

# Spectrum Bias or Spectrum Effect? Subgroup Variation in Diagnostic Test Evaluation

Stephanie A. Mulherin, MSPH, and William C. Miller, MD, PhD, MPH

Diagnostic tests must be evaluated in a clinically relevant population. However, test performance often varies across population subgroups. *Spectrum bias*, a term commonly used to describe this heterogeneity, is typically thought to occur when diagnostic test performance varies across patient subgroups and a study of that test's performance does not adequately represent all subgroups. Yet subgroup variation is not a bias if appropriate analyses are conducted. Failure to recognize and address heterogeneity will lead to estimates of test performance that are not generalizable to the relevant clinical populations. Heterogeneity can be addressed with relatively simple stratification procedures, limited primarily

by the sample size and the precision of the estimates. This paper proposes the use of the term *spectrum effect*, rather than *spectrum bias*, and outlines strategies for using stratified sensitivity and specificity estimates, likelihood ratios, and receiver-operating characteristic curves. Investigators of diagnostic tests should consider the potential for spectrum effect seriously and should address heterogeneity in their analyses. Furthermore, clinicians should consider study samples carefully to determine whether results are generalizable to their specific patient population.

*Ann Intern Med.* 2002;137:598-602.

www.annals.org

For author affiliations, see end of text.

When considering a new diagnostic test, clinicians must decide whether an evaluation of a new test is valid and whether the results are generalizable to their own patient population. An optimal study of a new diagnostic test includes a broad spectrum of persons who would normally undergo the test in a clinical setting. This broad spectrum is necessary to produce valid, precise, and generalizable estimates of test performance (1–3). However, test performance often varies with patient characteristics. For example, the sensitivity (the probability of positive test results given disease) ( $P[T+ | D+]$ ) and specificity (the probability of negative test results given no disease) ( $P[T- | D-]$ ) of a diagnostic test may be higher or lower in certain patient subgroups (4, 5). Evaluations of diagnostic tests must attempt to achieve valid, precise, and generalizable measurements of test accuracy, despite this heterogeneity. In this paper, we reexamine the issue of patient spectrum first described by Ransohoff and Feinstein (6). We provide guidance for clinicians interpreting reports of diagnostic test evaluations as well as suggestions for data analysis and presentation.

## ORIGINS OF THE CONCEPT OF SPECTRUM BIAS

Ransohoff and Feinstein introduced the concept of patient spectrum in a landmark article in 1978 (6). They observed that the performance of a test in practice may be misrepresented by clinical studies that include too narrow a range of diseased case-patients or too narrow a range of nondiseased controls. They highlighted characteristics to consider when designing a study that examines the efficacy of a test, including pathologic, clinical, and comorbid features of the case-patient and control groups. Their concern that the relationship between disease state and test performance may change according to characteristics of the patient sample has profoundly influenced diagnostic test evaluation.

Although Ransohoff and Feinstein did not use the

term *spectrum bias* explicitly, their discussion of patient spectrum in a paper examining biases affecting diagnostic test research fostered the use of the term to describe subgroup variation (5, 7–11). The term *spectrum bias* implies that patient spectrum is a problem requiring correction. Consequently, a wide variety of remedies has been proposed, including reporting the average sensitivity and specificity of the sample tested (3, 7, 8, 12), using covariate adjustment (13), reporting results of subgroup analyses (2–4, 7, 8, 10, 12, 14, 15), and reporting study population characteristics as a minimum (2, 3, 7, 8, 14). Others have argued that many estimates of sensitivity and specificity are useless because study samples are too dissimilar to actual patient populations (11, 16, 17).

## SPECTRUM EFFECT: AN ALTERNATIVE TERM FOR SUBGROUP VARIATION IN TEST PERFORMANCE

The term *spectrum bias* is commonly applied when the disease–test relationship is heterogeneous across patient subgroups and the study draws preferentially from a limited portion of the patient spectrum. Including a broad spectrum of patients has been the most common approach to addressing subgroup heterogeneity (7, 8, 14, 16, 18). However, spectrum bias is a misnomer because it falsely suggests a systematic error in study design, data collection, or analysis that compromises the validity of the results.

We suggest that the term *spectrum bias* be replaced with the term *spectrum effect*, which reflects the inherent variation in test performance among population subgroups. Subgroup variation is not a bias but is clinically relevant information to be identified and reported with appropriate analyses. If a spectrum effect is possible, heterogeneity should be assessed by subgroup analyses of test performance. The stability and validity of the subgroup estimates will be limited by the sample size of each category.

The most common approach for addressing patient

**Table 1. Performance of a Hypothetical Diagnostic Test under Different Spectrums of Age**

Examination of New Diagnostic Test	True-Positive Results	True-Negative Results	False-Positive Results	False-Negative Results	Sensitivity	Specificity
	←————— <i>n</i> —————→				%	
Hypothetical test performance in source sample						
Overall	850	850	150	150	85	85
Participants <50 y of age	475	375	125	25	95	75
Participants ≥50 y of age	375	475	25	125	75	95
First examination with preferentially recruited younger participants*						
Overall	450	400	100	50	90	80
Participants <50 y of age	356	281	94	19	95	75
Participants ≥50 y of age	94	119	6	31	75	95
Second examination with preferentially recruited older participants†						
Overall	400	450	50	100	80	90
Participants <50 y of age	119	94	31	6	95	75
Participants ≥50 y of age	281	356	19	94	75	95

\* Includes 75% of persons <50 years of age and 25% of persons ≥50 years of age.

† Includes 25% of persons <50 years of age and 75% of persons ≥50 years of age.

spectrum—including a variety of participants and reporting a single estimate of test performance—may actually produce an estimate with limited clinical utility. When heterogeneity is present, the overall population estimate is not generalizable to any specific patient population. Ignoring the spectrum effect will lead to a population estimate that is a weighted average of test performances across subgroups (7, 12, 13, 19, 20). This estimate will vary with the proportion of patients in each subgroup included in the weighted average.

The primary issue for the spectrum effect is generalizability. If test performance varies substantially by sex, there is little clinical value in using the population estimate of test performance that combines men and women to determine post-test probability for any individual patient. Instead, the subgroup estimates for men should be applied in men and the subgroup estimates for women should be used in women. Furthermore, estimating test performance in a clearly defined, narrow spectrum of patients is not biased but is instead a valid estimate for a specific patient subgroup. However, such estimates should not be generalized to other patient subgroups.

A simple example illustrates the spectrum effect (Table 1). Consider a new diagnostic test that performs heterogeneously across age. Among persons younger than 50 years of age, the test performs with a sensitivity and specificity of 95% and 75%, respectively. Among persons at least 50 years of age, the test performs with a sensitivity of 75% and a specificity of 95%. In an investigation of this test, the study sample is drawn preferentially from the source population so that 75% of persons younger than 50 years of age are included in the study and only 25% of persons 50 years of age or older are included. The sensitivity and specificity of the test in this study sample are found to be 90% and 80%, respectively. In a second study, only 25% of

persons younger than 50 years of age are included and 75% of persons 50 years of age or older are included. The sensitivity and specificity are determined to be 80% and 90%, respectively. Comparing these two studies, we find that both have included a reasonable spectrum of participants but provide differing population-average estimates of sensitivity and specificity. However, estimates of sensitivity and specificity within strata of age are unbiased, that is, identical to those in the source population.

## ASSESSING THE SPECTRUM EFFECT

### Stratified Analysis of Sensitivity, Specificity, Likelihood Ratios, and Receiver-Operating Characteristic Curves

Sensitivity, specificity, likelihood ratios, and receiver-operating characteristic (ROC) curves are commonly used indicators of test accuracy. Each of these measures can be evaluated for spectrum effect with stratified analyses of patient subgroups. Subgroup variation in sensitivity and specificity estimates can be examined analytically by stratifying on the characteristic defining the subgroup and by employing a simple chi-square test of association. An association between a patient characteristic and the result of a diagnostic test may indicate the need to report estimates for subgroups of the sample. The clinical significance of the difference in sensitivity or specificity must also be considered when the *P* value of the chi-square test is being examined. Small cell sizes may cause the test to miss a clinically meaningful difference. Conversely, very large cell sizes may result in a statistically significant association that is not clinically meaningful, as often occurs with specificity estimates arising from cross-sectional studies of populations with a low prevalence of disease.

The positive ( $P[T+ | D+]/P[T+ | D-]$ ) and negative ( $P[T- | D+]/P[T- | D-]$ ) likelihood ratios provide an

**Table 2. Sensitivity, Specificity, and Likelihood Ratios for Performance of an Enzyme Immunoassay for *Chlamydia trachomatis*, Stratified by Selected Patient Characteristics\***

Factor	Sensitivity, %	P Value†	Specificity, %	P Value‡	Positive Likelihood Ratio	P Value§	Negative Likelihood Ratio	P Value§
Overall	73.4 (69.9–76.9)		99.4 (99.2–99.6)		116.0 (84.1–160.0)		0.27 (0.23–0.31)	
Age		≤0.001		≤0.001		0.02		≤0.001
≤24 y	75.9 (72.1–79.7)		99.5 (99.3–99.7)		147.0 (92.5–233.7)		0.24 (0.21–0.28)	
>24 y	58.3 (47.7–68.9)		99.2 (98.8–99.6)		73.6 (45.9–118.0)		0.42 (0.33–0.54)	
Clinic type		>0.2		0.003		0.17		>0.2
Family planning	72.1 (66.4–77.8)		99.2 (98.8–99.6)		137.7 (90.4–209.9)		0.28 (0.23–0.35)	
Sexually transmitted disease	74.3 (69.6–79.0)		99.5 (99.3–99.7)		87.0 (53.0–142.5)		0.26 (0.22–0.31)	

\* Values in parentheses are 95% CIs.

† P value for comparison of sensitivities in the two strata.

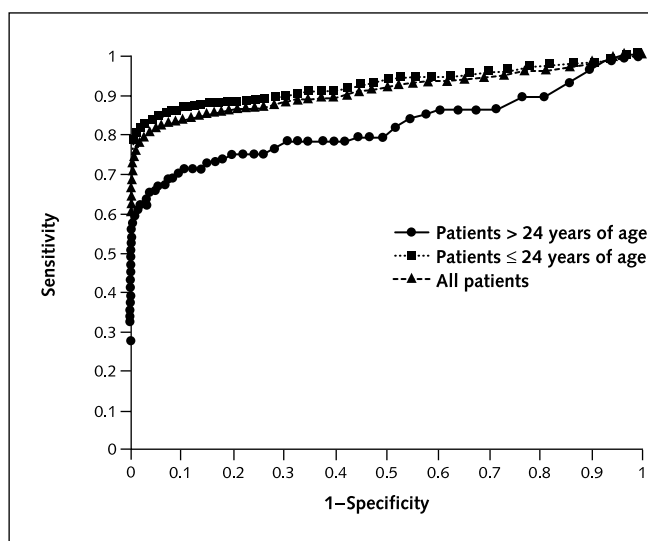
‡ P value for comparison of specificities in the two strata.

§ P value for the homogeneity test.

estimate of the change in the odds for disease given a particular test result. Because the algebraic construction of likelihood ratios is analogous to risk ratios (21), the spectrum effect can be evaluated in likelihood ratios by using a test of homogeneity, available in many software programs. Statistical tests of homogeneity have low power to detect an effect; therefore,  $\alpha$  values are often less conservative for these tests (for example,  $\alpha = 0.10$ ) (22).

Receiver-operating characteristic curves plot the false-positive rate ( $1 - \text{specificity}$ ) against the true-positive rate (sensitivity). The area under this curve indicates how accurately a test discriminates between true-positive and true-negative results, independent of the positive test threshold (23). Subgroup-specific ROC curves can be constructed

**Figure. Receiver-operating characteristic curve for performance of an enzyme immunoassay for *Chlamydia trachomatis*, stratified by age.**



The discriminatory accuracy for enzyme immunoassay is better for participants 24 years of age or younger (area under the curve [ $\pm$ SE],  $0.927 \pm 0.009$ ) than for those older than 24 years of age (area under the curve [ $\pm$ SE],  $0.819 \pm 0.033$ ). The areas under the stratified curves were significantly different ( $P \leq 0.001$ ).

and compared. The curve that lies farther toward the upper left quadrant has better discriminatory accuracy. Calculating the critical ratio and comparing it with a normal distribution can provide a quantitative comparison of these uncorrelated ROC curves (24).

### Stratified Analysis of Enzyme Immunoassay for *Chlamydia trachomatis*

As an example of addressing spectrum effect, we assessed the performance of an enzyme immunoassay for *Chlamydia trachomatis* (25). We examined age and clinic type as two representative factors with the potential for subgroup variation. Compared with a reference standard of ligase chain-reaction assay, the overall sensitivity and specificity for enzyme immunoassay were 73.4% and 99.4%, respectively (Table 2). When stratified by age, sensitivity and specificity varied significantly between older and younger patients; the test performed better in the younger subgroup ( $P \leq 0.001$ ) (Table 2). Positive and negative likelihood ratios also differed significantly after stratification by age. The enzyme immunoassay ROC curves and their areas revealed considerable heterogeneity after stratification by age (Figure). The enzyme immunoassay seemed to have the highest discriminatory accuracy among participants who were 24 years of age or younger. This finding was confirmed by formal comparison of the areas under the curve (critical ratio, 3.18;  $P \leq 0.001$ ), verifying our visual assessment.

In contrast, when clinic type was considered, only the specificity of enzyme immunoassay varied significantly ( $P = 0.003$ ) (Table 2). However, the qualitative differences in specificity were small for both age and clinic type; therefore, these differences in specificity may not be clinically meaningful. Sensitivity, positive and negative likelihood ratios, and ROC areas did not differ by clinic type.

### DISCUSSION

New diagnostic tests should be evaluated in a clinically relevant sample (12). However, even in such a sample, test performance is likely to vary across patient subgroups. The

term *spectrum bias* has been used to describe this heterogeneity and also has been applied when the study sample included only a clinically relevant subgroup. We believe the term *spectrum bias* falsely implies an error in study design and wrongly challenges the validity of study results. We suggest that *spectrum effect*, which reflects the different accuracies across subgroups of a sample, may be a better term. Subgroup variation is not a bias but must be evaluated by using appropriate methods and must be reported.

When reviewing a study of a diagnostic test, clinicians should examine the relevance of the study sample to their own clinical population. Clinicians should also consider whether spectrum effect has been evaluated. Furthermore, results in a narrow spectrum of patients should not be dismissed but rather generalized cautiously to an appropriately narrow clinical population.

One limitation of subcategorizing study samples is that the sample tested has to be large enough to support reasonably precise subgroup estimates. Diagnostic test evaluations are often performed with relatively small sample sizes, leading to imprecise estimates, particularly for sensitivity. Analysis of subgroups will be less precise than the overall population estimates. If possible, the impact of potentially modifying patient characteristics should be considered when power estimates are initially calculated. Oversampling characteristics that occur infrequently in clinical practice may help mediate the additional complexities introduced by considering the spectrum effect in the study-design phase. In the publication of diagnostic test evaluations, estimates of subgroup variables that seem to be meaningfully different should be presented even if they do not attain statistical significance. Inclusion of such results would greatly facilitate appropriate meta-analyses of diagnostic test performance.

In cases in which several characteristics of a clinical sample simultaneously modify the performance of a diagnostic test, categorical analysis of subpopulations may be impossible because of small cell sizes. In such circumstances, the spectrum effect can be assessed by modeling test performance with logistic regression (26). The logistic model can produce estimates of test performance when small cell sizes prohibit obtaining such estimates nonparametrically. Modeling also facilitates investigation of patient characteristics that are multicategorical or continuous and allows the investigator to model separately factors that affect sensitivity and specificity (4).

Subgroups of a clinically relevant sample should be carefully defined to provide relatively homogeneous estimates and to facilitate comparisons with other evaluations of diagnostic tests. However, patient characteristics are not necessarily dichotomous in nature and decisions to group similar patients together must be balanced against the need to maintain sufficiently large subgroups to estimate reasonably precise effects. If large subgroups are required to maintain precision, residual heterogeneity will persist. Reporting the characteristics of the sample tested, providing subgroup

estimates of test performance, and carefully avoiding speculation beyond the range of the data should be integral in every study of diagnostic tests (4, 12). Average effects will change based on the distribution of key modifiers in the study sample.

In summary, correct diagnosis is nearly as important as correct treatment. Therefore, the scientific rigor applied to the design and analysis of clinical trials should be mirrored in evaluations of diagnostic tests. We have outlined simple methods that can be easily executed to that end.

From University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

**Acknowledgments:** The authors thank David Ransohoff, MD; Michael Pignone, MD, MPH; and Jay Kaufman, PhD, for their thoughtful comments.

**Grant Support:** By the University of North Carolina STD Clinical Research Center (National Institute of Allergy and Infectious Diseases grant UO131496) and the Clinical Associate Physician Program of the General Clinical Research Center (RR00046), Division of Research Resources, National Institutes of Health.

**Requests for Single Reprints:** William C. Miller, MD, PhD, MPH, University of North Carolina at Chapel Hill, Department of Epidemiology, CB# 7435, 2105F McGavran-Greenberg, Chapel Hill, NC 27599-7435.

Current author addresses are available at [www.annals.org](http://www.annals.org).

## References

1. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. *JAMA*. 1994;271:389-91. [PMID: 8283589]
2. Philbrick JT, Horwitz RI, Feinstein AR, Langou RA, Chandler JP. The limited spectrum of patients studied in exercise test research. Analyzing the tip of the iceberg. *JAMA*. 1982;248:2467-70. [PMID: 7131702]
3. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, et al. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med*. 1994; 120:667-76. [PMID: 8135452]
4. Hlatky MA, Pryor DB, Harrell FE Jr, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. *Am J Med*. 1984;77:64-71. [PMID: 6741986]
5. Curtin F, Morabia A, Pichard C, Slosman DO. Body mass index compared to dual-energy x-ray absorptiometry: evidence for a spectrum bias. *J Clin Epidemiol*. 1997;50:837-43. [PMID: 9253396]
6. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978;299:926-30. [PMID: 692598]
7. Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann Intern Med*. 1992;117:135-40. [PMID: 1605428]
8. Mower WR. Evaluating bias and variability in diagnostic test reports. *Ann Emerg Med*. 1999;33:85-91. [PMID: 9867892]
9. O'Connor PW, Tansay CM, Detsky AS, Mushlin AI, Kucharczyk W. The effect of spectrum bias on the utility of magnetic resonance imaging and evoked potentials in the diagnosis of suspected multiple sclerosis. *Neurology*. 1996;47: 140-4. [PMID: 8710067]
10. Moons KG, van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and bayes' theorem in assessing

diagnostic probabilities: a clinical example. *Epidemiology*. 1997;8:12-7. [PMID: 9116087]

11. Reid MC, Lane DA, Feinstein AR. Academic calculations versus clinical judgments: practicing physicians' use of quantitative measures of test accuracy. *Am J Med*. 1998;104:374-80. [PMID: 9576412]

12. Hlatky MA, Mark DB, Harrell FE Jr, Lee KL, Califf RM, Pryor DB. Rethinking sensitivity and specificity. *Am J Cardiol*. 1987;59:1195-8. [PMID: 3554956]

13. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med*. 1987;6:411-23. [PMID: 3114858]

14. 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. [PMID: 7637146]

15. Sox HC. The evaluation of diagnostic tests: principles, problems, and new developments. *Annu Rev Med*. 1996;47:463-71. [PMID: 8712795]

16. Sox HC Jr. Probability theory in the use of diagnostic tests. An introduction to critical study of the literature. *Ann Intern Med*. 1986;104:60-6. [PMID: 3079637]

17. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. *JAMA*. 1994;271:703-7. [PMID: 8309035]

18. Sheps SB, Schechter MT. The assessment of diagnostic tests. A survey of

current medical research. *JAMA*. 1984;252:2418-22. [PMID: 6481928]

19. Harris JM Jr. The hazards of bedside Bayes. *JAMA*. 1981;246:2602-5. [PMID: 7299988]

20. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med*. 1992;116:78-84. [PMID: 1530753]

21. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol*. 1991;44:763-70. [PMID: 1941027]

22. Rothman KJ, Greenland S. *Modern Epidemiology*. 2nd ed. Philadelphia: Lippincott-Raven; 1998.

23. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143:29-36. [PMID: 7063747]

24. McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Med Decis Making*. 1984;4:137-50. [PMID: 6472062]

25. Miller WC, Hoffman IF, Owen-O'Dowd J, McPherson JT, Privette A, Schmitz JL, et al. Selective screening for chlamydial infection: which criteria to use? *Am J Prev Med*. 2000;18:115-22. [PMID: 10698241]

26. Coughlin SS, Trock B, Criqui MH, Pickle LW, Browner D, Tefft MC. The logistic modeling of sensitivity, specificity, and predictive value of a diagnostic test. *J Clin Epidemiol*. 1992;45:1-7. [PMID: 1738006]

---

**Current Author Addresses:** Ms. Mulherin and Dr. Miller: University of North Carolina at Chapel Hill, Department of Epidemiology, CB# 7435 McGavran-Greenberg, Chapel Hill, NC 27599-7435.