ELSEVIER

Journal of Clinical Epidemiology 126 (2020) 172−176

Journal of Clinical Epidemiology

COVID-19 ARTICLES

Testing COVID-19 tests faces methodological challenges

Patrick M. Bossuyt*

Department of Epidemiology & Data Science, Amsterdam Public Health, Amsterdam University Medical Centers, Room J1B-214, PO Box 22660, 1100 DD Amsterdam, The Netherlands

Accepted 30 June 2020; Published online 3 July 2020

Abstract

In battling the COVID-19 pandemic, testing is essential. The detection of viral RNA allows the identification of infected persons, whereas the detection of antibodies may reveal a response to a previous infection. Tests for coronavirus should be rigorously evaluated in terms of their analytical and clinical performance. This poses not only logistic challenges, but also methodological ones. Some of these are generic for the diagnostic accuracy paradigm, whereas others are more specific for tests for viruses. Problematic for evaluations of the clinical performance of tests for viral RNA is the absence of an independent reference standard. Many studies lack rigor in terms of the recruitment of study participants. Study reports are often insufficiently informative, which makes it difficult to assess the applicability of study findings. Attempts to summarize the performance of these tests in terms of a single estimate of the clinical sensitivity fail to do justice to the identifiable sources of the large heterogeneity in mechanisms for generating false negative results.

Keywords: COVID-19; SARS-CoV-2; Clinical performance; Diagnostic accuracy; Analytical performance; Medical tests

“We have a simple message to all countries: test, test, test,” said WHO Director General Tedros Adhanom Ghebreyesus at a news conference in Geneva, March 2020. “All countries should be able to test all suspected cases, they cannot fight this pandemic blindfolded.” The WHO director general called on all countries to ramp up their testing programs, to battle the corona pandemic.

But testing would be useless, maybe even dangerous, if the tests that we rely on are flawed. Infected individuals with a false negative result may continue to infect others, for example, posing a genuine health risk to their environment. Like any other test or intervention in health care, tests for COVID-19 should be rigorously evaluated before their use can be recommended.

The rapid spread of the pandemic created several challenges for test developers and regulatory agencies. In this commentary, I would like to focus on the methodological issues in the clinical evaluation of medical tests. After a brief reminder of the general principles, I will focus on some specific issues in COVID-19-related testing, discussing testing for SARS-CoV-2 RNA, for COVID-19 disease, and for SARS-CoV-2 antibodies.

1. The evaluation of medical tests

When evaluating medical tests, we can ask ourselves three different questions. Can I trust the results? Are the results clinically meaningful? Is testing clinically useful? These three questions refer to three concepts: the analytical (or technical) performance of a test, its clinical performance, and the clinical utility of using the test [1].

The analytical performance of a laboratory test refers to its ability to correctly detect or measure a particular measurand [2]. It can be expressed in a number of ways, such as trueness (corresponding to the true value, absence of bias), imprecision (repeatability and reproducibility), limit of detection (analytical sensitivity), and cross-reactivity (analytical specificity). Cross-reactivity studies are performed to demonstrate that the test does not react with related pathogens, high-prevalence disease agents, or normal or pathogenic flora that are reasonably likely to be encountered in the clinical specimen.

Epidemiologists will be more familiar with evaluations of clinical performance, especially for diagnostic tests.

* Corresponding author. Tel.: +31 20 566 3240; fax: +31 20 566 9004.
E-mail address: p.m.bossuyt@amsterdamumc.nl.

https://doi.org/10.1016/j.jclinepi.2020.06.037
0895-4356/© 2020 Elsevier Inc. All rights reserved.

Here clinical performance is typically expressed as the diagnostic accuracy of the test: its ability to correctly classify those with and without the target condition, based on comparisons between the index test result and the outcome of the clinical reference standard [3]. Evaluations of clinical utility explore whether testing offers benefits, to those being tested, to the health care system, or to public health [4]. Because testing in itself rarely improves patient outcomes directly, evaluations of clinical utility usually look at test-treatment strategies.

Evaluations of clinical utility will provide the most convincing evidence for building recommendations about using the test, but at present they are not required for regulatory approval of COVID-19 test; evidence of sufficient analytical and clinical performance suffices [2]. In the following, we explore what this means for COVID-related tests. We distinguish between testing for the virus, testing for the disease, and testing for the antibodies after a viral infection.

2. Testing for the virus

The first atypical pneumonia cases were observed in Hubei province, China, in December 2019. Bronchoalveolar lavage fluid and cultured isolates from nine inpatients, eight of whom had visited the Huanan seafood market in Wuhan, were used to isolate a novel coronavirus [5]. The ten genome sequences exhibited more than 99·98% sequence identity. The virus was initially named 2019-nCoV, but later classified by the Coronavirus Research Group of the International Committee for the classification of viruses as SARS-CoV-2 because of its similarities with the SARS-CoV virus that had swept China in 2003 [6].

Identification of the viral genome sequence opened the path for methods based on nucleic acid amplification to detect SARS-CoV-2 [7]. Reverse transcription polymerase chain reaction (RT-PCR) is a variation of PCR, which adds reverse transcription of RNA to DNA, to allow for amplification. Different RT-PCR tests have been developed, targeting different genes of the SARS-CoV-2 genome [8]. RT-PCR can detect the virus in nasal and pharyngeal swab specimens, bronchoalveolar lavage fluid, sputum, bronchial aspirates, anal swab, and other samples [9].

The evaluation of the limit of detection of RT-PCR methods is typically carried out with spiking RNA or inactivated virus into an artificial or real clinical matrix, such as bronchoalveolar lavage fluid or sputum. Zhen and colleagues, for example, used a dilution panel of SARS-CoV-2 synthetic RNA quantified control with concentrations ranging from 20,000 to 5 copies/mL. The limit of detection was defined as either the lowest dilution at which all replicates resulted positive with a 100% detection rate, or the lowest detectable dilution at which the synthetic RNA quantified control was positive with a 95% probability of detection.

More challenging is the evaluation of clinical performance. It has been well documented that serial RT-PCR testing in patients who are initially negative can produce positive results later [10]. It is also well known that if multiple samples are taken from the same patient, some can be positive, whereas others are negative: the virus can be found in some samples, while absent, or not detectable, in other samples [9,11].

Researchers trained in clinical epidemiology will be tempted to classify these negative test results as either true or false negatives. But doing so requires knowledge of the truth. What then is the truth? Unfortunately, there is no independent, separate gold standard for detecting the virus, or viral RNA. To address this absence of a gold standard, FDA accepts testing a minimum of 30 positive specimens and 30 negative specimens as determined by an authorized assay [12]. This pushes back the question: what was the evidence for authorizing that first assay? Does that authorized assay represent the truth?

The recognition that even an authorized assay might fail has consequences for the way findings are presented. Poljak and colleagues, for example, evaluated the cobas test from Roche Diagnostics against their own SARS-CoV-2 protocol in 502 clinical samples, but expressed the results of their analysis in terms of agreement and kappa statistics, not in terms of sensitivity and specificity [13]. Several other approaches have been used. Zhen and colleagues, for example, used a "consensus" reference standard in their comparison of four molecular methods: the result obtained by at least three of the four assays [14].

Several other features in clinical performance evaluations of molecular viral tests should worry methodologists. Most of the problems mentioned are not unique for COVID-19, but apply to other areas of RT-PCR testing as well [15].

One of these problems lies in the use of using contrived clinical samples for assessing clinical performance. In testing for SARS-CoV-2, these are typically leftover upper respiratory specimens, such as nasopharyngeal swabs, or lower respiratory tract specimens, such as sputum, spiked with RNA or inactivated virus. Initially, FDA accepted evaluations of performance for Emergency Use Authorization (EUA) based on such contrived samples. It is not difficult to see that these contrived samples are a poor proxy for actual clinical samples. It is unclear if the viral concentrations in such contrived samples are representative of the full range of material taken from patients' airways in the real world. In more recent guidance, the use of clinical specimens is required, either positive by an EUA-authorized assay, whereas specimens collected before the pandemic are acceptable as negatives [16].

A second reason for methodological concern is the poor description of the origins of the clinical samples. Commercially available assays always have a product insert, which includes a description of the clinical performance of the corresponding assay [17]. Even though the STARD guidelines for reporting diagnostic accuracy studies were first published in 2003, these summaries fail to include relevant details on how and where study participants were identified [18]. The Simplexa COVID-19 Direct rtRT-PCR test, for example, was evaluated "in 278 consecutive respiratory samples (nasal and nasopharyngeal swabs) "collected for COVID-19 diagnosis", but the study report fails to mention where the study was conducted, how eligible patients were identified and selected, or what their symptoms were [19]. Informative reporting is still far away.

A third problematic element is the characterization of evaluations of clinical performance in terms of sensitivity and specificity. Many courses in epidemiology teach students that sensitivity and specificity are fixed test-based properties, unlike the negative and predictive values, which vary with the prevalence of the target condition. This is a gross simplification, as becomes clear in the evaluation of SARS-CoV-2 tests.

In symptomatic COVID-19 infections, viral RNA becomes detectable in the nasopharyngeal swab as early as day 1 of symptoms, peaks within the first week of symptom onset, and declines thereafter [20]. The positivity timeline also differs depending on the specimen; positivity is assumed to decline more slowly in sputum samples, which may still be positive after nasopharyngeal swabs are negative [21]. Collecting and handling the samples also affects the chances of test positivity.

Given this variability, one can question whether it is clear what a proportion of test-positive findings in those diagnosed as COVID-19 with an authorized assay refers to. What is the population parameter? Is it the sensitivity in the universe of all potential patients? Or should we condition that probability further, considering the symptoms, the timing, the sample, preanalytical handling? That would give us not one, single sensitivity for a specific test, but a wide range of conditional sensitivities.

I believe the variability in the mechanisms that produce false negatives makes it also hazardous to characterize the chances of a true positive as a single conditional probability: "the" sensitivity of a particular assay. This also has consequences for those who try to help the community through the development of systematic reviews.

Several systematic reviews of evaluations of clinical performance have started to appear. Some include meta-analysis, generating single number summary estimates of sensitivity and specificity. Arevalo-Rodriguez and colleagues, for example, summarized results from five studies and presented a summary estimate of the proportion of false-negative initial RT-PCR tests, for all assays, of 0.085 [22]. This single summary estimate ignores the multiple and variables sources of false negatives, as well as the differences between assays. These authors therefore immediately—and rightfully—added that interpretation of that proportion should be avoided, given the large heterogeneity.

There is more at stake than a variability in sensitivity, as beyond all this is an even more fundamental question: what is the target condition one wants to detect? Yes, RT-PCR can detect viral RNA, but detection does not distinguish between the presence of live virus and noninfectious viral debris. Is one more interested in the (past) presence of the virus (as an explanation for the illness) or in infectability (based on risk of viral shedding) [21]? The questions are related, but different. In one study, scientists could not grow viruses from throat swabs or sputum specimens after day 8 of illness from people who had mild infections, which suggests they may no longer be infectious. Yet the duration of viral shedding seems to vary, likely depending on severity. Among 137 survivors of COVID-19, viral shedding based on testing of oropharyngeal samples ranged from 8 to 37 days, with a median of 20 days [23].

3. Testing for the disease

The World Health Organization has named the disease caused by the SARS-CoV-2 as "coronavirus disease 2019", or COVID-19. At present, the diagnosis of the COVID-19 is mainly based on clinical characteristics, epidemiological history, chest imaging, and viral detection [24]. Although understanding of the variability in manifestations of COVID-19 still grows, and manifold cardiovascular complications are rapidly emerging, many patients present with pneumonia-like symptoms. Consequently, chest CT is often used to evaluate patients with suspected COVID-19. The main CT feature of COVID-19 pneumonia is the presence of ground glass opacities, typically with a peripheral and subpleural distribution [25].

A growing number of evaluations of the clinical performance of chest CT for detecting COVID-19 in patients with respiratory problems have appeared [26]. A key challenge for these evaluations is, once again, the fallibility of the reference standard, which, in most cases, has been RT-PCR. This also challenges the development and evaluation of multimarker scoring systems and decision rules for evaluating patients with symptomatic COVID-19 [27].

4. Testing for antibodies

A different type of tests is based on the host immune response to SARS-CoV-2 infection. Targeted antibodies against SARS-CoV-2 can be detected one to weeks after infection [28]. Available serology tests differ in terms of the platform (lateral flow assays, enzyme-linked immunosorbent assays and chemiluminescent immunoassays), the type of antigens (spike proteins, nucleocapsid proteins, receptor-binding domain), and the type of antibody being detected (IgM, IgG, IgA) [29].

A plethora of serology tests have been brought to the market, both laboratory-based tests and point-of-care tools

[8]. The evaluation of these tests poses similar difficulties to the ones discussed earlier for molecular tests: problematic reference standards, poor reporting, and uninformative statistics.

Strictly speaking, the target condition of these tests is the presence of viral antigens. In the absence of a gold standard for antigens, molecular tests to detect viral RNA are used as the reference standard. Even though both target conditions are related to the virus, they are not interchangeable. As discussed earlier, the time positivity curves for viral RNA testing and antibody testing do not overlap. If no antibodies are found in a patient testing RT-PCR positive, this could be a false negative (the test failed to detect the antibodies) as well as a true negative (no antibodies have developed).

The language that FDA uses in its guidance and templates for manufacturers is particularly interesting in this respect [30,31]. For evaluating clinical performance of serology, FDA recommends PCR in nasal swab samples and a fingerstick or blood draw from the same patient as the comparator, with the results expressed as percentage positive agreement and negative percentage agreement: the proportion concordant positives in all PCR positives and the proportion concordant negatives in all PCR negatives. So, this language refers to ‘‘agreement’’ with a comparator, and not to ‘‘verification’’ by the reference standard, as one would expect in typical diagnostic accuracy research.

5. Call for action

Despite improvements, the development of methods for the evaluation of medical tests lags behind approaches for evaluating pharmaceuticals and other interventions. The area could benefit from the contribution of clinical epidemiologists: as researchers, as peer-reviewers, and as developers of stronger methods (Table 1).

Evaluations of analytical performance, not part of the standard curriculum of epidemiologists and clinical researchers, could benefit from more rigors and a better understanding of the importance of study designs, and clinical research could be stronger if more experts in laboratory medicine were involved [32,33].

For the evaluation of analytical and clinical performance, the intended use of the test should be clear, its purpose and role in the clinical pathway, and studies should be designed accordingly. Participants and samples should be collected in the target population, for example [1]. Minimally acceptable performance criteria should be defined accordingly, indicating the level of performance that is likely to generate clinical utility [34].

The diagnostic accuracy paradigm, the dominant approach when evaluating clinical performance, also needs further development. It is by now well understood that sensitivity and specificity are not fixed test properties; they describe the behavior of a test in specific circumstances, which can be described by the features of the population (symptoms, age, gender), the setting (community-based, hospital), and previous test results, among others [35]. This means that one should resist, especially in systematic reviews, the temptation to summarize test performance in a single number. Even reporting the mean sensitivity, as is typical in random-effects meta-analysis, can be misleading if one ignores the wide and identifiable variability.

A point of concern is the nature of the reference standard, and the definition of the target condition. To be fully informative, evaluations of clinical performance should be specific in describing what it is they are trying to detect: active virus, any viral debris, antibodies, type of antibodies, past infection, to name a few options. Measurand and target condition are not synonyms. For other purposes of testing—and maybe even for diagnostic ones—we should develop alternative approaches, to complement or replace the gold standard/clinical reference standard paradigm.

Above all, we need informative reporting of the recruitment of participants, methods of sampling, and assays used to make sense of the study results. Improved adherence to existing reporting guidelines would be helpful.

The worldwide efforts in developing tests for COVID-19 are impressive, and we all hope these will help to curb the pandemic. Epidemiologists should contribute to improve the level of current performance studies, make reporting more transparent, and develop stronger methods to evaluate clinical performance and clinical utility.

Table 1. Key problems and potential solutions when evaluating the clinical performance of tests for COVID-19

Phase	Current problem	Desired solution
General	Study design unrelated to intended use	Define target population Define clinical pathway Define consequences
Design	Struggle with clinical reference standard	Clarify comparator
	Dominance reference standard paradigm	Develop alternative approaches for evaluating clinical performance
Analysis	Single statistics for clinical performance	Better methods for characterizing heterogeneity in test performance
Reporting	Poor description of participants and samples	Better adherence to reporting guidelines

Acknowledgments

P.M.B. collected the ideas and wrote the manuscript. The opinions expressed are his and do not necessarily reflect the views of the organizations mentioned, unless expressed otherwise.

References

[1] Horvath AR, Lord SJ, StJohn A, Sandberg S, Cobbaert CM, Lorenz S, et al. From biomarkers to medical tests: the changing landscape of test evaluation. Clinica Chim Acta Int J Clin Chem 2014; 427:49−57.

[2] Regulation (EU) 2017/746 of the European Parliament and of the council of 5 April 2017 on in vitro diagnostic medical devices and repealing directive 98/79/EC and commission decision 2010/227/EU. Official Journal of the European Union 2017;117: 176-332. Available at http://data.europa.eu/eli/reg/2017/746/oj.

[3] Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 2016;6(11): e012799.

[4] Bossuyt PM, Reitsma JB, Linnet K, Moons KG. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. Clin Chem 2012;58: 1636−43.

[5] Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 2020;395:565−74.

[6] Gorbalenya AE, Baker SC, Baric RS, et al. Coronaviridae Study Group of the International Committee on Taxonomy of V. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020;5:536−44.

[7] Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DK, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro Surveill 2020;25(3). https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045.

[8] Covid-19 diagnostics resource centre. Available at https://www.finddx.org/covid-19/. Accessed June 16, 2020.

[9] Pan Y, Zhang D, Yang P, Poon LLM, Wang Q. Viral load of SARS-CoV-2 in clinical samples. Lancet Infect Dis 2020;20:411−2.

[10] Long C, Xu H, Shen Q, Zhang X, Fan B, Wang C, et al. Diagnosis of the coronavirus disease (COVID-19): rRT-PCR or CT? Eur J Radiol 2020;126:108961.

[11] Zhao Y, Xia Z, Liang W, Li J, Liu L, Huang D, et al. SARS-CoV-2 persisted in lung tissue despite disappearance in other clinical samples. Clin Microbiol Infect 2020. https://doi.org/10.1016/j.cmi.2020.05.013.].

[12] US Food and Drug Administration. Draft Guidance for Industry, Clincial Investigator, and FDA Staff - Design Considerations for Pivotal Clinical Investigations for Medical Devices. Rockville, MD: Center for Biologics Evaluation and Research; 2011.

[13] Poljak M, Korva M, Knap Gasper N, Fujs Komlos K, Sagadin M, Ursic T, et al. Clinical evaluation of the cobas SARS-CoV-2 test and a diagnostic platform Switch during 48 hours in the midst of the COVID-19 pandemic. J Clin Microbiol 2020;58(6).

[14] Zhen W, Manji R, Smith E, Berry GJ. Comparison of four molecular in vitro diagnostic assays for the detection of SARS-CoV-2 in nasopharyngeal specimens. J Clin Microbiol 2020. https://doi.org/10.1128/JCM.00743-20.

[15] Murphy J, Bustin SA. Reliability of real-time reverse-transcription PCR in clinical diagnostics: gold standard or substandard? Expert Rev Mol Diagn 2009;9(2):187−97.

[16] Food and Drug Administration. Molecular diagnostic template for manufacturers, [updated May 11, 2020]. Available at https://www.fda.gov/media/135900/download. Accessed June 16, 2020.

[17] Xpert® Xpress SARS-CoV-2. Instructions for use. Available at https://www.cepheid.com/Package%20Insert%20Files/Xpress-SARS-CoV-2/Xpert%20Xpress%20SARS-CoV-2%20Assay%20CE-IVD%20ENGLISH%20Package%20Insert%20302-3787%20Rev.%20A.pdf. Accessed June 16, 2020.

[18] Korevaar DA, Cohen JF, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, et al. Updating standards for reporting diagnostic accuracy: the development of STARD 2015. Res Integr Peer Rev 2016;1: 7. https://doi.org/10.1186/s41073-016-0014-7.

[19] Bordi L, Piralla A, Lalle E, Giardina F, Colavita F, Tallarita M, et al. Rapid and sensitive detection of SARS-CoV-2 RNA using the Simplexa COVID-19 direct assay. J Clin Virol 2020;128:104416.

[20] Sethuraman N, Jeremiah SS, Ryo A. Interpreting diagnostic tests for SARS-CoV-2. JAMA 2020. https://doi.org/10.1001/jama.2020.8259.

[21] Wolfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Muller MA, et al. Virological assessment of hospitalized patients with COVID-2019. Nature 2020;581:465−9. https://doi.org/10.1038/s41586-020-2196-x.

[22] Arevalo-Rodriguez I, Buitrago-Garcia D, Simancas-Racines D, Zambrano-Achig P, del Campo R, Ciapponi A, et al. False-negative results of initial RT-PCR assays for covid-19: a systematic review. medRxiv 2020. https://doi.org/10.1101/2020.04.16.20066787.

[23] Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet 2020;395:1054−62.

[24] Wang H, Li X, Li T, Zhang S, Wang L, Wu X, et al. The genetic sequence, origin, and diagnosis of SARS-CoV-2. Eur J Clin Microbiol Infect Dis 2020. https://doi.org/10.1007/s10096-020-03899-4.

[25] Hani C, Trieu NH, Saab I, Dangeard S, Bennani S, Chassagnon G, et al. COVID-19 pneumonia: a review of typical CT findings and differential diagnosis. Diagn Interv Imaging 2020;101(5):263−8.

[26] Adams HJA, Kwee TC, Yakar D, Hope MD, Kwee RM, et al. Systematic review and meta-analysis on the value of chest CT in the diagnosis of coronavirus disease (COVID-19): Sol Scientiae, Illustra nos. AJR Am J Roentgenol 20201−9.

[27] Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ 2020; 369:m1328.

[28] Lou B, Li T, Zheng S, Su Y, Li Z, Liu W, et al. Serology characteristics of SARS-CoV-2 infection since the exposure and post symptoms onset. medRxiv 2020. https://doi.org/10.1101/2020.03.23.20041707.

[29] World Health Organization. Laboratory testing for coronavirus disease (COVID-19) in suspected human cases: interim guidance, 19 March 2020. Geneva: World Health Organization; 2020.

[30] U.S. Department of Health and Human Services. Food and Drug Administration. Center for devices and radiological health. Policy for coronavirus disease-2019 tests during the public health emergency (revised). Immediately in effect guidance for clinical laboratories, commercial manufacturers, and food and Drug Administration Staff. [updated May 11, 2020. Available at https://www.fda.gov/media/135659/download. Accessed June 16, 2020.

[31] Food and Drug Administration. Serology template for manufacturers [updated May 11, 2020]. Available at https://www.fda.gov/media/137698/download. Accessed June 16, 2020.

[32] Sun Q, Welsh KJ, Bruns DE, Sacks DB, Zhao Z. Inadequate reporting of analytical characteristics of biomarkers used in clinical research: a threat to interpretation and replication of study findings. Clin Chem 2019;65:1554−62. https://doi.org/10.1373/clinchem.2019.309575.

[33] Bossuyt PM. Laboratory measurement's contribution to the replication and application crisis in clinical research. Clin Chem 2019;65:1479−80.

[34] Lord SJ, St John A, Bossuyt PM, Sandberg S, Monaghan PJ, O'Kane M, et al. Setting clinical performance specifications to develop and evaluate biomarkers for clinical use. Ann Clin Biochem 2019;56(5):527−35.

[35] Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. BMJ 2002;324:669−71.