Assessing Sensitivity and Specificity in New Diagnostic Tests: The Importance and Challenges of Study Populations

To the Editor—Hospital epidemiology frequently involves evaluating new diagnostic tests. This is particularly relevant in the example of active surveillance for multidrug-resistant organisms (MDROs) in high-risk populations. Determining the sensitivity and specificity, and therefore effectiveness, of potentially new diagnostic tests requires these surveillance methods to be epidemiologically and statistically assessed. Advice on study design and biostatistical methods for evaluating diagnostic tests has recently been published, but there are additional practical concerns for hospital epidemiology and other disciplines where the incidence of disease is low that are addressed only in this editorial.

Biostatistical advice on study design and reporting results for new diagnostic tests is limited. The STARD (Standards for the Reporting of Diagnostic Accuracy studies) statements discuss methods and provide recommendations on reporting and designing studies but do not delve into numerous and important biostatistical issues.[1,2] In order to help fill this gap we recently published an article discussing statistical methods for evaluating the sensitivity and specificity of novel dichotomous, diagnostic tests.[3] Our publication provides recommendations for sample size calculations during the design stage when assessing diagnostic tests, as well as detailing the importance of reporting confidence intervals around sensitivity and specificity results.

As briefly mentioned at the conclusion of the article, there is a trade-off between statistical rigor and the "practical realities of sample collection." In order to achieve narrow confidence intervals around estimates of sensitivity (and specificity), a minimum number of positive (and negative) patients must be drawn from the population to whom the test is going to be applied. Simply put, the sample size must include a large enough positive population so the confidence interval around the sensitivity can be precise. This is especially difficult when the prevalence is inherently low; as is often the case in hospital epidemiology.

When assessing the sensitivity of a new test, confidence intervals reflect both the sample size and the inherent sensitivity. Small sample sizes yield wide confidence intervals and therefore a greater statistical uncertainty of the results, even when the new diagnostic test has a high sensitivity. For example, the most recently reported prevalence of methicillin-resistant *Staphylococcus aureus* (MRSA) among hospital inpatients in US healthcare facilities is 66.4 per 1,000 patients (6.6%).[4] As shown in table 1, the low prevalence significantly affects the confidence intervals around the sensitivity when assessing a new MRSA surveillance test for hospitals.

These large sample sizes needed to achieve moderate to narrow confidence intervals can be daunting from a time and cost perspective. However, attempts to find alternatives to sampling large populations can prove difficult and come with their own set of epidemiological issues.

One tempting solution is to find known positive patients identified by clinical cultures and then test the new surveillance method to determine sensitivity. This practice of "cherry-picking" positive patients to get the population sizes needed for narrow confidence intervals potentially introduces even larger biases. Screening and diagnostic tests should be evaluated in the same population in which the test is going to be used.[5,6] Knowing the patient is positive can skew the interpretation of results and lead to an overestimation of sensitivity for the new test (known as diagnostic review bias).[2,7,8] As well, positive patients, as opposed to the true population in question, are likely to have different demographics, risk factors for the outcome, and different confounding variable distributions. These sometimes extreme differences between the population tested and the population in which the test is to be used lead to issues with generalizing results and, more importantly, yield incorrect sensitivity and specificity estimates for the test.[6,8]

Evaluating new diagnostic tests becomes even more complicated when there is no gold standard against which to compare the new test. Without a gold standard, additional methods are needed in order to maintain the validity of the

TABLE 1.   95% Confidence Intervals for 4 Different Predicted Sensitivities

Total sample size (estimated no. of positive patients)	Sensitivity of New MRSA Diagnostic Test			
	80%	85%	90%	95%
100 (7)	50.4%–100.9%	58.5%–111.4%	67.7%–112.2%	78.9%–111.1%
500 (33)	66.6%–93.4%	72.9%–97.0%	79.9%–100.1%	87.7%–102.3%
1,000 (65)	70.4%–89.6%	76.4%–93.6%	82.8%–97.2%	89.8%–100.2%

NOTE.   The 95% confidence intervals for 4 different predicted sensitivities use 3 different sample sizes and take into account 6.6% prevalence of MRSA.

analysis or risk incorrect sensitivity and specificity results. In these cases, options such as composite reference standard (recommended), latent class analysis, or discrepant analysis are needed in order to accurately assess the test.[3,7] All three of these methods require an additional sample to be collected or test to be performed and further add to the laboratory workup and analysis.[7]

Finding a balance between the need for narrow confidence intervals and the practical problems of finding a large and appropriate sample population is critical. Poorly designed or reported studies can lead to premature adoption of tests and the incorrect care of patients.[2,7,8] We believe that (a) studies should use larger sample sizes, and (b) the test should be evaluated in the population among which it is to be implemented, in order to provide the most accurate and meaningful results. Reporting confidence intervals and study population demographics will assist in this endeavor. Finding an appropriate balance will become increasingly more important as diagnostic tests, especially surveillance for MDROs, are used more routinely due to legislative requirements and the critical need to quickly and accurately test patients.

Paula Strassle, BS;[1] Aaron S. Hess, BS;[1]
Kerri A. Thom, MD, MS;[1]
Anthony D. Harris, MD, MPH;[1]

Affiliation: 1. Department of Epidemiology and Public Health, University of Maryland Baltimore, School of Medicine, Baltimore, Maryland.
Address correspondence to Anthony D. Harris, MD, MPH, MSTF 330, 685 W. Baltimore Street, Baltimore, MD 21201 (aharris@epi.umaryland.edu).

REFERENCES

1. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Ann Intern Med* 2003;138(1):40–44.
2. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med* 2003;138(1):W1–W12.
3. Hess AS, Shardell M, Johnson JK, et al. Methods and recommendations for evaluating and reporting a new diagnostic test. *Eur J Clin Microbiol Infect Dis* 2012;31(9):2111–2116.
4. Jarvis WR, Jarvis AA, Chinn RY. National prevalence of methicillin-resistant *Staphylococcus aureus* in inpatients at United States health care facilities, 2010. *Am J Infect Control* 2012;40:194–200.
5. Harper R, Henson D, Reeves BC. Appraising evaluations of screening/diagnostic tests: the importance of study populations. *Br J Ophthalmol* 2000;84:1198–1202.
6. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research: getting better but still not good. *JAMA* 1995;274(8):645–651.
7. Reitsma JB, Rutjes, AWS, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *J Clin Epidemiol* 2009; 62:797–806.
8. Whiting, P, Rutjes AWS, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004;140(3): 189–202.

How Long Is Long Enough? Determining the Optimal Surgical Site Infection Surveillance Period

To the Editor—Debate regarding the optimal postoperative surveillance period for detection of surgical site infections (SSIs) centers on the need for accurate case ascertainment balanced against efficient use of surveillance resources. Traditional surveillance definitions require a 1-year follow-up period for surgeries with an implantable device, the rationale being that indolent infections may not manifest for some time after the operative period. This prolonged duration for SSI surveillance places a burden on infection prevention and control resources and potentially delays reporting of adverse events in a timely manner to the surgical team.

A previous article retrospectively reviewed SSIs in total hip and knee replacements, coronary artery bypass grafts (CABGs), and mastectomies with implants for time to identification of a SSI.[1] Most deep or organ-space infections were captured within 90 days. This article describes an analysis of 10 years of SSI historical data at Vancouver Coastal Health to determine the proportion of infections identified within designated time frames of 1, 3, 6, 9, and 12 months.

Prospectively collected SSI data for cardiac, orthopedic, neurosurgical, spinal, thoracic, and vascular services were available from 2000 to 2010–2011 at our 3 facilities (1 tertiary care adult hospital and 2 community hospitals). The specific procedures followed included CABG, hip and knee replacements, craniotomies, spinal procedures with instrumentation/implants, thoracotomies, and vascular grafts. Standard definitions for SSI as described by the Centers for Disease Control and Prevention National Healthcare Surveillance Network were used.[2] Cases were identified by routine surveillance by infection preventionists, assessment of laboratory data, review of the surgical case list, voluntary surgeon reporting, and review of hospital readmissions with a diagnosis of infection; this methodology remained consistent throughout the 10-year analysis period. Cases where an SSI was detected beyond the standard 1-year follow-up were excluded from the analysis.

A total of 50,128 procedures were followed at our 3 facilities over the 10-year period, and 888 SSIs (1.7 SSI/100 proce-