*Clinical Infectious Diseases*

SPECIAL SECTION/INVITED ARTICLE

# Average Weighted Accuracy: Pragmatic Analysis for a Rapid Diagnostics in Categorizing Acute Lung Infections (RADICAL) Study

**Ying Liu,[1] Ephraim L. Tsalik,[2,3] Yunyun Jiang,[4] Emily R. Ko,[2] Christopher W. Woods,[2,5] Ricardo Henao,[2] and Scott R. Evans[4]**

[1]Biogen, Inc., Boston, MA, USA; [2]Center for Applied Genomics and Precision Medicine, Department of Medicine, Duke University, Durham, NC, USA; [3]Emergency Department Service, Durham VA Health Care System, Durham, NC, USA; [4]Biostatistics Center, George Washington Milken Institute School of Public Health, Rockville, MD, USA; and [5]Medicine Service, Durham VA Health Care System, Durham, NC, USA

Patient management relies on diagnostic information to identify appropriate treatment. Standard evaluations of diagnostic tests consist of estimating sensitivity, specificity, positive/negative predictive values, likelihood ratios, and accuracy. Although useful, these metrics do not convey the tests' clinical value, which is critical to informing decision-making. Full appreciation of the clinical impact of a diagnostic test requires analyses that integrate sensitivity and specificity, account for the disease prevalence within the population of test application, and account for the relative importance of specificity vs sensitivity by considering the clinical implications of false-positive and false-negative results. We developed average weighted accuracy (AWA), representing a pragmatic metric of diagnostic yield or global utility of a diagnostic test. AWA can be used to compare test alternatives, even across different studies. We apply the AWA methodology to evaluate a new diagnostic test developed in the Rapid Diagnostics in Categorizing Acute Lung Infections (RADICAL) study.

**Keywords.** diagnostic test; diagnostic yield; prevalence; relative importance; average weighted accuracy (AWA).

The management of infectious diseases relies heavily on diagnostic information. There are a tremendous number of disease-causing etiologies, specifically microorganisms, and a growing battery of drugs to treat them. Diagnostic testing is critical to correctly match the right treatment to the disease. In the absence of this information, management relies on empiricism and best practices. Enabled by recent technological advances, the number and capability of diagnostic tests have grown rapidly. In order to optimally use these diagnostic tests, new methodologies are needed to better inform their clinical utility.

Standard evaluations of diagnostic tests consist of estimating sensitivity, specificity, positive/negative predictive values, likelihood ratios, and accuracy [1]. Although useful, these evaluations do not effectively convey the clinical value of the test, which is critical to informing clinical decision-making. In order to fully appreciate the clinical impact of a diagnostic test, analyses must integrate test sensitivity and specificity, account for the prevalence of the disease within the population to which the diagnostic will be applied, and account for the relative importance of specificity vs sensitivity by considering the clinical implications of false-positive and false-negative results.

Typically in diagnostic studies, arbitrary goals for sensitivity and specificity are defined, for example, each having to exceed 90%. However, it is important to integrate the 2 measures. If sensitivity was extremely high (eg, 98%), then one may be willing to sacrifice on specificity to a degree. Whereas if sensitivity was only marginal (eg, 90%), then one may be less willing to sacrifice specificity. Consider choosing between 2 tests: one with a higher sensitivity and one with a higher specificity. Which diagnostic test will optimize clinical utility? Here, one must consider the relative importance of false-positive and false-negative errors based on the clinical consequences. False-positive results may result in overtreatment, while false-negative results may result in undertreatment. A test with false-positive and false-negative error rates of 1% and 11% would not satisfy the arbitrary goal noted above. However, if the negative error rate is much more important than the false-positive error rate, then the test may have more utility than, for example, a test with false-positive and false-negative error rates of 9% that would satisfy the goal. Addressing this challenge requires a global assessment accounting for both sensitivity and specificity.

The prevalence of disease in the population in which the diagnostic is applied should be considered since the resulting diagnostic yield depends on that prevalence. For example, consider application of a diagnostic test with a sensitivity and specificity of 90% and 80%, respectively, to 1000 patients when the

Received 23 January 2019; editorial decision 20 May 2019; accepted 1 June 2019; published online June 3, 2019.

Correspondence: S. R. Evans, George Washington University, Rockville, MD (sevans@bsc.gwu.edu).

**Clinical Infectious Diseases® 2019;XX(XX):1–6**



DOI: 10.1093/cid/ciz437

disease prevalence is 20%. The expected diagnostic yield is 20 false negatives and 160 false positives. If the prevalence was instead 2%, then the expected diagnostic yield for the same test is 2 false negatives and 196 false positives. Understanding this implication is important for clinical decision-making. If a disease is highly prevalent, then high sensitivity is increasingly important. Otherwise, the low sensitivity would result in many false negatives. However, if the disease is rare, then even a modest sensitivity would result in few false negatives while specificity becomes increasingly important to avoid a high false-positive rate.

Accuracy, defined as the number of correct test results/number of tests performed, combines sensitivity and specificity. However, accuracy cannot be compared across studies unless the prevalence of disease is similar. Another drawback of the accuracy statistic is that it does not distinguish different types of errors assuming that sensitivity and specificity are equally important. For example, a test with sensitivity of 0.2 and specificity of 0.8 will have the same accuracy (0.5) as a test with sensitivity of 0.8 and specificity of 0.2, given a disease prevalence of 0.5.

To address these limitations, we introduce the average weighted accuracy (AWA) representing a pragmatic metric of diagnostic yield or global utility of a diagnostic test accounting for the prevalence of disease and the relative importance of sensitivity and specificity. AWA can be used to compare test alternatives across different studies. We apply the AWA methodology to evaluate a diagnostic test developed in the Rapid Diagnostics in Categorizing Acute Lung Infections (RADICAL) study [2, 3].

## METHODOLOGY

### Weighted Accuracy

Weighted accuracy (WA) [4] is interpreted as accuracy adjusted for the relative importance (r) of false-positive vs false-negative errors. WA ranges from 0% to 100%, with higher values indicating better accuracy.

$$WA = \frac{[rp(specificity) + (1-p)sensitivity]}{rp + 1 - p}$$

where $p$ is the prevalence of the disease. A relative importance of $r = 0.25$ means that making 1 false-negative error is equivalent to making 4 false-positive errors (or specificity is 25% as important as sensitivity).

### Average Weighted Accuracy

The specific disease prevalence may be unknown, but the relevant range of the prevalence may be identifiable. The AWA is the average WA over the relevant prevalence range calculated as follows given a relevant prevalence $p \in [a,\ b]$, where $a$ and $b$ are the lower and upper boundary of the prevalence range, respectively.

$$AWA = c_1 Sensitivity + c_2 Specificity,$$

$$c_1 = \frac{1}{1-r} - \frac{r}{(b-a)(1-r)^2} ln\left(\frac{r + b(1-r)}{r + a(1-r)}\right),$$

$$c_2 = \frac{r}{(1-r)(b-a)} \ln\left(\frac{r + b(1-r)}{r + a(1-r)}\right) - \frac{r}{1-r} + \frac{r^2}{(1-r)^2(b-a)} \ln\left(\frac{r + b(1-r)}{r + a(1-r)}\right).$$

where $c_1$ and $c_2$ are constants that can be calculated from $a$, $b$ and $r$. The standard error for AWA can be calculated (Supplementary Materials, Part 1) and used for confidence interval (CI) estimation and hypothesis testing. AWA has 2 advantages compared with traditional statistics. First, the relative importance of sensitivity and specificity are incorporated. Second, it facilitates a flexible evaluation when the prevalence cannot be easily summarized by a single number. Utilizing AWA over a relevant range of prevalence takes into consideration the heterogeneity of the prevalence across subpopulations. A step-by-step algorithm implementing AWA analyses is described.

### Implementation of AWA Analyses

#### *Step 1: Determine the Relevant Range of the Prevalence*

The relevant prevalence range can often be obtained from existing literature or a pilot study of the population.

#### *Step 2: Determine the Relative Importance of False-positive and False–negative Test Results*

Relative importance can be formally determined through cost-effectiveness or benefit-to-risk analysis based on the clinical consequences of diagnostic test results. An outline of this approach is provided in the Supplementary Materials, Part 2. In lieu of data to conduct formal cost-effectiveness analyses, a survey of experts can be conducted.

#### *Step 3: Select a Cutoff for Defining a Diagnostic*

For biomarkers that are measured on a continuum, AWA can be used to select an optimal cutoff for defining a diagnostic test's threshold. Given the relative importance of the errors, what is the cutoff probability that would maximize diagnostic yield, that is, AWA? AWA can be plotted as a function of the selected probability to determine the cutoff that maximizes yield.

#### *Step 4: Determine the Hypothesis*

Formal hypothesis testing can be performed based on the AWA. To claim that the diagnostic provides utility or is superior to an alternative diagnostic, the AWA of a new diagnostic would have to be shown to exceed the AWA of a diagnostic alternative. If

an alternative does not exist, then the AWA for the test can be compared to that of the best random test (BRT).

#### The Best Random Test.

A random test is a test that indicates a positive test with a fixed probability regardless of the true disease status. An example of a random test is flipping a coin to determine disease status, which uses a fixed probability of 0.5. Other random tests can be created by varying the fixed probability between 0 and 1.

Given the relevant prevalence range and relative importance, the AWA for all random tests can be calculated by varying the fixed probability between 0 and 1. The BRT is the random test with the largest AWA. One can then compare the AWA of a new test vs the AWA of the BRT (Supplementary Materials, Part 3).

### Step 5: AWA Analyses

Analyses consist of calculating the point and CI estimates for the AWA of the new test and comparator test and calculating the point and CI estimates for the difference in AWA between the new test and comparator test. Sensitivity analyses to varying relative importance are recommended. A plot of (i) the AWA and associated confidence bands for each test can be plotted vs. the relative importance and (ii) the difference in AWA between the new test and comparator with associated confidence bands vs the relative importance.

### Example: Rapid Diagnostics in Categorizing Acute Lung Infections

RADICAL represents a series of studies designed to discover and validate host gene expression signatures for acute respiratory illness. RADICAL-2 is a single-visit, multisite clinical validation study designed to evaluate the performance of a host gene expression–based diagnostic test (TEST) that categorizes acute respiratory illness (ARI) as being due to bacterial infection, viral infection, coinfection, or no infection. Clinical adjudication is utilized as the reference standard. The study collects blood, nasal swabs, throat swabs, and urine from up to 1200 patients who present to the emergency department with ARI symptoms [2, 3].

The TEST is comprised of 2 host gene expression signatures; one measures the response to bacterial infection and the other measures the response to viral infection. In each case, the signature is trained to discriminate the specified condition vs alternatives. For example, the bacterial signature discriminates bacterial infection from nonbacterial infection (eg, viral or noninfectious illness). The viral signature discriminates viral infection from nonviral infection (eg, bacterial or noninfectious illness). The test works by measuring expression levels of prespecified mRNA targets in a peripheral whole blood sample. The mRNA targets on the TEST include previously described signatures [2, 3]. After quantifying the levels of each mRNA in the signature, a logistic regression model converts the data into a probability. Incorporation of both the bacterial and viral host response signatures into the TEST enables the identification of bacterial infection (bacterial signature detected), viral infection (viral signature detected), coinfection (both signatures detected), or no infection (neither signature detected). Here, we focus exclusively on the accuracy of the bacterial test given the primary interest in antibiotic stewardship.

#### *Step 1: Determination of the Relevant Range of the Prevalence of Bacterial Etiology*

Recent studies suggest that the prevalence of bacterial etiology is approximately 25% among patients with ARI. To account for the potentially lower prevalence in children, a relevant range of prevalence was selected to be 10% to 30% [5, 6].

#### *Step 2: Determination of the Relative Importance of False Bacterial vs False Nonbacterial Test Results*

An online survey was conducted among 88 clinicians to evaluate the distribution of perceived relative importance of a false bacterial (false-positive) vs a false nonbacterial (false-negative) determination. The clinicians represented multiple specialties including emergency medicine, hospital medicine, infectious diseases, pediatrics, and pulmonary/critical care medicine. Participants were presented with a clinical vignette about a patient with ARI for whom a hypothetical test could result in either a false-positive or a false-negative bacterial diagnosis. They were then asked, "How many patients with a viral infection would you be willing to unnecessarily treat with antibiotics to avoid missing 1 patient with a bacterial infection?" The median result was 4, indicating a relative importance of 0.25 (interquartile range, 0.125–0.5). Thus, 0.25 was selected as the value for r given its robustness to extreme perspectives.

#### *Step 3: Select a Cutoff for Defining the Diagnostic*

In the RADICAL study, the probability of bacterial etiology is estimated. Given the relative importance of the errors, the

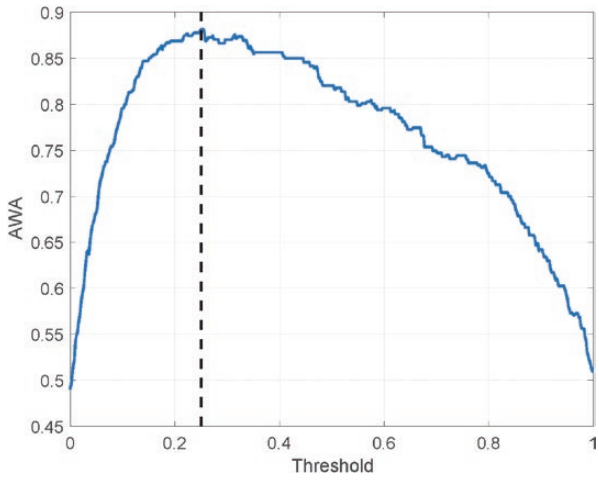


**Figure 1.** AWA as a function of test threshold for a test that diagnoses bacterial infection. In this manner, a threshold can be selected that maximizes the AWA of the test. Abbreviation: AWA, average weighted accuracy.

cutoff probability that maximizes AWA can be identified by plotting AWA as a function of the cutoff probability. Using data based on an earlier study of 385 samples, AWA is maximized at a cutoff probability of 0.25 (Figure 1).

***Step 4: Determination of the Hypotheses***

The AWA for the TEST can be compared with that of the BRT and a diagnostic alternative, procalcitonin (PCT), which is a marker frequently used to distinguish between a bacterial vs nonbacterial etiology [7].

***Step 5: Hypothetical AWA Analyses***

We conducted hypothetical analyses from the RADICAL study to illustrate AWA application. Calculations and R code are provided in the Supplement Materials, Parts 4 and 5.

The relevant prevalence range (10%–30%) and the relative importance of a false bacterial determination compared to a false nonbacterial determination (r = 0.25) were determined as above. Suppose that the sample size for RADICAL is 1200 and the observed prevalence is 25%; thus, there are 300 with a bacterial etiology and 900 without a bacterial etiology. Assuming that estimated sensitivity and specificity of the TEST for a bacterial etiology are 0.90 and 0.80, respectively, the estimated AWA for the TEST is 0.849 (0.828, 0.870).

The AWA of the BRT was calculated to be 0.51 (see Supplementary Figure 1 in Supplementary Materials, Part 4). The 95% CI for the difference between the AWA for the TEST and the BRT is 0.339 (0.318, 0.360), $P < .001$.

With respect to PCT as a diagnostic alternative, a threshold of 0.25 ng/mL generated an estimated sensitivity and specificity of $Se_{pct} = 66.9\%(60.6\% - 72.9\%)$ and $Sp_{pct} = 66.5\%(61.7\% - 71.1\%)$, respectively [7]. This results in an estimated AWA of 0.667 (0.629, 0.704) based on the sample size in the Self study [7]. The 95% CI for the difference between the AWA for the TEST and PCT is 0.182 (0.139, 0.225), $P < .001$.

The AWA for the TEST, the BRT, and PCT are presented as a function of the relative importance, r, in Figure 2A. The estimated difference (and 95% CI) in AWA between the TEST and the BRT (Figure 2B) and between the TEST and PCT (Figure 2C) are shown as a function of r. The TEST is superior to the BRT as long as the relative importance is greater than 2%. The TEST is consistently superior to PCT regardless of relative

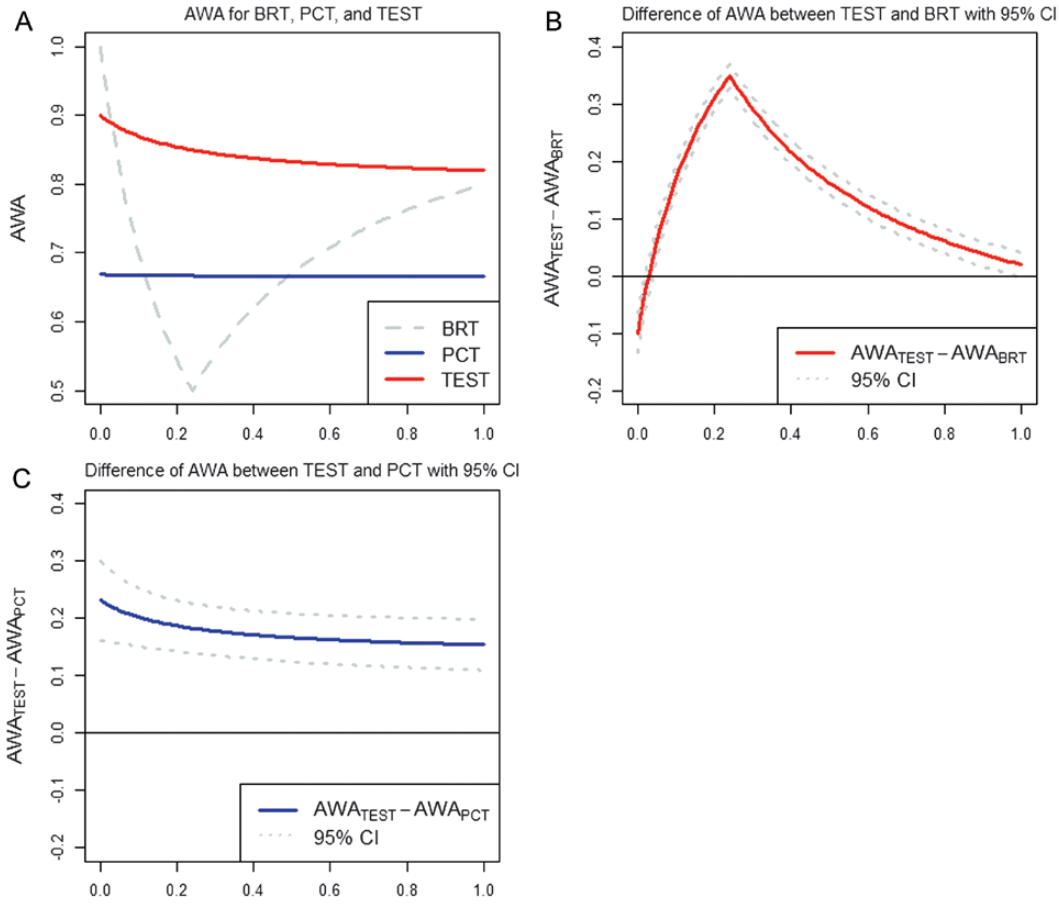


**Figure 2.** *A*, Estimated AWA for the TEST, BRT, and PCT as a function of the relative importance, r. Estimates with 95% confidence bands for the differences between the AWA for the TEST with the BRT (*B*) and PCT (*C*). Abbreviations: AWA, average weighted accuracy; BRT, best random test; CI, confidence interval; PCT, procalcitonin; TEST, host gene expression–based diagnostic test.

importance since the sensitivity and specificity of the TEST are higher than that of PCT.

Consider a second hypothetical example for which the performance of the test is less optimistic (Figure 3A). Here, we assume that the second TEST, TEST2, has a sensitivity of 90% but a specificity of 55%. Thus, it has higher sensitivity than PCT but lower specificity. In this case, TEST2 is superior to the BRT when the relative importance, r, is greater than 4% but less than 50% (Figure 3B). TEST2 is superior to PCT when the relative importance, r, is less than 50% (Figure 3C). Note a relative importance of 50% implies that clinicians would be willing to unnecessarily treat 2 patients with antibacterials to avoid missing 1 patient with a true bacterial infection.

## DISCUSSION

There has been a call for greater pragmatism in clinical trials, acknowledging that despite the rigor with which they are conducted, they suboptimally inform clinical practice [8]. Academics have been challenged to explore scholarly questions, not motivated by regulatory approval or profit, but from an interest in public health [9]. Greater pragmatism and leadership from academics in pursuing practical public health questions are needed in diagnostic studies as well.

The clinical impact of diagnostic application can be evaluated through evaluation of diagnostic yield. Diagnostic yield depends not only on the test's ability to discriminate disease from nondisease but also on the prevalence of disease and the relative importance of a false-positive vs false-negative result. AWA is a measure of diagnostic yield that incorporates these components, providing pragmatic evaluation of the diagnostics under investigation. AWA has been applied in the evaluation of rapid molecular diagnostics for infections caused by *Acinetobacter* spp. and *Pseudomonas aeruginosa* [10, 11]. We applied the methods to host response biomarkers for the differentiation of bacterial and viral infection.

When planning a pragmatic diagnostic study, the relevant range of disease prevalence, the relative importance of false-negative determination relative to a false-positive determination, and the hypothesis to be tested are evaluated and prespecified for transparency. A reasonable range of disease prevalence can often be derived from the medical literature.

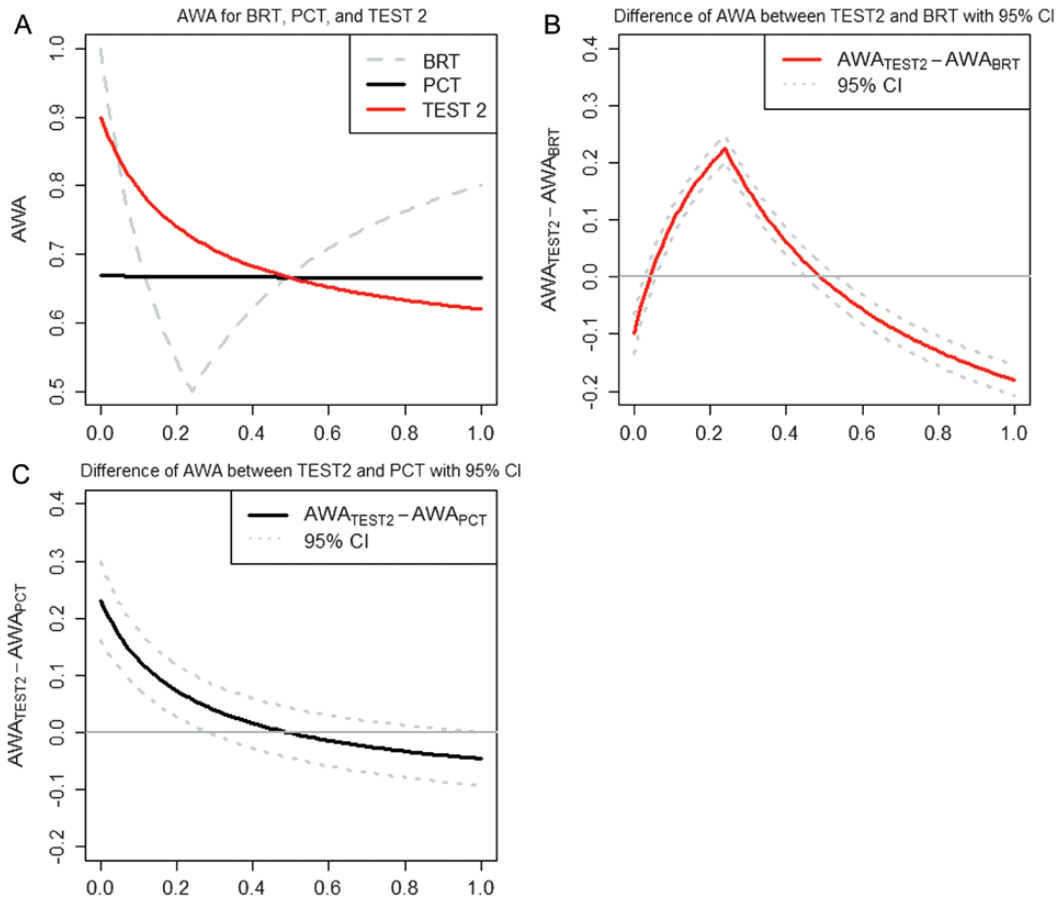


**Figure 3.** *A*, Estimated AWA for the second TEST (TEST2 with sensitivity of 90% and specificity of 55%), BRT, and PCT as a function of the relative importance, r. Estimates with 95% confidence bands for the differences between the AWA for TEST2 with the BRT (*B*) and PCT (*C*). Abbreviations: AWA, average weighted accuracy; BRT, best random test; CI, confidence interval; PCT, procalcitonin; TEST, host gene expression–based diagnostic test.

The relative importance of false negative vs false positive is evaluated based upon the clinical consequences of misdiagnoses via cost-effectiveness studies or surveys of experts. In the RADICAL study example, survey results suggested that failing to identify a bacterial etiology was a more important error than failing to identify nonbacterial etiologies (eg, viral or noninfectious). If there is a standard diagnostic to serve as a comparator, then a comparison of AWA of the investigational diagnostic can be compared to the standard. If no standard exists, then a comparison of AWA of the investigational diagnostic test can be compared to the BRT.

Typically, diagnostic studies are designed to test sensitivity and specificity. If the disease is rare, then the power for testing sensitivity may be limited, as the total number of disease-positive patients may be small. Researchers can enrich the population to recruit trial participants who are more likely to be disease-positive though at the expense of time and cost. AWA analyses may mitigate recruitment challenges since instead of segregating patients by disease status to estimate sensitivity and specificity separately, data from all patients are used to globally evaluate the test's performance.

When conducting AWA analyses, it is important to conduct supporting analyses including CI estimates of sensitivity/specificity (or positive/negative percent agreement) and positive and negative likelihood ratios, plots of the estimated positive/negative predicted values as a function of assumed disease prevalence with associated confidence bands, and a plot of the difference between AWA of the investigational test and the AWA of the reference with relevant confidence bands as a function of the relative importance.

AWA can be used as a tool to define diagnostic tests. Many diagnostic tests are based upon a biomarker measured on a continuum. A cutoff is selected whereby higher values are interpreted as either a positive or negative test, and lower values as the opposite. One challenge is selecting the optimal cutoff to balance sensitivity and specificity. If one defines the relevant range of prevalence and the relative importance, then the optimal cutoff in a pragmatic sense can be selected by identifying the cutoff that results in the largest AWA.

AWA provides a pragmatic evaluation of diagnostic utility. Additional research is needed to inform the relative importance of false-positive vs false-negative diagnostic results to help evaluate diagnostic utility and optimize clinical decision-making regarding diagnostic selection and application for different diseases.

## Supplementary Data

Supplementary materials are available at *Clinical Infectious Diseases* online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

## Notes

***Disclaimer.*** The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health (NIH).

***Financial support.*** This work was supported by the National Institute of Allergy and Infectious Diseases (NIAID) of the NIH (award UM1AI104681).

***Potential conflicts of interest.*** E. L. T. reports grants from NIH/Antibacterial Resistance Leadership Group (ARLG) during the conduct of the study; other from Predigen, outside the submitted work; In addition, E. L. T. has a patent Methods to Diagnose and Treat Acute Respiratory Infections (application PCT/US2016/040437) pending. Y. J. reports grants from NIAID/NIH, during the conduct of the study. H. G. reports grants from ARLG/NIH during the conduct of the study; In addition, H. G. has a patent Methods to Diagnose and Treat Acute Respiratory Infections (application PCT/US2016/040437) pending. C. W. W. reports grants from NIH/ARLG during the conduct of the study; other from Predigen, outside the submitted work; In addition, C. W. W. has a patent Methods to Diagnose and Treat Acute Respiratory Infections (application PCT/US2016/040437) pending. S. R. E. reports grants from NIAID/NIH during the conduct of the study; personal fees from Takeda/Millennium, personal fees from Pfizer, personal fees from Roche, personal fees from Novartis, personal fees from Achaogen, personal fees from Huntington's Study Group, personal fees from ACTTION, personal fees from Genentech, personal fees from Amgen, personal fees from GSK, personal fees from American Statistical Association, personal fees from FDA, personal fees from Osaka University, personal fees from Nationa Cerebral and Cardiovascular Center of Japan, personal fees from NIH, personal fees from Society for Clinical Trials, personal fees from Statistical Communications in Infectious Diseases (DeGruyter), personal fees from AstraZeneca, personal fees from Teva, personal fees from Austrian Breast & Colorectal Cancer Study Group (ABCSG)/Breast International Group (BIG) and the Alliance Foundation Trials (AFT) , personal fees from Zeiss, personal fees from Dexcom, personal fees from American Society for Microbiology, personal fees from Taylor and Francis, personal fees from Claret Medical, personal fees from Vir, personal fees from Arrevus, personal fees from Five Prime, personal fees from Shire, personal fees from Alexion, personal fees from Gilead, personal fees from Spark, personal fees from Clinical Trials Transformation Initiative, personal fees from Nuvelution, personal fees from Tracon, personal fees from Deming Conference, personal fees from Antimicrobial Resistance and Stewardship Conference, personal fees from World Antimicrobial Congress, personal fees from WAVE, personal fees from Advantagene, personal fees from Braeburn, personal fees from Cardinal Health, personal fees from Lipocine, personal fees from Microbiotix, personal fees from Stryker, outside the outside the submitted work. All other authors report no potential conflicts. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

## References

1. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open **2016**; 6. doi:10.1136/bmjopen-2016-012799.
2. Zaas AK, Burke T, Chen M, et al. A host-based RT-PCR gene expression signature to identify acute respiratory viral infection. Sci Transl Med **2013**; 5:203ra126.
3. Tsalik EL, Henao R, Nichols M, et al. Host gene expression classifiers diagnose acute respiratory illness etiology. Sci Transl Med **2016**; 8:322ra11.
4. Evans SR, Pennello G, Pantoja-Galicia N, et al; Antibacterial Resistance Leadership Group. Benefit-risk evaluation for diagnostics: a framework (BED-FRAME). Clin Infect Dis **2016**; 63:812–7.
5. Jain S, Self WH, Wunderink RG, et al; CDC EPIC Study Team. Community-acquired pneumonia requiring hospitalization among U.S. adults. N Engl J Med **2015**; 373:415–27.
6. Oved K, Cohen A, Boico O, et al. A novel host-proteome signature for distinguishing between acute bacterial and viral infections. PLoS One **2015**; 10:e0120012.

7. Self WH, Balk RA, Grijalva CG, et al. Procalcitonin as a marker of etiology in adults hospitalized with community-acquired pneumonia. Clin Infect Dis **2017**; 65:183–90.
8. Couzin-Frankel J. Medical research. Clinical trials get practical. Science **2015**; 348:382.
9. DeMets DL, Califf RM. A historical perspective on clinical trials innovation and leadership: where have the academics gone? JAMA **2011**; 305:713–4.
10. Evans SR, Tran TTT, Hujer AM, et al; Antibacterial Resistance Leadership Group. Rapid molecular diagnostics to inform empiric use of ceftazidime/avibactam and ceftolozane/tazobactam against *Pseudomonas aeruginosa*: PRIMERS IV. Clin Infect Dis **2018**.
11. Pennello G, Pantoja-Galicia N, Evans SR. Benefit-risk comparisons of diagnostic tests based on diagnostic yield and expected utility. J of Biopharm Stats **2016**; 26:1083–97.