PLOS ONE

RESEARCH ARTICLE

Diagnostic test evaluation methodology: A systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard – An update

Chinyereugo M. Umemneku Chikere[1☯]*, Kevin Wilson[2☯], Sara Graziadio[3], Luke Vale[1☯], A. Joy Allen[4☯]

1 Institute of Health & Society, Faculty of Medical Sciences Newcastle University, Newcastle upon Tyne, England, United Kingdom, **2** School of Mathematics, Statistics and Physics, Newcastle University, Newcastle upon Tyne, England, United Kingdom, **3** National Institute for Health Research, Newcastle In Vitro Diagnostics Co-operative, Newcastle upon Tyne Hospitals National Health Services Foundation Trust, Newcastle upon Tyne, England, United Kingdom, **4** National Institute for Health Research, Newcastle In Vitro Diagnostics Co-operative, Newcastle University, Newcastle upon Tyne, England, United Kingdom

☯ These authors contributed equally to this work.
* C.Umemneku2@newcastle.ac.uk

OPEN ACCESS

Citation: Umemneku Chikere CM, Wilson K, Graziadio S, Vale L, Allen AJ (2019) Diagnostic test evaluation methodology: A systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard – An update. PLoS ONE 14(10): e0223832. https://doi.org/10.1371/journal.pone.0223832

Editor: Gianni Virgili, Universita degli Studi di Firenze, ITALY

Received: June 3, 2019

Accepted: September 29, 2019

Published: October 11, 2019

Data Availability Statement: All relevant data are within the manuscript and its supporting information files.

Funding: This work is supported by the Newcastle University Research Excellence; the School of Mathematics, Statistics and Physics Newcastle University; the Institute of Health & Society Newcastle University; and the National Institute for Health Research (NIHR) [NIHR Newcastle In Vitro Diagnostics Co-operative]. The view and opinions

Abstract

Objective

To systematically review methods developed and employed to evaluate the diagnostic accuracy of medical test when there is a missing or no gold standard.

Study design and settings

Articles that proposed or applied any methods to evaluate the diagnostic accuracy of medical test(s) in the absence of gold standard were reviewed. The protocol for this review was registered in PROSPERO (CRD42018089349).

Results

Identified methods were classified into four main groups: methods employed when there is a missing gold standard; correction methods (which make adjustment for an imperfect reference standard with known diagnostic accuracy measures); methods employed to evaluate a medical test using multiple imperfect reference standards; and other methods, like agreement studies, and a mixed group of alternative study designs. Fifty-one statistical methods were identified from the review that were developed to evaluate medical test(s) when the true disease status of some participants is unverified with the gold standard. Seven correction methods were identified and four methods were identified to evaluate medical test(s) using multiple imperfect reference standards. Flow-diagrams were developed to guide the selection of appropriate methods.

expressed are those of the authors and do not necessary reflect those of the NIHR Newcastle In Vitro Diagnostics Co-operative, Newcastle University and Newcastle upon Tyne NHS Foundation Trust, the NHS or Newcastle Research Academy. The views expressed are those of the authors and not necessarily those of the NIHR, the NHS or the Department of Health and Social Care. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interest exist.

Conclusion

Various methods have been proposed to evaluate medical test(s) in the absence of a gold standard for some or all participants in a diagnostic accuracy study. These methods depend on the availability of the gold standard, its' application to the participants in the study and the availability of alternative reference standard(s). The clinical application of some of these methods, especially methods developed when there is missing gold standard is however limited. This may be due to the complexity of these methods and/or a disconnection between the fields of expertise of those who develop (e.g. mathematicians) and those who employ the methods (e.g. clinical researchers). This review aims to help close this gap with our classification and guidance tools.

Introduction

Before a new medical test can be introduced into clinical practice, it should be evaluated for analytical validity (does the test work in the laboratory?), clinical validity (does the test work in the patient population of interest?) and clinical utility (is the test useful–can it lead to improvement in health outcomes?) [1, 2]. Clinical validity studies, also called diagnostic accuracy studies, evaluate the test's accuracy in discriminating between patients with or without the target condition (disease) [3]. The characteristics of the test (e.g. sensitivity and specificity) may inform what role the index test (the new test under evaluation) plays in the diagnostic pathway; is it a triage, add-on or replacement test? [4] Sensitivity (the proportion of participants correctly identified by the index test as having the target condition e.g. those with the disease) and specificity (the proportion of participants correctly identified by the index as not having the target condition) [5–7] are basic measures of the diagnostic accuracy of a test. Other common measures are predictive values, likelihood values, overall accuracy [8, 9], receiver operating characteristic (ROC) curve, area under the ROC curve (AUROC) [10], ROC surface, and volume under the ROC surface (VUS) [11–13]. These measures are obtained by comparing the index test results with the results of the best currently available test for diagnosing the same target condition in the same participants; both tests are supposedly applied to all participants of the study [14]. The test employed as the benchmark to evaluate the index test is called the reference standard [15]. The reference standard could be a gold standard (GS), with sensitivity and specificity equal to 100%. This means that the gold standard perfectly discriminates between participants with or without the target conditions and provides unbiased estimates of the diagnostic accuracy measure of the index test as describe in Fig 1. The term "bias" in this review is defined as the difference between the estimated value and the true value of the parameter of interest [16].

It is also expected that when evaluating the diagnostic accuracy of a medical test, the participants undertake both the index and reference tests within a short time-period if not simultaneously. This is to avoid biases caused by changes in their true disease status, which can also affect the diagnostic accuracy of the index test.

In addition to the common aforementioned diagnostic accuracy measures, there are other ways to evaluate the test performance of an index test. These include studies of agreement or concordance [17] between the index test and the reference standard and test positivity (or negativity) rate; that is the proportion of diagnostic tests that are positive (or negative) to the target condition [18].

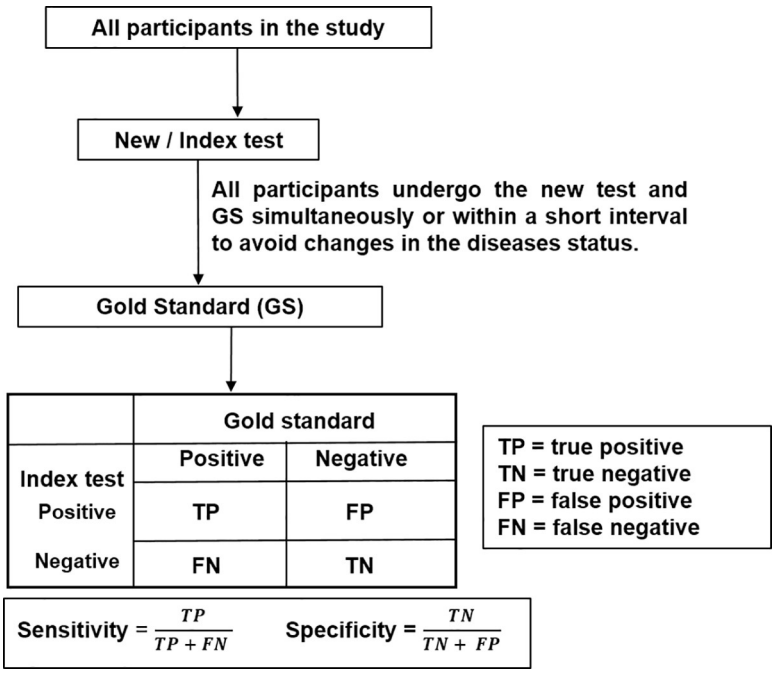

Fig 1. Classical method of evaluating the diagnostic accuracy of a medical test with binary test result and dichotomized disease status.

https://doi.org/10.1371/journal.pone.0223832.g001

In practice, there are deviations from the classical method (Fig 1). These deviations are:

1. Scenarios where the gold standard is not applied to all participants in the study (i.e. there is a missing gold standard) because it is expensive, or invasive, or patients do not consent to it, or the clinicians decided not to give the gold test to some patients for medical reasons [19, 20]. Evaluating the new test using data only from participants whose disease status was confirmed with the gold standard can produce work-up or verification bias [21].
2. Scenarios where the reference standard is not a gold standard (i.e. it is an imperfect reference standard) because it has a misclassification error or because there is no generally accepted reference standard for the target condition. Using an imperfect reference standard produces reference standard bias [22, 23].

Several methods have been developed and used to evaluate the test performance of a medical test in these two scenarios.

Reviews of some of these methods have been undertaken previously. The reviews by Zhou [24], Alonzo [25] and the report by Naaktgeboren et al [26] focused on methods when the gold standard or reference standard is not applied to all participants in the study; Van Smeden et al [27] and Collins and Huynh [28] focused on the latent class models (LCMs); and Hui and Zhou [29], Trikalinos and Balion [30] and Enoe et al [31] focused on methods employed when the reference standard is imperfect. Zaki et al [32] focused on the agreement between medical tests whose results are reported as a continuous response. Branscum et al [33] focused on Bayesian approaches; and the reviews by Walsh [23], Rutjes et al [14] and Reitsma et al [34] focused around methods for evaluating diagnostic tests when there is a missing or imperfect reference standard.

The existing comprehensive reviews on this topic were published about 11 years ago [14, 34]; knowledge, ideas, and research in this field has evolved significantly since then. Several

new methods have been proposed and some existing methods have been modified. It is also possible that some previously identified methods may now be obsolete. Therefore, one of the aims of this systematic review is to review new and existing methods employed to evaluate the test performance of medical test(s) in the absence of gold standard for all or some of the participants in the study. It also aims to provide easy to use tools (flow-diagrams) for the selection of methods to consider when evaluating medical tests when sub-sample of the participants do not undergo the gold standard. The review builds upon the earlier reviews by Rutjes et al and Reitsma et al [14, 34]. This review sought to identify methods developed to evaluate a medical test with continuous results in the presence of verification bias and when the diagnostic outcome (disease status) is classified into three or more groups (e.g. diseased, intermediate and non-diseased). This is a gap identified in the review conducted by Alonzo [25] in 2014.

The subsequent sections discuss the method employed to undertake the review, the results, the discussion of the findings and guidance to researchers involved in test accuracy studies.

Methodology

A protocol for this systematic review was developed, peer-reviewed and registered on PROSPERO (CRD42018089349).

Eligibility criteria

The review includes methodological articles (that is papers that proposed or developed a method) and application articles (that is papers where any of the proposed methods) were applied.

Inclusion.

- Articles published in English language in a peer-reviewed journal.
- Articles that focus on evaluating the diagnostic accuracy of new (index) test when there is a missing gold standard, no gold standard or imperfect reference standard.

Exclusion.

- Articles that assumed that the reference standard was a gold standard and the gold standard was applied to all participants in the study.
- Books, dissertations, thesis, conference abstracts, and articles not published in a peer reviewed journal.
- Systematic reviews and meta-analyses of the diagnostic accuracy of medical test(s) for a target condition (disease) in the absence of gold standard for some or all of the participants. However, individual articles included in these reviews that met the inclusion criteria were included.

Search strategies and selection of articles

The PRISMA statement [35] was used as a guideline when conducting this systematic review. The PRISMA checklist for this review, S1 Checklist, is included as one of the supplementary materials. The following bibliographic databases were searched: EMBASE, MEDLINE, SCOPUS, WILEY online library (which includes Cochrane library, EBM), PSYCINFO, Web of Science, and CINAHL. The details of the search strategies at reported in the S1 Appendix. The search dates were from January 2005 –February 2019. This is because, this review is an update

of a review by Rutjes et al and Reitsma et al whose searched up to 2005. However, original methodological articles that proposed and described a method to evaluate medical test(s) when there is a missing or no gold standard published before 2005 were also included in the review. These original articles were identified by "snowballing" [36] from the references of some articles. All articles obtained from the electronic databases were imported to Endnote X8.0.2. The selection of articles to be included in this review were done by three people (CU, AJA, and KW). The sifting process was in two-stages: by title and abstract and then by full text against the inclusion and exclusion criteria. Any discrepancies between reviewers were resolved in a group meeting.

Data synthesis

A data collection form was developed for this review (S1 Data), which was piloted on seven studies and remodified to fit the purpose of this review. Information extracted from the included articles were synthesized narratively.

Results

A total of 6127 articles were identified; 5472 articles were left after removing the duplicated articles; 5071 articles were excluded after sifting by title and abstract; 401 articles went forward to full text assessment; and a total of 209 articles were included in the review. The search and selection procedure are depicted using the PRISMA [35] flow-diagram (Fig 2).

The articles included in this review used a wide variety of different study designs, like cross-sectional studies, retrospective studies, cohort studies, prospective studies and simulation studies.

The identified methods were categorized into four groups based on the availability and/or application of the gold standard to the participants in the study. These group are:

- Group 1: Methods employed when there is a missing gold standard.
- Group 2: Correction methods which adjust for using an imperfect reference standard whose diagnostic accuracy is known.

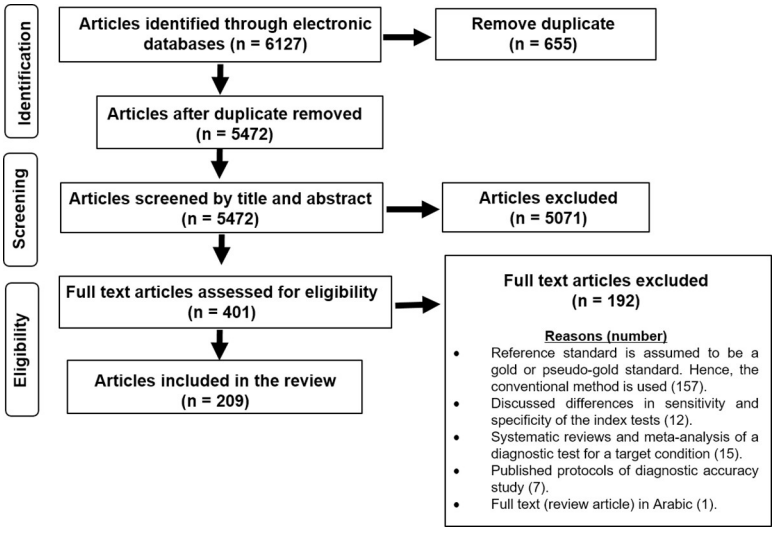

Fig 2. PRISMA flow-diagram of articles selected and included in the systematic review.

https://doi.org/10.1371/journal.pone.0223832.g002

- Group 3: Methods employed when using multiple imperfect reference standards.
- Group 4: ***"other methods"***. This group includes methods like study of agreement, test positivity rate, and considering alternative study design like validation.

Methods in groups 2, 3 and 4 are employed when there is no gold standard to evaluate the diagnostic accuracy of the index test; while methods in group 1 are employed when there is a gold standard to evaluate the diagnostic accuracy of the index test(s). However, the gold standard is applied to only a sub-sample of the participants.

A summary of all methods identified in the review, their key references and the clinical applications of these methods are reported on Table 1.

Methods employed when gold standard is missing

Fifty-one statistical methods were identified from the review that were developed to evaluate the diagnostic accuracy of index test(s) when the true disease status of some participants is not verified with the gold standard. These methods are divided into two subgroups:

- **Imputation and bias-correction methods**: This includes methods to correct for verification bias while the disease-status of the unverified participants are left unverified. Forty-eight

Table 1. Summary of classification of methods employed when there is missing or no gold standard.

Main Classification	Main Characteristics	Key references	Clinical Application
Group 1: **Method employed when there is missing gold standard:** • Imputation and bias-correction methods • Differential verification	The true disease status is verified with the gold standard only in a subsample of the study participants. The methods are grouped into ***imputation and bias-correction methods*** (Fig 3: Imputation and bias–correction methods in binary diagnostic outcomes. and Fig 4: Imputation and bias–correction methods in three- classes diagnostic outcomes where ROC surface and VUS are estimated.) and ***differential verification*** approach.	***Imputation and bias correction methods*** [10], [11], [13], [21], [37–81] ***Differential verification*** [82–84]	***Imputation & Bias-correction methods*** [85–89] ***Differential verification*** [90]
Group 2: **Correction methods**	The reference standard is imperfect. However, there is available information about the sensitivity and specificity of the reference standard which is used to correct or adjust the estimated sensitivity and specificity of the index test.	***Correction methods*** [91–96]	***Correction methods*** [97–99]
Group 3: Methods employed when using multiple imperfect reference standards or tests. • Discrepancy analysis • Latent class analysis • Composite reference standard (CRS) • Expert or panel or consensus diagnosis	A gold standard that diagnoses a target condition or accurate information on the diagnostic accuracy of an imperfect reference standard that diagnoses same condition may not be available. Thus, multiple imperfect tests may be employed to evaluate the index test. Methods in this group include discrepancy analysis, latent class analysis, composite reference standard, and panel or consensus diagnosis.	***Discrepancy analysis*** [100], [101] ***Latent class analysis*** ***Frequentist LCA***: [29], [102–112] ***Bayesian LCA***: [33], [113–119] ***ROC (NGS)***: [120–130] **Composite reference standard** [131–134] **Panel or consensus diagnosis** [135]	***Discrepancy analysis*** [136–139] ***Latent class analysis*** ***Frequentist LCA***: [140–152] ***Bayesian LCA***: [153–174] ***ROC (NGS)***: [175, 176] ***CRS***: [20, 177–184] ***Consensus diagnosis***: [185–189]
Group 4: **Other methods** • Considering an alternative study design like a validation study • Study of agreement • Test positivity rate	***Analytic validation*** of a medical test is the process of verifying the test based on what it is designed to do. Experimental or case-control are common designs for these studies. ***Studies of agreement*** aim to investigate the concordance between two or more tests (probably an index test and a reference standard). ***Test positivity rate***: is the proportion of participants who have positive results on a test. This approach was used by Van Dyck et al [18] to reduce the number of tests subjected to further evaluation.	***Validation*** [190, 191] ***Study of agreement***: [32], [192] ***Test positivity rate***: [18]	***Validation***: [193, 194] ***Study of agreement***: [165, 195–199] ***Test positivity rate*** [18, 192]

LCA: latent class analysis; CRS is composite reference standard. ROC is receiver operating characteristics; NGS is no gold standard

https://doi.org/10.1371/journal.pone.0223832.t001

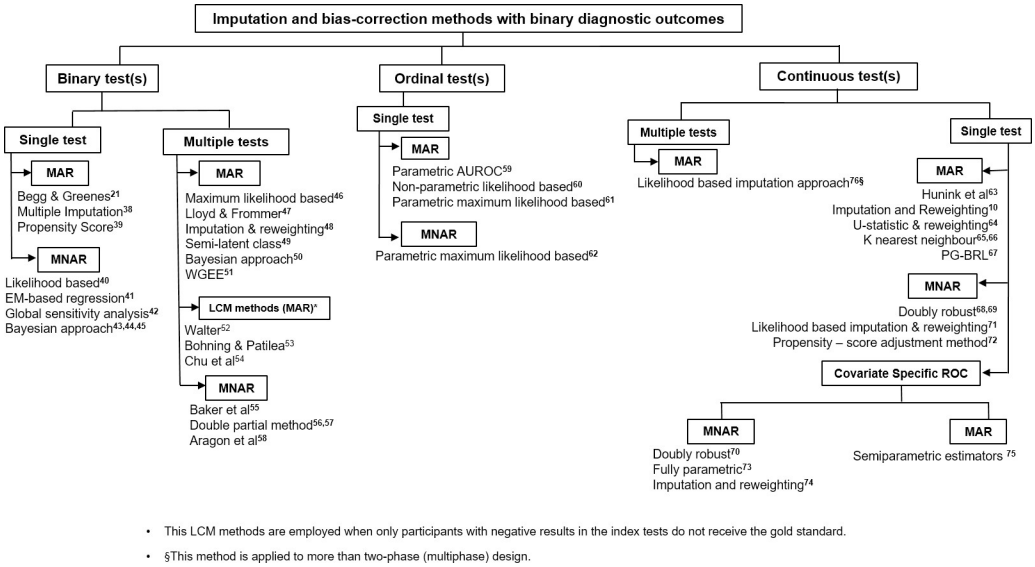

Fig 3. Imputation and bias–correction methods in binary diagnostic outcomes.

https://doi.org/10.1371/journal.pone.0223832.g003

statistical methods were identified in this group. These methods are further classified based on the result of the index test (binary, ordinal or continuous), the number of index tests evaluated (single or multiple), the assumptions made about verification (ignorable or missing at random–MAR) or non-ignorable or missing not at random–MNAR), and the classification of the diagnostic outcomes (disease-status). The identified methods in this subgroup are displayed Figs 3 and 4.

- **Differential verification approach**: Participants whose disease status was not verified with the gold standard could undergo another reference standard (that is imperfect or less invasive than the gold standard [84]) to ascertain their disease status. This is known as ***differential verification*** [200]. Differential verification has been explored Alonzo et al, De Groot

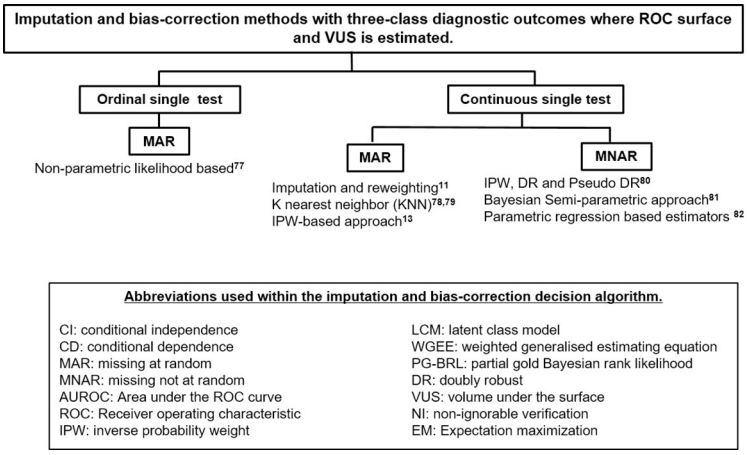

Fig 4. Imputation and bias–correction methods in three- classes diagnostic outcomes where ROC and VUS is estimated.

https://doi.org/10.1371/journal.pone.0223832.g004

et al and Naaktgeboren et al [200–202]. They discussed the bias associated with differential verification, and how results using this approach could be presented. There are three identified statistical methods in this group. They are: a Bayesian latent class approach proposed by De Groot et al [82], a Bayesian method proposed by Lu et al [203] and a ROC approach proposed by Glueck et al [16]. These three methods aim to simultaneously adjust for differential verification bias and reference standard bias that arises from using an alternative reference standard (i.e. imperfect reference standard) for participants whose true disease status was not verified with the gold standard.

Correction methods

This group includes algebraic methods developed to correct the estimated sensitivity and specificity of the index test when the sensitivity and specificity of the imperfect reference standard is known. There are seven statistical methods in this group described in five different articles [91–95]. The methods by Emerson et al [95] does not estimate a single value for sensitivity or specificity, unlike the other correction methods [91–94] but produces an upper bound value and a lower bound value for the sensitivity and specificity of the index test. These bounded values are used to explain the uncertainty around the estimated sensitivity and specificity of the index test.

Methods with multiple imperfect reference standards

A gold standard or accurate information about the diagnostic accuracy of the imperfect reference standard are often not available to evaluate the index test. In these situations, multiple imperfect reference standards can be employed to evaluate the index test. Methods in this group include:

- **Discrepancy analysis**: this compares the index test with an imperfect reference standard. Participants with discordant results undergo another imperfect test, called the resolver test, to ascertain their disease status. Discrepancy analysis is typically not recommended because it produces biased estimates [100, 204]. Modifications of this approach have been proposed [18, 101, 136]. In these, some of the participants with concordant responses (true positives and true negatives) are sampled to undertake the resolver test alongside participants with discordant responses (false negative–FN and false positive–FP). However, further research is needed to explore if these modified approaches are adequate to remove or reduce the potential bias.

- **Latent class analysis (LCA)**: The test performance of all the tests employed in the study are evaluated simultaneously using probabilistic models with the basic assumption that the disease status is latent or unobserved. There are frequentist LCAs and Bayesian LCAs. The frequentist LCAs use only the data from the participants in the study to estimate the diagnostic accuracy measures of the tests; while the Bayesian LCAs employ external information (e.g. expert opinion or estimates from previous research study) on the diagnostic accuracy measures of the tests evaluated in additional to the empirical data obtained from participants within the study. The LCAs assume that the tests (new test and reference standards) are either conditionally independent given the true disease status or the tests are conditionally dependent. To model the conditional dependence among the tests, various latent class model (LCM) with different dependence structure have been developed such as the Log-linear LCM [102], Probit LCM [103], extended log-linear and Probit LCM [108], Gaussian Random Effect LCM [105] and two-crossed random effect LCM [107] among others. However, some studies [205],[206] have shown that latent class models with different conditional

dependence structures produce different estimates of sensitivities and specificities and each model still has a good fit. Thus, further research could be carried out to explore if each of the conditional dependence LCM are case specific.

- **Construct composite reference standard**: this method combines results from multiple imperfect tests (excluding the index test) with a predetermined rule to construct a reference standard that is used to evaluate the index test. By excluding the index test as part of the composite reference standard, incorporation bias can be avoided [131]. A novel method identified under the composite reference standard is the "dual composite reference standard" proposed by Tang et al [134].
- **Panel or consensus diagnosis**: this method uses the decision from a panel of experts to ascertain the disease status of each participant, which is then used to evaluate the index test.

Other methods

This group includes methods that fit the inclusion criteria but could not be placed into the other three groups. They include study of agreement, test positivity rate and the use of an alternative study design such as analytical validation. Study of agreement and test positivity rate are best used as exploratory tools alongside other methods [152, 178] because they are not robust enough to assess the diagnostic ability of the medical test. Validation of a medical test cut across different disciplines in medicine such as psychology, laboratory or experimental medicine. With this approach, the medical test is assessed based on what it is designed to do [191]. Other designs include case-control designs (where the participants are known to have or not have the target condition) [207, 208], laboratory based studies or experimental studies which are undertaken with the aim to evaluate the analytical sensitivity and specificity of the index test [190, 209, 210].

Guidance to researchers

The guidance flowchart (Fig 5) is a modification and extension of the guidance for researchers flow-diagram developed by Reitsma et al [34].

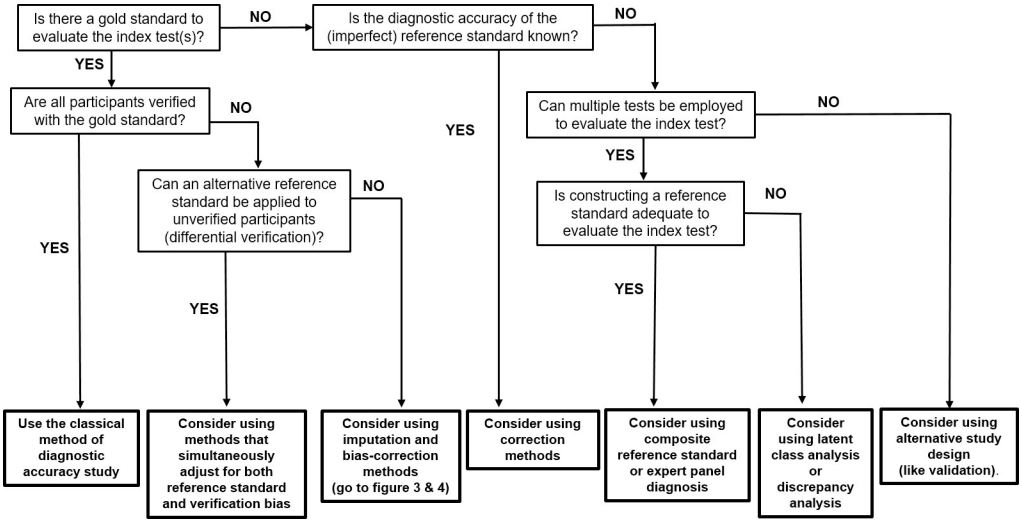

Fig 5. Guidance flowchart of methods employed to evaluate medical test in missing and no gold standard scenarios.

https://doi.org/10.1371/journal.pone.0223832.g005

Box 1: Suggestions when designing a diagnostic accuracy study.

- **Design a protocol**: The protocol describes every step of the study. It states the problem and how it will be addressed.
- **Selection of participants from target population**: The target population determines the criteria for including participants in the study. Also, the population is important in selecting the appropriate setting for the study.
- **Selection of appropriate reference standard:** The reference standard should diagnose same target condition as the index test. The choice of reference standard (gold or non-gold) determines the methods to apply when evaluating the index test (see Fig 5).
- **Sample size**: Having adequate sample size is necessary to make precise inference from the statistical analysis that will be carried out. Studies that discuss the appropriate sample size to consider when planning test accuracy are [211–215].
- **Selection of accuracy measure to estimate**: The researchers need to decide which accuracy measures they wish to estimate, and this is often determined by the test's response (binary or continuous).
- **Anticipate and eliminate possible bias**: multiple forms of bias may exist [26, 216–218]. Exploring how to avoid or adjust for these bias (if they are unavoidable) is important.
- **Validation of results**: Is validation of the results from the study on an independent sample feasible? Validation ensures an understanding of the reproducibility, strengths, and limitations of the study.

Since, evaluating the accuracy measures of the index test is the focus of any diagnostic accuracy study, the flowchart starts with asking the first question "Is there a gold standard to evaluate the index test?" Following the responses from each question box (not bold); methods are suggested (bold boxes at the bottom of the flowchart) to guide clinical researchers, test evaluators, and researchers as to the different methods to consider.

Although, this review aims to provide up-to-date approaches that have been proposed or employed to evaluate the diagnostic accuracy of an index test in the absence of a gold standard for some or all of the participants in the accuracy study; some things researchers can consider when designing an accuracy study aside from the aim of their studies, are outlined in Box 1 ([26, 211–218]).

Some guidelines and tools have been developed to assist in designing, conducting and reporting diagnostic accuracy studies such as the STARD [219–223] guidelines, GATE [224] framework, QUADAS [225] tools; which can aid the design of a robust test accuracy study.

Discussion

This review sought to identify and review new and existing methods employed to evaluate the diagnostic accuracy of a medical test in the absence of gold standard. The identified methods are classified into four main groups based on the availability and/or the application of the gold

standard on the participants in the study. The four groups are: methods employed when only a sub-sample of the participants have their disease status verified with the gold standard (group 1); correction methods (group 2); methods using multiple imperfect reference standards (group 3) and other methods (group 4) such as study of agreement, test positivity rate and alternative study designs like validation.

In this review additional statistical methods have been identified that were not included in the earlier reviews on this topic by Reitsma et al [34] and Alonzo [25]. A list of all the methods identified in this review are presented in the supplementary material (S1 Supplementary Information). This includes a brief description of the methods and a discussion of their strengths and weaknesses and any identified case studies where the methods have been clinically applied. Only a small number of the methods we have identified have applied clinically and published [38, 63]. This may be due to the complexity of these methods (in terms of application and interpretation of results), and/or a disconnection between the fields of expertise of those who develop (e.g. mathematicians or statisticians) and those who employ the methods (e.g. clinical researchers). For example, the publication of such method in specialist statistical journals may not be readily accessible to clinical researchers designing the study. In order to close this gap, two flow-diagrams (Figs 3 and 4) were constructed in addition to the modified guidance flowchart, (Fig 5) as guidance tools to aid clinical researchers and test evaluators in the choice of methods to consider when evaluating medical test in the absence of gold standard. Also, an R package (*bcROCsurface*) and an interactive web application (Shiny app) that estimates the ROC surface and VUS in the presence of verification bias have been developed by To Duc [78] to help bridge the gap.

One of the issues not addressed in this current review was on methods that evaluate the differences in diagnostic accuracy of two or more tests in the presence of verification bias. Some published articles that consider this issue are Nofuentes and Del Castillo [226–230], Marin-Jimenez and Nofuentes [231], Harel and Zhou [232] and Zhou and Castelluccio [233]. This review also did not consider methods employed to estimate the time-variant sensitivity and specificity of diagnostic test in absence of a gold standard. This issue has recently been addressed by Wang et al [234].

In terms of the methodology, a limitation of this review is the exclusion of books, dissertations, thesis, conference abstract and articles not published in English language (such as the review by Masaebi et al [235] which was published in 2019), which could imply that there could still be some methods not identified by this review.

Regarding the methods identified in this review, further research could be carried to explore the different modification to the discrepancy analysis approaches to understand if these modifications reduce or remove the potential bias. In addition, further research is needed to determine if the different methods developed to evaluate an index test in the presence of verification bias are robust methods. Given the large numbers of statistical methods that have been developed especially to evaluate medical tests when there is a missing gold standard and the complexity of some of these methods; more interactive web application (e.g. Shiny package in R [236]) could be developed to implement these methods in addition to the Shiny app developed by To Duc [78] and Lim et al [237]. The development of such interactive web tools will expedite the clinical applications of these developed methods and help bridge the gap between the method developers and the clinical researchers or tests evaluators who are the end users of these methods.

Conclusion

Various methods have been proposed and applied in the evaluation of medical tests when there is a missing gold standard result for some participants, or no gold standard at all. These

methods depend on the availability of the gold standard, its application to all or subsample of participants in the study, the availability of alternative reference standard(s), and underlying assumption(s) made with respect to the index test(s) and / or participants in the study.

Knowing the appropriate method to employ when analysing the data from participants of a diagnostic accuracy studies in the absence of gold standard, help to make statistically robust inference on the accuracy of the index test. This, in addition to data on cost-effectiveness, utility and usability of the test will support clinicians, policy makers and stake holders to decide the adoption of the new test in practice or not.

Supporting information

S1 Checklist. PRISMA checklist.
(DOC)

S1 Data. Data extraction form.
(DOCX)

S1 Appendix.
(DOCX)

S1 Supplementary Information.
(DOCX)

Acknowledgments

The authors will like to acknowledge Professor Patrick Bossuyt from the Department of Clinical Epidemiology and Biostatistics, Academic Medical centre, University of Amsterdam, the Netherlands, for giving the consent to update his review, reviewing the protocol and his continued advice throughout this work. Also we will like to acknowledge the authors of the previous review, Dr Anne Rutjes in University of Bern, Switzerland; Professor Johannes Reitsma in the Department of Epidemiology, Julius Center Research Program Methodology UMC Utrecht, The Netherlands; Professor Arri Coomarasamy in the College of Medical and Dental Sciences, University of Birmingham, UK; and Professor Khalid Saeed Khan in Queen Mary, University of London for the guidance flowchart which was modified and extended. AJA, SG, and LV are supported by the National Institute for Health Research (NIHR) Newcastle In Vitro Diagnostics Co-operative. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Author Contributions

Conceptualization: Chinyereugo M. Umemneku Chikere, Kevin Wilson, Luke Vale, A. Joy Allen.

Data curation: Chinyereugo M. Umemneku Chikere, Kevin Wilson, Luke Vale, A. Joy Allen.

Formal analysis: Chinyereugo M. Umemneku Chikere, Kevin Wilson, Luke Vale, A. Joy Allen.

Funding acquisition: Kevin Wilson, Luke Vale, A. Joy Allen.

Investigation: Chinyereugo M. Umemneku Chikere, Kevin Wilson, Sara Graziadio, Luke Vale, A. Joy Allen.

Methodology: Chinyereugo M. Umemneku Chikere, Kevin Wilson, Luke Vale, A. Joy Allen.

Supervision: Chinyereugo M. Umemneku Chikere, Kevin Wilson, Sara Graziadio, Luke Vale, A. Joy Allen.

Writing – original draft: Chinyereugo M. Umemneku Chikere.

Writing – review & editing: Chinyereugo M. Umemneku Chikere, Kevin Wilson, Sara Graziadio, Luke Vale, A. Joy Allen.

References

1. Bossuyt PM, Reitsma JB, Linnet K, Moons KG. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. Clinical chemistry. 2012; 58(12):1636–43. https://doi.org/10.1373/clinchem.2012.182576 PMID: 22730450
2. Burke W. Genetic tests: clinical validity and clinical utility. Current protocols in human genetics. 2014; 81(1):9.15. 1–9. 8.
3. Mallett S, Halligan S, Matthew Thompson GP, Collins GS, Altman DG. Interpreting diagnostic accuracy studies for patient care. BMJ (Online). 2012; 345(7871). https://doi.org/10.1136/bmj.e3999 PMID: 22750423
4. Bossuyt PMI L.; Craig J.; Glasziou P. Comparative accuracy: Assessing new tests against existing diagnostic pathways. British Medical Journal. 2006; 332(7549):1089–92. https://doi.org/10.1136/bmj.332.7549.1089 PMID: 16675820
5. Altman DG, Bland JM. Diagnostic tests 1: Sensitivity and specificity. British Medical Journal. 1994; 308 (6943):1552. https://doi.org/10.1136/bmj.308.6943.1552 PMID: 8019315
6. Eusebi P. Diagnostic Accuracy Measures. Cerebrovascular Diseases. 2013; 36(4):267–72. https://doi.org/10.1159/000353863 WOS:000326935800004. PMID: 24135733
7. Šimundić A-M. Measures of diagnostic accuracy: basic definitions. Ejifcc. 2009; 19(4):203. PMID: 27683318
8. Altman DG, Bland JM. Diagnostic tests 2: Predictive values. British Medical Journal. 1994; 309 (6947):102. https://doi.org/10.1136/bmj.309.6947.102 PMID: 8038641
9. Wong HB, Lim GH. Measures of diagnostic accuracy: Sensitivity, specificity, PPV and NPV. Proceedings of Singapore Healthcare. 2011; 20(4):316–8. https://doi.org/10.1177/201010581102000411
10. Alonzo TA, Pepe MS. Assessing accuracy of a continuous screening test in the presence of verification bias. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2005; 54(1):173–90. https://doi.org/10.1111/j.1467-9876.2005.00477.x
11. Duc KT, Chiogna M, Adimari G. Bias–corrected methods for estimating the receiver operating characteristic surface of continuous diagnostic tests. Electronic Journal of Statistics. 2016; 10(2):3063–113. https://doi.org/10.1214/16-EJS1202
12. Chi YY, Zhou XH. Receiver operating characteristic surfaces in the presence of verification bias. Journal of the Royal Statistical Society Series C: Applied Statistics. 2008; 57(1):1–23. https://doi.org/10.1111/j.1467-9876.2007.00597.x
13. Zhang Y, Alonzo TA, for the Alzheimer's Disease Neuroimaging I. Inverse probability weighting estimation of the volume under the ROC surface in the presence of verification bias. Biometrical Journal. 2016; 58(6):1338–56. https://doi.org/10.1002/bimj.201500225 PMID: 27338713
14. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health technology assessment (Winchester, England). 2007; 11(50):iii, ix–51.
15. Kohn MA, Carpenter CR, Newman TB. Understanding the Direction of Bias in Studies of Diagnostic Test Accuracy. Academic Emergency Medicine. 2013; 20(11):1194–206. https://doi.org/10.1111/acem.12255 WOS:000327026400017. PMID: 24238322
16. Glueck DHL M. M.; O'Donnell C. I.; Ringham B. M.; Brinton J. T.; Muller K. E.; Lewin J. M.; Alonzo T. A.; Pisano E. D. Bias in trials comparing paired continuous tests can cause researchers to choose the wrong screening modality. BMC medical research methodology. 2009; 9:4. https://doi.org/10.1186/1471-2288-9-4 PMID: 19154609
17. Theel ES, Hilgart H, Breen-Lyles M, McCoy K, Flury R, Breeher LE, et al. Comparison of the QuantiFERON-TB gold plus and QuantiFERON-TB gold in-tube interferon gamma release assays in patients at risk for tuberculosis and in health care workers. Journal of Clinical Microbiology. 2018;56(7). https://doi.org/10.1128/JCM.00614-18 PMID: 29743310
18. Van Dyck E, Buvé A, Weiss HA, Glynn JR, Brown DWG, De Deken B, et al. Performance of commercially available enzyme immunoassays for detection of antibodies against herpes simplex virus type 2

in African populations. Journal of Clinical Microbiology. 2004; 42(7):2961–5. https://doi.org/10.1128/JCM.42.7.2961-2965.2004 PMID: 15243045

19. Naaktgeboren CA, De Groot JAH, Rutjes AWS, Bossuyt PMM, Reitsma JB, Moons KGM. Anticipating missing reference standard data when planning diagnostic accuracy studies. BMJ (Online). 2016;352. https://doi.org/10.1136/bmj.i402 PMID: 26861453

20. Karch AK A.; Zapf A.; Zerr I.; Karch A. Partial verification bias and incorporation bias affected accuracy estimates of diagnostic studies for biomarkers that were part of an existing composite gold standard. Journal of Clinical Epidemiology. 2016; 78:73–82. https://doi.org/10.1016/j.jclinepi.2016.03.022 WOS:000389615400010. PMID: 27107877

21. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics. 1983; 39(1):207–15. PMID: 6871349

22. Thompson M, Van den Bruel A. Sources of Bias in Diagnostic Studies. Diagnostic Tests Toolkit: Wiley-Blackwell; 2011. p. 26–33.

23. Walsh T. Fuzzy gold standards: Approaches to handling an imperfect reference standard. Journal of Dentistry. 2018; 74:S47–S9. https://doi.org/10.1016/j.jdent.2018.04.022 PMID: 29929589

24. Zhou XH. Correcting for verification bias in studies of a diagnostic test's accuracy. Statistical Methods in Medical Research. 1998; 7(4):337–53. https://doi.org/10.1177/096228029800700403 PMID: 9871951

25. Alonzo TA. Verification bias-impact and methods for correction when assessing accuracy of diagnostic tests. Revstat Statistical Journal. 2014; 12(1):67–83.

26. Naaktgeboren CA, de Groot JA, Rutjes AW, Bossuyt PM, Reitsma JB, Moons KG. Anticipating missing reference standard data when planning diagnostic accuracy studies. bmj. 2016; 352:i402. https://doi.org/10.1136/bmj.i402 PMID: 26861453

27. 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. American Journal of Epidemiology. 2014; 179(4):423–31. https://doi.org/10.1093/aje/kwt286 WOS:000331264100003. PMID: 24272278

28. Collins J, Huynh M. Estimation of diagnostic test accuracy without full verification: a review of latent class methods. Statistics in Medicine. 2014; 33(24):4141–69. https://doi.org/10.1002/sim.6218 PMID: 24910172

29. Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. Statistical Methods in Medical Research. 1998; 7(4):354–70. https://doi.org/10.1177/096228029800700404 PMID: 9871952

30. Trikalinos TA, Balion CM. Chapter 9: Options for summarizing medical test performance in the absence of a "gold standard". Journal of General Internal Medicine. 2012; 27(SUPPL.1):S67–S75.

31. Enøe C, Georgiadis MP, Johnson WO. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. Preventive Veterinary Medicine. 2000; 45(1):61–81. https://doi.org/10.1016/S0167-5877(00)00117-3.

32. Zaki R, Bulgiba A, Ismail R, Ismail NA. Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: a systematic review. PloS one. 2012; 7(5):e37908. https://doi.org/10.1371/journal.pone.0037908 PMID: 22662248

33. Branscum AJ, Gardner IA, Johnson WO. Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. Preventive veterinary medicine. 2005; 68(2–4):145–63. https://doi.org/10.1016/j.prevetmed.2004.12.005 PMID: 15820113

34. Reitsma JBR A. W. S.; Khan K. S.; Coomarasamy A.; Bossuyt P. M. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. Journal of Clinical Epidemiology. 2009; 62(8):797–806. https://doi.org/10.1016/j.jclinepi.2009.02.005 PMID: 19447581

35. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ (Clinical research ed). 2009;339. https://doi.org/10.1136/bmj.b2700 PMID: 19622552

36. Tips Sayers A. and tricks in performing a systematic review. Br J Gen Pract. 2008; 58(547):136–.

37. Harel OZ X. H. Multiple imputation for correcting verification bias. Statistics in Medicine. 2006; 25(22):3769–86. https://doi.org/10.1002/sim.2494 WOS:000242429400001. PMID: 16435337

38. He H, McDermott MP. A robust method using propensity score stratification for correcting verification bias for binary tests. Biostatistics. 2012; 13(1):32–47. https://doi.org/10.1093/biostatistics/kxr020 PMID: 21856650

39. Zhou XH. Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. Communications in Statistics—Theory and Methods. 1993; 22(11):3177–98. https://doi.org/10.1080/03610929308831209

40. Kosinski AS, Barnhart HX. Accounting for nonignorable verification bias in assessment of diagnostic tests. Biometrics. 2003; 59(1):163–71. https://doi.org/10.1111/1541-0420.00019 PMID: 12762453

41. Kosinski AS, Barnhart HX. A global sensitivity analysis of performance of a medical diagnostic test when verification bias is present. Statistics in Medicine. 2003; 22(17):2711–21. https://doi.org/10.1002/sim.1517 PMID: 12939781

42. Martinez EZAA J.; Louzada-Neto F. Estimators of sensitivity and specificity in the presence of verification bias: A Bayesian approach. Computational Statistics and Data Analysis. 2006; 51(2):601–11. https://doi.org/10.1016/j.csda.2005.12.021

43. Buzoianu M, Kadane JB. Adjusting for verification bias in diagnostic test evaluation: A Bayesian approach. Statistics in Medicine. 2008; 27(13):2453–73. https://doi.org/10.1002/sim.3099 PMID: 17979150

44. Hajivandi A, Shirazi HRG, Saadat SH, Chehrazi M. A Bayesian analysis with informative prior on disease prevalence for predicting missing values due to verification bias. Open Access Macedonian Journal of Medical Sciences. 2018; 6(7):1225–30. https://doi.org/10.3889/oamjms.2018.296 PMID: 30087725

45. Zhou XH. Comparing accuracies of two screening tests in a two-phase study for dementia. Journal of the Royal Statistical Society Series C: Applied Statistics. 1998; 47(1):135–47.

46. Lloyd CJ, Frommer DJ. An application of multinomial logistic regression to estimating performance of a multiple-screening test with incomplete verification. Journal of the Royal Statistical Society Series C-Applied Statistics. 2008; 57:89–102. https://doi.org/10.1111/j.1467-9876.2007.00602.x WOS:000252330800006.

47. Albert PS. Imputation approaches for estimating diagnostic accuracy for multiple tests from partially verified designs. Biometrics. 2007; 63(3):947–57. https://doi.org/10.1111/j.1541-0420.2006.00734.x PMID: 17825024

48. Albert PS, Dodd LE. On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. Journal of the American Statistical Association. 2008; 103(481):61–73. https://doi.org/10.1198/016214507000000329 WOS:000254311500014. PMID: 19802353

49. Martinez EZ, Achcar JA, Louzada-Neto F. Bayesian estimation of diagnostic tests accuracy for semi-latent data with covariates. Journal of Biopharmaceutical Statistics. 2005; 15(5):809–21. https://doi.org/10.1081/BIP-200067912 PMID: 16078387

50. Xue X, Kim MY, Castle PE, Strickler HD. A new method to address verification bias in studies of clinical screening tests: Cervical cancer screening assays as an example. Journal of Clinical Epidemiology. 2014; 67(3):343–53. https://doi.org/10.1016/j.jclinepi.2013.09.013 PMID: 24332397

51. Walter SD. Estimation of test sensitivity and specificity when disease confirmation is limited to positive results. Epidemiology. 1999:67–72. PMID: 9888282

52. Böhning D, Patilea V. A capture–recapture approach for screening using two diagnostic tests with availability of disease status for the test positives only. Journal of the American Statistical Association. 2008; 103(481):212–21.

53. Chu HZ, Yijie; Cole, Stephen R.; Ibrahim, Joseph G. On the estimation of disease prevalence by latent class models for screening studies using two screening tests with categorical disease status verified in test positives only. Statistics in Medicine. 2010; 29(11):1206–18. https://doi.org/10.1002/sim.3862 PMID: 20191614

54. Baker SG. Evaluating multiple diagnostic tests with partial verification. Biometrics. 1995; 51(1):330–7. https://doi.org/10.2307/2533339 PMID: 7539300

55. Van Geloven NB K. A.; Opmeer B. C.; Mol B. W.; Zwinderman A. H. How to deal with double partial verification when evaluating two index tests in relation to a reference test? Statistics in Medicine. 2012; 31(11–12):1265–76. https://doi.org/10.1002/sim.4440 PMID: 22161741

56. Van Geloven N, Broeze KA, Opmeer BC, Mol BW, Zwinderman AH. Correction: How to deal with double partial verification when evaluating two index tests in relation to a reference test? Statistics in Medicine. 2012; 31(28):3787–8.

57. Aragon DC, Martinez EZ, Alberto Achcar J. Bayesian estimation for performance measures of two diagnostic tests in the presence of verification bias. Journal of biopharmaceutical statistics. 2010; 20(4):821–34. https://doi.org/10.1080/10543401003618868 PMID: 20496208

58. Gray R, Begg CB, Greenes RA. Construction of receiver operating characteristic curves when disease verification is subject to selection bias. Medical Decision Making. 1984; 4(2):151–64. https://doi.org/10.1177/0272989X8400400204 PMID: 6472063

59. Zhou XH. A nonparametric maximum likelihood estimator for the receiver operating characteristic curve area in the presence of verification bias. Biometrics. 1996; 52(1):299–305. https://doi.org/10.2307/2533165 PMID: 8934599

60. Rodenberg C, Zhou XH. ROC curve estimation when covariates affect the verification process. Biometrics. 2000; 56(4):1256–62. https://doi.org/10.1111/j.0006-341X.2000.01256.x PMID: 11129488

61. Zhou XH, Rodenberg CA. Estimating an ROC curve in the presence of non-ignorable verification bias. Communications in Statistics—Theory and Methods. 1998; 27(3):635–57. https://doi.org/10.1080/03610929808832118

62. Hunink MG, Richardson DK, Doubilet PM, Begg CB. Testing for fetal pulmonary maturity: ROC analysis involving covariates, verification bias, and combination testing. Medical Decision Making. 1990; 10 (3):201–11. https://doi.org/10.1177/0272989X9001000307 PMID: 2370827

63. He HL, Jeffrey M.; McDermott, Michael P. Direct estimation of the area under the receiver operating characteristic curve in the presence of verification bias. Statistics in Medicine. 2009; 28(3):361–76. https://doi.org/10.1002/sim.3388 PMID: 18680124

64. Adimari G, Chiogna M. Nearest-neighbor estimation for ROC analysis under verification bias. International Journal of Biostatistics. 2015; 11(1):109–24. https://doi.org/10.1515/ijb-2014-0014 PMID: 25781712

65. Adimari G, Chiogna M. Nonparametric verification bias-corrected inference for the area under the ROC curve of a continuous-scale diagnostic test. Statistics and its Interface. 2017; 10(4):629–41. https://doi.org/10.4310/SII.2017.v10.n4.a8

66. Gu J, Ghosal S, Kleiner DE. Bayesian ROC curve estimation under verification bias. Statistics in Medicine. 2014; 33(29):5081–96. https://doi.org/10.1002/sim.6297 PMID: 25269427

67. Fluss RR, Benjamin; Faraggi, David; Rotnitzky, Andrea. Estimation of the ROC Curve under Verification Bias. Biometrical Journal. 2009; 51(3):475–90. https://doi.org/10.1002/bimj.200800128 PMID: 19588455

68. Rotnitzky A, Faraggi D, Schisterman E. Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias. Journal of the American Statistical Association. 2006; 101(475):1276–88.

69. Fluss R, Reiser B, Faraggi D. Adjusting ROC curves for covariates in the presence of verification bias. Journal of Statistical Planning and Inference. 2012; 142(1):1–11.

70. Liu DZ, Xiao-Hua. A Model for Adjusting for Nonignorable Verification Bias in Estimation of the ROC Curve and Its Area with Likelihood-Based Approach. Biometrics. 2010; 66(4):1119–28. https://doi.org/10.1111/j.1541-0420.2010.01397.x PMID: 20222937

71. Yu W, Kim JK, Park T. Estimation of area under the ROC Curve under nonignorable verification bias. Statistica Sinica. 2018; 28(4):2149–66. https://doi.org/10.5705/ss.202016.0315 PMID: 31367164

72. Page JH, Rotnitzky A. Estimation of the disease-specific diagnostic marker distribution under verification bias. Computational Statistics and Data Analysis. 2009; 53(3):707–17. https://doi.org/10.1016/j.csda.2008.06.021 PMID: 23087495

73. Liu DZ, Xiao-Hua. Covariate Adjustment in Estimating the Area Under ROC Curve with Partially Missing Gold Standard. Biometrics. 2013; 69(1):91–100. https://doi.org/10.1111/biom.12001 PMID: 23410529

74. Liu D, Zhou XH. Semiparametric Estimation of the Covariate-Specific ROC Curve in Presence of Ignorable Verification Bias. Biometrics. 2011; 67(3):906–16. https://doi.org/10.1111/j.1541-0420.2011.01562.x PMID: 21361890

75. Yu BZ, Chuan. Assessing the accuracy of a multiphase diagnosis procedure for dementia. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2012; 61(1):67–81. https://doi.org/10.1111/j.1467-9876.2011.00771.x

76. Chi Y-YZ, Xiao-Hua. Receiver operating characteristic surfaces in the presence of verification bias. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2008; 57(1):1–23. https://doi.org/10.1111/j.1467-9876.2007.00597.x

77. Duc KT, Chiogna M, Adimari G. Nonparametric Estimation of ROC Surfaces Under Verification Bias. 2016.

78. To Duc K. bcROCsurface: An R package for correcting verification bias in estimation of the ROC surface and its volume for continuous diagnostic tests. BMC Bioinformatics. 2017;18(1). https://doi.org/10.1186/s12859-016-1415-9

79. Zhang Y, Alonzo TA, for the Alzheimer's Disease Neuroimaging I. Estimation of the volume under the receiver-operating characteristic surface adjusting for non-ignorable verification bias. Statistical Methods in Medical Research. 2018; 27(3):715–39. https://doi.org/10.1177/0962280217742541 PMID: 29338546

80. Zhu R, Ghosal S. Bayesian Semiparametric ROC surface estimation under verification bias. Computational Statistics and Data Analysis. 2019; 133:40–52. https://doi.org/10.1016/j.csda.2018.09.003

81. To Duc K, Chiogna M, Adimari G, for the Alzheimer's Disease Neuroimaging I. Estimation of the volume under the ROC surface in presence of nonignorable verification bias. Statistical Methods and Applications. 2019. https://doi.org/10.1007/s10260-019-00451-3

82. De Groot JAH, Dendukuri N, Janssen KJM, Reitsma J, Bossuyt PM, Moons KGM. Adjusting for differential verification bias in diagnostic accuracy studies: A bayesian approach. American Journal of Epidemiology. 2010; 11):S140.

83. Lu YD, Nandini; Schiller Ian; Joseph Lawrence. A Bayesian approach to simultaneously adjusting for verification and reference standard bias in diagnostic test studies. Statistics in Medicine. 2010; 29 (24):2532–43. https://doi.org/10.1002/sim.4018 PMID: 20799249

84. Glueck DH, Lamb MM, O'Donnell CI, Ringham BM, Brinton JT, Muller KE, et al. Bias in trials comparing paired continuous tests can cause researchers to choose the wrong screening modality. Bmc Medical Research Methodology. 2009; 9. https://doi.org/10.1186/1471-2288-9-4 WOS:000264155400001. PMID: 19154609

85. Capelli GN A.; Nardelli S.; di Regalbono A. F.; Pietrobelli M. Validation of a commercially available cELISA test for canine neosporosis against an indirect fluorescent antibody test (IFAT). Preventive Veterinary Medicine. 2006; 73(4):315–20. https://doi.org/10.1016/j.prevetmed.2005.10.001 WOS:000236336000007. PMID: 16293328

86. Ferreccio C, Barriga MI, Lagos M, Ibáñez C, Poggi H, González F, et al. Screening trial of human papillomavirus for early detection of cervical cancer in Santiago, Chile. International Journal of Cancer. 2012; 132(4):916–23. https://doi.org/10.1002/ijc.27662 PMID: 22684726

87. Iglesias-Garriz I, Rodríguez MA, García-Porrero E, Ereño F, Garrote C, Suarez G. Emergency Non-traumatic Chest Pain: Use of Stress Echocardiography to Detect Significant Coronary Artery Stenosis. Journal of the American Society of Echocardiography. 2005; 18(11):1181–6. https://doi.org/10.1016/j.echo.2005.07.020 PMID: 16275527

88. Cronin AM, Vickers AJ. Statistical methods to correct for verification bias in diagnostic studies are inadequate when there are few false negatives: A simulation study. BMC Medical Research Methodology. 2008; 8. https://doi.org/10.1186/1471-2288-8-75 PMID: 19014457

89. de Groot JAH, Janssen KJM, Zwinderman AH, Bossuyt PMM, Reitsma JB, Moons KGM. Correcting for Partial Verification Bias: A Comparison of Methods. Annals of Epidemiology. 2011; 21(2):139–48. https://doi.org/10.1016/j.annepidem.2010.10.004 WOS:000286348200009. PMID: 21109454

90. Heida A, Van De Vijver E, Van Ravenzwaaij D, Van Biervliet S, Hummel TZ, Yuksel Z, et al. Predicting inflammatory bowel disease in children with abdominal pain and diarrhoea: Calgranulin-C versus calprotectin stool tests. Archives of Disease in Childhood. 2018; 103(6):565–71. https://doi.org/10.1136/archdischild-2017-314081 PMID: 29514815

91. Brenner H. Correcting for exposure misclassification using an alloyed gold standard. Epidemiology. 1996; 7(4):406–10. https://doi.org/10.1097/00001648-199607000-00011 PMID: 8793367

92. Gart JJ, Buck AA. COMPARISON OF A SCREENING TEST AND A REFERENCE TEST IN EPIDEMIOLOGIC STUDIES .2. A PROBABILISTIC MODEL FOR COMPARISON OF DIAGNOSTIC TESTS. American Journal of Epidemiology. 1966; 83(3):593–&. https://doi.org/10.1093/oxfordjournals.aje.a120610 WOS:A19667894500018. PMID: 5932703

93. Staquet M, Rozencweig M, Lee YJ, Muggia FM. Methodology for the assessment of new dichotomous diagnostic tests. Journal of Chronic Diseases. 1981; 34(12):599–610. https://doi.org/10.1016/0021-9681(81)90059-x PMID: 6458624

94. Albert PS. Estimating diagnostic accuracy of multiple binary tests with an imperfect reference standard. Statistics in Medicine. 2009; 28(5):780–97. https://doi.org/10.1002/sim.3514 PMID: 19101935

95. Emerson SC, Waikar SS, Fuentes C, Bonventre JV, Betensky RA. Biomarker validation with an imperfect reference: Issues and bounds. Statistical Methods in Medical Research. 2018; 27(10):2933–45. https://doi.org/10.1177/0962280216689806 PMID: 28166709

96. Thibodeau L. Evaluating diagnostic tests. Biometrics. 1981:801–4.

97. Hahn AL, Marc; Landt Olfert; Schwarz, Norbert Georg; Frickmann Hagen. Comparison of one commercial and two in-house TaqMan multiplex real-time PCR assays for detection of enteropathogenic, enterotoxigenic and enteroaggregative Escherichia coli. Tropical Medicine & International Health. 2017; 22(11):1371–6. https://doi.org/10.1111/tmi.12976 PMID: 28906580

98. Matos RN, T. F.; Braga M. M.; Siqueira W. L.; Duarte D. A.; Mendes F. M. Clinical performance of two fluorescence-based methods in detecting occlusal caries lesions in primary teeth. Caries Research. 2011; 45(3):294–302. https://doi.org/10.1159/000328673 PMID: 21625126

99. Mathews WC, Cachay ER, Caperna J, Sitapati A, Cosman B, Abramson I. Estimating the accuracy of anal cytology in the presence of an imperfect reference standard. PLoS ONE. 2010; 5(8). https://doi.org/10.1371/journal.pone.0012284 PMID: 20808869

100. Hadgu A, Dendukuri N, Hilden J. Evaluation of nucleic acid amplification tests in the absence of a perfect gold-standard test: a review of the statistical and epidemiologic issues. Epidemiology. 2005:604–12. https://doi.org/10.1097/01.ede.0000173042.07579.17 PMID: 16135935

101. Hawkins DMG J. A.; Stephenson B. Some issues in resolution of diagnostic tests using an imperfect gold standard. Statistics in Medicine. 2001; 20(13):1987–2001. https://doi.org/10.1002/sim.819 PMID: 11427955

102. Hagenaars JA. Latent structure models with direct effects between indicators: local dependence models. Sociological Methods & Research. 1988; 16(3):379–405.

103. Uebersax JS. Probit latent class analysis with dichotomous or ordered category measures: Conditional independence/dependence models. Applied Psychological Measurement. 1999; 23(4):283–97.

104. Yang I, Becker MP. Latent variable modeling of diagnostic accuracy. Biometrics. 1997:948–58. PMID: 9290225

105. Qu Y, Tan M, Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. Biometrics. 1996; 52(3):797–810. PMID: 8805757

106. Albert PS, McShane LM, Shih JH, Network USNCIBTM. Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors. Biometrics. 2001; 57(2):610–9. PMID: 11414591

107. Zhang BC Z.; Albert P. S. Estimating Diagnostic Accuracy of Raters Without a Gold Standard by Exploiting a Group of Experts. Biometrics. 2012; 68(4):1294–302. https://doi.org/10.1111/j.1541-0420.2012.01789.x PMID: 23006010

108. Xu HB, Michael A.; Craig, Bruce A. Evaluating accuracy of diagnostic tests with intermediate results in the absence of a gold standard. Statistics in Medicine. 2013; 32(15):2571–84. https://doi.org/10.1002/sim.5695 PMID: 23212851

109. Wang Z, Zhou X-H, Wang M. Evaluation of diagnostic accuracy in detecting ordered symptom statuses without a gold standard. Biostatistics. 2011; 12(3):567–81. https://doi.org/10.1093/biostatistics/kxq075 PMID: 21209155

110. Wang ZZ, Xiao-Hua. Random effects models for assessing diagnostic accuracy of traditional Chinese doctors in absence of a gold standard. Statistics in Medicine. 2012; 31(7):661–71. https://doi.org/10.1002/sim.4275 PMID: 21626532

111. Liu WZ B.; Zhang Z. W.; Chen B. J.; Zhou X. H. A pseudo-likelihood approach for estimating diagnostic accuracy of multiple binary medical tests. Computational Statistics & Data Analysis. 2015; 84:85–98. https://doi.org/10.1016/j.csda.2014.11.006 WOS:000348263200007.

112. Xue X, Oktay M, Goswami S, Kim MY. A method to compare the performance of two molecular diagnostic tools in the absence of a gold standard. Statistical Methods in Medical Research. 2019; 28(2):419–31. https://doi.org/10.1177/0962280217726804 PMID: 28814156

113. Nérette P, Stryhn H, Dohoo I, Hammell L. Using pseudogold standards and latent-class analysis in combination to evaluate the accuracy of three diagnostic tests. Preventive veterinary medicine. 2008; 85(3–4):207–25. https://doi.org/10.1016/j.prevetmed.2008.01.011 PMID: 18355935

114. Dendukuri N, Hadgu A, Wang L. Modeling conditional dependence between diagnostic tests: a multiple latent variable model. Statistics in medicine. 2009; 28(3):441–61. https://doi.org/10.1002/sim.3470 PMID: 19067379

115. Johnson WO, Gastwirth JL, Pearson LM. Screening without a "gold standard": The Hui-Walter paradigm revisited. American Journal of Epidemiology. 2001; 153(9):921–4. https://doi.org/10.1093/aje/153.9.921 PMID: 11323324

116. Martinez EZL-N F.; Derchain S. F. M.; Achcar J. A.; Gontijo R. C.; Sarian L. O. Z.; Syrjänen K. J. Bayesian estimation of performance measures of cervical cancer screening tests in the presence of covariates and absence of a gold standard. Cancer Informatics. 2008; 6:33–46. PMID: 19259401

117. Zhang J, Cole SR, Richardson DB, Chu H. A Bayesian approach to strengthen inference for case-control studies with multiple error-prone exposure assessments. Statistics in medicine. 2013; 32(25):4426–37. https://doi.org/10.1002/sim.5842 PMID: 23661263

118. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2002; 64(4):583–639.

119. Pereira da Silva HD, Ascaso C, Gonçalves AQ, Orlandi PP, Abellana R. A Bayesian approach to model the conditional correlation between several diagnostic tests and various replicated subjects measurements. Statistics in Medicine. 2017; 36(20):3154–70. https://doi.org/10.1002/sim.7339 PMID: 28543307

120. Zhou X-HC, Pete; Zhou Chuan. Nonparametric Estimation of ROC Curves in the Absence of a Gold Standard. Biometrics. 2005; 61(2):600–9. https://doi.org/10.1111/j.1541-0420.2005.00324.x PMID: 16011710

121. Henkelman RM, Kay I, Bronskill MJ. Receiver operator characteristic (ROC) analysis without truth. Medical Decision Making. 1990; 10(1):24–9. https://doi.org/10.1177/0272989X9001000105 PMID: 2325524

122. Beiden SV, Campbell G, Meier KL, Wagner RF, editors. The problem of ROC analysis without truth: The EM algorithm and the information matrix. Medical Imaging 2000: Image Perception and Performance; 2000: International Society for Optics and Photonics.

123. Choi YK, Johnson WO, Collins MT, Gardner IA. Bayesian inferences for receiver operating characteristic curves in the absence of a gold standard. Journal of Agricultural, Biological, and Environmental Statistics. 2006; 11(2):210–29. https://doi.org/10.1198/108571106X110883

124. Wang C, Turnbull BW, Gröhn YT, Nielsen SS. Nonparametric estimation of ROC curves based on Bayesian models when the true disease state is unknown. Journal of Agricultural, Biological, and Environmental Statistics. 2007; 12(1):128–46. https://doi.org/10.1198/108571107X178095

125. Branscum AJJ, Wesley O.; Hanson, Timothy E.; Gardner, Ian A. Bayesian semiparametric ROC curve estimation and disease diagnosis. Statistics in Medicine. 2008; 27(13):2474–96. https://doi.org/10.1002/sim.3250 PMID: 18300333

126. Erkanli AS, Minje; Jane Costello E.; Angold Adrian. Bayesian semi-parametric ROC analysis. Statistics in Medicine. 2006; 25(22):3905–28. https://doi.org/10.1002/sim.2496 PMID: 16416403

127. García Barrado L, Coart E, Burzykowski T. Development of a diagnostic test based on multiple continuous biomarkers with an imperfect reference test. Statistics in Medicine. 2016; 35(4):595–608. https://doi.org/10.1002/sim.6733 PMID: 26388206

128. Coart E, Barrado LG, Duits FH, Scheltens P, Van Der Flier WM, Teunissen CE, et al. Correcting for the Absence of a Gold Standard Improves Diagnostic Accuracy of Biomarkers in Alzheimer's Disease. Journal of Alzheimer's Disease. 2015; 46(4):889–99. https://doi.org/10.3233/JAD-142886 PMID: 25869788

129. Jafarzadeh SR, Johnson WO, Gardner IA. Bayesian modeling and inference for diagnostic accuracy and probability of disease based on multiple diagnostic biomarkers with and without a perfect reference standard. Statistics in Medicine. 2016; 35(6):859–76. https://doi.org/10.1002/sim.6745 PMID: 26415924

130. Hwang BS, Chen Z. An Integrated Bayesian Nonparametric Approach for Stochastic and Variability Orders in ROC Curve Estimation: An Application to Endometriosis Diagnosis. Journal of the American Statistical Association. 2015; 110(511):923–34. https://doi.org/10.1080/01621459.2015.1023806 PMID: 26839441

131. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. Statistics in Medicine. 1999; 18(22):2987–3003. https://doi.org/10.1002/(sici)1097-0258(19991130)18:22<2987::aid-sim205>3.0.co;2-b PMID: 10544302

132. Schiller IvS M.; Hadgu A.; Libman M.; Reitsma J. B.; Dendukuri N. Bias due to composite reference standards in diagnostic accuracy studies. Statistics in Medicine. 2016; 35(9):1454–70. https://doi.org/10.1002/sim.6803 PMID: 26555849

133. Naaktgeboren CA, Bertens LC, van Smeden M, de Groot JA, Moons KG, Reitsma JB. Value of composite reference standards in diagnostic research. Bmj. 2013; 347:f5605. https://doi.org/10.1136/bmj.f5605 PMID: 24162938

134. Tang S, Hemyari P, Canchola JA, Duncan J. Dual composite reference standards (dCRS) in molecular diagnostic research: A new approach to reduce bias in the presence of Imperfect reference. Journal of Biopharmaceutical Statistics. 2018; 28(5):951–65. https://doi.org/10.1080/10543406.2018.1428613 PMID: 29355450

135. Bertens LC, Broekhuizen BD, Naaktgeboren CA, Rutten FH, Hoes AW, van Mourik Y, et al. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. PLoS medicine. 2013; 10(10):e1001531. https://doi.org/10.1371/journal.pmed.1001531 PMID: 24143138

136. Juhl DV A.; Luhm J.; Ziemann M.; Hennig H.; Görg S. Comparison of the two fully automated anti-HCMV IgG assays: Abbott Architect CMV IgG assay and Biotest anti-HCMV recombinant IgG ELISA. Transfusion Medicine. 2013; 23(3):187–94. https://doi.org/10.1111/tme.12036 PMID: 23578169

137. Rostami MNR B. H.; Aghsaghloo F.; Nazari R. Comparison of clinical performance of antigen based-enzyme immunoassay (EIA) and major outer membrane protein (MOMP)-PCR for detection of genital Chlamydia trachomatis infection. International Journal of Reproductive Biomedicine. 2016; 14(6):411–20. WOS:000388374300007. PMID: 27525325

138. Spada EP Daniela; Baggiani Luciana; Bagnagatti De Giorgi Giada; Perego Roberta; Ferro Elisabetta. Evaluation of an immunochromatographic test for feline AB system blood typing. Journal of Veterinary Emergency and Critical Care. 2016; 26(1):137–41. https://doi.org/10.1111/vec.12360 PMID: 26264678

139. Brocchi E, Bergmann IE, Dekker A, Paton DJ, Sammin DJ, Greiner M, et al. Comparative evaluation of six ELISAs for the detection of antibodies to the non-structural proteins of foot-and-mouth disease virus. Vaccine. 2006; 24(47):6966–79. https://doi.org/https://doi.org/10.1016/j.vaccine.2006.04.050

140. Williams GJM, Petra; Kerr Marianne; Fitzgerald Dominic A.; Isaacs David; Codarini Miriam; McCaskill Mary; Prelog Kristina; Craig Jonathan C. Variability and accuracy in interpretation of consolidation on chest radiography for diagnosing pneumonia in children under 5 years of age. Pediatric Pulmonology. 2013; 48(12):1195–200. https://doi.org/10.1002/ppul.22806 PMID: 23997040

141. Asselineau J, Paye A, Bessède E, Perez P, Proust-Lima C. Different latent class models were used and evaluated for assessing the accuracy of campylobacter diagnostic tests: Overcoming imperfect reference standards? Epidemiology and Infection. 2018; 146(12):1556–64. https://doi.org/10.1017/S0950268818001723 PMID: 29945689

142. Sobotzki CR M.; Kennerknecht N.; Hulsse C.; Littmann M.; White A.; Von Kries R.; Von Kotnig C. H. W. Latent class analysis of diagnostic tests for adenovirus, Bordetella pertussis and influenza virus infections in German adults with longer lasting coughs. Epidemiology and Infection. 2016; 144(4):840–6. https://doi.org/10.1017/S0950268815002149 WOS:000369712100021. PMID: 26380914

143. Poynard TDL V.; Zarski J. P.; Stanciu C.; Munteanu M.; Vergniol J.; France J.; Trifan A.; Le Naour G.; Vaillant J. C.; Ratziu V.; Charlotte F. Relative performances of FibroTest, Fibroscan, and biopsy for the assessment of the stage of liver fibrosis in patients with chronic hepatitis C: A step toward the truth in the absence of a gold standard. Journal of Hepatology. 2012; 56(3):541–8. https://doi.org/10.1016/j.jhep.2011.08.007 PMID: 21889468

144. De La Rosa GDV M. L.; Arango C. M.; Gomez C. I.; Garcia A.; Ospina S.; Osorno S.; Henao A.; Jaimes F. A. Toward an operative diagnosis in sepsis: A latent class approach. BMC Infectious Diseases. 2008; 8 (no pagination)(18).

145. Xie YC, Zhen; Albert Paul S. A crossed random effects modeling approach for estimating diagnostic accuracy from ordinal ratings without a gold standard. Statistics in Medicine. 2013; 32(20):3472–85. https://doi.org/10.1002/sim.5784 PMID: 23529923

146. See CWA W.; Melese M.; Zhou Z.; Porco T. C.; Shiboski S.; Gaynor B. D.; Eng J.; Keenan J. D.; Lietman T. M. How reliable are tests for trachoma?—A latent class approach. Investigative Ophthalmology and Visual Science. 2011; 52(9):6133–7. https://doi.org/10.1167/iovs.11-7419 PMID: 21685340

147. Nérette P, Dohoo I, Hammell L. Estimation of specificity and sensitivity of three diagnostic tests for infectious salmon anaemia virus in the absence of a gold standard. Journal of Fish Diseases. 2005; 28(2):89–99. https://doi.org/10.1111/j.1365-2761.2005.00612.x PMID: 15705154

148. Pak SIK D. Evaluation of diagnostic performance of a polymerase chain reaction for detection of canine Dirofilaria immitis. Journal of Veterinary Clinics. 2007; 24(2):77–81.

149. Jokinen J, Snellman M, Palmu AA, Saukkoriipi A, Verlant V, Pascal T, et al. Testing Pneumonia Vaccines in the Elderly: Determining a Case Definition for Pneumococcal Pneumonia in the Absence of a Gold Standard. American Journal of Epidemiology. 2018; 187(6):1295–302. https://doi.org/10.1093/aje/kwx373 PMID: 29253067

150. Santos FLN, Campos ACP, Amorim LDAF, Silva ED, Zanchin NIT, Celedon PAF, et al. Highly accurate chimeric proteins for the serological diagnosis of chronic chagas disease: A latent class analysis. American Journal of Tropical Medicine and Hygiene. 2018; 99(5):1174–9. https://doi.org/10.4269/ajtmh.17-0727 PMID: 30226130

151. Mamtani M, Jawahirani A, Das K, Rughwani V, Kulkarni H. Bias-corrected diagnostic performance of the naked eye single tube red cell osmotic fragility test (NESTROFT): An effective screening tool for β-thalassemia. Hematology. 2006; 11(4):277–86. https://doi.org/10.1080/10245330600915875 PMID: 17178668

152. Karaman BF, Açıkalın A, Ünal İ, Aksungur VL. Diagnostic values of KOH examination, histological examination, and culture for onychomycosis: a latent class analysis. International Journal of Dermatology. 2019; 58(3):319–24. https://doi.org/10.1111/ijd.14255 PMID: 30246397

153. Yan Q, Karau MJ, Greenwood-Quaintance KE, Mandrekar JN, Osmon DR, Abdel MP, et al. Comparison of diagnostic accuracy of periprosthetic tissue culture in blood culture bottles to that of prosthesis sonication fluid culture for diagnosis of prosthetic joint infection (PJI) by use of Bayesian latent class modeling and IDSA PJI criteria for classification. Journal of Clinical Microbiology. 2018; 56(6). https://doi.org/10.1128/JCM.00319-18 PMID: 29643202

154. Lurier T, Delignette-Muller ML, Rannou B, Strube C, Arcangioli MA, Bourgoin G. Diagnosis of bovine dictyocaulosis by bronchoalveolar lavage technique: A comparative study using a Bayesian approach. Preventive Veterinary Medicine. 2018; 154:124–31. https://doi.org/10.1016/j.prevetmed.2018.03.017 PMID: 29685436

155. Falley BN, Stamey JD, Beaujean AA. Bayesian estimation of logistic regression with misclassified covariates and response. Journal of Applied Statistics. 2018; 45(10):1756–69. https://doi.org/10.1080/02664763.2017.1391182

156. Dufour SD J.; Dubuc J.; Dendukuri N.; Hassan S.; Buczinski S. Bayesian estimation of sensitivity and specificity of a milk pregnancy-associated glycoprotein-based ELISA and of transrectal

ultrasonographic exam for diagnosis of pregnancy at 28–45 days following breeding in dairy cows. Preventive Veterinary Medicine. 2017; 140:122–33. https://doi.org/10.1016/j.prevetmed.2017.03.008 PMID: 28460745

157. Bermingham MLH I. G.; Glass E. J.; Woolliams J. A.; Bronsvoort B. M. D. C.; McBride S. H.; Skuce R. A.; Allen A. R.; McDowell S. W. J.; Bishop S. C. Hui and Walter's latent-class model extended to estimate diagnostic test properties from surveillance data: A latent model for latent data. Scientific Reports. 2015; 5. https://doi.org/10.1038/srep11861 PMID: 26148538

158. Busch EL, Don PK, Chu H, Richardson DB, Keku TO, Eberhard DA, et al. Diagnostic accuracy and prediction increment of markers of epithelial-mesenchymal transition to assess cancer cell detachment from primary tumors. BMC Cancer. 2018; 18(1). https://doi.org/10.1186/s12885-017-3964-3 PMID: 29338703

159. de Araujo Pereira GL F.; de Fatima Barbosa V.; Ferreira-Silva M. M.; Moraes-Souza H. A general latent class model for performance evaluation of diagnostic tests in the absence of a gold standard: an application to Chagas disease. Computational and mathematical methods in medicine. 2012; 2012:487502. https://doi.org/10.1155/2012/487502 PMID: 22919430

160. Hubbard RA, Huang J, Harton J, Oganisian A, Choi G, Utidjian L, et al. A Bayesian latent class approach for EHR-based phenotyping. Statistics in Medicine. 2019; 38(1):74–87. https://doi.org/10.1002/sim.7953 PMID: 30252148

161. Caraguel C, Stryhn H, Gagné N, Dohoo I, Hammell L. Use of a third class in latent class modelling for the diagnostic evaluation of five infectious salmon anaemia virus detection tests. Preventive Veterinary Medicine. 2012; 104(1):165–73. https://doi.org/10.1016/j.prevetmed.2011.10.006.

162. De Waele V, Berzano M, Berkvens D, Speybroeck N, Lowery C, Mulcahy GM, et al. Age-Stratified Bayesian Analysis To Estimate Sensitivity and Specificity of Four Diagnostic Tests for Detection of Cryptosporidium Oocysts in Neonatal Calves. Journal of Clinical Microbiology. 2011; 49(1):76–84. https://doi.org/10.1128/JCM.01424-10 WOS:000285787100010. PMID: 21048012

163. Dendukuri N, Wang LL, Hadgu A. Evaluating Diagnostic Tests for Chlamydia trachomatis in the Absence of a Gold Standard: A Comparison of Three Statistical Methods. Statistics in Biopharmaceutical Research. 2011; 3(2):385–97. https://doi.org/10.1198/sbr.2011.10005 WOS:000292680800023.

164. Habib IS I.; Uyttendaele M.; De Zutter L.; Berkvens D. A Bayesian modelling framework to estimate Campylobacter prevalence and culture methods sensitivity: application to a chicken meat survey in Belgium. Journal of Applied Microbiology. 2008; 105(6):2002–8. https://doi.org/10.1111/j.1365-2672.2008.03902.x PMID: 19120647

165. Vidal EM A.; Bertolini E.; Cambra M. Estimation of the accuracy of two diagnostic methods for the detection of Plum pox virus in nursery blocks by latent class models. Plant Pathology. 2012; 61 (2):413–22. https://doi.org/10.1111/j.1365-3059.2011.02505.x

166. Aly SSA R. J.; Whitlock R. H.; Adaska J. M. Sensitivity and Specificity of Two Enzyme-linked Immunosorbent Assays and a Quantitative Real-time Polymerase Chain Reaction for Bovine Paratuberculosis Testing of a Large Dairy Herd. International Journal of Applied Research in Veterinary Medicine. 2014; 12(1):1–7. WOS:000348617300001.

167. Rahman AKMA, Saegerman C, Berkvens D, Fretin D, Gani MO, Ershaduzzaman M, et al. Bayesian estimation of true prevalence, sensitivity and specificity of indirect ELISA, Rose Bengal Test and Slow Agglutination Test for the diagnosis of brucellosis in sheep and goats in Bangladesh. Preventive Veterinary Medicine. 2013; 110(2):242–52. https://doi.org/10.1016/j.prevetmed.2012.11.029 PMID: 23276401

168. Praet NV, Jaco J.; Mwape, Kabemba E.; Phiri Isaac K.; Muma John B.; Zulu Gideon; van Lieshout Lisette; Rodriguez-Hidalgo Richar; Benitez-Ortiz Washington; Dorny Pierre; Gabriël Sarah. Bayesian modelling to estimate the test characteristics of coprology, coproantigen ELISA and a novel real-time PCR for the diagnosis of taeniasis. Tropical Medicine & International Health. 2013; 18(5):608–14. https://doi.org/10.1111/tmi.12089 PMID: 23464616

169. Espejo LA, Zagmutt FJ, Groenendaal H, Munoz-Zanzi C, Wells SJ. Evaluation of performance of bacterial culture of feces and serum ELISA across stages of Johne's disease in cattle using a Bayesian latent class model. Journal of dairy science. 2015; 98(11):8227–39. https://doi.org/10.3168/jds.2014-8440 PMID: 26364104

170. Haley C, Wagner B, Puvanendiran S, Abrahante J, Murtaugh MP. Diagnostic performance measures of ELISA and quantitative PCR tests for porcine circovirus type 2 exposure using Bayesian latent class analysis. Preventive veterinary medicine. 2011; 101(1–2):79–88. https://doi.org/10.1016/j.prevetmed.2011.05.001 PMID: 21632130

171. Menten JB Marleen; Lesaffre Emmanuel. Bayesian latent class models with conditionally dependent diagnostic tests: A case study. Statistics in Medicine. 2008; 27(22):4469–88. https://doi.org/10.1002/sim.3317 PMID: 18551515

172. Tasony-Wagener EA. Evaluation of Antigen Detection Assays for the Avian Influenza Virus [Ph.D.]. Ann Arbor: University of Prince Edward Island (Canada); 2012.

173. Weichenthal S, Joseph L, Bélisle P, Dufresne A. Bayesian Estimation of the Probability of Asbestos Exposure from Lung Fiber Counts. Biometrics. 2010; 66(2):603–12. https://doi.org/10.1111/j.1541-0420.2009.01279.x PMID: 19508240

174. Jafarzadeh SR, Warren DK, Nickel KB, Wallace AE, Fraser VJ, Olsen MA. Bayesian estimation of the accuracy of ICD-9-CM- and CPT-4-based algorithms to identify cholecystectomy procedures in administrative data without a reference standard. Pharmacoepidemiology and Drug Safety. 2016; 25 (3):263–8. https://doi.org/10.1002/pds.3870 WOS:000371825900004. PMID: 26349484

175. García Barrado L, Coart E, Burzykowski T. Estimation of diagnostic accuracy of a combination of continuous biomarkers allowing for conditional dependence between the biomarkers and the imperfect reference-test. Biometrics. 2017; 73(2):646–55. https://doi.org/10.1111/biom.12583 PMID: 27598904

176. Jafarzadeh SR, Johnson WO, Utts JM, Gardner IA. Bayesian estimation of the receiver operating characteristic curve for a diagnostic test with a limit of detection in the absence of a gold standard. Statistics in Medicine. 2010; 29(20):2092–106.

177. Saugar JM, Merino FJ, Martin-Rabadan P, Fernandez-Soto P, Ortega S, Garate T, et al. Application of real-time PCR for the detection of Strongyloides spp. in clinical samples in a reference center in Spain. Acta tropica. 2015; 142:20–5. https://doi.org/10.1016/j.actatropica.2014.10.020 PMID: 25447829

178. Peterson LRY S. A.; Davis T. E.; Wang Z. X.; Duncan J.; Noutsios C.; Liesenfeld O.; Osiecki J. C.; Lewinski M. A. Evaluation of the cobas cdiff test for detection of toxigenic clostridium difficile in stool samples. Journal of Clinical Microbiology. 2017; 55(12):3426–36. https://doi.org/10.1128/JCM.01135-17 PMID: 28954901

179. Fiebrich HBB A. H.; Kerstens M. N.; Pijl M. E. J.; Kema I. P.; De Jong J. R.; Jager P. L.; Elsinga P. H.; Dierckx R. A. J. O.; Van Der Wal J. E.; Sluiter W. J.; De Vries E. G. E.; Links T. P. 6-[F-18]fluoro-L-dihydroxyphenylalanine positron emission tomography is superior to conventional imaging with123I-metaiodobenzylguanidine scintigraphy, computer tomography, and magnetic resonance imaging in localizing tumors causing catecholamine excess. Journal of Clinical Endocrinology and Metabolism. 2009; 94(10):3922–30. https://doi.org/10.1210/jc.2009-1054 PMID: 19622618

180. Wu HM, Cordeiro SM, Harcourt BH, Carvalho M, Azevedo J, Oliveira TQ, et al. Accuracy of real-time PCR, Gram stain and culture for Streptococcus pneumoniae, Neisseria meningitidis and Haemophilus influenzae meningitis diagnosis. BMC Infectious Diseases. 2013; 13 (1) (no pagination)(26).

181. Dendukuri N, Schiller I, De Groot J, Libman M, Moons K, Reitsma J, et al. Concerns about composite reference standards in diagnostic research. BMJ (Online). 2018;360. https://doi.org/10.1136/bmj.j5779 PMID: 29348126

182. Driesen M, Kondo Y, de Jong BC, Torrea G, Asnong S, Desmaretz C, et al. Evaluation of a novel line probe assay to detect resistance to pyrazinamide, a key drug used for tuberculosis treatment. Clinical Microbiology and Infection. 2018; 24(1):60–4. https://doi.org/10.1016/j.cmi.2017.05.026 PMID: 28587904

183. Bessède E, Asselineau J, Perez P, Valdenaire G, Richer O, Lehours P, et al. Evaluation of the diagnostic accuracy of two immunochromatographic tests detecting campylobacter in stools and their role in campylobacter infection diagnosis. Journal of Clinical Microbiology. 2018; 56(4). https://doi.org/10.1128/JCM.01567-17 PMID: 29436423

184. Alcántara R, Fuentes P, Antiparra R, Santos M, Gilman RH, Kirwan DE, et al. MODS-Wayne, a colorimetric adaptation of the Microscopic-Observation Drug Susceptibility (MODS) assay for detection of mycobacterium tuberculosis pyrazinamide resistance from sputum samples. Journal of Clinical Microbiology. 2019;57(2). https://doi.org/10.1128/JCM.01162-18 PMID: 30429257

185. Ziswiler HR, Reichenbach S, Vögelin E, Bachmann LM, Villiger PM, Jüni P. Diagnostic value of sonography in patients with suspected carpal tunnel syndrome: A prospective study. Arthritis and Rheumatism. 2005; 52(1):304–11. https://doi.org/10.1002/art.20723 PMID: 15641050

186. Taylor SA, Mallett S, Bhatnagar G, Baldwin-Cleland R, Bloom S, Gupta A, et al. Diagnostic accuracy of magnetic resonance enterography and small bowel ultrasound for the extent and activity of newly diagnosed and relapsed Crohn's disease (METRIC): a multicentre trial. The Lancet Gastroenterology and Hepatology. 2018; 3(8):548–58. https://doi.org/10.1016/S2468-1253(18)30161-4 PMID: 29914843

187. Eddyani M, Sopoh GE, Ayelo G, Brun LVC, Roux JJ, Barogui Y, et al. Diagnostic accuracy of clinical and microbiological signs in patients with skin lesions resembling buruli ulcer in an endemic region. Clinical Infectious Diseases. 2018; 67(6):827–34. https://doi.org/10.1093/cid/ciy197 PMID: 29538642

188. Lerner EB, McKee CH, Cady CE, Cone DC, Colella MR, Cooper A, et al. A consensus-based gold standard for the evaluation of mass casualty triage systems. Prehospital Emergency Care. 2015; 19 (2):267–71. https://doi.org/10.3109/10903127.2014.959222 PMID: 25290529

189. van Houten CB, de Groot JAH, Klein A, Srugo I, Chistyakov I, de Waal W, et al. A host-protein based assay to differentiate between bacterial and viral infections in preschool children (OPPORTUNITY): a double-blind, multicentre, validation study. The Lancet Infectious Diseases. 2017; 17(4):431–40. https://doi.org/10.1016/S1473-3099(16)30519-9 PMID: 28012942

190. Elliott DG, Applegate LJ, Murray AL, Purcell MK, McKibben CL. Bench-top validation testing of selected immunological and molecular Renibacterium salmoninarum diagnostic assays by comparison with quantitative bacteriological culture. Journal of Fish Diseases. 2013; 36(9):779–809. https://doi.org/10.1111/jfd.12079 PMID: 23346868

191. Bland JM, Altman DG. Validating scales and indexes. Bmj. 2002; 324(7337):606–7. https://doi.org/10.1136/bmj.324.7337.606 PMID: 11884331

192. Hsia ECS Neil; Cush John J.; Chaisson Richard E.; Matteson Eric L.; Xu Stephen; Beutler Anna; Doyle Mittie K.; Hsu Benjamin; Rahman Mahboob U. Interferon-γ release assay versus tuberculin skin test prior to treatment with golimumab, a human anti-tumor necrosis factor antibody, in patients with rheumatoid arthritis, psoriatic arthritis, or ankylosing spondylitis. Arthritis & Rheumatism. 2012; 64 (7):2068–77. https://doi.org/10.1002/art.34382 PMID: 104469597. Language: English. Entry Date: 20120717. Revision Date: 20150711. Publication Type: Journal Article.

193. Itza F, Zarza D, Salinas J, Teba F, Ximenez C. Turn-amplitude analysis as a diagnostic test for myofascial syndrome in patients with chronic pelvic pain. Pain Research and Management. 2015; 20 (2):96–100. https://doi.org/10.1155/2015/562349 PMID: 25848846

194. Booi ANM Jerome; Norton H. James; Anderson William E.; Ellis Amy C. Validation of a Screening Tool to Identify Undernutrition in Ambulatory Patients With Liver Cirrhosis. Nutrition in Clinical Practice. 2015; 30(5):683–9. https://doi.org/10.1177/0884533615587537 PMID: 26024676

195. von Heymann W, Moll H, Rauch G. Study on sacroiliac joint diagnostics: Reliability of functional and pain provocation tests. Manuelle Medizin. 2018; 56(3):239–48. https://doi.org/10.1007/s00337-018-0405-6

196. Schliep KC, Stanford JB, Chen Z, Zhang B, Dorais JK, Boiman Johnstone E, et al. Interrater and intrarater reliability in the diagnosis and staging of endometriosis. Obstetrics and Gynecology. 2012; 120(1):104–12. https://doi.org/10.1097/AOG.0b013e31825bc6cf PMID: 22914398

197. Teresa Pérez-Warnisher MTG-G; Giraldo-Cadavid Luis Fernando; Troncoso Acevedo Maria Fernanda; Rodríguez Rodríguez Paula; Carballosa de Miguel Pilar; González Mangado Nicolás. Diagnostic accuracy of nasal cannula versus microphone for detection of snoring. The Laryngoscope. 2017; 127(12):2886–90. https://doi.org/10.1002/lary.26710 PMID: 28731530

198. Soltan MA, Tsai YL, Lee PYA, Tsai CF, Chang HFG, Wang HTT, et al. Comparison of electron microscopy, ELISA, real time RT-PCR and insulated isothermal RT-PCR for the detection of Rotavirus group A (RVA) in feces of different animal species. Journal of Virological Methods. 2016; 235:99–104. https://doi.org/10.1016/j.jviromet.2016.05.006 PMID: 27180038

199. Palit ST N.; Knowles C. H.; Lunniss P. J.; Bharucha A. E.; Scott S. M. Diagnostic disagreement between tests of evacuatory function: a prospective study of 100 constipated patients. Neurogastroenterology & Motility. 2016; 28(10):1589–98. https://doi.org/10.1111/nmo.12859 PMID: 27154577

200. Alonzo TA, Brinton JT, Ringham BM, Glueck DH. Bias in estimating accuracy of a binary screening test with differential disease verification. Statistics in Medicine. 2011; 30(15):1852–64. https://doi.org/10.1002/sim.4232 PMID: 21495059

201. Naaktgeboren CAdG J. A.; van Smeden M.; Moons K. G.; Reitsma J. B. Evaluating diagnostic accuracy in the face of multiple reference standards. Annals of Internal Medicine. 2013; 159(3):195–202. https://doi.org/10.7326/0003-4819-159-3-201308060-00009 PMID: 23922065.

202. De Groot JAHB P. M. M.; Reitsma J. B.; Rutjes A. W. S.; Dendukuri N.; Janssen K. J. M.; Moons K. G. M. Verification problems in diagnostic accuracy studies: Consequences and solutions. BMJ (Online). 2011;343(7821). https://doi.org/10.1136/bmj.d4770 PMID: 21810869

203. Lu Y, Dendukuri N, Schiller I, Joseph L. A Bayesian approach to simultaneously adjusting for verification and reference standard bias in diagnostic test studies. Statistics in Medicine. 2010; 29(24):2532–43. https://doi.org/10.1002/sim.4018 PMID: 20799249

204. Dendukuri N, Wang L, Hadgu A. Evaluating diagnostic tests for Chlamydia trachomatis in the absence of a gold standard: A comparison of three statistical methods. Statistics in Biopharmaceutical Research. 2011; 3(2):385–97. https://doi.org/10.1198/sbr.2011.10005

205. 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(2):427–35. https://doi.org/10.1111/j.0006-341X.2004.00187.x PMID: 15180668

206. Pepe MS, Janes H. Insights into latent class analysis of diagnostic test performance. Biostatistics. 2006; 8(2):474–84. https://doi.org/10.1093/biostatistics/kxl038 PMID: 17085745

207. Nortunen T, Puustinen J, Luostarinen L, Huhtala H, Hänninen T. Validation of the finnish version of the montreal cognitive assessment test. Acta Neuropsychologica. 2018; 16(4):353–60. https://doi.org/10.5604/01.3001.0012.7964

208. Cheng MF, Guo YL, Yen RF, Chen YC, Ko CL, Tien YW, et al. Clinical Utility of FDG PET/CT in Patients with Autoimmune Pancreatitis: A Case-Control Study. Scientific Reports. 2018; 8(1). https://doi.org/10.1038/s41598-018-21996-5 PMID: 29483544

209. Gorman SLR S.; Melnick M. E.; Abrams G. M.; Byl N. N. Development and validation of the function in sitting test in adults with acute stroke. Journal of Neurologic Physical Therapy. 2010; 34(3):150–60. https://doi.org/10.1097/NPT.0b013e3181f0065f PMID: 20716989. Language: English. Entry Date: 20101004. Revision Date: 20150818. Publication Type: Journal Article.

210. Young GP, Senore C, Mandel JS, Allison JE, Atkin WS, Benamouzig R, et al. Recommendations for a step-wise comparative approach to the evaluation of new screening tests for colorectal cancer. Cancer. 2016; 122(6):826–39. https://doi.org/10.1002/cncr.29865 PMID: 26828588

211. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. Journal of clinical epidemiology. 2005; 58(8):859–62. https://doi.org/10.1016/j.jclinepi.2004.12.009 PMID: 16018921

212. Cheng D, Branscum AJ, Johnson WO. Sample size calculations for ROC studies: parametric robustness and Bayesian nonparametrics. Statistics in Medicine. 2012; 31(2):131–42. https://doi.org/10.1002/sim.4396 PMID: 22139729

213. Branscum AJ, Johnson WO, Gardner IA. Sample size calculations for studies designed to evaluate diagnostic test accuracy. Journal of agricultural, biological, and environmental statistics. 2007; 12 (1):112.

214. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. Journal of biomedical informatics. 2014; 48:193–204. https://doi.org/10.1016/j.jbi.2014.02.013 PMID: 24582925

215. Dendukuri N, Rahme E, Bélisle P, Joseph L. Bayesian sample size determination for prevalence and diagnostic test studies in the absence of a gold standard test. Biometrics. 2004; 60(2):388–97. https://doi.org/10.1111/j.0006-341X.2004.00183.x PMID: 15180664

216. Schmidt RL, Factor RE. Understanding sources of bias in diagnostic accuracy studies. Archives of pathology & laboratory medicine. 2013; 137(4):558–65.

217. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Grp Q-S. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. Journal of Clinical Epidemiology. 2013; 66 (10):1093–104. https://doi.org/10.1016/j.jclinepi.2013.05.014 WOS:000324086000005. PMID: 23958378

218. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy. A systematic review. Annals of internal medicine. 2004; 140(3):189–202. https://doi.org/10.7326/0003-4819-140-3-200402030-00010 PMID: 14757617

219. 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). https://doi.org/10.1136/bmjopen-2016-012799 PMID: 28137831

220. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Bmj-British Medical Journal. 2015;351. https://doi.org/10.1136/bmj.h5527 WOS:000364152500002. PMID: 26511519

221. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. Croatian Medical Journal. 2003; 44(5):639–50. WOS:000186027500024. PMID: 14515429

222. Kostoulas P, Nielsen SS, Branscum AJ, Johnson WO, Dendukuri N, Dhand NK, et al. Reporting guidelines for diagnostic accuracy studies that use Bayesian latent class models (STARD-BLCM). Statistics in Medicine. 2017; 36(23):3603–4. https://doi.org/10.1002/sim.7316 PMID: 28675923

223. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Croatian Medical Journal. 2003; 44(5):635–8. WOS:000186027500023. PMID: 14515428

224. Jackson R, Ameratunga S, Broad J, Connor J, Lethaby A, Robb G, et al. The GATE frame: critical appraisal with pictures. BMJ Evidence-Based Medicine. 2006; 11(2):35–8.

225. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC medical research methodology. 2003; 3(1):25.

226. Nofuentes JAR, Del Castillo JDL. Comparing the likelihood ratios of two binary diagnostic tests in the presence of partial verification. Biometrical Journal. 2005; 47(4):442–57. https://doi.org/10.1002/bimj.200410134 PMID: 16161803

227. Nofuentes JAR, Del Castillo JdDL. Comparison of the likelihood ratios of two binary diagnostic tests in paired designs. Statistics in Medicine. 2007; 26(22):4179–201. https://doi.org/10.1002/sim.2850 PMID: 17357992

228. Nofuentes JAR, Del Castillo JDL. EM algorithm for comparing two binary diagnostic tests when not all the patients are verified. Journal of Statistical Computation and Simulation. 2008; 78(1):19–35. https://doi.org/10.1080/10629360600938102

229. Nofuentes JARDC J. D. L.; Marzo P. F. Computational methods for comparing two binary diagnostic tests in the presence of partial verification of the disease. Computational Statistics. 2009; 24(4):695–718. https://doi.org/10.1007/s00180-009-0155-y WOS:000271540200008.

230. Nofuentes JARDC J. D. L.; Jimenez A. E. M. Comparison of the accuracy of multiple binary tests in the presence of partial disease verification. Journal of Statistical Planning and Inference. 2010; 140 (9):2504–19. https://doi.org/10.1016/j.jspi.2010.02.026 WOS:000278398300016.

231. Marin-Jimenez AE, Roldan-Nofuentes JA. Global hypothesis test to compare the likelihood ratios of multiple binary diagnostic tests with ignorable missing data. Sort-Statistics and Operations Research Transactions. 2014; 38(2):305–23. WOS:000346689100011.

232. Harel O, Zhou XH. Multiple imputation for the comparison of two screening tests in two-phase Alzheimer studies. Statistics in Medicine. 2007; 26(11):2370–88. https://doi.org/10.1002/sim.2715 PMID: 17054089

233. Zhou XH, Castelluccio P. Nonparametric analysis for the ROC areas of two diagnostic tests in the presence of nonignorable verification bias. Journal of Statistical Planning and Inference. 2003; 115 (1):193–213. https://doi.org/10.1016/S0378-3758(02)00146-5

234. Wang C, Turnbull BW, Nielsen SS, Grohn YT. Bayesian analysis of longitudinal Johne's disease diagnostic data without a gold standard test. Journal of Dairy Science. 2011; 94(5):2320–8. https://doi.org/10.3168/jds.2010-3675 WOS:000289789000017. PMID: 21524521

235. Masaebi F, Zayeri F, Nasiri M, Azizmohammadlooha M. Contrastive analysis of diagnostic tests evaluation without gold standard: Review article. Tehran University Medical Journal. 2019; 76(11):708–14.

236. Beeley C. Web application development with R using Shiny: Packt Publishing Ltd; 2013.

237. Lim C, Wannapinij P, White L, Day NP, Cooper BS, Peacock SJ, et al. Using a web-based application to define the accuracy of diagnostic tests when the gold standard is imperfect. PloS one. 2013; 8(11): e79489. https://doi.org/10.1371/journal.pone.0079489 PMID: 24265775