Statistical Comparison of Initial and Confirmatory (Final) Diagnostic Screening Test Results for a Condition in a Population

Precious O. Ibeakuzie[1]*, Cyprian A. Oyeka[1], Jane O. Odum[1], Chilota C. Efobi[2], Adaora A. Onyiaorah[3]

[1]Department of Statistics, Faculty of Physical Sciences, Nnamdi Azikiwe University, Awka, Nigeria

[2]Consultant Haematologist, Department of Haematology and Blood Banking, Nnamdi Azikiwe University Teaching Hospital, Nnewi, Nigeria

[3]Consultant Ophthalmologist, Department of Ophthalmology, Nnamdi Azikiwe University Teaching Hospital, Nnewi, Nigeria

***Corresponding Author**

DOI: https://doi.org/10.47772/IJRISS.2026.100800249

Received: 21 August 2026; Accepted: 26 August 2026; Published: 01 September 2026

ABSTRACT

Background and Objectives: Diagnostic screening tests are routinely used to detect a condition in a population. When the same subjects undergo an initial screen and a confirmatory test, the consistency of the two sets of results must be assessed. Existing methods answer this only in part: the McNemar test detects marginal heterogeneity without quantifying the direction of discordance, while Cohen's kappa measures overall agreement without isolating directional asymmetry. This paper proposes a nonparametric method that explicitly addresses directional discordance between initial and confirmatory screening results.

Methods: Trichotomous scores of 1, 0 and -1 are assigned to subjects according to their response patterns across the two occasions. An aggregate statistic W is derived with its variance, and a chi-square statistic with one degree of freedom is obtained. The method is compared analytically with the McNemar test and Cohen's kappa, and a Monte Carlo study of 10,000 replications evaluates its Type I error rate and power against the McNemar test for sample sizes from 50 to 2000.

Results: The method was applied to published data on 305 hospitalised patients in whom conjunctival pallor was the initial screen for severe anaemia and laboratory haemoglobin below 7 g/dL the confirmatory test. The proportions shifting from positive-to-negative and negative-to-positive were 27.9% and 0.7%, a 42-fold imbalance. The chi-square statistic was 106.950 ($df = 1$, $p < 0.001$) against 79.184 for the McNemar test, while Cohen's kappa of 0.230 disclosed neither the direction nor the magnitude of the imbalance. Simulation confirmed nominal Type I error for $n \geq 100$ and power at least matching McNemar.

Conclusions: Conjunctival pallor is a dependable rule-out but unreliable rule-in sign for severe anaemia: it rarely misses a case yet over-classifies heavily, so haemoglobin confirmation remains indispensable. The method fills a gap in paired categorical data analysis by explicitly quantifying directional discordance.

Keywords: diagnostic screening; directional discordance; trichotomous scoring; McNemar test; Cohen's kappa

INTRODUCTION

The assessment of agreement and consistency between repeated diagnostic measurements is a fundamental challenge in clinical medicine, epidemiology and public health. When a diagnostic screening test for a condition or disease is administered to the same group of subjects on two separate but closely spaced occasions, it is necessary to determine whether the results of the initial test and those of the subsequent confirmatory test are in statistical agreement or show a systematic directional departure from one another.

Diagnostic screening tests occupy a central position in modern healthcare delivery. They are used not only to identify individuals who may be affected by a condition but also to assist clinicians in making treatment decisions, monitoring disease progression and planning public health interventions. The reliability of a diagnostic test, expressed partly through the consistency of its results across repeated administrations, is therefore a matter of both scientific and clinical importance (1, 2).

A central concern in this context is not merely whether two sets of test results agree, but whether the pattern of disagreement is directionally asymmetric. That is, one may ask: among subjects who produced discordant results, do more subjects shift from a positive initial result to a negative confirmatory result than from negative to positive, or vice versa? This question of directional discordance has direct implications for clinical decision-making. If an initial test systematically over-diagnoses a condition (positive initial, negative confirmatory), a clinician who relies solely on the initial test may expose patients to unnecessary treatment or anxiety. Conversely, if the confirmatory test systematically identifies cases missed initially (negative initial, positive confirmatory), failure to perform confirmatory testing may result in missed diagnoses. Quantifying this directional asymmetry is therefore essential.

The McNemar test (3) is the most widely used statistical procedure for comparing two related proportions arising from a 2×2 contingency table. It evaluates whether the two off-diagonal cell frequencies are significantly different, testing the null hypothesis of marginal homogeneity. While the McNemar test is well established and broadly applied, it has a notable limitation: it does not directly quantify the direction or the probabilistic magnitude of the discordance. It answers the question of whether discordance is significant but provides no explicit framework for characterising how much more likely a subject is to shift in one direction than in the other.

Cohen's kappa (4) is another widely used measure of agreement, which adjusts for chance agreement and provides an overall index ranging from 0 to 1. However, kappa is a scalar summary of agreement and does not distinguish between the two types of directional discordance. A kappa of 0.70, for example, does not indicate whether the disagreements arise predominantly from over-classification or from under-classification.

This paper proposes, develops and applies a nonparametric statistical method that explicitly addresses directional discordance. The method assigns trichotomous scores of 1, 0 and -1 to subjects based on their response patterns, derives a directional test statistic W, and constructs a chi-square test of the null hypothesis that the probability of a positive-to-negative shift equals the probability of a negative-to-positive shift. The method complements the McNemar test and Cohen's kappa by providing a richer, probabilistically grounded assessment of directional asymmetry in diagnostic screening test results.

The proposed method is illustrated using published secondary data on 305 hospitalised patients in whom conjunctival pallor was assessed as an initial bedside screen for severe anaemia and laboratory haemoglobin served as the confirmatory test (5). This application is chosen deliberately, because it sits at the meeting point of two clinical disciplines: the screening observation is an ocular sign, examined by inspection of the palpebral conjunctiva, while the condition being screened for and the confirmatory measurement are haematological. It is also a setting in which the direction of discordance carries asymmetric consequences, since a pallor-positive patient who proves not to be anaemic is subjected to unnecessary investigation, whereas a pallor-negative patient who is severely anaemic may be sent away untreated. A Monte Carlo simulation study is also conducted to evaluate the Type I error rate and statistical power of the proposed test relative to the McNemar test. The paper is structured as follows: Section 2 presents a review of related literature; Section 3 describes the proposed method and its theoretical derivation; Section 4 presents the simulation study; Section 5 illustrates the method with real data; Section 6 provides a discussion, including explicit comparisons with the McNemar test and Cohen's kappa; and Section 7 presents the conclusions.

LITERATURE REVIEW

The statistical comparison of paired categorical measurements has a long and well-established history in the biomedical and social sciences. The foundation of much of this literature rests on the work of McNemar (3), who proposed a test of marginal homogeneity for 2×2 contingency tables arising from matched pairs. The

McNemar test evaluates whether the two off-diagonal frequencies, representing subjects who changed their response category from one occasion to the other, are significantly different from one another. It has since become a standard tool in diagnostic medicine, psychology and clinical trial design.

Stuart (6) extended McNemar's approach to larger square contingency tables, developing a test of marginal homogeneity accommodating polytomous categorical responses. Bowker (7) proposed a test of symmetry for square tables, complementing the marginal homogeneity approach by examining whether the off-diagonal cells are symmetric in a more general sense. Bhapkar (8) introduced a powerful generalisation for testing marginal homogeneity in square contingency tables of arbitrary dimension; Bhapkar's test is asymptotically equivalent to the Stuart test under the null hypothesis but tends to exhibit better power in practice. The relationship between these tests has been discussed by Agresti (9) and Fleiss, Levin and Paik (10).

In diagnostic medicine, the comparison of paired diagnostic test results has been approached from several angles. Pepe (1) provided a comprehensive treatment of the statistical evaluation of medical diagnostic tests, including sensitivity, specificity, predictive values and receiver operating characteristic (ROC) curves, emphasising that the paired nature of the data must be accounted for when a confirmatory test is compared with an initial test on the same subjects. Zhou, Obuchowski and McClish (2) extended this framework to cover a wide range of methods for comparing correlated diagnostic tests, including the comparison of proportions based on paired binary data.

Cohen's kappa coefficient (4) is widely used to quantify the degree of agreement between two raters or two test occasions beyond chance. Landis and Koch (11) provided interpretive benchmarks for the kappa statistic. However, as Fleiss et al. (10) noted, kappa measures overall agreement and does not specifically address directional inconsistency, that is, whether subjects are more likely to shift from positive to negative or from negative to positive across test occasions. This limitation is particularly important in clinical diagnostic settings where the direction of discordance carries different clinical consequences.

The concept of test-retest reliability is closely related to the question addressed in this paper. Carry-over effects in crossover designs and repeated measurement studies have been studied extensively. Senn (12) provided a thorough discussion of the statistical analysis of crossover trials, emphasising the importance of adequate washout periods. Jones and Kenward (13) discussed general strategies for the design and analysis of crossover experiments, including diagnostic settings.

In the field of haematology, the consistency of test results across repeated measurements has important clinical implications. Studies have documented variability in diagnostic classifications for conditions such as anaemia, leucocytosis and various lesion types when subjects are re-tested after short intervals (14). This variability may arise from biological fluctuation, operator differences or instrument calibration, and it underscores the importance of statistical methods for formally assessing whether observed discordances lie within the bounds of chance variation.

In ophthalmology, where confirmatory testing is routinely employed to verify initial findings related to refractive errors, glaucoma, macular degeneration and lesion classification, the statistical assessment of agreement between repeated measurements is of direct clinical utility. Bland-Altman analysis (15) is commonly used for continuous measurements, but analogous methods for binary or categorical diagnostic outcomes are less standardised.

The nonparametric tradition in biomedical statistics offers a range of methods for analysing paired categorical or ordinal data without strong distributional assumptions (16, 17). The sign test evaluates whether the number of positive signs is significantly greater or smaller than expected under the null hypothesis of symmetry. The proposed method bears a conceptual resemblance to the sign test but is adapted specifically to the context of comparing initial and confirmatory diagnostic screening test results through a trichotomous scoring framework.

Oyeka and Ebuh (18) employed a similar trichotomous approach in the development of modified nonparametric tests for paired data, assigning scores of 1, 0 and -1 based on the direction of change. Such scores allow the separation of concordant pairs, which contribute a score of 0, from directionally discordant pairs, which contribute scores of 1 or -1 depending on the direction of change.

Clinical pallor provides one of the oldest and most widely used instances of the initial-then-confirmatory design, and one that spans haematology and ophthalmology. Inspection of the palpebral conjunctiva for pallor is a standard bedside screen for anaemia, particularly where laboratory access is limited, and is followed wherever possible by a confirmatory haemoglobin or packed cell volume measurement. Its accuracy has been examined repeatedly. Chalco, Huicho, Alamo, Carreazo and Bada (19) meta-analysed eleven studies of clinical pallor in children and concluded that its diagnostic value varies widely with the site examined and the haemoglobin threshold adopted. Kalantri, Karambelkar, Joshi, Kalantri and Jajoo (20) assessed pallor at several sites in 390 hospitalised adults and reported both its accuracy and the inter-observer reliability of the sign. Butt, Ashfaq, Sherazi, Jan and Shahbaz (5) examined pallor at four sites, including the conjunctiva, in 305 hospitalised patients against a haemoglobin reference standard. What these studies share is a reliance on sensitivity, specificity and likelihood ratios, which describe how the sign behaves within each true-status group but say nothing about whether the net movement between screen and confirmation is directionally balanced across the screened population as a whole. That is precisely the quantity the present paper isolates.

The application of statistical methods for comparing diagnostic test results in African clinical settings has likewise received relatively limited attention. Ogunfowokan, Ogunfowokan and Nwajei (21), for example, examined the performance of a rapid diagnostic test for malaria in a Nigerian primary care clinic using measures of sensitivity and specificity, but did not address the specific issue of directional consistency between initial and confirmatory test results. The present study contributes to this underrepresented area of biostatistical application.

In summary, this paper builds upon a well-established tradition of statistical methods for paired categorical data, draws on the theoretical framework of trichotomous scoring and nonparametric hypothesis testing, and applies the resulting method to a clinically relevant diagnostic problem. The proposed approach complements existing methods such as the McNemar test, Cohen's kappa and Bhapkar's test, while offering a more direct and intuitive quantification of directional discordance.

METHODS

Study Design and Data Structure

Suppose a research scientist or clinician collects a random sample of $n = n_{..}$ subjects from a population for a condition or disease of interest. Each subject is classified twice: once by an initial diagnostic screening test and once by a confirmatory (final) test. Two situations are covered by this design and by the method developed below. In the first, the same diagnostic instrument is applied on two separate occasions, in which case the two administrations must be separated by a washout period sufficient to avoid carry-over effects (12, 13). In the second, which is the more common arrangement in screening programmes, a rapid or low-cost initial screen is followed by a distinct confirmatory assay of higher accuracy applied to the same subjects; here no washout is required, but the two procedures must be performed independently, so that the confirmatory result is not influenced by knowledge of the initial one. In either situation the data consist of n pairs of binary outcomes on the same subjects, which is all the method requires. Let A denote a positive result on the initial test and $\bar{A}$ a negative result on the initial test; let B denote a positive result on the confirmatory test and $\bar{B}$ a negative result on the confirmatory test. The results are arranged in a 2×2 contingency table as shown in Table 1.

Table 1. Format for the presentation of initial and confirmatory diagnostic screening test results for a condition in a population, with the trichotomous score assigned to each cell

Initial test result	**Confirmatory: positive** (B)	**Confirmatory: negative** ($\bar{B}$)	**Total** ($n_{i.}$)
Positive (A)	n_{11} (score 0)	n_{12} (score 1)	$n_{1.}$
Negative ($\bar{A}$)	n_{21} (score -1)	n_{22} (score 0)	$n_{2.}$
Total $(n_{.j})$	$n_{.1}$	$n_{.2}$	$n = n_{..}$

In Table 1, n_{11} denotes the number of subjects who test positive on both occasions (concordant positive); n_{12} denotes the number who test positive initially but negative at confirmatory testing; n_{21} denotes the number who test negative initially but positive at confirmatory testing; and n_{22} denotes the number who test negative on both

occasions (concordant negative). The score entered in each cell makes explicit the direction of discordance, which is the central innovation of the proposed method.

Trichotomous Scoring Scheme

To capture directional discordance, a trichotomous score u_i is assigned to each subject i as follows:

$$u_i = \begin{cases} 1, & \text{if subject i is positive (A) initially and negative } (\bar{B}) \text{ at confirmation, that is } A\bar{B} \\ 0, & \text{if subject i gives the same result on both occasions, that is AB or } \bar{A}\bar{B} \text{ (concordant)} \\ -1, & \text{if subject i is negative } (\bar{A}) \text{ initially and positive (B) at confirmation, that is } \bar{A}B \end{cases}$$

The rationale for this scoring scheme is directly tied to the notion of directional discordance. A score of 1 indicates that the initial test classified the subject as a case but the confirmatory test did not, suggesting potential over-classification at the initial stage. A score of -1 indicates the opposite: the confirmatory test identified the subject as a case while the initial test did not, suggesting potential under-classification at the initial stage. Concordant pairs, which carry no directional information, receive a score of 0 and do not contribute to the test of directional asymmetry.

Probability Framework

Let the corresponding probabilities be defined as

$$\pi^+ = P(u_i = 1), \qquad \pi^0 = P(u_i = 0), \qquad \pi^- = P(u_i = -1),$$

$$\text{with} \qquad \pi^+ + \pi^0 + \pi^- = 1.$$

Here π^+ is the probability that a randomly selected subject tests positive initially and negative at the confirmatory test; π^- is the probability that a randomly selected subject tests negative initially and positive at the confirmatory test; and π^0 is the probability that a randomly selected subject produces the same result on both occasions.

Since $u_i^2 = 1$ for every discordant subject and $u_i^2 = 0$ for every concordant subject, it follows that

$E(u_i^2) = \pi^+ + \pi^-$. The expected value and variance of u_i are therefore

$$E(u_i) = \pi^+ - \pi^-, \qquad Var(u_i) = (\pi^+ + \pi^-) - (\pi^+ - \pi^-)^2.$$

The Test Statistic

Define the aggregate statistic W as the sum of all trichotomous scores:

$$W = \sum_{i=1}^{n} u_i.$$

Because the n subjects are drawn independently, the expected value and variance of W are

$$E(W) = n(\pi^+ - \pi^-), \qquad Var(W) = n[(\pi^+ + \pi^-) - (\pi^+ - \pi^-)^2].$$

Replacing the population probabilities by their sample counterparts, the sample estimate of W and of its variance are

$$\widehat{W} = f^+ - f^- = n_{12} - n_{21} = n(\hat{\pi}^+ - \hat{\pi}^-),$$

$$\widehat{Var}(W) = n[(\hat{\pi}^+ + \hat{\pi}^-) - (\hat{\pi}^+ - \hat{\pi}^-)^2],$$

where the sample probability estimates are

$$\hat{\pi}^{+} = \frac{n_{12}}{n}, \qquad \hat{\pi}^{0} = \frac{n_{11} + n_{22}}{n}, \qquad \hat{\pi}^{-} = \frac{n_{21}}{n}.$$

Hypothesis Test

The null hypothesis of primary research interest is that the proportion of subjects who shift from positive to negative equals the proportion who shift from negative to positive:

$$H_0: \pi^{+} - \pi^{-} = 0 \qquad \text{versus} \qquad H_1: \pi^{+} - \pi^{-} \neq 0.$$

This null hypothesis is tested using the chi-square test statistic

$$\chi^2 = \frac{\widehat{W}^2}{\widehat{Var}(W)} = \frac{n(\hat{\pi}^{+} - \hat{\pi}^{-})^2}{(\hat{\pi}^{+} + \hat{\pi}^{-}) - (\hat{\pi}^{+} - \hat{\pi}^{-})^2}.$$

Under H_0 this statistic follows approximately a chi-square distribution with one degree of freedom for sufficiently large n. The null hypothesis is rejected at the α level of significance if $\chi^2 \geq \chi^2_{\alpha,1}$. For $\alpha = 0.05$ the critical value is $\chi^2_{0.05,1} = 3.841$.

Comparison with the McNemar Test and Cohen's Kappa

It is instructive to compare the proposed method formally with the McNemar test and Cohen's kappa, as this comparison clarifies both the contribution and the advantage of the proposed approach.

The McNemar test statistic is computed as

$$\chi^2_{McN} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}.$$

The proposed test statistic can be written in terms of the same cell frequencies as

$$\chi^2_{prop} = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21}) - \dfrac{(n_{12} - n_{21})^2}{n}}.$$

The two statistics share the same numerator $(n_{12} - n_{21})^2$, which represents the squared net directional shift. Their denominators differ. The McNemar denominator is simply the total number of discordant pairs, $n_{12} + n_{21}$. The proposed denominator additionally subtracts a term that reflects the squared net shift scaled by the sample size, namely

$$\frac{(n_{12} - n_{21})^2}{n}.$$

Because this term is always non-negative, the proposed denominator is smaller than or equal to that of the McNemar test, and the proposed test statistic is therefore always greater than or equal to the McNemar statistic, with equality only when the net shift is zero. Structurally, the proposed statistic is a Wald-type statistic in which the variance is estimated under the alternative hypothesis, whereas the McNemar statistic is a score-type statistic in which the variance is evaluated under the null hypothesis. This is the source of the difference between the two denominators, and it is verified empirically in the simulation study of Section 4.

Furthermore, the proposed method operates within an explicit probabilistic framework that allows the researcher to estimate and interpret π^{+} and π^{-} directly, providing information that the McNemar test does not supply. The McNemar test answers the question: "Are the discordant frequencies equal?" The proposed method answers the more targeted question: "What is the probability that a randomly selected subject shifts in each direction, and is this asymmetry statistically significant?" This distinction is not merely semantic. In clinical practice, a finding

that discordant results are significantly different is less actionable than a finding that, for example, 12% of subjects shift from positive to negative while only 4% shift in the opposite direction, and that this asymmetry is statistically significant. The proposed method thus enables the clinician to estimate the magnitude of directional discordance and to assess its clinical significance beyond the binary reject or fail-to-reject output of the McNemar test.

Cohen's kappa for a 2×2 table is

$$\kappa = \frac{P_o - P_e}{1 - P_e},$$

where

$$P_o = \frac{n_{11} + n_{22}}{n} \qquad \text{and} \qquad P_e = \frac{n_{1\cdot}\, n_{\cdot 1} + n_{2\cdot}\, n_{\cdot 2}}{n^2}$$

are respectively the observed proportion of agreement and the proportion of agreement expected by chance. Cohen's kappa is a global measure of agreement that treats both types of disagreement symmetrically. It does not distinguish between subjects who shift from positive to negative and those who shift from negative to positive, so a moderate kappa may mask a clinically important directional imbalance. Table 2 summarises the key distinctions between the three approaches.

Table 2. Comparative overview of the proposed method, the McNemar test and Cohen's kappa

Feature	**Proposed method**	**McNemar test**	**Cohen's kappa**
Primary question	Are the directional shift probabilities equal?	Are the discordant frequencies equal?	What is the overall agreement beyond chance?
Directional information	Yes; explicitly estimates π^+ and π^-	No; tests the difference only	No; symmetric measure
Probabilistic framework	Trichotomous probability model	Binomial (discordant pairs)	None (descriptive index)
Discordant pair focus	Yes, with direction	Yes, without direction	Indirectly, through agreement
Sensitivity to asymmetry	$\geq$ McNemar when the net shift $\neq 0$	Standard	Not designed for this
Hypothesis test	Chi-square, 1 df	Chi-square, 1 df	No formal test (descriptive)
Clinical actionability	High; estimates the magnitude of each shift direction	Moderate; yes or no signal	Moderate; overall agreement score

Simulation Study

Design

A Monte Carlo simulation study was conducted to evaluate the finite-sample performance of the proposed test statistic under the null and the alternative hypotheses. The simulation was designed to assess (i) whether the test maintains the nominal Type I error rate of 5% under H_0, and (ii) the statistical power of the test across a range of sample sizes and levels of directional discordance, in each case alongside the McNemar test computed on the same simulated samples.

Under H_0, that is $\pi^+ = \pi^-$, five representative configurations were considered, spanning concordance rates from 0.60 to 0.955: (a) $\pi^+ = \pi^- = 0.0225$ with $\pi^0 = 0.955$; (b) $\pi^+ = \pi^- = 0.05$ with $\pi^0 = 0.90$; (c) $\pi^+ = \pi^- = 0.10$ with $\pi^0 = 0.80$; (d) $\pi^+ = \pi^- = 0.1425$ with $\pi^0 = 0.715$; and (e) $\pi^+ = \pi^- = 0.20$ with $\pi^0 = 0.60$. Configuration (d) reproduces the concordance rate observed in the application of Section 5, and configuration (a) the opposite extreme, in which fewer than one subject in twenty is discordant. Under H_1, that is $\pi^+ \neq \pi^-$, four levels of directional discordance were examined, namely $|\pi^+ - \pi^-| = 0.02$, 0.05, 0.10 and 0.15, with π^0

held constant at 0.80 so that $\pi^+ + \pi^- = 0.20$ throughout. A second power study was then run in the regime of the application, with π^0 held at 0.715 so that $\pi^+ + \pi^- = 0.285$, at $|\pi^+ - \pi^-| = 0.05$, 0.10, 0.20 and 0.272; the last of these reproduces the directional imbalance actually observed in Section 5. Sample sizes of $n = 50$, 100, 200, 500 and 1000 were considered in the main study, and $n = 25$, 50, 100, 200, 305 and 500 in the second, the value $n = 305$ being the sample size of the application. For each configuration, 10,000 replicate samples were generated from a multinomial distribution with parameters (π^+, π^0, π^-), and both the proposed chi-square statistic and the McNemar chi-square statistic were computed for each replicate. The simulated Type I error rate was estimated as the proportion of replicate samples for which H_0 was rejected at the 5% level of significance, and the simulated power was estimated as the proportion of replicate samples for which H_0 was correctly rejected. All computations were carried out in Python using the NumPy pseudo-random number generator with a fixed seed to ensure reproducibility.

Results

Table 3 presents the simulated Type I error rates for the proposed test and for the McNemar test under H_0. For all configurations with $n \geq 100$, the estimated Type I error rates of the proposed test lay within the acceptable range of 0.04 to 0.06, consistent with the nominal 5% level. At $n = 50$ a modest inflation of the Type I error rate of the proposed test was observed, reaching 0.070 when $\pi^+ = \pi^- = 0.10$ and 0.064 when $\pi^+ = \pi^- = 0.05$, whereas the McNemar test remained at or below the nominal level throughout. This reflects the fact that the proposed statistic estimates its variance under the alternative hypothesis and is therefore mildly liberal when the number of discordant pairs is small. The sparse configuration behaves in the opposite manner: at $\pi^0 = 0.955$ and $n = 50$ both tests are markedly conservative, rejecting in only 1.7% of replicates, because the expected number of discordant pairs is barely two and a great many samples contain too few discordant pairs to reach the critical value at all. In both of these small-sample situations exact methods or a continuity correction are recommended in preference to the chi-square approximation.

Table 3. Simulated Type I error rates for the proposed test and the McNemar test under H_0 ($\alpha = 0.05$; 10,000 replications per cell)

Configuration	Test	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
$\pi^+ = \pi^- = 0.0225, \pi^0 = 0.955$	Proposed	0.017	0.046	0.049	0.051	0.048
	McNemar	0.017	0.046	0.046	0.051	0.048
$\pi^+ = \pi^- = 0.05, \pi^0 = 0.90$	Proposed	0.064	0.046	0.050	0.051	0.046
	McNemar	0.050	0.040	0.048	0.051	0.046
$\pi^+ = \pi^- = 0.10, \pi^0 = 0.80$	Proposed	0.070	0.051	0.047	0.049	0.047
	McNemar	0.044	0.050	0.044	0.049	0.047
$\pi^+ = \pi^- = 0.1425, \pi^0 = 0.715$	Proposed	0.058	0.055	0.055	0.051	0.050
	McNemar	0.044	0.055	0.051	0.050	0.050
$\pi^+ = \pi^- = 0.20, \pi^0 = 0.60$	Proposed	0.051	0.056	0.048	0.049	0.050
	McNemar	0.049	0.051	0.048	0.049	0.050

Table 4 presents the simulated power of the two tests under H_1. Power increased consistently with the sample size and with the magnitude of directional discordance. At $|\pi^+ - \pi^-| = 0.10$ the proposed test achieved power of 0.909 at $n = 200$, and at $|\pi^+ - \pi^-| = 0.15$ it exceeded 0.80 at $n = 50$ and reached 0.965 at $n = 100$. The power of the proposed test was at least as high as that of the McNemar test in every configuration examined, as the analytical relationship between the two statistics in Section 3.6 predicts. For $n \geq 100$, where both tests hold the nominal size, the advantage of the proposed test is small, of the order of one percentage point; the larger apparent advantage at $n = 50$ is attributable in part to the mild size inflation reported in Table 3 and should not be interpreted as a genuine gain in efficiency. These results confirm that the proposed test is adequately powered to detect clinically relevant levels of directional discordance at sample sizes routinely available in clinical diagnostic studies.

Table 4. Simulated power of the proposed test and the McNemar test under H_1 ($\alpha = 0.05$; $\pi^0 = 0.80$; 10,000 replications per cell)

$\lvert\pi^+ - \pi^-\rvert$	**Test**	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
0.02	Proposed	0.083	0.082	0.099	0.168	0.299
	McNemar	0.053	0.079	0.094	0.167	0.298
0.05	Proposed	0.153	0.211	0.360	0.720	0.944
	McNemar	0.105	0.208	0.350	0.719	0.944
0.10	Proposed	0.422	0.645	0.909	0.999	1.000
	McNemar	0.333	0.638	0.905	0.999	1.000
0.15	Proposed	0.782	0.965	0.999	1.000	1.000
	McNemar	0.703	0.961	0.999	1.000	1.000

A second power study was conducted in the regime of the application, with π^0 fixed at 0.715. The results are given in Table 5. At the imbalance actually observed, $|\pi^+ - \pi^-| = 0.272$, both tests are effectively certain to detect it from $n = 100$ upward, and the proposed test already attains power 0.909 at $n = 25$; the study analysed in Section 5, with $n = 305$, was therefore very heavily powered for the effect it found. The table is more informative about what such a design could not have detected: a moderate imbalance of $|\pi^+ - \pi^-| = 0.05$ would have been detected with probability only 0.382 at $n = 305$, so an absence of significance in a study of this size would have been weak evidence of directional balance. The advantage of the proposed test over the McNemar test is again concentrated at the smallest sample sizes, where its size is also least reliable.

Table 5. Simulated power of the proposed test and the McNemar test in the regime of the application ($\alpha = 0.05$; $\pi^0 = 0.715$; 10,000 replications per cell)

$\lvert\pi^+ - \pi^-\rvert$	**Test**	$n = 25$	$n = 50$	$n = 100$	$n = 200$	$n = 305$	$n = 500$
0.050	Proposed	0.100	0.117	0.163	0.262	0.382	0.554
	McNemar	0.066	0.098	0.162	0.254	0.380	0.551
0.100	Proposed	0.192	0.281	0.481	0.776	0.916	0.989
	McNemar	0.136	0.244	0.479	0.768	0.915	0.989
0.200	Proposed	0.578	0.837	0.985	1.000	1.000	1.000
	McNemar	0.483	0.796	0.985	1.000	1.000	1.000
0.272	Proposed	0.909	0.998	1.000	1.000	1.000	1.000
	McNemar	0.885	0.995	1.000	1.000	1.000	1.000

Results: Application To Conjunctival Pallor Screening For Severe Anaemia

Data

The proposed method is illustrated using published secondary data reported by Butt, Ashfaq, Sherazi, Jan and Shahbaz (5), who examined 305 consecutive hospitalised patients. Each patient was assessed twice. The initial screen was a clinical examination of the palpebral conjunctiva for pallor, recorded as present or absent; the confirmatory test was a laboratory haemoglobin measurement, with severe anaemia defined by the original authors as haemoglobin below 7 g/dL. The two assessments were made on the same 305 patients, so the data form exactly the paired binary structure of Table 1.

The design is of particular interest here because the two stages belong to different clinical disciplines. The screening observation is an ocular sign, elicited by inspecting the conjunctiva, and its reliability is a question for ophthalmology; the target condition and the confirmatory measurement are haematological. A statistical statement about the direction in which the two stages disagree is therefore a statement about how far an ocular sign can be trusted to stand in for a blood test.

The original authors reported their findings as counts of patients with pallor present and absent, cross-classified against the presence of severe anaemia; these four counts are, by construction, the four cells n_{11}, n_{12}, n_{21} and

n_{22} of the paired table. Recast in the notation of this paper, the data are as shown in Table 6. Of the 305 patients, 23 were severely anaemic by the haemoglobin criterion and 282 were not, in agreement with the marginal totals reported in the source. The four cells also reproduce the sensitivity of 91.3% and specificity of 69.9% reported there.

Table 6. Conjunctival pallor as initial screen and haemoglobin below 7 g/dL as confirmatory test for severe anaemia in hospitalised patients ($n = 305$); data from Butt et al. (5)

Initial screen: conjunctival pallor	**Confirmatory: severe anaemia** (B)	**Confirmatory: not severely anaemic** $(\bar{B})$	**Total** $(n_{i.})$
Pallor present (A)	21	85	106
Pallor absent $(\bar{A})$	2	197	199
Total $(n_{.j})$	23	282	305

Computation

From Table 6, the sample probability estimates are

$$\hat{\pi}^{+} = \frac{n_{12}}{n} = \frac{85}{305} = 0.2787 \quad (27.9\%),$$

$$\hat{\pi}^{0} = \frac{n_{11} + n_{22}}{n} = \frac{21 + 197}{305} = \frac{218}{305} = 0.7148 \quad (71.5\%),$$

$$\hat{\pi}^{-} = \frac{n_{21}}{n} = \frac{2}{305} = 0.0066 \quad (0.7\%).$$

The sum $\hat{\pi}^{+} + \hat{\pi}^{0} + \hat{\pi}^{-} = 0.2787 + 0.7148 + 0.0066 = 1.0000$, confirming the constraint (the components are rounded to four decimal places). The sample estimate of W is

$$\widehat{W} = n_{12} - n_{21} = 85 - 2 = 83,$$

and the estimated variance of W is

$$\widehat{Var}(W) = (n_{12} + n_{21}) - \frac{(n_{12} - n_{21})^2}{n} = 87 - \frac{6889}{305} = 87 - 22.5869 = 64.4131.$$

The chi-square test statistic is therefore

$$\chi^2 = \frac{\widehat{W}^2}{\widehat{Var}(W)} = \frac{(83)^2}{64.4131} = \frac{6889}{64.4131} = 106.950.$$

For comparison, the McNemar test statistic is

$$\chi^2_{McN} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} = \frac{(85 - 2)^2}{85 + 2} = \frac{6889}{87} = 79.184,$$

and Cohen's kappa is obtained as

$$P_o = \frac{21 + 197}{305} = 0.7148,$$

$$P_e = \frac{(106)(23) + (199)(282)}{(305)^2} = \frac{2{,}438 + 56{,}118}{93{,}025} = \frac{58{,}556}{93{,}025} = 0.6295,$$

$$\kappa = \frac{P_o - P_e}{1 - P_e} = \frac{0.7148 - 0.6295}{1 - 0.6295} = \frac{0.0853}{0.3705} = 0.230.$$

This application shows the analytical relationship of Section 3.6 at its most visible. The two statistics share the numerator 6889, but the proposed denominator is 64.4131 against the McNemar denominator of 87, so the proposed statistic is larger by a factor of 1.35. Both are far beyond the critical value, so the conclusion is unaffected; the comparison is reported because the gap between the two denominators widens as the net shift grows relative to the sample size, and this is the largest such gap encountered in this paper.

SUMMARY OF RESULTS

Table 7. Summary of statistical results for the conjunctival pallor screening data

Statistic	Value	Interpretation
Proposed χ^2 $(df = 1)$	106.950 $(p < 0.001)$	Highly significant at $\alpha = 0.05$
McNemar χ^2 $(df = 1)$	79.184 $(p < 0.001)$	Highly significant at $\alpha = 0.05$
Cohen's κ	0.230	Fair agreement (11)
$\hat{\pi}^+$ (positive-to-negative shift)	0.279 (27.9%)	Proportion with pallor but not severely anaemic
$\hat{\pi}^-$ (negative-to-positive shift)	0.007 (0.7%)	Proportion without pallor but severely anaemic
$\hat{\pi}^0$ (concordant)	0.715 (71.5%)	Proportion with consistent results
Ratio $\frac{\hat{\pi}^+}{\hat{\pi}^-}$	42.5	Positive-to-negative shifts are 42.5 times as frequent

The chi-square statistic of 106.950 overwhelmingly exceeds the critical value of 3.841 at the 5% level of significance with one degree of freedom. The null hypothesis H_0 of equal directional shift probabilities is therefore decisively rejected.

DISCUSSION

Main Findings

The primary finding is that conjunctival pallor and the confirmatory haemoglobin measurement are profoundly discordant in one direction. Of the 305 patients, 27.9% showed pallor without severe anaemia, while only 0.7% were severely anaemic without showing pallor. Positive-to-negative shifts were therefore 42.5 times as frequent as negative-to-positive shifts, and the asymmetry is decisive ($\chi^2 = 106.950$, $df = 1$, $p < 0.001$).

The clinical translation is precise and useful. Conjunctival pallor is an excellent rule-out sign and a poor rule-in sign for severe anaemia. A clinician who finds no pallor may be reasonably confident the patient is not severely anaemic, since only about seven patients in a thousand fall into that gap. A clinician who finds pallor learns much less: of the 106 patients with pallor, only 21 proved severely anaemic, so roughly four in five positive screens were not confirmed.

It is worth emphasising how little of this is conveyed by the conventional summaries. The overall concordance rate is 71.5%; the sensitivity of the sign is $\frac{21}{23} = 91.3\%$ and its specificity $\frac{197}{282} = 69.9\%$; and Cohen's kappa is 0.230, indicating only fair agreement while saying nothing about which direction the disagreement runs. A reader given the sensitivity alone might well conclude that the sign performs well. Only the directional decomposition into $\hat{\pi}^+ = 0.279$ and $\hat{\pi}^- = 0.007$, together with the formal test of their equality, makes the asymmetry visible and quantifies it on the scale of the screened population rather than within diagnostic subgroups.

Directional Discordance: Why the Proposed Method Is Needed

The central methodological contribution of this paper lies in its explicit treatment of directional discordance. In clinical diagnostic practice it is not sufficient to know that two sets of test results disagree; it is equally important

to know in which direction the disagreement runs, and how large that directional imbalance is. The application in Section 5 illustrates the point directly.

Faced with the data of Table 6, the McNemar test correctly registers a significant discordance, since

$$\chi^2_{\mathrm{McN}} = \frac{(85-2)^2}{85+2} = \frac{6889}{87} = 79.184 \qquad (p < 0.001),$$

but this single number is the whole of its output. It does not provide estimates of π^+ and π^-, nor does it quantify how much more likely a positive-to-negative shift is than a negative-to-positive shift. Cohen's kappa, at 0.230, indicates only that agreement is fair, and by construction cannot attribute that disagreement to one particular direction. The proposed method, by contrast, delivers

$$\hat{\pi}^+ = \frac{85}{305} = 0.279, \qquad \hat{\pi}^- = \frac{2}{305} = 0.007, \qquad \frac{\hat{\pi}^+}{\hat{\pi}^-} = 42.5,$$

$$\chi^2_{\mathrm{prop}} = \frac{(85-2)^2}{(85+2) - \frac{(85-2)^2}{305}} = \frac{6889}{64.4131} = 106.950 \qquad (p < 0.001),$$

and thus states, and formally tests, the substantive finding: some twenty-eight patients in every hundred are flagged by the sign and then cleared by the blood test, against seven in every thousand cleared by the sign and later found severely anaemic.

The distinction between this and the conventional accuracy measures deserves emphasis, because it is easy to conflate them. Sensitivity and specificity are conditional on true status: they answer "given that the patient is severely anaemic, how often is pallor seen?" That is the right question for a diagnostician who already knows the answer, which is to say for nobody. The quantities π^+ and π^- are unconditional: they answer "in a hundred patients walking into this ward, how many will be flagged and then cleared, and how many will be cleared and then flagged?" That is the question a screening programme must answer in order to plan, because it determines how many confirmatory tests will be consumed and how many cases will slip through. The two framings can point in opposite directions. Here, sensitivity of 91.3% looks reassuring while $\hat{\pi}^+ = 0.279$ is alarming, and both are correct descriptions of the same table.

This information is directly actionable in a way that a bare rejection of marginal homogeneity is not. It tells a hospital that adopting conjunctival pallor as a triage screen would refer roughly 35% of all patients for confirmatory haemoglobin testing in order to capture 91% of severe anaemia cases, and it puts a number on both sides of that trade. Where haemoglobin testing is freely available, the over-referral is a modest inefficiency and the high rule-out value makes the sign worth using. Where laboratory capacity is scarce, the same numbers argue for a more specific triage rule, since four in five referrals will be unnecessary.

The importance of this directional information extends beyond any one screening programme. Systematic over-classification at the initial stage leads to unnecessary follow-up procedures, patient anxiety and wasted resources. Systematic under-classification at the initial stage leads to missed diagnoses and delayed treatment. A method that can reliably detect and quantify directional discordance between initial and confirmatory tests therefore has direct public health implications, particularly in resource-limited settings where optimising the use of diagnostic resources is a pressing concern.

How the Proposed Method Improves on the McNemar Test

The McNemar test and the proposed method are closely related: both focus on the discordant pairs and share the same numerator, $(n_{12} - n_{21})^2$. Their key differences are threefold.

First, the McNemar test does not provide estimates of the individual directional shift probabilities π^+ and π^-. The proposed method does, and these estimates are interpretable as the proportion of the screened population

that can be expected to shift in each direction. This probabilistic framing is natural and intuitive for clinicians and public health practitioners.

Second, the denominator of the proposed test statistic accounts for the squared net directional shift, so that the proposed statistic is at least as large as the McNemar statistic whenever asymmetry exists, and the two coincide when $\hat{\pi}^+ = \hat{\pi}^-$. The simulation results in Section 4 confirm that the power of the proposed test is correspondingly at least as high as that of the McNemar test in every configuration examined. It should be stated plainly, however, that for $n \geq 100$ this advantage is small, and that at $n = 50$ part of the apparent advantage is offset by a mild inflation of the Type I error rate. The practical case for the proposed method therefore rests principally on the directional information it supplies rather than on a large gain in power.

Third, the proposed method operates within an explicit probabilistic framework derived from the trichotomous scoring scheme. This framework provides a natural decomposition of the information in the 2×2 table into concordant information (π^0) and directional discordance information (π^+ and π^-). This decomposition is absent from the McNemar test, which conflates both discordant categories into a single count ($n_{12} + n_{21}$).

How the Proposed Method Complements Cohen's Kappa

Cohen's kappa is fundamentally a measure of overall agreement and is not designed to test hypotheses about directional discordance. The value obtained here, $\kappa = 0.230$, indicates only fair agreement according to the benchmarks of Landis and Koch (11). It signals that something is amiss, but it does not tell the clinician whether the disagreement arises predominantly from over-classification or from under-classification by the initial screen, nor does it provide a formal test of directional asymmetry.

The present data also expose a well-known awkwardness of kappa. Because severe anaemia is uncommon in this series, only 23 of 305 patients, the chance-agreement term is large at $P_e = 0.6295$, and kappa is driven down accordingly. Read on its own, $\kappa = 0.230$ suggests the two procedures barely agree at all, whereas 71.5% of patients were in fact classified identically. Kappa is thus reporting a mixture of two things, the prevalence of the condition and the extent of disagreement, and cannot separate them. The proposed method is unaffected by this, because it conditions attention on the discordant cells and reports each direction separately as a proportion of the whole sample. Used together, kappa as a measure of overall agreement and the proposed test as a measure of directional discordance give a more complete picture of the relationship between initial and confirmatory test results than either provides alone.

Clinical Implications

The clinical reading of these results is unambiguous, and it differs for the two disciplines that meet in this screening pair.

For the examining clinician, the finding supports a specific and limited use of the sign. Absence of conjunctival pallor is strong reassurance: only 2 of the 199 patients without pallor proved severely anaemic. Presence of pallor is a weak signal that must not be acted on alone, since 85 of the 106 patients with pallor were not severely anaemic. Treating on the strength of the sign, without haemoglobin confirmation, would therefore mean treating roughly four patients unnecessarily for every one treated appropriately, with the attendant costs of unwarranted haematinics, investigation and, in some settings, transfusion referral.

For the service planner, the same two numbers are the inputs to a capacity calculation. A screening policy based on conjunctival pallor would send about 35% of admissions for confirmatory haemoglobin measurement. Whether that is acceptable depends on local laboratory capacity, and the point of estimating π^+ and π^- separately is that it makes the trade explicit rather than burying it in a single agreement index. The asymmetry of consequences matters here too: an unnecessary confirmatory test costs a reagent, whereas a missed case of severe anaemia may cost a life, and the very small $\hat{\pi}^- = 0.007$ is what justifies using an imperfect sign at all.

It should be noted that these estimates are conditional on the laboratory haemoglobin being treated as the reference, and on the particular threshold of 7 g/dL adopted in the source study. A less stringent threshold would

reclassify many of the 85 discordant patients as anaemic and would shrink $\hat{\pi}^{+}$ substantially, so the estimates reported here describe the sign as a screen for severe anaemia specifically, not for anaemia in general. Both quantities are also properties of the screened population as much as of the sign, and would differ in a population with a different underlying prevalence.

Limitations

Several limitations warrant acknowledgement. First, the large-sample chi-square approximation may be unreliable when the number of discordant pairs is small, as confirmed by the simulation results at $n = 50$ in Table 3; in such settings exact methods based on the binomial or conditional distribution of the discordant pairs should be preferred. This is not a concern in the present application, where there are 87 discordant pairs and the test statistic exceeds the critical value by more than an order of magnitude. Second, the estimated variance $\widehat{\mathrm{Var}}(W)$ degenerates to zero in the extreme case in which every subject is discordant in the same direction, that is when $n_{12} = n$ and $n_{21} = 0$ or the reverse, so the statistic is undefined at that boundary; such a configuration is not encountered in practice but should be handled explicitly in any software implementation. Third, where the design involves the same instrument administered twice, the method assumes the two occasions are sufficiently separated in time to preclude carry-over effects; where it involves a screen followed by a distinct confirmatory test, as in Section 5, it assumes instead that the confirmatory result was obtained independently of knowledge of the screening result. That assumption is worth scrutiny in the present case, since a clinician who has just recorded pallor may not be blind to that finding when the haemoglobin is requested and interpreted; the source report does not state whether blinding was enforced, and this cannot be verified from a published summary table. Fourth, the analysis uses secondary data extracted from a published report, so the authors of the present paper could not audit the primary records, and the cell counts are taken as reported. Fifth, the method is currently formulated for binary outcomes; the underlying condition here is in truth continuous, since haemoglobin was dichotomised at 7 g/dL, and extensions to ordinal or continuous outcomes remain an important direction for future methodological work. Sixth, the application uses data from a single centre and a hospitalised population, in which the prevalence and the spectrum of severity differ from those of a community screening programme, so the estimates should not be transported uncritically to other settings.

Future Directions

Future research should explore extensions of the proposed method in several directions. First, extensions to ordinal and polytomous diagnostic outcomes would broaden the applicability of the trichotomous scoring framework. Second, the development of exact conditional tests for small samples would address the limitation noted above regarding the chi-square approximation. Third, Bayesian extensions of the trichotomous framework, in which prior information about the directional shift probabilities is incorporated, could be valuable in settings where historical data are available. Fourth, the method could be extended to stratified or clustered data, such as multi-centre diagnostic studies, using methods analogous to the stratified McNemar test. Fifth, the relationship between the proposed test statistic and the broader class of tests for symmetry and marginal homogeneity (6, 8) deserves further theoretical investigation.

CONCLUSION

This paper proposed and applied a nonparametric statistical method for formally comparing initial and confirmatory diagnostic screening test results obtained from the same subjects in a population. The central innovation of the proposed method is its explicit treatment of directional discordance, achieved through a trichotomous scoring scheme that assigns scores of 1, 0 and -1 to subjects according to the direction of their shift between the initial and the confirmatory test result. A chi-square test statistic with one degree of freedom was derived and shown to be approximately chi-square distributed for sufficiently large samples.

The proposed method was compared analytically with the McNemar test and Cohen's kappa. The McNemar test detects marginal heterogeneity but does not provide estimates of the directional shift probabilities π^{+} and π^{-} and does not quantify the direction of discordance. Cohen's kappa is a global measure of agreement that is not designed to test directional asymmetry. The proposed method addresses these limitations by operating within an

explicit probabilistic framework that separates directional discordance from concordance, estimates the magnitude of each directional shift and provides a formal test of their equality.

A Monte Carlo simulation study confirmed that the proposed test maintains the nominal 5% Type I error rate for samples of $n \geq 100$ across concordance rates from 0.60 to 0.955, is mildly liberal at $n = 50$ under moderate discordance and conservative at $n = 50$ when discordance is very sparse, and achieves power at least as high as that of the McNemar test at every level of directional discordance examined.

Applied to published secondary data on 305 hospitalised patients in whom conjunctival pallor served as an initial bedside screen for severe anaemia and laboratory haemoglobin as the confirmatory test, the proposed method found decisive directional discordance between the two stages ($\chi^2 = 106.950$, $df = 1$, $p < 0.001$), with 27.9% of patients shifting from positive to negative against only 0.7% shifting from negative to positive, a 42.5-fold imbalance. The substantive conclusion is that conjunctival pallor is a dependable rule-out but an unreliable rule-in sign for severe anaemia, so that haemoglobin confirmation is indispensable before treatment is begun. Neither the McNemar test nor Cohen's kappa, applied to the same table, would have delivered that conclusion: the former reports only that the discordant frequencies differ, and the latter, at $\kappa = 0.230$, reports fair agreement without direction and is in any case depressed by the low prevalence of the condition. The method is simple to compute and to interpret, makes no strong distributional assumptions, and is applicable in any clinical, epidemiological or public health setting in which paired binary diagnostic outcomes are available.

Ethics Statement

This study is a secondary statistical analysis of aggregated, fully anonymised data previously published in the peer-reviewed literature (5). No new data were collected from human participants by the authors of the present paper, no individual-level records were accessed, and no identifiable personal information was available at any stage; the analysis was performed entirely on the four summary cell counts of a published contingency table. Under the guidelines governing research involving human subjects in Nigeria, and consistent with the principles of the Declaration of Helsinki, secondary analysis of published aggregate data of this kind does not constitute human subjects research and does not require separate ethical approval. Ethical oversight of the original data collection, including participant consent, was the responsibility of the investigators of the primary study and is described in their report.

Data Availability Statement

All data analyzed in this study are contained within the published article of Butt et al. (5), which is available from the journal's open archive, and are reproduced in full in Table 6 of the present paper. No additional data were generated. The Python code used to compute the test statistics reported in Section 5 and to reproduce the Monte Carlo results reported in Tables 3, 4 and 5 is available from the corresponding author on request.

Declaration Of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGEMENTS

The authors thank Butt and colleagues for making the screening data used in Section 5 openly available, and the anonymous reviewers for comments that improved the manuscript.

REFERENCES

1. Pepe, M. S. (2003). The statistical evaluation of medical tests for classification and prediction. Oxford University Press.
2. Zhou, X. H., Obuchowski, N. A., & McClish, D. K. (2011). Statistical methods in diagnostic medicine (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/9780470906514

3. McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2), 153-157. https://doi.org/10.1007/BF02295996
4. Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46. https://doi.org/10.1177/001316446002000104
5. Butt, Z., Ashfaq, U., Sherazi, S. F. H., Jan, N. U., & Shahbaz, U. (2010). Diagnostic accuracy of "pallor" for detecting mild and severe anaemia in hospitalized patients. Journal of the Pakistan Medical Association, 60(9), 762-765.
6. Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. Biometrika, 42(3-4), 412-416. https://doi.org/10.1093/biomet/42.3-4.412
7. Bowker, A. H. (1948). A test for symmetry in contingency tables. Journal of the American Statistical Association, 43(244), 572-574. https://doi.org/10.1080/01621459.1948.10483284
8. Bhapkar, V. P. (1966). A note on the equivalence of two test criteria for hypotheses in categorical data. Journal of the American Statistical Association, 61(313), 228-235. https://doi.org/10.1080/01621459.1966.10502021
9. Agresti, A. (2002). Categorical data analysis (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/0471249688
10. Fleiss, J. L., Levin, B., & Paik, M. C. (2003). Statistical methods for rates and proportions (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/0471445428
11. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174. https://doi.org/10.2307/2529310
12. Senn, S. (2002). Cross-over trials in clinical research (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/0470854596
13. Jones, B., & Kenward, M. G. (2014). Design and analysis of cross-over trials (3rd ed.). CRC Press.
14. Bain, B. J., Bates, I., & Laffan, M. A. (2017). Dacie and Lewis practical haematology (12th ed.). Elsevier.
15. Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, 327(8476), 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8
16. Conover, W. J. (1999). Practical nonparametric statistics (3rd ed.). John Wiley & Sons.
17. Hollander, M., Wolfe, D. A., & Chicken, E. (2014). Nonparametric statistical methods (3rd ed.). John Wiley & Sons.
18. Oyeka, I. C. A., & Ebuh, G. U. (2012). Modified Wilcoxon signed-rank test. Open Journal of Statistics, 2(2), 172-176. https://doi.org/10.4236/ojs.2012.22019
19. Chalco, J. P., Huicho, L., Alamo, C., Carreazo, N. Y., & Bada, C. A. (2005). Accuracy of clinical pallor in the diagnosis of anaemia in children: A meta-analysis. BMC Pediatrics, 5, 46. https://doi.org/10.1186/1471-2431-5-46
20. Kalantri, A., Karambelkar, M., Joshi, R., Kalantri, S., & Jajoo, U. (2010). Accuracy and reliability of pallor for detecting anaemia: A hospital-based diagnostic accuracy study. PLoS ONE, 5(1), e8545. https://doi.org/10.1371/journal.pone.0008545
21. Ogunfowokan, O., Ogunfowokan, B. A., & Nwajei, A. I. (2020). Sensitivity and specificity of malaria rapid diagnostic test (mRDT CareStat) compared with microscopy amongst under five children attending a primary care clinic in southern Nigeria. African Journal of Primary Health Care & Family Medicine, 12(1), a2212. https://doi.org/10.4102/phcfm.v12i1.2212