Clinical Infectious Diseases

VIEWPOINTS

Clinically Adjudicated Reference Standards for Evaluation of Infectious Diseases Diagnostics

Robin Patel,[1,2,©] Ephraim L. Tsalik,[3,4,5] Scott Evans,[6] Vance G. Fowler,[4,7] and Sarah B. Doernberg,[8] for The Antibacterial Resistance Leadership Group

[1]Division of Clinical Microbiology, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, USA; [2]Division of Public Health, Infectious Diseases and Occupational Medicine, Department of Medicine, Mayo Clinic, Rochester, Minnesota, USA; [3]Emergency Medicine Service, Durham VA Health Care System, Durham, North Carolina, USA; [4]Division of Infectious Diseases, Department of Medicine, Duke University School of Medicine, Durham, North Carolina, USA; [5]Danaher Diagnostics, Washington, District of Columbia, USA; [6]Biostatistics Center and the Department of Biostatistics and Bioinformatics, George Washington Milken Institute School of Public Health, George Washington University, Washington, District of Columbia, USA; [7]Duke Clinical Research Institute, Durham, North Carolina, USA; and [8]Division of Infectious Diseases, Department of Medicine, University of California, San Francisco, California, USA

Lack of a gold standard can present a challenge for evaluation of diagnostic test accuracy of some infectious diseases tests, particularly when the test's accuracy potentially exceeds that of its predecessors. This approach may measure agreement with an imperfect reference, rather than correctness, because the right answer is unknown. Solutions consist of multitest comparators, including those that involve a test under evaluation if multiple new tests are being evaluated together, using latent class modeling, and clinically adjudicated reference standards. Clinically adjudicated reference standards may be considered as comparator methods when no predefined test or composite of tests is sufficiently accurate; they emulate clinical practice in that multiple data pieces are clinically assessed together.

Keywords. clinically adjusted reference standards.

CLINICAL VIGNETTE

An adult with cellulitis is enrolled in a research study of a new rapid diagnostic test intended to detect microorganisms in blood. Whereas blood cultures are negative, the new diagnostic detects *Streptococcus pyogenes*. Is this a false-positive or true-positive result?

An ideal diagnostic test used for infectious diseases establishes the existence of an infection, its etiology, and potential treatment options accurately and in a clinically actionable timeframe. One such example is the GeneXpert TB test, which simultaneously identifies the presence of *Mycobacterium tuberculosis* and select genotypic susceptibility markers directly from sputum samples. In recent years, accelerated by the global response to the coronavirus disease 2019 pandemic, new and promising diagnostic assays for infectious diseases have proliferated. Notable areas of advancement include molecular diagnostics that detect a growing number of pathogens, including highly multiplexed syndromic panels; deep sequencing approaches based on targeted or shotgun metagenomic sequencing; rapid antimicrobial susceptibility testing; metabolomics-based assays; and host response analysis, including multianalyte protein or messenger RNA assessment. These new tests may facilitate more rapid transition to antimicrobial therapy targeted to the causative pathogen(s), provide information to support personalized durations of treatment, and enable antibiotic avoidance, when appropriate. Although many new tests offer the potential for improved accuracy compared with currently available methods, such as standard culture with phenotypic susceptibility testing, assessing accuracy can be challenging. This article reviews issues that arise in this situation and discusses options for how clinical researchers designing trials, regulatory bodies considering new tests for approval, and the scientific community weighing diagnostic options may evaluate new tests.

The lack of a gold standard presents a common challenge in the evaluation of diagnostic test accuracy, particularly when a new test's accuracy exceeds that of its predecessors. A newer, potentially more accurate test may appear less sensitive or specific (or both) compared with the ultimate (often unmeasurable) truth when evaluated against an error-prone comparator. (The terms "sensitivity" and "specificity" are usually reserved for situations in which the reference standard is perfect/correct/the truth; when the reference standard is imperfect, what is measured is agreement, and the terms positive and negative percent agreement are used.) For example, in the current case vignette, if blood culture were considered the reference standard, detection of *S pyogenes* would be classified as a false-positive result, despite a clinical diagnosis suggesting that this organism may be a true pathogen. Assuming *S pyogenes* caused the cellulitis, the use of blood cultures as the reference standard would make the new diagnostic appear to have a lower specificity by classifying the result as false positive rather than true positive. In addition,

Received 31 August 2022; editorial decision 13 October 2022; published online 20 October 2022

Correspondence: R. Patel, Division of Clinical Microbiology, Department of Laboratory Medicine and Pathology, Mayo Clinic, 200 First St SW, Rochester, MN 55905, USA (patel.robin@mayo.edu).

Clinical Infectious Diseases® 2023;76(5):938–43

https://doi.org/10.1093/cid/ciac829

interpretation of blood cultures may be challenged by detection of contaminants. In the case vignette, if blood cultures but not the new diagnostic test had detected *Staphylococcus epidermidis* (a blood culture contaminant in this example), and blood cultures were considered the reference standard, the novel test would appear less sensitive, again, because of misclassification based on the reference standard [1].

When there is a suitable extant test to use as the reference standard, some have argued for use of the term, "interchangeability," rather than accuracy; a test that is interchangeable with a comparator can replace the comparator without changing diagnostic accuracy or utility, even if imperfect [2, 3]. Interchangeability is assessed based on simple agreement between the test and comparator. For many new diagnostics, however, there is not a suitable extant test to use as an interchangeable comparator.

For diagnostics aiming to detect organisms directly from blood, using the term "bloodstream infection" may be connotatively challenging. In most cases, a microorganism (especially a bacterium or fungus) detected in blood originates from another site of infection (eg, pneumonia, soft-tissue infection, urinary tract infection, skin/skin structure [as in the vignette]). Therefore, what is being identified is not solely a "bloodstream infection" but rather an organism causing infection elsewhere that has translocated into the blood. Newer, more sensitive tests (eg, those that detect microbial components such as nucleic acids or proteins) may detect organisms originating from nonbloodstream sources at levels lower than would be detected by blood culture. Because of this, it might be necessary to evaluate "blood-based" diagnostics for specific infectious syndromes or microorganisms in comparison to more conventional microbial detection at nonblood sites of infection. For example, in native joint septic arthritis, a novel diagnostic may detect a microorganism in blood when blood cultures are negative; isolation of the same microorganism from synovial fluid might confirm the accuracy of the novel diagnostic test in the absence of confirmatory blood cultures. Evaluating performance of blood-based microorganism detection tests in syndromic-focused diagnoses may be limited by lack of comprehensive (eg, pneumonia) or suitable (eg, blood culture-negative endocarditis) testing, or by difficulty in differentiating pathogens from commensals (eg, rhinovirus in a nasopharyngeal swab). These are gaps that can potentially be closed by new diagnostics.

The absence of appropriate standards may affect novel tests at multiple stages, beginning with test development. Tests using machine learning algorithms, such as multianalyte host biomarker tests, may be prone to errors when trained on incorrectly phenotyped samples. This may lead test developers to abandon potentially promising tests because diagnostic utility is obscured by a poor reference or to develop tests that perform well during discovery and training but fail to validate. Once a test has been developed, the regulatory approval process requires a comparator by which to assess performance of the new test. Such comparators should provide interpretable information to clinical and regulatory communities while accounting for inherent problems associated with a poor reference. Finally, real-world evaluations performed by independent researchers may also use reference standards inconsistently, resulting in varying assessments of diagnostic utility.

Alternative Reference Standards

Multitest Comparator

Faced with imperfect current standard tests, there are several alternative approaches to a single comparator test (Table 1). One strategy combines results of other tests to serve as the comparator method, directed at measuring agreement. This approach can be used in multiple circumstances, including true absence of a gold standard, such as with acute respiratory illness [8]. This is illustrated by a study by Self et al, in which a multitest algorithmic approach was developed by combining culture, nucleic acid amplification tests, serological tests, procalcitonin, and white blood cell count to categorize patients with upper respiratory tract infection as having bacterial infection, viral infection, or neither [4]. By incorporating multiple tests, this approach can mitigate, albeit probably not completely overcome, limitations of any single test.

Comparator Comprising Combination of Tests, Including New Tests

In a scenario in which a comparator recognized by regulatory bodies is lacking but several new tests have strong biological plausibility to identify the microorganism of interest, combining the new tests to create a comparator directed at measuring agreement can facilitate estimates of diagnostic performance. For example, the MASTER-GC study validated three investigational nucleic acid amplification tests for detection of extragenital *Neisseria gonorrhoeae* and *Chlamydia trachomatis* infection, despite the insensitivity of traditional microbial culture and the absence of a Food and Drug Administration–cleared test to serve as a predicate or comparator [5]. The

Table 1. Reference Standards for Infectious Diseases Diagnostics in the Absence of a Suitable Single Comparator Test[a]

Possible Reference Standard	Example
Multitest comparator	Acute respiratory illness [4]
Comparator comprised of combination of tests, including new tests	Extragenital gonorrhea and *Chlamydia trachomatis* [5]
Latent class modeling	Latent tuberculosis infection [6]
Clinically adjudicated reference standard	Sepsis diagnostic distinguishing systemic inflammatory response (SIRS) without infection from sepsis (or SIRS with infection) [7]

[a]Standardized reference material may be used for evaluation of some quantitative infectious diseases diagnostics.

Table 2. Types of Technology and Possible Reference Standards

New Technology	Current Comparators	Limitations to Current Comparator[a]	Proposed Comparators
Pathogen detection in blood	Standard culture; nucleic acid amplification test; sequencing; serologic test	Incomplete sensitivity (eg, because of antibiotic use); Incomplete specificity (eg, contaminants, detection of latent infections); lack of specificity of some serologic tests; poorly defined cutoff values of some serologic tests and accuracy of some nucleic acid amplification tests	Combination of pathogen detection tests; clinical adjudication
Pathogen detection directly from body fluids/tissues, respiratory specimens, swabs	Standard culture; nucleic acid amplification test; sequencing; serologic test	Deficient sensitivity of culture (eg, because of antibiotic use); inadequate specificity (eg, contaminants, commensals); lack of specificity of some serologic tests; poorly defined cutoff values of some serologic tests and accuracy of some nucleic acid amplification tests	Combination of pathogen detection tests; clinical adjudication
Pathogen identification from culture growth	Standard culture	None	Likely does not need adjudication
Genotypic antimicrobial resistance from cultured isolate	Standard phenotypic susceptibility test	None	Likely does not need adjudication[b]
Genotypic antimicrobial resistance direct from specimen	Standard phenotypic susceptibility testing; nucleic acid amplification test; sequencing	Poorly defined accuracy of some nucleic acid amplification tests	Likely does not need adjudication
Host response biomarkers	None	NA	Clinical adjudication; latent class modeling

Abbreviation: NA, not applicable.

[a]If the comparators are not gold standards (ie, correct), sensitivity and specificity should not be reported, but positive and negative percent agreement measured.

[b]Handling situations in which genotypic resistance is detected but an isolate is phenotypically susceptible provides specific challenges and may depend on the specific research question. If the goal is successful detection of resistance genes, regardless of phenotype, application of a combination of tests using other mechanisms for gene identification is one possibility. If the goal is accurate prediction of phenotypic susceptibility, options include clinical adjudication (assessing whether clinical response to the antibiotic of interest was favorable), consideration of the results of future culture and phenotypic susceptibility testing, using alternative methods of phenotypic susceptibility testing, or relying solely on phenotypic susceptibility testing results as the source of truth.

investigators evaluated results of each experimental test against a composite of the other 2 using an additional test as a tiebreaker; each test had distinct molecular targets and used distinct methods and instruments. This combination of tests served as the "anatomic site infected status." Because the assays under consideration had track records for diagnosis of infection at other anatomic sites (eg, the genitourinary tract), there was strong biological plausibility that the tests would perform well at rectal and pharyngeal locations. This allowed use of each test in the composite for the other tests despite some risk of correlated testing errors, mitigated by each test using different methods and molecular targets.

Latent Class Modeling

Latent class modeling [9] is a statistical method that uses observed test results to estimate disease prevalence and diagnostic performance when no gold standard exists. Latent class models view the true disease status as an unmeasured (latent) variable reflecting the underlying disease status (eg, bacterial vs other etiology). A statistical model is developed relating the investigational test and the reference test considered to be imperfect to the latent disease status. The model allows estimation of sensitivity and specificity of the investigational test. Validity depends on the model's correctness (which is unknown) and the assumption of the independence of observations conditional on the true disease status [10]. Stout et al used latent class modeling to estimate diagnostic performance of tests for latent tuberculosis infection [6]. This approach functioned well for the example of latent tuberculosis infection, in which there is no gold standard diagnostic test and there are clear downsides to using development of active tuberculosis as the comparator because this is an uncommon event that does not occur in all testing positive for latent tuberculosis infection and depends on uptake of latent tuberculosis infection treatment in the population. Similarly, Peel et al used latent class modeling to estimate the sensitivity and specificity of a novel periprosthetic joint infection diagnostic approach involving inoculation of tissues into blood culture bottles for diagnosis of periprosthetic joint infection [11]. This approach may have limitations, particularly when conditional dependence between tests is misspecified, resulting in biased estimation of sensitivity and specificity [12].

Clinically Adjudicated Reference Standard

When no predefined test or composite of tests is sufficiently accurate, expert clinical panel adjudication can be considered for use as the comparator method directed at measuring agreement [13, 14]. Examples of tests for which this might be useful include pathogen detection in blood or body fluids/tissues, and host response discrimination (Table 2). Clinically adjudicated reference standards, sometimes referred to as "consensus diagnosis" or "adjudicated diagnosis," have been extensively used outside of infectious diseases (eg, psychiatric, cardiovascular, respiratory disorders), and are exemplified by studies of

troponin for diagnosis of cardiac ischemia [15]. For infectious diseases, clinically adjudicated reference standards have been featured in the evaluation of tests for sepsis [7, 16], vaccines [17], bloodstream infection [18, 19] (https://clinicaltrials.gov/ct2/show/NCT03138733), periprosthetic joint infection [20], and bacterial/viral discrimination [21, 22].

A clinically adjudicated reference standard emulates clinical practice in that multiple data pieces are clinically assessed together, essentially extending the concept of using a combination of tests for the comparator. Structured methods to adjudicate diagnoses that address potential bias and account for performance limitations of current microbiological testing are applied. Adjudication categories can be clean (YES [confirmed, present]/NO [ruled out, absent]), or may include "indeterminate" or lower confidence categories, such as "possible," although calculation of traditional agreement measures such as positive and negative percent agreement requires classifying intermediate categories as being positive or negative. Sensitivity analyses can help to examine how reclassifying "indeterminate" or "possible" categories or removing these cases impacts estimates of test performance.

Information provided to adjudicators may include the electronic medical record (if possible), additional testing for the causative etiology, and a case report form, but should not initially include results of the test being evaluated, to avoid incorporation bias. Certain additional tests, such as serologic testing or radiographs, may be required for adjudicators to have available as part of the protocol. When assigning a diagnostic label, adjudicators should consider strength of the microbiologic evidence based on test characteristics in a pathogen- and patient-specific manner. For example, a positive urinary pneumococcal antigen test in an adult is good evidence of invasive pneumococcal infection [23], but may be more likely to reflect colonization in a child where carriage rates are as high as 59% [24]. When etiologic testing is negative, but a case is clinically consistent with a particular pathogen type, it may be adjudicated as "suspected." For example, a subject with a fever and unilateral leg erythema might be designated as having suspected bacterial cellulitis. Assessing long-term outcomes can mitigate limitations of imperfect reference standards. For example, in assessing breast imaging for cancer diagnosis, results of biopsy combined with long-term follow-up have been used as a reference standard for the presence or absence of cancer [25–27]. For infectious diseases diagnostics, short- and long-term follow-up may help to bolster clinical assessment, such as observing resolution of suspected bacterial cellulitis with antibacterial treatment. However, caution is needed with this approach because many infections, including bacterial infections, resolve without specific treatment, and conditions that mimic infections, such as dermatitis, may also resolve with time. Other infections, such as fungal pneumonias or brain abscesses, may necessitate treatment before progression to a point at which "definitive" culture-based diagnosis is possible.

Although adjudication panel sizes have varied, 3 or more expert adjudicators typically independently (ie, in a blinded way) assess each case [15]. Adjudicators should be clinicians experienced in managing patients with the disease under study. In addition to infectious diseases, relevant areas of expertise may include hospital medicine, emergency medicine, family practice, surgery, pulmonary/critical care medicine, or others, depending on the test and disease under study. For pediatric subjects, training or experience in pediatrics should be included. If each adjudicator does not review every case, cases may be randomly assigned, while stratifying for adjudicator characteristics, such as experience and area of expertise (eg, fellow vs attending, infectious diseases vs noninfectious diseases specialty) because these characteristics may bring different perspectives to the classification. Discordance may be reconciled either by further (tiebreaker) experts or using a panel approach.

As an alternative to independent review, a panel of multiple experts, again typically 3 or more, may review cases together to provide a final determination. However, this approach may be impacted by "groupthink," or dominant group members' opinions [28]. As mentioned previously, a panel can also be used to resolve discordance.

Despite applying a rigorous methodology to clinical adjudications, there will be challenging cases that limit this approach, such as cases with lower diagnostic confidence (eg, suspected bacterial infection, possible aspergillosis), or diseases that are simply not well characterized, such as emerging infections. Another limitation is that novel findings may be made by new infectious diseases diagnostic tests (eg, detecting bunyavirus in cerebrospinal fluid or *Metamycoplasma salivarium* in periprosthetic joint infection) [29, 30]. It would likely be impossible for adjudicators to anticipate such findings absent results of the new tests themselves or additional directed confirmatory tests, though the adjudicators may well have been able to classify a case of suspected viral encephalitis or bacterial periprosthetic joint infection. These findings may indeed represent unanticipated positive results or may represent false positives because one might detect using a broad shotgun metagenomic sequencing test. One option is to identify such cases as "possible" infections and then readjudicate them using results of the new test or to confirm findings with further testing using established or investigational approaches. Ultimately, performance of truly novel tests should be evaluated by studies performed in clinical practice to assess clinical utility, recognizing that design and execution of these types of studies can pose distinct methodological challenges [31, 32].

Given the challenges described here, there should be discretion in interpreting results reported in studies of novel tests that do not have a gold standard comparator. When reporting results of these studies, there are steps that can be taken to enable interpretation. For example, the primary analysis population could be restricted to those with the highest confidence adjudication or reference.

However, reporting results restricted to high confidence reference cases limits generalizability. This is also a misleading approach as more challenging situations are ignored, limiting practicability, and, arguably, diagnostic tests play their most important role when there is significant uncertainty.

In the end, all diagnostics, and especially novel diagnostics, need to be evaluated for clinical utility to determine whether they change patient management, outcomes, cost of care, or yield societal benefits. This can be evaluated using randomized controlled trials that use the novel diagnostic as an intervention, noting that such studies are not easy to execute because of financial and logistical challenges [31, 32]. Although this type of information has not been required for regulatory approval, it may inform reimbursement strategies alongside how and where such novel tests should be adopted in clinical practice (eg, informing development of diagnostic algorithms and clinical guidelines).

CONCLUSION

Standardized approaches for evaluation of novel infectious diseases diagnostics are needed in scenarios in which no gold standard exists. Clinically adjudicated reference standards, albeit imperfect, are a solution to this challenge in some scenarios.

Notes

Financial support. This study was supported in part by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award number UM1AI104681. The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

Potential conflicts of interest. V. G. F. reports personal fees from Novartis, Novadigm, Durata, Debiopharm, Genentech, Achaogen, Affinium, Medicines Co., Cerexa, Tetraphase, Trius, MedImmune, Bayer, Theravance, Basilea, Affinergy, Janssen, xBiotech, Contrafect, Regeneron, Basilea, Destiny, Amphliphi Biosciences, Integrated Biotherapeutics, C3J, Armata, Valanbio, Akagera, Aridis, Roche, and Pfizer; grants from the National Institutes of Health (NIH), MedImmune, Cerexa/Forest/Actavis/Allergan, Pfizer, Advanced Liquid Logics, Theravance, Novartis, Cubist/Merck, Medical Biosurfaces, Locus, Affinergy, Contrafect, Karius, Genentech, Regeneron, Basilea, Janssen, from Green Cross, Cubist, Cerexa, Durata, Theravance, and Debiopharm. Royalties from UpToDate. V. G. F. also reports support for attending meetings and/or travel from Contrafect to present phase 2 data at 2019 ECCMID; patents planned, issued, or pending (Sepsis Diagnostics); a role as Associate Editor, *Clinical Infectious Diseases* (2017–2022); and stock or stock options from ArcBio and Valanbio. R. P. reports grants from ContraFect, TenNor Therapeutics Limited, and BioFire. R. P. is a consultant to Curetis, PathoQuest, Selux Diagnostics, 1928 Diagnostics, PhAST, Torus Biosystems, Day Zero Diagnostics, Mammoth Biosciences, and Qvella; monies are paid to Mayo Clinic. Mayo Clinic and R. P. have a relationship with Pathogenomix. R. P. has research supported by Adaptive Phage Therapeutics. Mayo Clinic has a royalty-bearing know-how agreement and equity in Adaptive Phage Therapeutics. R. P. is also a consultant to Netflix, Abbott Laboratories, and CARB-X. In addition, R. P. has a patent on *Bordetella pertussis/parapertussis* polymerase chain reaction issued, a patent on a device/method for sonication with royalties paid by Samsung to Mayo Clinic, and a patent on an antibiofilm substance issued. R. P. receives honoraria from the NBME, Up-to-Date, and the Infectious Diseases Board Review Course. E. L. T. reports grants or contracts from NIAID, DTRA, Sanofi, NIGMS, DARPA, Karius, Department of Veterans Affairs, and Department of Defense. E. L. T. has patents issued or pending for methods for characterizing infections and methods for developing tests for the same; host based molecular signatures of human infection with severe acute respiratory syndrome coronavirus 2; transcriptional signature for candidemia, gene expression signature useful to predict or diagnose sepsis and methods of using the same; methods to diagnose and treat acute respiratory infections; and biomarkers for the molecular classification of bacterial infection; E. L. T. is an employee of Danaher Corp and previously received consulting fees from Biomeme and has stock or stock options for Biomeme and Danaher. S. R. E. reports grant support from NIH; a contract with DeGruyter; royalties from Taylor & Francis; honoraria from the Analgesic, Anesthetic, and Addiction Clinical Trial Translations, Innovations, Opportunities, and Networks (ACTTION); meeting support from the Food and Drug Administration, Deming Conference on Applied Statistics, Clinical Trial Transformation Initiative, Council for International Organizations of Medical Sciences, International Chinese Statistical Association Applied Statistics Symposium, Antimicrobial Resistance and Stewardship Conference; unpaid board membership from the American Statistical Association, Society for Clinical Trials, Frontier Science Foundation; consulting fees from Genentech, AstraZeneca, Takeda, Microbiotix, Johnson & Johnson, Endologix, ChemoCentryx, Becton Dickenson, Atricure, Roivant, Neovasc, Nobel Pharma, Horizon, International Drug Development Institute, and SVB Leerink; data and safety monitoring board or advisory board service for NIH, Breast International Group, University of Pennsylvania, Washington University, Duke University, Roche, Pfizer, Takeda, Akouos, Apellis, Teva, Vir, DayOneBio, Alexion, Tracon, Rakuten, Abbvie, GSK, Eli Lilly, Nuvelution, Clover, FHI Clinical, Lung Biotech, SAB Biopharm, and Advantagene/Candel. S. B. D. reports research funding related to coronavirus disease 2019 (COVID-19) from Gilead, Regeneron, Shinogi, and the NIH and non–COVID-19-related consulting and research funding from Genentech, Basilea, Shinogi, J + J, and the NIH; support for travel from IDSA, and patents issued (WO2006116688A3). 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. Glasziou P, Irwig L, Deeks JJ. When should a new test become the current reference standard? Ann Intern Med **2008**; 149:816–22.
2. Obuchowski NA, Subhas N, Schoenhagen P. Testing for interchangeability of imaging tests. Acad Radiol **2014**; 21:1483–9.
3. Barnhart HX, Kosinski AS, Haber MJ. Assessing individual agreement. J Biopharm Stat **2007**; 17:697–719.
4. Self WH, Rosen J, Sharp SC, et al. Diagnostic accuracy of FebriDx: a rapid test to detect immune responses to viral and bacterial upper respiratory infections. J Clin Med **2017**; 6:94.
5. Doernberg SB, Komarow L, Tran TTT, et al. Simultaneous evaluation of diagnostic assays for pharyngeal and rectal *Neisseria gonorrhoeae* and *Chlamydia trachomatis* using a master protocol. Clin Infect Dis **2020**; 71:2314–22.
6. Stout JE, Wu Y, Ho CS, et al. Evaluating latent tuberculosis infection diagnostics using latent class analysis. Thorax **2018**; 73:1062–70.
7. Miller RR III, Lopansri BK, Burke JP, et al. Validation of a host response assay, SeptiCyte LAB, for discriminating sepsis from systemic inflammatory response syndrome in the ICU. Am J Respir Crit Care Med **2018**; 198:903–13.
8. Fleming-Dutra KE, Hersh AL, Shapiro DJ, et al. Prevalence of inappropriate antibiotic prescriptions among US ambulatory care visits, 2010–2011. JAMA **2016**; 315:1864–73.
9. Collins J, Huynh M. Estimation of diagnostic test accuracy without full verification: a review of latent class methods. Stat Med **2014**; 33:4141–69.
10. van Smeden M, Naaktgeboren CA, Reitsma JB, Moons KGM, de Groot JAH. Latent class models in diagnostic studies when there is no reference standard-A systematic review. Am J Epidemiol **2014**; 179:423–31.
11. Peel TN, Dylla BL, Hughes JG, et al. Improved diagnosis of prosthetic joint infection by culturing periprosthetic tissue specimens in blood culture bottles. mBio **2016**; 7:e01776–15.
12. Albert PS, Dodd LE. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. Biometrics **2004**; 60:427–35.

13. Reitsma JB, Rutjes AW, 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.
14. Jones J, Hunter D. Consensus methods for medical and health services research. BMJ **1995**; 311:376–80.
15. Bertens LC, Broekhuizen BD, Naaktgeboren CA, et al. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. PLoS Med **2013**; 10:e1001531.
16. Gámez-Díaz LY, Enriquez LE, Matute JD, et al. Diagnostic accuracy of HMGB-1, sTREM-1, and CD64 as markers of sepsis in patients recently admitted to the emergency department. Acad Emerg Med **2011**; 18:807–15.
17. Fowler VG, Allen KB, Moreira ED, et al. Effect of an investigational vaccine for preventing *Staphylococcus aureus* infections after cardiothoracic surgery: a randomized trial. JAMA **2013**; 309:1368–78.
18. Fowler VG Jr, Boucher HW, Corey GR, et al. Daptomycin versus standard therapy for bacteremia and endocarditis caused by *Staphylococcus aureus*. N Engl J Med **2006**; 355:653–65.
19. Holland TL, Raad I, Boucher HW, et al. Effect of algorithm-based therapy vs usual care on clinical success and serious adverse events in patients with staphylococcal bacteremia: a randomized clinical trial. JAMA **2018**; 320:1249–58.
20. Ivy MI, Sharma K, Greenwood-Quaintance KE, et al. Synovial fluid α defensin has comparable accuracy to synovial fluid white blood cell count and polymorphonuclear percentage for periprosthetic joint infection diagnosis. Bone Joint J **2021**; 103-b:1119–26.
21. Ko ER, Henao R, Frankey K, et al. Prospective validation of a rapid host gene expression test to discriminate bacterial from viral respiratory infection. JAMA Netw Open **2022**; 5:e227299.
22. van Houten CB, Naaktgeboren CA, Ashkenazi-Hoffnung L, et al. Expert panel diagnosis demonstrated high reproducibility as reference standard in infectious diseases. J Clin Epidemiol **2019**; 112:20–7.
23. Sinclair A, Xie X, Teltscher M, Dendukuri N. Systematic review and meta-analysis of a urine-based pneumococcal antigen test for diagnosis of community-acquired pneumonia caused by *Streptococcus pneumoniae*. J Clin Microbiol **2013**; 51: 2303–10.
24. Vancikova Z, Trojanek M, Zemlickova H, et al. Pneumococcal urinary antigen positivity in healthy colonized children: is it age dependent? Wien Klin Wochenschr **2013**; 125:495–500.
25. Pisano ED. Digital compared with screen-film mammography: measures of diagnostic accuracy among women screened in the Ontario breast screening program–evidence that direct radiography is superior to computed radiography for cancer detection. Radiology **2016**; 278:311–2.
26. Pisano ED, Hendrick RE, Yaffe MJ, et al. Diagnostic accuracy of digital versus film mammography: exploratory analysis of selected population subgroups in DMIST. Radiology **2008**; 246:376–83.
27. Lewin JM, Hendrick RE, D'Orsi CJ, et al. Comparison of full-field digital mammography with screen-film mammography for cancer detection: results of 4,945 paired examinations. Radiology **2001**; 218:873–80.
28. Bankier AA, Levine D, Halpern EF, Kressel HY. Consensus interpretation in imaging research: is there a better way? Radiology **2010**; 257:14–7.
29. Rodino KG, Toledano M, Norgan AP, et al. Retrospective review of clinical utility of shotgun metagenomic sequencing testing of cerebrospinal fluid from a U. S. Tertiary care medical center. J Clin Microbiol **2020**; 58:e01729–20.
30. Thoendel M, Jeraldo P, Greenwood-Quaintance KE, et al. A novel prosthetic joint infection pathogen, *Mycoplasma salivarium*, identified by metagenomic shotgun sequencing. Clin Infect Dis **2017**; 65:332–5.
31. Banerjee R, Komarow L, Virk A, et al. Randomized trial evaluating clinical impact of RAPid IDentification and susceptibility testing for gram-negative bacteremia: rAPIDS-GN. Clin Infect Dis **2021**; 73:e39–46.
32. Banerjee R, Teng CB, Cunningham SA, et al. Randomized trial of rapid multiplex polymerase chain reaction-based blood culture identification and susceptibility testing. Clin Infect Dis **2015**; 61:1071–80.