ischemic heart disease. Heretofore, our training and backgrounds have been deficient in interpreting the accuracy of diagnostic testing. We have been misled by our ignorance. The data have not been false, but the interpretation and inferences have been.

What Went Wrong

Nomori and colleagues[1] provide important details that give us not only insight into the value of their study but also a clue about the trap they have innocently set for the unsuspecting. The 44 patients presented are a highly selected subset of patients who had (1) major lung resection with mediastinal lymph node dissection and pathologic classification of disease (*gold standard* or *reference standard*), (2) tumors of specified size (large enough to be resolved by FDG-PET scanning) and characteristics (<3 cm, noncalcified nodule), and (3) a specific clinical diagnosis of cancer stage based at least in part on the very test they evaluated. Figure 1 shows a patient flow diagram formatted as suggested by the recently published Standards for Reporting of Diagnostic Accuracy (STARD) Initiative.[2] Note the many question marks accompanying various n values. What is apparent is that the 44 cases belong to a large group of noncalcified malignant tumors less than 3 cm in diameter on computed tomography, and that these were themselves a subset of 223 patients, probably most of whom did not have a gold standard (reference) diagnosis. A diagram like this shows the many ways bias can be introduced and lead to unjustified inferences.

Narrowing down a study to a defined patient subset is necessary to examine the relationship of diagnostic test results to particular kinds of pathology, as in Nomori and colleagues' article.[1] But it makes extrapolation of test results to the more general population (including our very next patient), whose pathology is not yet known, treacherous. Specifically, it is now known that if a test has been used to select patients—that is, has been used for its intended diagnostic purpose—and it is predominantly only patients with positive test results for whom gold standard verification of disease is obtained, then for that group of patients, sensitivity of the

From the Department of Thoracic and Cardiovascular Surgery,[a] the Department of Biostatistics and Epidemiology,[b] and the Department of Cardiovascular Medicine,[c] The Cleveland Clinic Foundation, Cleveland, Ohio.

Received for publication March 22, 2004; accepted for publication March 26, 2004.

Address for reprints: Eugene H. Blackstone, MD, The Cleveland Clinic Foundation, 9500 Euclid Ave, Desk F24, Cleveland, OH 44195 (E-mail: blackse@ccf.org).

J Thorac Cardiovasc Surg 2004;128:341-4

0022-5223/$30.00

doi:10.1016/j.jtcvs.2004.03.039

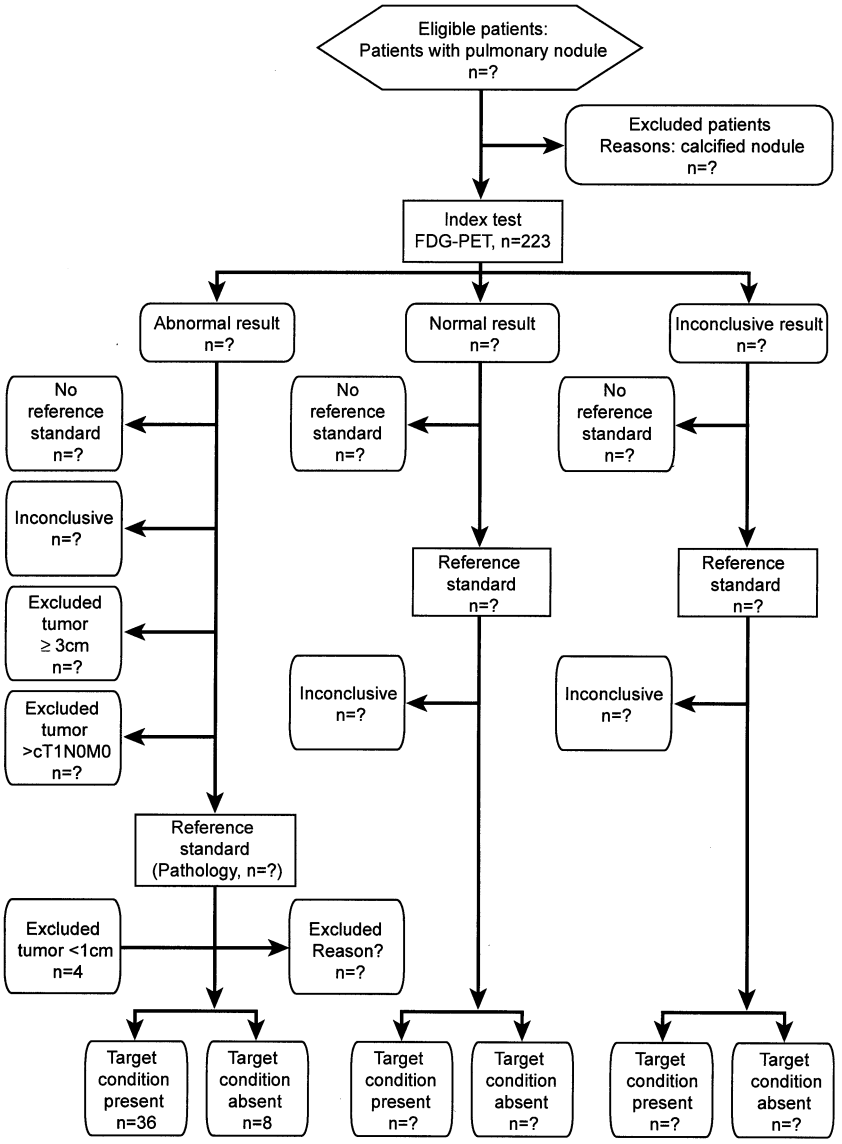

Figure 1. Partial reconstruction of patient flow diagram for study group of Nomori and colleagues in the fashion recommended in the STARD Initiative.[2,25] Many details and boxes are missing from the diagram because they could not be reconstructed from data supplied in the manuscript.

test is artificially inflated, often by vast amounts, and specificity is simultaneously deflated.[3] Thus, the unsuspecting reader may conclude that a contrast ratio of less than 0.5 on FDG-PET scanning is 100% sensitive and therefore useful for making the kind of surgical decisions suggested (specifically, not needing to perform a gold standard operation). In fact, it would be surprising if the test were more than 40% to 60% sensitive if it follows the pattern unfolding for diagnostic testing that affects cardiac surgery decisions!

The particular problem here, and the only one we dwell on in this editorial, is known as *work-up bias*.

Work-Up Bias

Ransohoff and Feinstein[4] coined the term work-up bias for their 1978 *New England Journal of Medicine* exposé of bias in diagnostic testing. Work-up bias occurs whenever a test is performed and a gold standard (reference) validation is not performed for each patient, and accuracy of the test is reported for only patients with reference validation. This is particularly apt to occur when the gold standard involves an invasive procedure, such as obtaining pathologic tissue in lung cancer. It also occurs when patients with a positive result go on to further testing (sequential-ordering bias[2]).

Work-up bias, or slight variants of it, has been called *verification bias*,[5,6] *validation bias*,[7] *referral bias*,[8,9] *sampling bias*,[10] and *selection bias*.[10-13]

The effect of work-up bias on purported accuracy of a diagnostic test is illustrated in Figure 2.[14] Patients with a positive test result are likely to undergo a procedure for tissue pathologic verification, resulting in a disproportionately large share of patients undergoing verification having a positive test. Sensitivity (positive test when disease is present) appears to be high. As a corollary, because few patients undergoing an invasive procedure will have had a negative test result, few of the patients found not to have pathologic disease will have had a negative test. Thus, specificity will appear poor (negative pathology in patients with negative test results).

What is important for readers to grasp is the magnitude of work-up bias on what we have been led to believe is the accuracy of the test. Take prostate-specific antigen (PSA) screening for prostate cancer, for which work-up bias is introduced from selective biopsy (gold standard). It has been suggested that the threshold for biopsy, a PSA level of 4.1 ng $\cdot$ mL^{-1} or greater, should be lowered to improve test sensitivity.[15] For men under the age of 60 years, it is thought that sensitivity of PSA screening is 57%, with 60% specificity. Punglia and colleagues[15] found that after adjustment for work-up bias, sensitivity was only 18%; that is, 82% of cancers are missed! However, specificity was 98%; that is, only 2% of men without cancer have a positive test result. In diagnosis of ischemic heart disease, simple stress testing is thought to have a sensitivity of 67% and a specificity of 73%.[16] In a Veterans Affairs study, in which patients referred for stress testing were required to undergo gold standard cardiac catheterization (eliminating work-up bias), the test was found to have a true sensitivity of only 44% but a reasonable specificity of 87%. Thus, better tests—imaging tests—were developed, such as stress echocardiography. This test was once thought to have 80% sensitivity for coronary artery disease and 45% specificity, but after accounting for work-up bias, these were about 40% and 85%, respectively![17] In the early days of exercise radionuclide testing, Rozanski and colleagues[18] found over a 5-year period that specificity of the test decreased from 84% to 27% as the spectrum of cases (spectrum bias) narrowed and its use as a screening test for angiography increased (work-up bias). Just the inverse (low to high) happened to sensitivity.

Why Are We Misled?

Of all diagnostic testing biases, work-up bias is the most counterintuitive.[19] Logically, a test's reference values, such as sensitivity and specificity, should be computed by using the subgroup of patients for whom a gold standard test has been made. However, we fail to appreciate that the results of the diagnostic test have themselves determined which patients will receive a gold standard test and which will not. Thus, we have observed lack of work-up bias only in settings in which a surgeon does not believe in the test or ignores it for purposes of decision making, always gets the test results "after the fact," or follows a protocol that requires gold standard testing no matter what is found in diagnostic testing. Otherwise, there is a strong correlation between the test results and performance of gold standard testing,[20-24] hence bias.

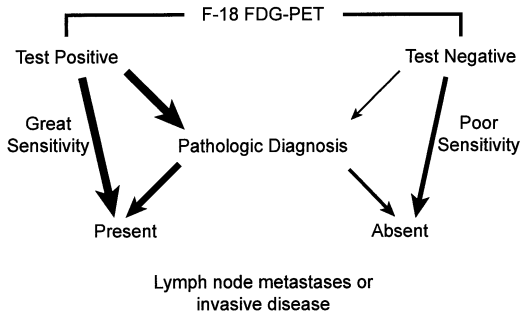

Figure 2. Simplified illustration of work-up bias in interpreting FDG-PET imaging in lung cancer. Few patients with a negative test result undergo gold standard pathologic verification of the test. Because patients with positive test results preferentially undergo invasive pathologic verification, test sensitivity is artificially inflated.

What to Do

Faith in diagnostic tests is being shattered just as "shopping mall diagnostics" are taking off! Although shoppers who submit to such testing are probably a somewhat biased group, they are more likely than known ill patients to represent the general population. Therefore, without work-up bias, one will find these tests rather insensitive in picking up existing disease, but considerably more specific (fewer false-negative results) than we are accustomed to thinking.

So alarming is the present state of diagnostic testing reporting that journals are adopting the STARD checklist.[2,25] The STARD Initiative was an international effort stimulated by growing recognition of biases that have fooled us all. Group members developed a 25-item checklist with cryptic explanation. Work-up bias is included in item 16: "The number of participants satisfying the criteria for inclusion that did or did not undergo the index test and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended)." This deceptively simple statement hardly seems to address biases, but it is absolutely fundamental because it is the nature of the patients tested and the influence of the test on whether the diagnosis is verified that introduce bias.

With respect to the article by Nomori and colleagues,[1] the STARD statement seems not to preclude publishing such articles.[26] Rather, it encourages authors to state carefully all subsets of their population and to consider the many sources of bias. It is presumed that authors (and readers) will use that information in interpreting their data, being particularly careful not to extrapolate conclusions to patients with yet unknown extent of disease.

Is warning, awareness, or even a 25-point checklist sufficient? We would suggest that as a minimum, such articles acknowledge that accuracy of testing has not been corrected for bias. Perhaps in the face of the rampant misinterpretation of test accuracy, whenever it is possible to estimate magnitude of the bias, correction of referent values for bias should be required.[12,13]

All Is Not Lost

If the reader's appropriate profound disillusion with diagnostic testing has now reached the level of despair, we suggest that just because a test performs poorly diagnostically (once work-up bias is accounted for) does not necessarily mean it is useless clinically. It may be that the test still has substantial prognostic value. This has been found to be the case, for example, with stress testing.[27] Schröder and Kranse[28] suggest that new recommendations for prostate cancer screening should arise from the European Screening for Prostate Cancer trial and the Prostate, Lung, Colorectal and Ovary trial, which focus on whether screening reduces mortality. That is, they seem to be suggesting that screening tests should focus on long-term results rather than accuracy of diagnosis. Screening tests may also be of value for identifying patients most likely to respond to therapy, particularly those therapies that carry important morbidity, such as chemoradiotherapy. Of course, study of prognostic importance requires long-term clinical studies and well-designed clinical trials, which are clearly more difficult and expensive to perform than studies of diagnostic accuracy.

Further Reading

For cardiothoracic surgeons, we highly recommend the article by Kelly and associates,[8] who review a large number of sources of bias in diagnostic imaging for esophageal cancer. The Mayo Clinic group provides an appendix that illustrates Begg and Greenes' method for correcting work-up bias.[9]

References

1. Nomori H, Watanabe K, Ohtsuka T, Naruke T, Suemasu K, Kobayashi T, et al. Fluorine 18–tagged fluorodeoxyglucose positron emission tomographic scanning to predict lymph node metastasis, invasiveness, or both, in clinical T1 N0 M0 lung adenocarcinoma. *J Thorac Cardiovasc Surg.* 2004;128:396-401.
2. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Ann Intern Med.* 2003;138: 40-4.
3. Choi BC. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *J Clin Epidemiol.* 1992;45:581-6.
4. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med.* 1978;299:926-30.
5. Kosinski AS, Barnhart HX. Accounting for nonignorable verification bias in assessment of diagnostic tests. *Biometrics.* 2003;59:163-71.
6. Cecil MP, Kosinski AS, Jones MT, Taylor A, Alazraki NP, Pettigrew RI, 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;49:735-42.
7. Green MS. The effect of validation group bias on screening tests for coronary artery disease. *Stat Med.* 1985;4:53-61.
8. Kelly S, Berry E, Roderick P, Harris KM, Cullingworth J, Gathercole L, et al. The identification of bias in studies of the diagnostic performance of imaging modalities. *Br J Radiol.* 1997;70:1028-35.
9. 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;112:290-7.
10. Diamond GA. "Work-up bias." *J Clin Epidemiol.* 1993;46:207-8.
11. Diamond GA. Reverend Bayes' silent majority: an alternative factor affecting sensitivity and specificity of exercise electrocardiography. *Am J Cardiol.* 1986;57:1175-80.
12. Diamond GA, Rozanski A, Forrester JS, Morris D, Pollock BH, Staniloff HM, et al. A model for assessing the sensitivity and specificity of tests subject to selection bias. *J Chronic Dis.* 1986;39:343-55.
13. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics.* 1983;39:207-15.
14. Lauer MS. Role of stress testing and cardiac imaging in patients who have undergone previous coronary revascularization. *Cardiol Rev.* 2000;8:158-65.
15. 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;349:335-42.
16. Froelicher VF, Lehmann KG, Thomas R, Goldman S, Morrison D, Edson 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: 965-74.
17. 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;95:405-10.
18. Rozanski A, Diamond GA, Berman D, Forrester JS, Morris D, Swan HJC. The declining specificity of exercise radionuclide ventriculography. *N Engl J Med.* 1983;309:518-22.
19. Begg CB. Advances in statistical methodology for diagnostic medicine in the 1980's. *Stat Med.* 1991;10:1887-95.
20. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med.* 1987;6:411-23.
21. Drum DE, Christacopoulos JS. Hepatic scintigraphy in clinical decision making. *J Nucl Med.* 1972;13:908-15.
22. McNeil BJ, Sanders R, Alderson PO, Hessel SJ, Finberg H, Siegelman SS, et al. A prospective study of computed tomography, ultrasound and gallium imaging in patients with fever. *Radiology.* 1981;139:647-53.
23. Marshall V, Williams DC, Smith KD. Diaphanography as a means of detecting breast cancer. *Radiology.* 1984;150:339-43.
24. Barr JT, Schumaker GE. Applying decision analysis in therapeutic drug monitoring: using receiver-operating characteristic curves in comparative evaluations. *Clin Pharm.* 1986;5:239-46.
25. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med.* 2003;138:W1-12.
26. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research: getting better but still not good. *JAMA.* 1995;274:645-51.
27. Lauer MS. Exercise electrocardiogram testing and prognosis: novel markers and predictive instruments. *Cardiol Clin.* 2001;19:401-14.
28. Schroder FH, Kranse R. Verification bias and the prostate-specific antigen test—is there a case for a lower threshold for biopsy? *N Engl J Med.* 2003;349:393-5.