BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Comparative diagnostic accuracy studies with an imperfect reference standard – a comparison of correction methods

Chinyereugo M. Umemneku Chikere[1*], Kevin J. Wilson[2], A. Joy Allen[3] and Luke Vale[1]

Abstract

Background: Staquet et al. and Brenner both developed correction methods to estimate the sensitivity and specificity of a binary-response index test when the reference standard is imperfect and its sensitivity and specificity are known. However, to our knowledge, no study has compared the statistical properties of these methods, despite their long application in diagnostic accuracy studies.

Aim: To compare the correction methods developed by Staquet et al. and Brenner.

Methods: Simulations techniques were employed to compare the methods under assumptions that the new test and the reference standard are conditionally independent or dependent given the true disease status of an individual. Three clinical datasets were analysed to understand the impact of using each method to inform clinical decision-making.

Results: Under the assumption of conditional independence, the Staquet et al. correction method outperforms the Brenner correction method irrespective of the prevalence of disease and whether the performance of the reference standard is better or worse than the index test. However, when the prevalence of the disease is high (> 0.9) or low (< 0.1), the Staquet et al. correction method can produce illogical results (i.e. results outside [0,1]). Under the assumption of conditional dependence; both methods failed to estimate the sensitivity and specificity of the index test especially when the covariance terms between the index test and the reference standard is not close to zero.

Conclusion: When the new test and the imperfect reference standard are conditionally independent, and the sensitivity and specificity of the imperfect reference standard are known, the Staquet et al. correction method outperforms the Brenner method. However, where the prevalence of the target condition is very high or low or the two tests are conditionally dependent, other statistical methods such as latent class approaches should be considered.

Keywords: Diagnostic accuracy, Imperfect reference standard, Sensitivity, Specificity, Staquet, Brenner, Correction method

* Correspondence: cmuc1@leicester.ac.uk
[1]Population Health Science Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK
Full list of author information is available at the end of the article

Background

The diagnostic accuracy measures (sensitivity and specificity) of a new test are traditionally estimated through comparison with the best available reference standard. The reference standard is often assumed to be a *"gold standard"*, that is, "error free". However, no test is perfect and ignoring this imperfection can result in either over or underestimating the accuracy of a new test (the index test) [1].

Following the reviews by Rutjes et al. [2] and Chikere et al. [3], three statistical methods (Gart and Buck [4], Staquet et al. [5], and Brenner [6]) were identified as being appropriate to evaluate the sensitivity and specificity of a binary response index test when the sensitivity and specificity of the imperfect reference standard are known and the index test and reference standard are conditionally independent. The estimates of the sensitivity and specificity of the imperfect reference standard can be obtained from previous validation studies, experimental or field studies. The three statistical methods are referred to as *"correction methods"*, because they aim to correct the estimated sensitivity and specificity of the index test using the available information (sensitivity and specificity) of the imperfect reference standard via algebraic functions. In addition, these correction methods do not require probabilistic modelling like latent class models [7, 8]. Both the correction methods and latent class models assume that the true disease status of the participants are unknown (latent). However, the latent class models are "*probabilistic*" approaches that estimate the accuracy measures of the index test and / or the reference standard via a statistical model. In addition, the Bayesian latent class models [9] incorporate other sources of information about the parameters of interest aside the from observation to make inference about the parameters of interest.

It is possible that there are certain scenarios where one correction method is more appropriate or may outperform the other. Hence, we decided to explore these correction methods to provide recommendations to test evaluators. To our knowledge, no study has directly compared these correction methods.

Methods

Notation

Let IT and RS denote index test and reference standard respectively. The results from both tests are considered to be binary (diseased and non-diseased). The results from the participants are often classified into a two-by-two contingency table (Table 1), which displays the number of participants with each combination of the test results. Notation used in this paper is reported in Table 2.

The classical estimates of the sensitivity and specificity of the index test, assuming that the reference standard is a gold standard, are:

$$Sn_{IT} = \frac{a}{e} \qquad Sp_{IT} = \frac{d}{f} \qquad Prr = \frac{e}{N} \tag{1}$$

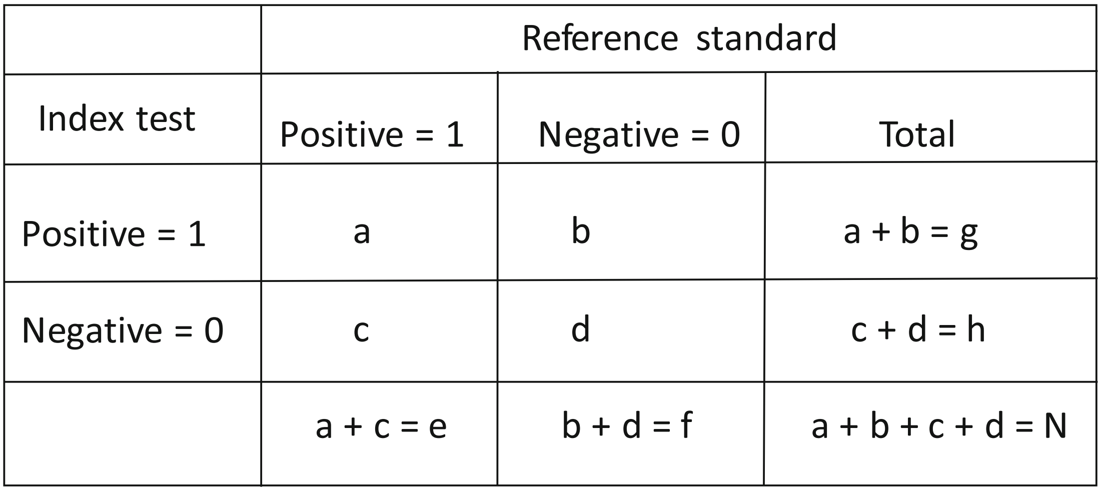
Correction methods

The correction methods identified from the systematic reviews [2, 3] were the Gart and Buck [4] correction method, the Staquet et al. [5] correction method and the Brenner [6] correction method.

Gart and Buck [4] correction method

The pair of estimators proposed by Gart and Buck [4] to estimate the sensitivity and specificity of the IT are:

Table 1 2 by 2 contingency table of the index test and imperfect reference standard



	Reference standard		
Index test	Positive = 1	Negative = 0	Total
Positive = 1	a	b	a + b = g
Negative = 0	c	d	c + d = h
	a + c = e	b + d = f	a + b + c + d = N

Table 2 Table of Notation

Notation	Meaning
Sn_{IT}	Sensitivity of the index test
Sp_{IT}	Specificity of the index test
Sn_{RS}	Sensitivity of the reference standard
Sp_{RS}	Specificity of the reference standard
Sn_{cor}	Corrected sensitivity of index test
Sp_{cor}	Corrected specificity of index test
$\hat{P}$	Estimated population prevalence
J	Youden's index, $Sn_{RS} + Sp_{RS} - 1$
Prr	Sample prevalence

$$Sn_{cor}^{GB} = \frac{Sp_{RS} \times Prr \times Sn_{IT} + (1 - Sp_{RS})(1 - Prr) \times Sp_{IT} - (1 - Sp_{RS})(Sp_{RS} - \hat{P}J)}{\hat{P}J} \tag{2}$$

$$Sp_{cor}^{GB} = \frac{Sn_{RS} \times (1 - Prr) \times Sp_{IT} + (1 - Sn_{RS}) \times Prr \times Sn_{IT} - (1 - Sn_{RS})(1 - Sp_{RS} + \hat{P}J)}{J(1 - \hat{P})} \tag{3}$$

Staquet et al. [5] correction method

Staquet et al. [5] proposed two pairs of estimators to estimate the sensitivity and specificity of the IT under two scenarios. The first pair of estimators (to estimate the sensitivity and specificity of the IT) is proposed for when the IT and RS are conditionally independent and the sensitivity and specificity of the RS are known. A second pair of estimators (to estimate the sensitivities of the IT and RS) is proposed when the specificities of the IT and RS are perfect (100%). In this paper, we focus on the first pair of estimators. This pair of estimators is employed to estimate the sensitivity and specificity of the IT given that the IT and the RS are conditionally independent and the sensitivity and specificity of the RS are known. These estimators are:

$$Sn_{cor}^{sq} = \frac{gSp_{RS} - b}{N(Sp_{RS} - 1) + e}; \quad Sp_{Cor}^{sq} = \frac{hSn_{RS} - c}{NSn_{RS} - e}; \quad \hat{P} = \frac{N(Sp_{RS} - 1) + e}{N(Sn_{RS} + Sp_{RS} - 1)} \tag{4}$$

The Staquet et al. [5] correction method is equivalent to the Gart and Buck [4] correction method (see Additional file 1).

Brenner [6] correction method

Brenner [6] proposed two pairs of estimators to estimate the sensitivity and specificity of the IT. The first pair of estimators assumes that the IT and the RS are conditionally independent and the second pair of estimators assumes that the IT and RS are conditionally dependent (positively correlated) given the true disease status of the individuals. In both pairs of estimators, the sensitivity and specificity of RS are assumed known. However, in this paper, we focus on the first pair of estimators, where the IT and RS are assumed to be conditionally independent. The first pair of estimators is expressed as:

$$Sn_{cor}^{B1} = \frac{Prr \times Sn_{RS} \times Sn_{IT} + (1 - \mathrm{Prr})(1 - \mathrm{Sp_{RS}})(1 - Sp_{IT})}{Prr \times Sn_{RS} + (1 - Prr)(1 - Sp_{RS})} \tag{5}$$

$$Sp_{cor}^{B1} = \frac{Prr \times (1 - Sn_{RS})(1 - Sn_{IT}) + (1 - Prr) \times Sp_{RS} \times Sp_{IT}}{Prr \times (1 - Sn_{RS}) + (1 - Prr) \times Sp_{RS}} \tag{6}$$

The two estimators (5) & (6) can be re-written as (7) & (8) respectively (see Additional file 1)

$$Sn_{cor}^{B1} = \frac{aSn_{RS} + b(1 - Sp_{RS})}{eSn_{RS} + f(1 - Sp_{RS})} \tag{7}$$

$$Sp_{cor}^{B1} = \frac{c(1 - Sn_{RS}) + dSp_{RS}}{e(1 - Sn_{RS}) + fSp_{RS}} \tag{8}$$

Simulation study

The correction methods were compared using simulation techniques and analysis of clinical datasets. Since the Staquet et al. [5] approach is equivalent to the Gart and Buck [4] correction method (see Additional file 1), only the Staquet et al. [5] approach was compared to the Brenner correction method. The simulation was conducted following the guidelines by Morris et al. [10] which include **P**lanning for the simulation, **C**oding and execution, **A**nalysis and **R**eporting the simulation study appropriately (PCAR), and using R-Studio statistical software [11]. In the simulation, the Staquet et al. [5] and Brenner [6] approaches were compared with the classical method [12]. The classical method assumes that the reference standard is a gold standard (Eq. (1)). The estimates obtained from the classical method will be called unadjusted estimates of sensitivity and specificity.

The performance measures in the simulation are the basic statistical properties used to ascertain a good estimator. These properties are unbiasedness, mean square error (MSE) and consistency. Further notes on these properties are presented in Additional file 1.

The fixed effects [1, 13] modelling approach was employed to simulate the different datasets using the multinomial distribution ("rmulti" function in R [14]). This approach models the pairwise conditional dependence (or correlation) between two tests among the diseased and non-diseased groups using covariance terms which are fixed across participants [15]. In the simulation process, the sensitivity and specificity of the IT and RS are known, and the prevalence of the target condition is known. The template showing how the cell probabilities were calculated using the prevalence, sensitivity and specificity of the IT and RS, and covariance terms is reported in Additional file 1 – Table S1.

The simulation study was carried out under two assumptions given that the RS is imperfect. Firstly, the IT and RS were assumed to be conditionally independent and secondly, the tests were assumed to be conditionally dependent. Theoretically, when the RS is error-free (perfect), the classical and correction methods estimate the sensitivity and specificity of the IT accurately. This is shown algebraically in Additional file 1.

Under the assumption that the RS and IT are conditionally independent, multiple (200) random samples of different sample sizes from 50 to 1000 were simulated using the multinomial distribution under three scenarios, which are:

- Scenario one: The RS is assumed to be better than the IT. This implies that the sensitivity and specificity of the RS are higher than the sensitivity and specificity of the IT.
- Scenario two: The IT is assumed to be better than the RS. This implies that the sensitivity and specificity of the IT are higher than the sensitivity and specificity of the RS.
- Scenario three: The sensitivity and specificity of RS and IT are assumed to be the same.

Scenario one

The sensitivity and specificity of the RS are assumed to be 0.9 and the sensitivity and specificity of index test are 0.8 and 0.7 respectively. The prevalence is assumed to be 0.3. The unadjusted and corrected estimates are presented in Fig. 1.

Scenario two

The sensitivity and specificity of the index test are 0.9 and the sensitivity and specificity of the reference standard are 0.8 and 0.7 respectively. The prevalence is 0.3. The unadjusted and corrected estimates are presented in Fig. 2.

Scenario three

The sensitivity and specificity of the index and reference tests are all 0.9, and the prevalence of the target condition is 0.3. The unadjusted and corrected estimates are presented in Fig. 3.

The yellow dashed lines on the plots of the mean sensitivities and mean specificities in Fig. 1 – Fig. 5 are the simulated true values of the sensitivity and specificity of the IT. The choice of parameters for the sensitivities and specificities of the IT and RS, and prevalence of the target condition were informed by clinical case studies identified from the review previously conducted by Chikere et al. [3].

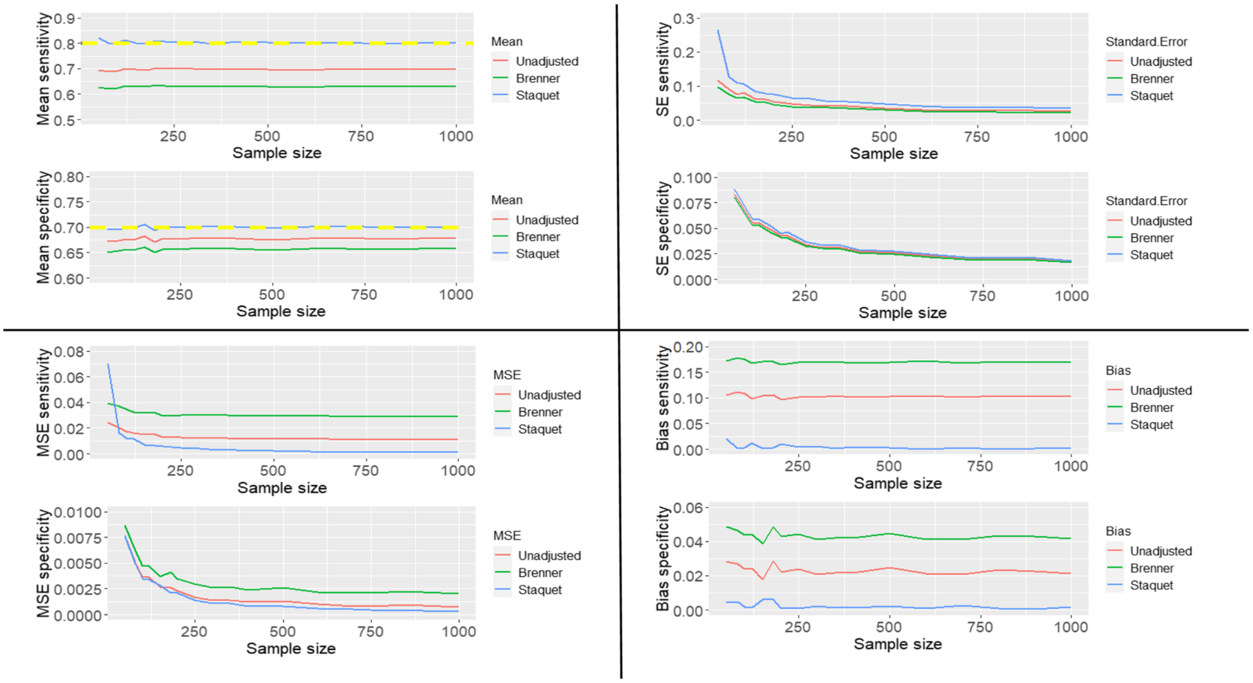


Fig. 1 The mean, standard error, mean square error and bias of the unadjusted and corrected sensitivity and specificity of the index test when the reference standard is imperfect and better than the index test

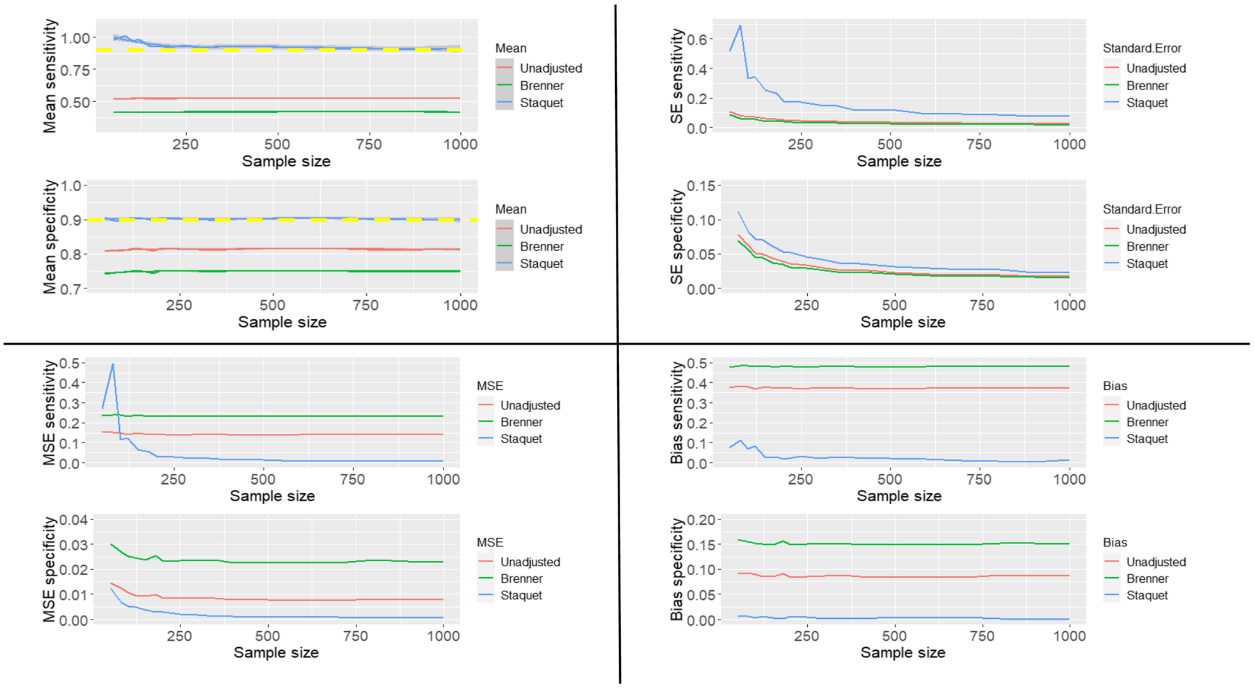


Fig. 2 The mean, standard error, mean square error and bias of the unadjusted and corrected sensitivity and specificity of the index test when the reference standard is imperfect and worse than the index test

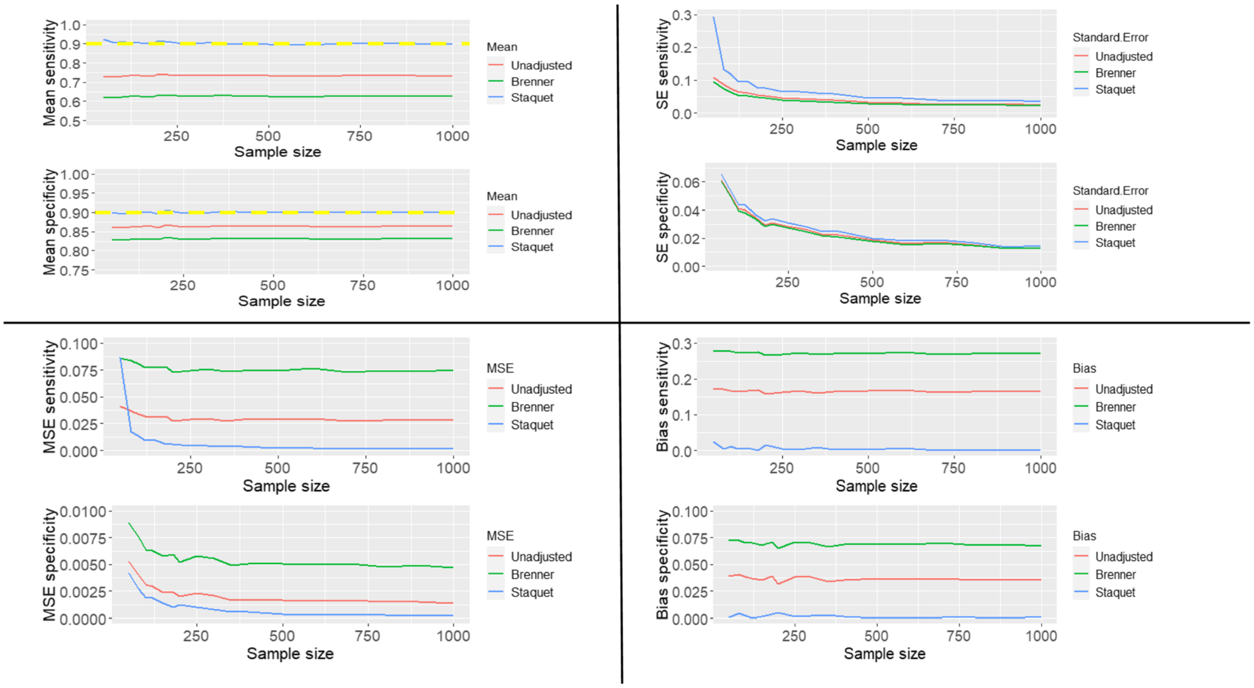


Fig. 3 The mean, standard error, mean square error and bias of the unadjusted and corrected sensitivity and specificity of the index test when the reference standard is imperfect and has same diagnostic accuracy measures as the index test

Based on the three scenarios simulated, the estimates obtained from the Staquet et al. [5] method are accurate irrespective of which test is better or worse than the other. However, when the accuracy measures of the index test are better than the sensitivity and specificity of the reference standard (Fig. 2), a relatively large sample size ($n > 200$) would be recommended; as using small sample sizes produced mean sensitivities that were slightly above the simulated true value (0.9). Practically, information about the index test is usually unknown, so using relatively a large sample size in diagnostic accuracy study is typical. The unadjusted and Brenner corrected sensitivities are consistently lower than the simulated true value and the bias is consistently greater than 0.1. The unadjusted specificities are typically slightly below the simulated true values, and the bias is typically relatively small, below 0.05, except the bias from scenario two (Fig. 2), which is larger than 0.05 but below 0.1. The Brenner corrected specificities are consistently below the simulated true values and the bias is consistently above 0.1, except in scenario one (Fig. 1), in which is below 0.05. Further scenarios explored include cases where the sensitivity (or specificity) of the index test was better than the sensitivity (or specificity) of the RS. The results of these simulations are reported in Additional file 1. With the simulated scenarios explored, (given that the IT and RS are conditionally independent), the Staquet et al. [5] correction method outperforms the Brenner correction method. There could be other possible scenarios where the sensitivity (or specificity) of the IT and RS, and prevalence is not equivalent to the values explored in this paper. These scenarios can be explored using the R-Code written by the Authors. The R-Code employed to generate and analyse the simulated and clinical datasets is presented in the appendix (Additional file 2).

In the three scenarios above, the simulation process only looked at a single prevalence ($p = 0.3$). Further analyses were carried out to explore the correction methods under varying prevalences ranging from 0 to 1 (in increments of 0.01) using scenario one, where the RS is better than the IT. The unadjusted and corrected sensitivity and specificity of the IT are presented in Fig. 4. Only scenario one was explored further because, in each scenario above, the Staquet et al. [5] correction method outperforms the Brenner correction and classical methods.

From Fig. 4 the unadjusted and Brenner corrected sensitivities tend towards the simulated true value as the prevalence tends to one. The Brenner corrected and unadjusted specificities tend to the simulated truth as the prevalence tends to zero. The Staquet et al. [5] sensitivity and specificity of the IT are approximately unbiased and equivalent to the simulated true value irrespective of the prevalence indicating a constant sensitivity and specificity across populations with different prevalences. However, when the prevalence is very low (< 0.1) or very high (> 0.9), there is the possibility of obtaining illogical estimated sensitivity or specificity via the Staquet et al. [5] approach. Illogical results imply that the estimated sensitivity or specificity is greater than one or less than zero. In the simulated datasets generated to produce Fig. 4, when the prevalence was 0.01 and 0.02 the estimated mean sensitivities were 5.38×10^{12} and 4.33×10^{12} respectively. These values are illogical and excluded from the plot. In addition, in the simulated datasets employed to plot Fig. 3, when the sample size is 50 and 80, the mean estimated sensitivities were -2.04×10 [13] and -1.94×10 [13], these values were also excluded from the plot. The simulated datasets are reproducible using the R-Code reported in

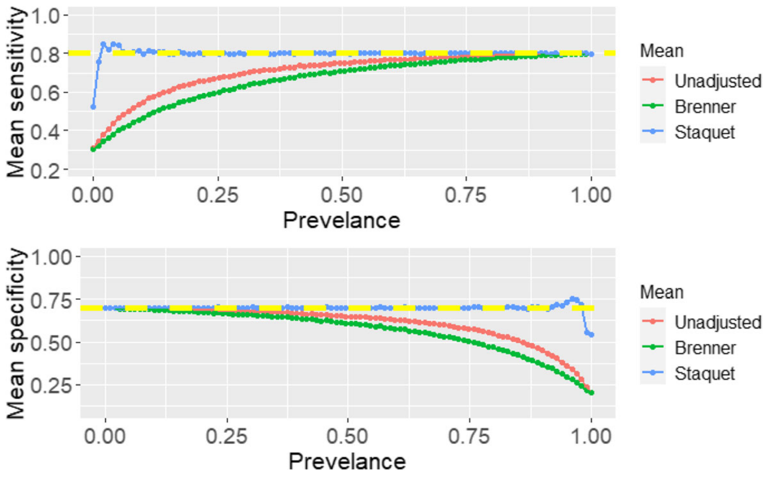


Fig. 4 Unadjusted and corrected sensitivity and specificity of the index test under varying prevalence

Additional file 2. Further exploration of illogical estimates obtained via the Staquet et al. [5] correction method are discussed in Additional file 3.

Under the assumption that the IT and RS are conditionally dependent, the covariance terms among the disease and non-disease groups are non-zero and they are varied to represent a selection of the diverse possible scenarios. A possible scenario is a case where the IT and the RS are positively correlated. In such case, the covariance terms among the diseased and non-diseased groups are positive. With the sensitivity and specificity of IT as 0.8 and the sensitivity and specificity of RS as 0.9, a scenario explored is where the covariance term among the diseased and non-diseased group is 0.05. The choice of covariance terms is constrained inequality constraint employed in generating the simulated datasets using the fixed effects modelling approach [1, 13]. The unadjusted and corrected estimates are presented in Fig. 5.

Further scenarios were explored and are reported in Additional file 4. Under the assumption of conditional dependence between IT and RS, all the correction methods performed poorly in the scenarios explored (see Additional file 4). This is expected, as the approaches were not developed to estimate the accuracy measures of the index test when the index test and the reference standard are conditionally dependent. However, when the covariance term between the disease groups is relatively small (close to zero), the Staquet et al. [5] correction method outperforms the Brenner correction method. In addition, when the IT and RS are conditionally dependent, the estimated sensitivity and specificity of the index test obtained via the Staquet et al. [5] correction method are not constant across different populations with varying prevalences, compared to the estimates obtained when the IT and RS are conditionally independent.

Analysis of three clinical datasets

Three clinical datasets from two published articles (Mathews et al. [16] and Matos et al. [17]) were analysed to understand the impact of the choice of method in clinical decision making, and to support the findings from the simulation studies. The 95% confidence intervals of the estimates obtained were calculated using the Wilson score interval approach [18].

Analysis of the clinical dataset from Mathews et al. [16] (case-study one)

The extracted clinical dataset from Mathews et al. [16] (Table *3*) aims to estimate the sensitivity and specificity of high resolution anoscopy (HRA) cytology in discriminating HIV patients into high grade squamous intraepithelial lesion (HSIL) and atypical squamous cells cannot rule out high grade (ASC-H) or not.

The punch biopsy was employed as the RS but it is known to be imperfect. According to Mathews et al. [16], the sensitivity and specificity of punch biopsy were extracted from Byrom et al. [19] and are 0.74 and 0.91 respectively. The study employed the Staquet et al. [5] approach to correct for the sensitivity and specificity of HRA cytology given that the accuracy measures of punch biopsy are known and assuming the tests (IT and RS) are conditionally independent. The dataset (Table 3) was reanalysed using the Brenner correction method. The estimated prevalence is 0.27 and the corrected and unadjusted sensitivity and specificity estimates of HRA cytology are presented in Table 4.

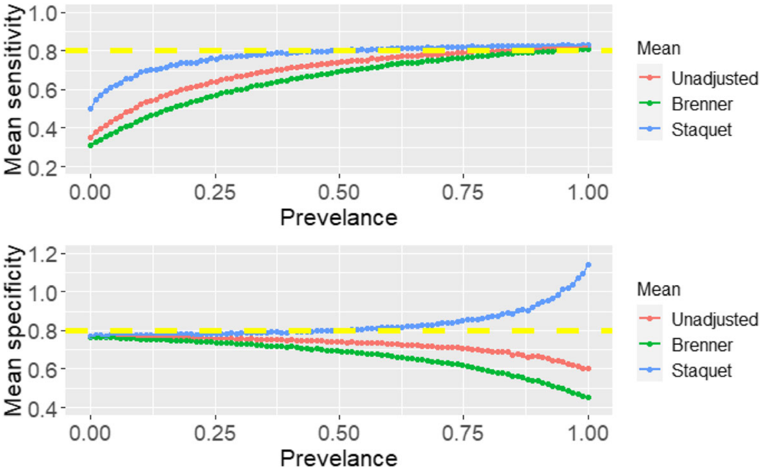


Fig. 5 The mean unadjusted and corrected sensitivity and specificity of the index test when the reference standard is imperfect and the index test and reference standard are positively correlated

Table 3 Results of HRA cytology and punch biopsy in classifying patients into high grade and non-high grade squamous intraepithelial lesion

	Biopsy ≥*AIN*2	Biopsy <*AIN*2	Total
Cytology HSIL or ASC-H	40	22	62
Cytology < HSIL	22	177	199
	62	199	261

HSIL High grade squamous intraepithelial lesion, *ASC-H* Atypical squamous cells cannot rule out high grade, *AIN* Anal intraepithelial neoplasia

From Table 4, the estimated sensitivity and specificity of HRA cytology obtained via the Staquet et al. [5] approach are higher than the estimates obtained via the classical and Brenner correction methods. In addition, no illogical estimates were obtained via the Staquet et al. [5] approach. Furthermore, the confidence intervals from the Brenner and Staquet et al. [5] correction methods do not overlap. In this clinical application, correcting the sensitivity and specificity of HRA cytology using the Brenner approach would underestimate the sensitivity of HRA cytology; thus, discouraging its use to rule out the diagnosis of HSIL as the sensitivity would appear poor (0.5). However, correcting the diagnostic accuracy measures of HRA cytology using the Staquet et al. [5] approach encourages the use of HRA cytology in clinical practice to rule in the diagnosis of HSIL as it has a specificity that is close to one.

Analysis of two clinical datasets from Matos et al. [17] dataset – (case-study two)

The extracted datasets from Matos et al. [17] reported in Table *5* and Table *6* are from Examiner 1 and the aim is to estimate the sensitivity and specificity of fluorescence – based devices (Fluorescence camera – FC and DIAG NOdent – a pen type laser fluorescence abbreviated as LFpen) used in detecting occlusal caries lesions in primary teeth. The study used the Brenner [6] correction method to estimate the sensitivities and specificities under the assumption that the sensitivity and specificity of the RS (visual inspection) are known and that the fluorescence devices are conditionally independent of the RS. The two different target conditions are non-cavitated caries lesions (NC) and dentine caries lesions (D3).

Table 5 reports the classification of the results from the index tests (FC and LFpen) and reference standard when the target condition is NC and Table 6 reports the classification of the results from the index tests (FC and LFpen) and reference standard when the target condition is D3.

Matos et al. [17] obtained the diagnostic accuracy of the reference standards from previous studies [20–24]. For the NC detection, the sensitivity and specificity of the RS were 0.796 and 0.799 respectively. For the D3, the sensitivity and specificity of the RS were 0.786 and 0.995 respectively. In addition, the teeth were assumed to be independent.

The unadjusted and corrected sensitivities and specificities of the LFpen and FC in discriminating between teeth with NC (Table 5) are presented in Table 7.

The sample prevalence for NC (0.92) and the estimated prevalence via the Staquet et al. [5] approach is 1.2 (which is illogical). An illogical prevalence is explored in Additional file 3; it was observed that when the sensitivity of the reference standard is less than the sample prevalence, illogical prevalence is likely to be obtained. The estimated sensitivities for LFpen (≅ 0.7) and FC (0.44 or 0.45) are consistent across all methods (Table 7), and the confidence intervals from the corrected and unadjusted sensitivities overlap. The specificities of LFpen and FC differ across the methods, with the Staquet et al. [5] corrected specificities (LFpen is 0.04, and FC is 0.36) being the lowest of all. At so a high prevalence and in particular when the estimated prevalence is illogical (1.2), the estimated specificity via the Staquet et al. [5] should be treated with scepticism.

The second dataset from Matos et al. [17] (Table 6) was analysed; the estimated prevalence is 0.06 and the sample prevalence is 0.052. The unadjusted and corrected sensitivity and specificity of the LFpen and FC in discriminating between teeth D3 are presented in Table 8.

From Table 8, the estimated and sample prevalence of D3 are very low (< 0.1), hence the specificities of LFpen and FC (≅ 0.9) are consistent across all methods and are estimated accurately, and the confidence intervals from the corrected and unadjusted specificities overlap. However, the sensitivities differ, with Staquet et al. [5] providing illogical estimates (estimates greater than 1). Given the illogical results, the 95% confidence intervals cannot be estimated. Obtaining illogical estimates for

Table 4 Unadjusted and corrected sensitivities and specificities of HRA cytology

Accuracy measures	Methods		
	Unadjusted (95% CI)	Brenner (95% CI)	Staquet et al (95% CI)
Sensitivity	0.65 (0.52, 0.75)	0.50(0.38, 0.62)	0.89 (0.79, 0.95)
Specificity	0.89 (0.84, 0.93)	0.85 (0.79, 0.89)	0.96 (0.92, 0.98)

CI Confidence interval

Table 5 Results of the visual inspection (reference standard) and fluorescence - based devices (LFpen and FC) in discriminating teeth with non-cavitated lesions

	Reference standard (NC – Examiner 1)			Reference standard (NC – Examiner 1)	
Index test	Positive	Negative	Index test	Positive	Negative
LFpen positive	241	6	FC positive	156	3
LFpen negative	110	26	FC positive	195	29
	351	32		351	32

the sensitivity of FC and LFpen via the Staquet et al. [5] approach are in line with the observations from the simulation study.

Discussion

Simulation study

Firstly, when the RS is perfect ($Sn_{RS} = Sp_{RS} = 1$), the estimates obtained from all the correction methods and the classical method are the same (this is expressed algebraically in Additional file 1). Secondly, when the RS is imperfect and the RS is conditionally independent of the IT, the Staquet et al. [5] correction method outperforms the Brenner correction method irrespective of which test is better. In addition, the estimates obtained via the Staquet et al. [5] correction method uphold the assumption of constant sensitivity and specificity across populations with different prevalence. This implies that the disease prevalence in the population does not affect the estimates obtained via the Staquet et al. [5] method unlike the classical and the Brenner correction methods. At a low prevalence, the estimated sensitivity from the classical and Brenner methods are often underestimated and at high prevalence the estimated specificity from the classical and Brenner correction method are typically underestimated. Thus, when there is a high prevalence, the sensitivity is more likely to be accurately estimated by all methods and at low prevalence, the specificity is likely to be accurately estimated by all methods. This is consistent with findings reported by other researchers that in a high prevalence population the sensitivity of the index test is often likely to be estimated accurately and in a low prevalence population the specificity is likely to be estimated accurately [1, 25]. Furthermore, not correcting for the imperfection of the IT using the classical method yield estimates that are closer to the simulated truth than correcting for the imperfection of the IT using the Brenner correction method. Thus, the classical method performs better than the Brenner correction method. Furthermore, when the IT and RS are conditionally dependent, both the Staquet et al. [5] and Brenner correction methods perform poorly.

Analysis of the clinical datasets

The clinical datasets explored had varying prevalences, which aid in the exploration of the methods in clinical applications. The Mathews et al. [16] dataset had a sample prevalence of 0.23, and the two datasets from Matos et al. [17] had very low (0.052) and very high (approximately 0.92) sample prevalences. Using clinical datasets with varying prevalences supported the findings from the simulation study.

The analysis of the clinical datasets alongside the simulation study have shown that the prevalence of the target condition can cause illogical estimates for the sensitivity and specificity of the IT via the Staquet et al. [5] approach. However, there could be alternative rationale that could be considered if this occurs, for example that the two tests (IT and RS) are mathematically conditionally dependent, even though the IT and RS do not use the same biological component. However, we cannot conclude that obtaining an illogical estimate via the Staquet et al. [5] approach is a sufficient condition to establish that the IT and RS are conditionally dependent given the true disease status. The information in Additional file 3 – Table S2 shows that

Table 6 Results of the visual inspection (reference standard) and fluorescence - based devices (LFpen and FC) in discriminating teeth with Dentine lesions

	Reference standard (D3 – Examiner 1)			Reference standard (D3 - Examiner 1)	
Index test	Positive	Negative	Index test	Positive	Negative
LFpen positive	20	45	FC positive	21	38
LFpen negative	1	341	FC positive	0	348
	21	386		21	386

Table 7 Unadjusted and corrected sensitivities and specificities of LFpen and FC in detection of NC

	Methods		
Accuracy measures	**Unadjusted (95% CI)**	**Brenner (95% CI)**	**Staquet et al (95% CI)**
Non-cavitated caries lesion (NC) – LFpen			
Sensitivity	0.69 (0.64, 0.73)	0.68 (0.63, 0.73)	0.70 (0.65, 0.75)
Specificity	0.81 (0.65, 0.91)	0.44 (0.28, 0.61)	0.04 (0.01, 0.17)
Non-cavitated caries lesion (NC) – FC			
Sensitivity	0.44 (0.39, 0.50)	0.44 (0.39, 0.49)	0.45 (0.40, 0.50)
Specificity	0.91 (0.76, 0.97)	0.65 (0.48, 0.79)	0.36 (0.22, 0.53)

CI Confidence interval, *LFpen* Laser florescence pen, *FC* Fluorescence camera

illogical estimates can be obtained via the Staquet et al. [5] approach when the tests are conditionally independent.

In scenarios where the estimates obtained via the Staquet et al. [5] approach are illogical, the traditional latent class model [7, 26–29] could be employed and the known sensitivity and specificity of the RS would be used as the priors (deterministic or probabilistic) in the model to estimate the accuracy measures of the IT. The latent class model has the advantage over the Staquet et al. [5] approach in that it does not produce illogical estimates. In addition, if the tests are conditionally dependent a Bayesian latent class model [15] could be considered.

One of the limitations of this study is that the traditional latent class model is not explored, as it is a probabilistic modelling approach and is not the focus of this work. In addition, the coverage probability of confidence intervals was not explored because it is a property of the procedure producing the confidence interval and not the estimators themselves, and as such is outside the scope of this study. Furthermore, there could be other scenarios (possible combinations of sensitivities and specificities of RS and IT, and prevalence) not explored in this paper; however, the R-Code (Additional file 2) written by the Authors would aid researchers who wish to explore more. A further area of research is to explore the Staquet et al. [5] approach to understand if there is a way it can indicate conditional dependence between two tests. Furthermore, as observed from the analysis of the clinical datasets (Table 5, Table 7 and Additional file 3), an illogical value was obtained for the estimated prevalence, which could have impacted the estimated specificities of the index tests. Thus, the Staquet et al. [5] approach could be further explored to ascertain other conditions that can make the Staquet et al. [5] approach produce illogical estimates, as well as possible implications where multiple conditions are satisfied simultaneously.

Conclusions

From the simulation study (using the scenarios explored in this paper) and the analysis of the clinical datasets, the Staquet et al. [5] correction method outperforms the Brenner correction method. However, when the prevalence of the target condition is very high (> 0.9) or low (< 0.1), or the tests employed in the diagnostic accuracy study are correlated (conditionally dependent) other statistical methods should be considered such as the latent class model (frequentist or Bayesian) to avoid obtaining illogical or inaccurate estimates. Furthermore, using poor estimates of the accuracy measures for the reference standard would affect the estimated corrected sensitivity and specificity of the index test.

Table 8 Unadjusted and corrected sensitivities and specificities of LFpen and FC in detecting D3

	Methods		
Accuracy measures	**Unadjusted (95% CI)**	**Brenner (95% CI)**	**Staquet et al (95% CI)**
Dentine caries lesion (D3) – LFpen			
Sensitivity	0.95 (0.77, 0.99)	0.86 (0.66, 0.95)	1.04 (NaN)
Specificity	0.88 (0.85, 0.91)	0.87 (0.83, 0.90)	0.90 (0.87, 0.93)
Dentine caries lesion (D3) – FC			
Sensitivity	1.00 (0.85, 1.00)	0.91 (0.72, 0.98)	1.09 (NaN)
Specificity	0.90 (0.87, 0.93)	0.89 (0.86, 0.92)	0.92 (0.89, 0.94)

CI Confidence interval, *LFpen* Laser florescence pen, *FC* Fluorescence camera, *NaN* Not available or cannot be estimated

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-021-01255-4.

Additional file 1.

Additional file 2.

Additional file 3.

Additional file 4.

Acknowledgements

Not applicable.

Authors' contributions

Conception and design of the work by CMUC, JAA, KW and LV. Manuscript drafted by CMUC. Simulation and analysis of dataset by CMUC. Interpretation of results by CMUC, KW, JAA and LV. Project Supervision and Revision of manuscript by JAA, KW and LV. The authors read and approved the final manuscript.

Authors' information

Not applicable.

Funding

- Newcastle University Research Excellence Award
- Population Health Science Institute, Faculty of Medical Sciences, Newcastle University
- School of Mathematics, Statistics and Physics, Newcastle University
- National Institute for Health Research, Newcastle In Vitro Diagnostics Co-operative, Newcastle University

This research is supported by the National Institute for Health Research (NIHR) Newcastle In Vitro Diagnostics Co-operative and funded by the NIHR MedTech In Vitro Diagnostic Co-operatives scheme (ref MIC-2016-014). The views expressed are those of the author(s) and not necessarily those of the NIHR, the NHS or the Department of Health and Social Care.

Availability of data and materials

Datasets generated and analysed are included in this published article [and R-Code in Additional file 2].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Population Health Science Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK. [2]School of Mathematics, Statistics and Physics, Newcastle University, Newcastle upon Tyne, UK. [3]National Institute for Health Research, Newcastle In Vitro Diagnostics Co-operative, Newcastle University, Newcastle upon Tyne, UK.

Received: 18 January 2021 Accepted: 16 March 2021
Published online: 12 April 2021

References

1. Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics. 1985;41(4):959–68. https://doi.org/10.2307/2530967.
2. Rutjes AW, Reitsma JB, Coomarasamy A, et al. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess (Winch Eng). 2007;11:iii, ix-51 Review.
3. Chikere CMU, Wilson K, Graziadio S, et al. Diagnostic test evaluation methodology: a systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard–an update. PLoS One. 2019;14:e0223832.
4. 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. Am J Epidemiol. 1966;83:593. https://doi.org/10.1093/oxfordjournals.aje.a120610.
5. Staquet M, Rozencweig M, Lee YJ, et al. Methodology for the assessment of new dichotomous diagnostic tests. J Chronic Dis. 1981;34:599–610. https://doi.org/10.1016/0021-9681(81)90059-X.
6. Brenner H. Correcting for exposure misclassification using an alloyed gold standard. Epidemiology. 1996;7:406–10 Article.
7. Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. Stat Methods Med Res. 1998;7(4):354–70Review. https://doi.org/10.1177/096228029800700404.
8. 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. https://doi.org/10.2307/2533043.
9. Branscum AJ, Gardner IA, Johnson WO. Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. Prev Vet Med. 2005; 68(2-4):145–63. https://doi.org/10.1016/j.prevetmed.2004.12.005.
10. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074–102. https://doi.org/10.1002/sim.8086.
11. Allaire J. RStudio: integrated development environment for R. Boston; 2012. p. 770.
12. Altman DG, Bland JM. Diagnostic tests 1: sensitivity and specificity. Br Med J. 1994;308:1552 Note.
13. Wang Z, Dendukuri N, Zar HJ, Joseph L. Modeling conditional dependence among multiple diagnostic tests. Stat Med. 2017;36(30):4843–59. https://doi.org/10.1002/sim.7449.
14. Team R. RStudio: integrated development for R, vol. 42. Bostonhttp://www.rstudio com: RStudio, Inc; 2015. p. 14.
15. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics. 2001;57:158–67Article. https://doi.org/10.1111/j.0006-341X.2001.00158.x.
16. Mathews WC, Cachay ER, Caperna J, et al. Estimating the accuracy of anal cytology in the presence of an imperfect reference standard. PLoS One. 2010;5Article. https://doi.org/10.1371/journal.pone.0012284.
17. Matos RN, T. F, Braga MM, Siqueira WL, Duarte DA, Mendes FM. Clinical performance of two fluorescence-based methods in detecting occlusal caries lesions in primary teeth. Caries Res. 2011;45:294–302Article. https://doi.org/10.1159/000328673.
18. Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc. 1927;22(158):209–12. https://doi.org/10.1080/01621459.1927.10502953.
19. Byrom J, Douce G, Jones P, et al. Should punch biopsies be used when high-grade disease is suspected at initial colposcopic assessment? A prospective study. Int J Gynecol Cancer. 2006;16(1):253–6. https://doi.org/10.1111/j.1525-1438.2006.00344.x.
20. Jablonski-Momeni A, Stachniss V, Ricketts D, et al. Reproducibility and accuracy of the ICDAS-II for detection of occlusal caries in vitro. Caries Res. 2008;42(2):79–87. https://doi.org/10.1159/000113160.
21. Braga M, Mendes F, Martignon S, et al. In vitro comparison of Nyvad's system and ICDAS-II with lesion activity assessment for evaluation of severity and activity of occlusal caries lesions in primary teeth. Caries Res. 2009;43(5):405–12. https://doi.org/10.1159/000239755.
22. Rodrigues J, Hug I, Diniz M, et al. Performance of fluorescence methods, radiographic examination and ICDAS II on occlusal surfaces in vitro. Caries Res. 2008;42(4):297–304. https://doi.org/10.1159/000148162.
23. Diniz MB, Rodrigues JA, Hug I, de Cássia Loiola Cordeiro R, Lussi A. Reproducibility and accuracy of the ICDAS-II for occlusal caries detection. Community Dent Oral Epidemiol. 2009;37(5):399–404. https://doi.org/10.1111/j.1600-0528.2009.00487.x.
24. Bader JD and Shugars DA. A systematic review of the performance of a laser fluorescence device for detecting caries. *J Am Dent Assoc* 2004; 135: 1413–1426. Review. DOI: https://doi.org/10.14219/jada.archive.2004.0051.
25. Enøe C, Georgiadis MP, Johnson WO. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. Prev Vet Med. 2000;45:61–81. https://doi.org/10.1016/S0167-5877(00)00117-3.
26. Albert PS, McShane LM, Shih JH, et al. 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. https://doi.org/10.1111/j.0006-341X.2001.00610.x.

27. Asselineau J, Paye A, Bessède E, et al. Different latent class models were used and evaluated for assessing the accuracy of campylobacter diagnostic tests: overcoming imperfect reference standards? Epidemiol Infect. 2018;146: 1556–64Article. https://doi.org/10.1017/S0950268818001723.
28. Garrett ES, Zeger SL. Latent class model diagnosis. Biometrics. 2000;56(4): 1055–67. https://doi.org/10.1111/j.0006-341X.2000.01055.x.
29. Pereira GD, Louzada F, Barbosa VD, et al. A general latent class model for performance evaluation of diagnostic tests in the absence of a gold standard: an application to Chagas disease. Comput Math Methods Med. 2012;2012:1–12. https://doi.org/10.1155/2012/487502.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.