Statistical Papers (2022) 63:995–1012
https://doi.org/10.1007/s00362-021-01266-9

REGULAR ARTICLE

Asymptotic analysis of reliability measures for an imperfect dichotomous test

Alla Slynko[1]

Received: 5 May 2021 / Revised: 5 September 2021 / Accepted: 17 September 2021 /
Published online: 5 October 2021

Abstract

To access the reliability of a new dichotomous test and to capture the random variability of its results in the absence of a gold standard, two measures, the inconsistent acceptance probability (IAP) and inconsistent rejection probability (IRP), were introduced in the literature. In this paper, we first analyze the limiting behavior of both measures as the number of test repetitions increases and derive the corresponding accuracy estimates and rates of convergence. To overcome possible limitations of IRP and IAP, we then introduce a one-parameter family of refined reliability measures, $\Delta(k, s)$. Such measures characterize the consistency of the results of a dichotomous test in the absence of a gold standard as the threshold for a positive aggregate test result varies. Similar to IRP and IAP, we also derive corresponding accuracy estimates and rates of convergence for $\Delta(k, s)$ as the number k of test repetitions increases.

Keywords Dichotomous test · Inconsistent acceptance probability · Inconsistent rejection probability · Reliability measures · Testing without a gold standard

Mathematics Subject Classification 62J15 · 62F12

1 Introduction

Sensitivity and specificity, two measures which characterize the probabilities of correct classification, are often used when accessing the performance of a given diagnostic test. The estimation of both accuracy measures is straightforward when the true disease status is observable, for instance, through the availability of a gold standard

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00362-021-01266-9.

✉ Alla Slynko
alla.a.slynko@gmail.com

[1] Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada

(Yerushalmy 1947). On the other hand, a gold standard may become unavailable due to a number of reasons such as cost constraints, practicability, disease prevalence, or even ethical grounds. As examples one can consider a definite diagnosis of such diseases as depression (Gelaye et al. 2014), childhood pulmonary tuberculosis (Wang et al. 2017), Alzheimer's disease as well as myocardial infarction (Hui and Zhou 1988), or the determination of the hydroxymethylation status of a given cytosine (Slynko and Benner 2019).

There is a whole range of statistical procedures that can be used for assessing the accuracy of a new diagnostic test in the absence of a gold standard. For instance, one can either use one or several imperfect reference tests or apply a latent variables model as proposed in Alonzo and Pepe (1999), Hui and Zhou (1988), Hui and Walter (1980), Pepe and Janes (2007). Note that in many cases such procedures implement the idea of assessing the reliability of a new test by means of additional reference tests.

In this paper we access the performance of a given test by means of its reliability and introduce reliability measures which do not require any knowledge about the true disease status of a patient, as those measures rely on the results of the given test's repetitions. The main assumption is that the disease status remains the same over the time of test repetitions. Note that multiple repetitions of imperfect tests are common in practice. Quick, inexpensive, and non-invasive tests such as HPV tests or rapid SARS-CoV-2 tests (Guglielmi 2020) are just some prominent examples. Theoretical results on the performance of the diagnostic tests under multiple test repetitions can be found, among many others, in Albert (2007), Chiang (1951), Hui and Zhou (1988), Lachenbruch (1988), Nissen-Meyer (1964), Politser (1982), Wang et al. (2017), Wang and Hanson (2019).

The idea for the corresponding reliability measure comes from Akkerhuis et al. (2019). The two reliability measures, IRP and IAP, introduced in that paper, are defined as the probability to obtain a particular outcome that is different from the most frequent result of the test sequence and are designed to describe a random measurement error of a new test. Acknowledging the fact that it is impossible to capture the closeness of the test outcomes to the true value if a gold standard is unavailable, IRP and IAP measure the closeness of the repeated results to each other and thus characterize the extent to which the new test produces reproducible results.

The first purpose of this paper is to analyze the asymptotic behavior of $\mathrm{IRP}(k)$ and $\mathrm{IAP}(k)$ as the number k of test repetitions increases. In particular, we demonstrated that in general the convergence of $\mathrm{IRP}(k)$ and $\mathrm{IAP}(k)$ to their limiting value is of the order $O(1/\sqrt{k})$, with even a geometric decay in some specific cases. The limiting values from our analysis already appear in Eq. (5) of Akkerhuis et al. (2019) and were there taken to be equal to $\mathrm{IRP}(k)$ and $\mathrm{IAP}(k)$ for each finite value of k. While we point out that such an identity does not hold in general, our analysis shows that $\mathrm{IRP}(k)$ and $\mathrm{IAP}(k)$ at least approach those values asymptotically as the number k of test repetitions increases.

A central limitation of the IRP and IAP measures is that both measures assume that the test result can be derived symmetrically, based on the mode of the obtained k test outcomes. In many situations, however, it might be desirable to place the cut-off for the aggregate test result at a level different from the 50% level of the mode-based cut-off. For instance, the test might be designed in a way that leads to skewed outcomes, or one

of the outcomes might have more serious consequences than the other. For example, one might wish to avoid false negatives in repeated rapid SARS-CoV-2 testing or to have a high degree of certainty of a positive result before engaging in invasive treatment. This leads us very naturally to a family of refined reliability measures $\Delta(k, s)$, which allow for a flexible choice of the percentage at which the aggregate test result is interpreted as positive.

Our proposed measures address the reliability of a given test and apply also to situations in which we cannot expect to find reasonable estimates for sensitivity and specificity, because the true disease status is unknown. Thus there is no direct way to compare our approach with the approaches that estimate sensitivity and specificity under multiple test repetitions, even though a high repeatability is obviously essential for a test with high sensitivity and specificity (Nissen-Meyer 1964). In this sense, the proposed reliability measures should complement the estimation of the sensitivity and specificity.

The paper is organized as follows. In Sect. 2 we perform an extended analysis of the limiting behavior of the reliability measures IRP and IAP as introduced in Akkerhuis et al. (2019) when the number k of test repetitions becomes large, and illustrate our findings by means of a numerical example. In that section, we also discuss the limitations of the considered reliability measures. As a result, in Sect. 3 we then propose an improvement of the measures IRP and IAP, a one-parameter family of reliability measures $\Delta(k, s)$, and address the limiting behavior of $\Delta(k, s)$ as k goes to infinity. Section 4 provides some crucial results on the rates of convergence of the measures IRP, IAP and $\Delta(k, s)$ as the number k of test repetitions increases. Section 5 illustrates the obtained results by means of a real-data example. The proofs of our mathematical results can be found in Supporting Information.

2 IRP and IAP: limiting behavior

We consider $k \geq 2$ repetitions of a new dichotomous test, completed for a given patient. For $m = 1, 2, \ldots, k$, let $Y_m \in \{0, 1\}$ denote the result of the m^{th} test repetition. Further, let Mode_k denote the mode of a sequence $Y_1, Y_2, \ldots, Y_k$. In cases in which the number of test repetitions k is even and the number of ones equals the number of zeros, we set $\text{Mode}_k = 0$.

To access the reliability of a new dichotomous test in the absence of a gold standard, two measures, the inconsistent acceptance probability (IAP) and inconsistent rejection probability (IRP) were introduced in Akkerhuis et al. (2019) as follows,

$$\text{IRP}(k) = \mathbb{P}(Y_m = 0 | \text{Mode}_k = 1) \quad \text{and} \quad \text{IAP}(k) = \mathbb{P}(Y_m = 1 | \text{Mode}_k = 0).$$

According to that paper, both measures quantify the probability of random errors for a new test if a gold standard is unavailable. If IRP and IAP yield large values, then the new test is considered to provide highly random results and therefore to be uninformative. Otherwise, the test is considered to be capable to discriminate (at least) between positive and negative states. In order to estimate IRP and IAP from real data,

Akkerhuis et al. (2019) suggests to fit an adaptive polynomial model to the distribution of the probability that Y_m is equal to zero.

Let us address the limiting behavior of both measures IRP(k) and IAP(k) as the number k of test repetitions becomes large. For the sake of consistency, in our notation we will follow the conventions in Akkerhuis et al. (2019). In the following discussion, let $Y_1, Y_2, \ldots$ and R be random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $Y_1, Y_2, \ldots$ independent and Bernoulli$(1-r)$-distributed under the conditional distribution $\mathbb{P}(\cdot|R = r)$, so that R is the probability that Y_m is equal to zero. Note that $Y_1, Y_2, \ldots$ are exchangeable but not independent under the unconditional probability $\mathbb{P}$.

Let $S_k = \sum_{j=1}^{k} Y_j$ and denote by $\lceil \cdot \rceil$ the ceiling function. We first write

$$\begin{aligned} \mathrm{IRP}(k) &= \mathbb{P}\left(Y_m = 0 | \mathrm{Mode}_k = 1\right) \\ &= \frac{\mathbb{P}\left(Y_m = 0, S_k \geq \lceil \frac{k}{2} \rceil\right)}{\mathbb{P}\left(S_k \geq \lceil \frac{k}{2} \rceil\right)} = \frac{\mathbb{P}\left(Y_k = 0, S_{k-1} \geq \lceil \frac{k}{2} \rceil\right)}{\mathbb{P}\left(S_k \geq \lceil \frac{k}{2} \rceil\right)} \end{aligned} \tag{1}$$

Note that Y_k and S_{k-1} are not independent under $\mathbb{P}$. Also, note that IRP(k) actually does not depend on the index m in Y_m since $Y_1, Y_2, \ldots, Y_k$ are exchangeable.

With F as the cumulative distribution function of R, and the corresponding density f, we obtain the following result.

Proposition 1 *As the number k of test repetitions tends to infinity, we get*

$$\mathrm{IRP}(k) \longrightarrow \frac{\int_0^{1/2} x f(x)\, dx}{\int_0^{1/2} f(x)\, dx} = \mathbb{E}[R | R \leq 1/2] \tag{2}$$

as well as

$$\mathrm{IAP}(k) \longrightarrow \frac{\int_{1/2}^{1} (1-x) f(x)\, dx}{\int_{1/2}^{1} f(x)\, dx} = \mathbb{E}[1 - R | R \geq 1/2]. \tag{3}$$

The proof of this proposition is presented in Sect. 1.1 of Supporting Information.

Proposition 1 takes issue with Eq. (5) in Akkerhuis et al. (2019) where it is claimed that $\mathrm{IRP}(k) = \mathbb{E}[R|R \leq 1/2]$ and $\mathrm{IAP}(k) = \mathbb{E}[1 - R|R \geq 1/2]$, for any given value of k. While it is clear from (1), (2) and (3) that these identities cannot hold for all f and k, as claimed in Akkerhuis et al. (2019), Proposition 1 shows that the claim does hold (at least) asymptotically as $k \uparrow \infty$.

The following example illustrates the limiting results (2) and (3).

Example 1 Following the ideas of Nissen-Meyer (1964), we assume that R follows a beta distribution with the density

$$f(x; a, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a, b)} \tag{4}$$

and parameters a and b. As stated in Lachenbruch (1988), a beta distribution is the most commonly used distribution for representation of variation among binomial probabilities. Let us now consider the data presented in Table 1 of Chiang (1951), with $k = 5$ repetitions of a given test. Using that data and the maximum likelihood estimation procedure described in that paper, we first estimate the parameters a and b in (4) and get $\hat{a} = 0.0347$ and $\hat{b} = 0.5837$. Under the assumption that $f(\cdot)$ remains unchanged as the number k of test repetitions increases, we then calculate both measures IRP(k) and IAP(k) for increasing values of k, together with their limiting values (2) and (3); see Fig. 1 for an illustration.

Later in this paper we will analyze the rates of convergence in (2) and (3). But before we start with that analysis, let us discuss possible limitations of the proposed reliability measures, IRP(k) and IAP(k). In particular, both measures assume that the test result can be derived symmetrically, based on the mode of the obtained k test outcomes. This approach is clearly not optimal when either a positive or a negative test result can carry significant consequences, e.g., when the positive test result may lead to further invasive tests or treatment such as biopsy, an X-ray or CT scan or even surgery. In such situations, a certain skewness of aggregate test results might be desirable. Thus we might want to place the cut-off for the aggregate test result at a level different from the 50% level of the mode-based cut-off. This leads us to a family of refined reliability measures $\Delta(k, s)$ introduced in Sect. 3, with the percentages of the positive test results as the corresponding cut-off. Note that such refined measures also help to overcome the other limitation of the IRP(k) and IAP(k) measures, namely that those reliability measures are based on central tendency of the obtained test results and thus do not discriminate between two given sequences of test repetitions if those sequences have the same mode.

3 Refined reliability measures

Let us start this section with the following definition.

Definition: For k independent test repetitions $Y_1, Y_2, \ldots, Y_k$ and $S_k = \sum_{j=1}^{k} Y_j$, we first define the *relative score* S_k^{rel} as $S_k^{\text{rel}} = \frac{S_k}{k}$. Let s denote the observed value of S_k^{rel}. Then, if $s \geq \frac{1}{2}$, we assess the reliability of a given test by means of a reliability measure $\Delta(k, s)$, with

$$\Delta(k, s) := \begin{cases} \mathbb{P}(Y_m = 0 | S_k^{\text{rel}} \geq s) & \text{for } k \geq \frac{1}{1-s}, \\ 0 \quad \text{otherwise} \end{cases}$$

If $s < \frac{1}{2}$, then we define

$$\Delta(k, s) := \begin{cases} \mathbb{P}(Y_m = 1 | S_k^{\text{rel}} < s) & \text{for } k > \frac{1}{s} \\ 0 \quad \text{otherwise} \end{cases}$$

Note that the bounds on s are introduced in order to ensure that the measure $\Delta(s, k)$ is well-defined.

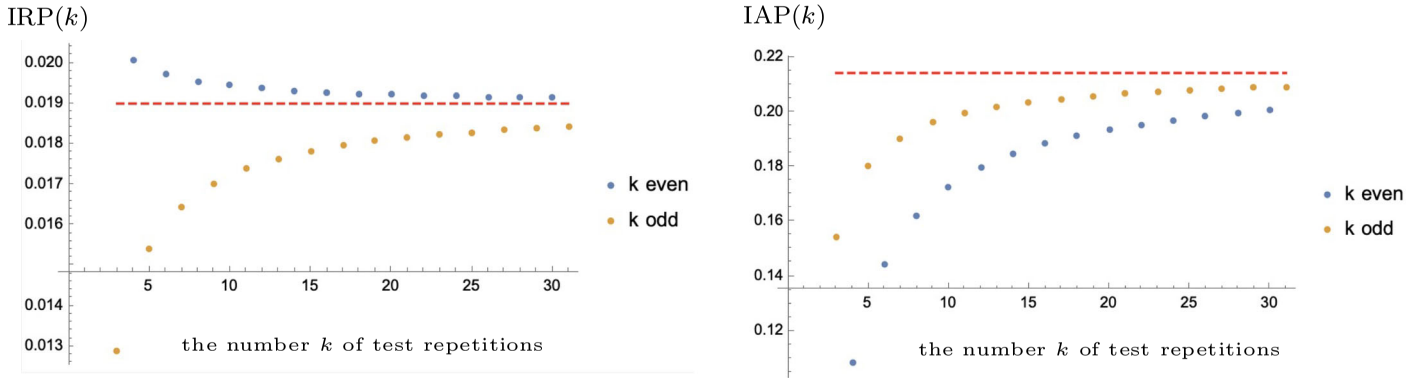

Fig. 1 The limiting behavior of the measures IRP(k) (left-hand panel) and IAP(k) (right-hand panel) for increasing number k of test repetitions. Blue dots represent values of IRP(k) and IAP(k) for even values of k, orange dots represent values of IRP(k) and IAP(k) for odd values of k, and the dashed red lines show the corresponding limiting values $\mathbb{E}[R|R \leq 1/2]$ and $\mathbb{E}[1-R|R \geq 1/2]$ as derived in Proposition 1

For $k \uparrow \infty$, with arguments similar to those for IRP(k) and IAP(k), we further obtain

Proposition 2 *As the number k of test repetitions tends to infinity, we get*

$$\mathbb{P}\left(Y_m = 0 | S_k^{rel} \geq s\right) \longrightarrow \frac{\int_0^{1-s} x f(x)\, dx}{\int_0^{1-s} f(x)\, dx} = \mathbb{E}[R|R \leq 1-s]. \tag{5}$$

as well as

$$\mathbb{P}\left(Y_m = 1 | S_k^{rel} < s\right) \to \frac{\int_{1-s}^{1} (1-x) f(x)\, dx}{\int_{1-s}^{1} f(x)\, dx} = \mathbb{E}[1-R|R \geq 1-s]. \tag{6}$$

We use the data from Example 1 to address the limiting behavior of the measure $\Delta(k, s)$ when the number k of test repetitions increases; an illustration of that behavior for different values of s is presented in Fig. 2 below.

4 Rates of convergence

The limiting results (2) and (3) suggest a convergence of IRP(k) and IAP(k) to a fixed value as the number k of test repetitions increases. Theorem 1 provides upper bounds and corresponding rates for this convergence. In what follows, let f be the density function for the law of the random variable R as introduced above, $F(x) = \int_0^x f(y)\, dy$ be its cumulative distribution function, $G(y) = \int_0^y x f(x)\, dx$, $\|f\|_q^q := \int_0^1 (f(x))^q\, dx$ and $\|g\|_q^q := \int_0^1 (x f(x))^q\, dx$.

Theorem 1 (Convergence of the IRP and IAP measures)
Let $k \geq 2$.

(a) *If f vanishes in an open interval (u, v) containing $\frac{1}{2}$, with $0 \leq u < \frac{1}{2} < v \leq 1$, then*

$$\left|\mathrm{IRP}(k) - \mathbb{E}[R|R \leq 1/2]\right| \leq \frac{2}{F(1/2)}\Big(G(1) + 4\Big)(4\theta(1-\theta))^{k/2} \tag{7}$$

where $\theta = \max\{u, 1-v\}$.

(b) *If f does not vanish in any open interval containing $\frac{1}{2}$, but is bounded in a neighborhood of $\frac{1}{2}$, i.e., there are $u, v \in [0, 1]$ such that $u < \frac{1}{2} < v$ and $|f(x)| \leq c_{u,v}$ for all $x \in (u, v)$ and a constant $c_{u,v} > 0$, then*

$$\begin{aligned} &\left|\mathrm{IRP}(k) - \mathbb{E}[R|R \leq 1/2]\right| \\ &\leq \frac{2}{F(1/2)}\left(2c_{u,v}\sqrt{2\pi}e^{1/6}\frac{1}{\sqrt{k}} + 2^{-k} + G(1)(4\theta(1-\theta))^{k/2}\right) \end{aligned} \tag{8}$$

where $\theta = \max\{u, 1-v\}$.

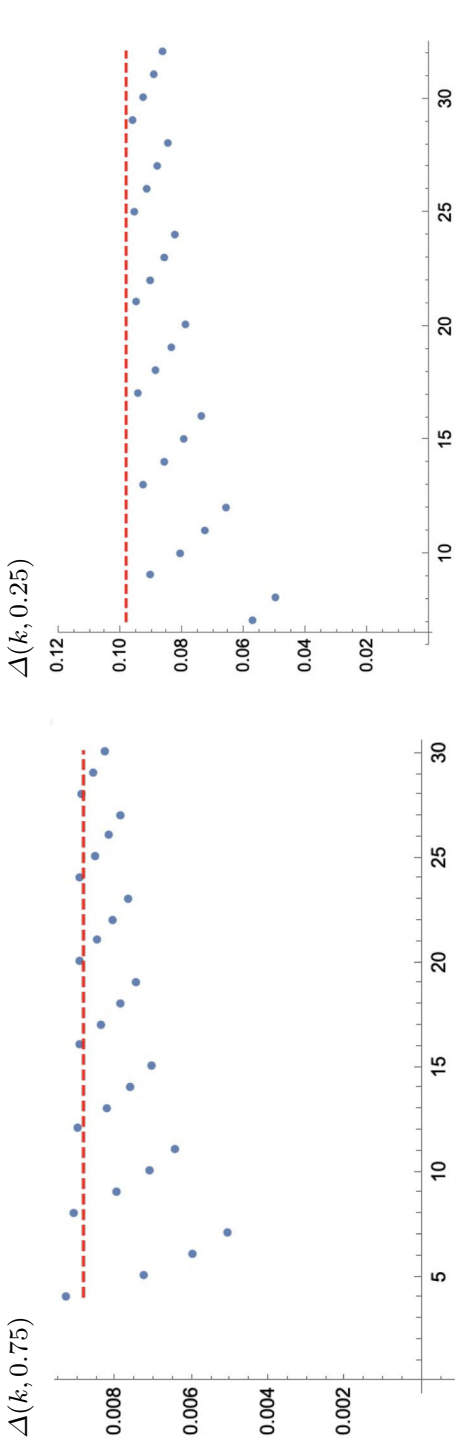

Fig. 2 The limiting behavior of the reliability measure $\Delta(k, s)$ for increasing number k of test repetitions, and the relative score $s = 0.75$ (left-hand panel) as well as $s = 0.25$ (right-hand panel). Horizontal red lines represent the corresponding limiting values as obtained in Proposition 2

(c) *If f is unbounded in every neighborhood of $\frac{1}{2}$, but satisfies $\|f\|_q^q < \infty$ for some $q \geq 1$, then*

$$\begin{aligned}&\left|\mathrm{IRP}(k) - \mathbb{E}[R|R \leq 1/2]\right| \\ &\leq \frac{2}{F(1/2)}\left(\|f\|_q^q + \|g\|_q^q\right) \cdot \left(2\pi e^{1/(3p)}\right)^{1/(2p)} \cdot (kp)^{-1/(2p)}\end{aligned} \tag{9}$$

where p is such that $\frac{1}{q} + \frac{1}{p} = 1$.

(d) *The upper bounds analogous to (7), (8) and (9) also hold for*

$$\left|\mathrm{IAP}(k) - \mathbb{E}[1 - R| \leq 1/2]\right|.$$

For the proof of this theorem see Sect. 1.3 of Supporting Information.

Next, we present the following generalization of Theorem 1 for the refined measures $\Delta(k, s)$.

Theorem 2 (Convergence of the reliability measure $\Delta(k, s)$)
For $s \geq \frac{1}{2}$ and k test repetitions satisfying $k \geq \frac{1}{1-s}$.

(a) *Suppose that f is locally bounded in a neighborhood of $1 - s$, i.e., there are $u, v \in [0, 1]$ so that $u < 1 - s < v$ and $|f(x)| \leq c_{u,v}$ for all $x \in (u, v)$ and a constant $c_{u,v} > 0$. Then*

$$\begin{aligned}&\left|\Delta(k, s) - \mathbb{E}[R|R \leq 1 - s]\right| \\ &\qquad \leq \frac{2}{1 - F(1/2)}\left[C_u\gamma_u^k + C_v\gamma_v^k + 2c_{u,v}\sqrt{2\pi}e^{1/(12s)}\frac{1}{\sqrt{k}}\right]\end{aligned} \tag{10}$$

with the constants

$$\begin{aligned}&\gamma_x = \exp\left((1 - s)\log\frac{x}{1 - s} + s\log\frac{1 - x}{s}\right), \\ &C_u = G(u) + F(u)\gamma_u \quad \text{and} \quad C_v = G(1) - G(v) + (1 - F(v))\gamma_v\end{aligned} \tag{11}$$

(b) *Suppose that f is unbounded in every neighborhood of $1 - s$, but satisfies $\|f\|_q^q < \infty$ for some $q \geq 1$. Then*

$$\begin{aligned}&\left|\Delta(k, s) - \mathbb{E}[R|R \leq 1 - s]\right| \\ &\leq \frac{2}{1 - F(1/2)} \cdot \left(\|g\|_q^q + \|f\|_q^q\right) \cdot \left(2\pi e^{1/(6ps)}\right)^{1/(2p)} \cdot (kp)^{-1/2p}\end{aligned} \tag{12}$$

where p is such that $\frac{1}{q} + \frac{1}{p} = 1$.

(c) *The upper bounds (10) and (12) also hold for $\left|\Delta(k, s) - \mathbb{E}[1 - R|R \geq 1 - s]\right|$, with $s < \frac{1}{2}$ and $k > \frac{1}{s}$.*

For the proof of this theorem see Sect. 1.2 of Supporting Information.

Remark: Note that the function $x \mapsto \gamma_x$ as defined in (11) has a strict global maximum in $x = 1 - s$ and satisfies $\gamma_{1-s} = 1$. In particular, we have $\gamma_u < 1$ and $\gamma_v < 1$ for u and v as in Theorem 2. It follows that the last term on the right-hand side of (10) dominates the upper bound if $c_{u,v} > 0$, and so the upper bound in (10) will be of the order $O(1/\sqrt{k})$ in that case. If, however, $c_{u,v} = 0$, then the fact that γ_u and γ_v are strictly less than 1 implies geometric decay of the upper bound (10).

5 A numerical study

We start our numerical studies by considering a real data example on hydroxymethylcytosine (5hmC) methylation as one of the well-known epigenetic marks involved in gene regulation. In particular, Slynko and Benner (2019) considers the following measure for the 5hmC detection at a single base resolution

$$\Delta\beta = \beta_{BS} - \beta_{oxBS} = \frac{M_{BS}}{M_{BS} + U_{BS} + 100} - \frac{M_{oxBS}}{M_{oxBS} + U_{oxBS} + 100}, \quad (13)$$

with M_{BS} (M_{oxBS}) as the intensity of the methylated allele obtained from the *BS-seq* (*oxBS-seq*) method, U_{BS} (U_{oxBS}) as the intensity of the unmethylated allele obtained from the *BS-seq* (*oxBS-seq*) method, β_{BS} as the methylation level obtained from the *BS-seq* method, and β_{oxBS} as the methylation level derived by means of the *oxBS-seq* method. In the context of our discussion, we can definitely interpret $\Delta\beta$ as a diagnostic test that classifies all CpG sites into hydroxymethylated (with $\Delta\beta > 0$) and non-hydroxymethylated (with $\Delta\beta \leq 0$). Note that such classification is to be performed in the absence of a gold standard what makes the application of our reliability measures particularly relevant.

Let us apply the results obtained in Sects. 2 and 4 to $\Delta\beta$ by using the data from Field et al. (2015) for an illustration. That data is 5hmC and 5mC data for a single, commercially available, cerebellum DNA sample, with 4 replicates that we interpret as 4 different test repetitions. Originally, Field et al. (2015) provides the data for 438,016 probes. To eliminate possible dependencies among those probes, we selected 10,317 probes with absolute value of correlation less than 0.001 and interpreted those probes as independent subjects. Then we calculated the number of probes with 0, 1, 2, 3 or 4 positive values of $\Delta\beta$. The results are presented in Table 1.

Assuming that the random variable R follows a beta distribution Beta(a, b) as in (4), we first estimate the parameters a and b of that distribution and get $\hat{a} = 0.186$ and $\hat{b} = 0.148$. The corresponding density $f_\beta(\cdot)$ is presented in Fig. 3. For $f_\beta(\cdot)$ remaining unchanged as the number k of test repetitions increases, we then use the results of Sect. 2 to calculate both measures IRP(k) and IAP(k), for increasing values of k, together with their limiting values (2) and (3); see Fig. 4 for an illustration. To demonstrate the usability of both reliability measures when k is relatively small, Table 2 provides the values of those measures for fixed number k of test repetitions; the data points from that table can also be found on Fig. 4 below.

Table 1 Observed frequency distribution for $k = 4$ readings of $n = 10,317$ probes

Number of positive $\Delta\beta$ readings, k	Number of probes
0	3459
1	611
2	741
3	965
4	4541
Total number of probes	10,317

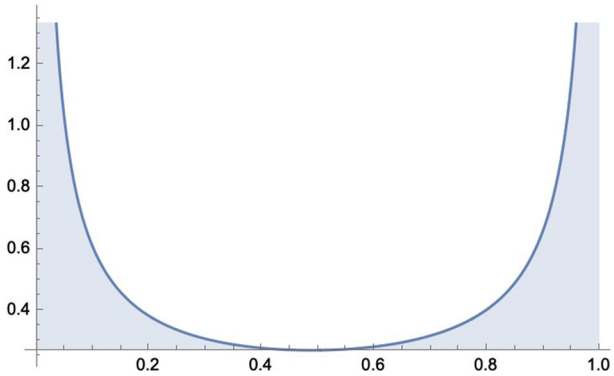

Fig. 3 Density f_β of Beta(0.186, 0.148) distribution, fitted to the data in Table 1

Table 2 Reliability measures IRP(k) and IAP(k) for $\Delta\beta$ and fixed (odd) number k of test repetitions

k	3	5	7	9	11	13	15	17	19	21	Limiting value
IRP(k)	0.069	0.081	0.085	0.088	0.090	0.091	0.092	0.092	0.093	0.093	0.097
IAP(k)	0.056	0.066	0.070	0.072	0.074	0.075	0.075	0.076	0.076	0.077	0.080

The results of Theorem 2, as applied for the 5hmC measure $\Delta\beta$, are illustrated by Fig. 5 below.

We can also use our results to compare two 5hmC measures with respect to their reliability. To do so, we consider a different 5hmC measure, Δm, earlier introduced in Field et al. (2015) as

$$\Delta m = \text{logit}(\beta_{BS}) - \text{logit}(\beta_{oxBS})$$

Following the lines of our analysis for $\Delta\beta$, we first estimate the parameters c and d of the corresponding beta distribution Beta(c, d) and obtain $\hat{c} = 0.172$ and $\hat{d} = 0.139$. Then we estimates the values of both reliability measures, IRP(k) and IAP(k), for increasing number k of test repetitions, together with their limiting values (2) and (3). We also compare those values with the corresponding IRP(k) and IAP(k) values as derived in case of the $\Delta\beta$ measure; see Tables 3 and 4 for more detail. After reviewing

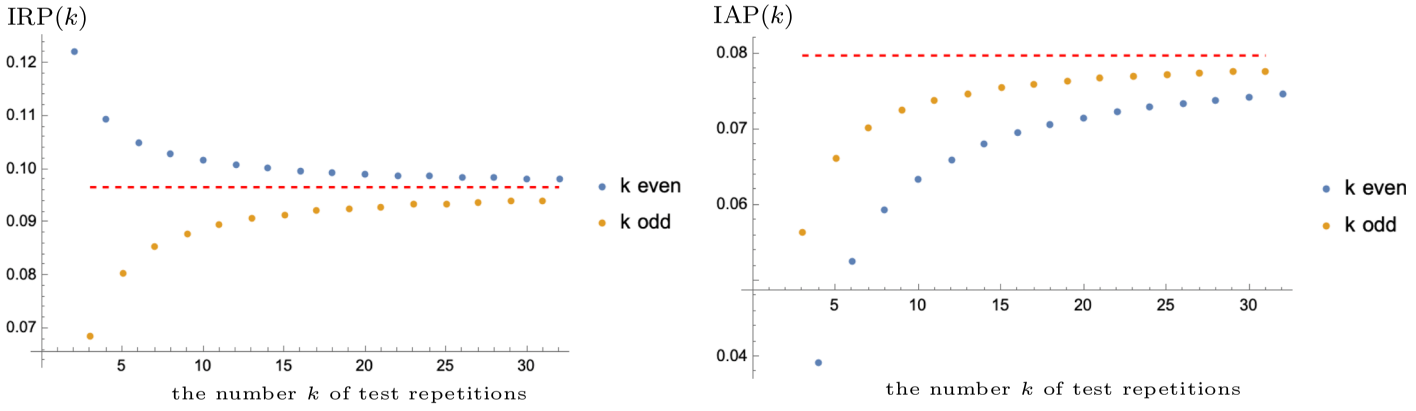

Fig. 4 The limiting behavior of the measures IRP(k) (left-hand panel) and IAP(k) (right-hand panel) for increasing number k of test repetitions and $\Delta\beta$. Blue dots represent values of IRP(k) and IAP(k) for even values of k, orange dots represent values of IRP(k) and IAP(k) for odd values of k, and the dashed red lines show the corresponding limiting values $\mathbb{E}[R|R \leq 1/2]$ and $\mathbb{E}[1-R|R \geq 1/2]$ as derived in Proposition 1

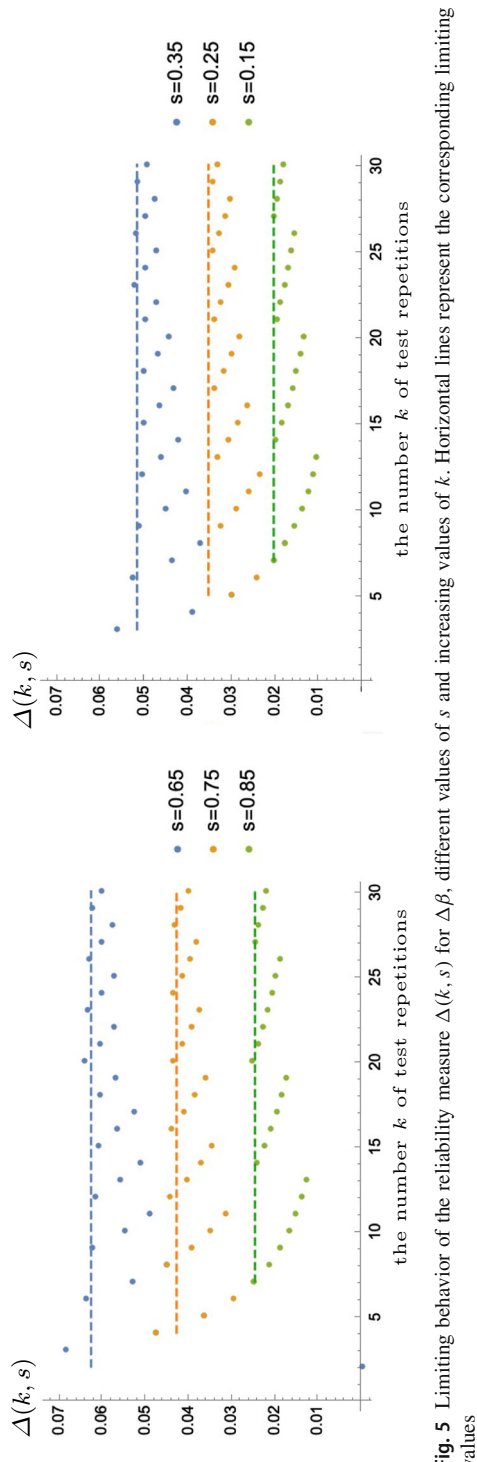

Fig. 5 Limiting behavior of the reliability measure $\Delta(k,s)$ for $\Delta\beta$, different values of s and increasing values of k. Horizontal lines represent the corresponding limiting values

Table 3 Reliability measure IRP(k) for $\Delta\beta$ and Δm, and fixed (odd) number k of test repetitions

k	3	5	7	9	11	13	15	17	19	21	Limiting value
IRP(k) for $\Delta\beta$	0.069	0.081	0.085	0.088	0.090	0.091	0.092	0.092	0.093	0.093	0.097
IRP(k) for Δm	0.065	0.076	0.081	0.083	0.085	0.086	0.086	0.087	0.087	0.088	0.091

Table 4 Reliability measure IAP(k) for $\Delta\beta$ and Δm, and fixed (odd) number k of test repetitions

k	3	5	7	9	11	13	15	17	19	21	Limiting value
IAP(k) for $\Delta\beta$	0.056	0.066	0.070	0.072	0.074	0.075	0.075	0.076	0.076	0.077	0.080
IAP(k) for Δm	0.054	0.063	0.067	0.069	0.070	0.071	0.072	0.072	0.073	0.073	0.076

Table 5 The refined measure $\Delta(k, s)$ for $\Delta\beta$ and Δm, $s \geq \frac{1}{2}$, and the fixed number k of test repetitions

k	5	6	7	8	9	10	11	12	13	14	Limiting value
$\Delta(k, 0.5)$ for $\Delta\beta$	0.081	0.105	0.085	0.103	0.088	0.102	0.090	0.101	0.091	0.100	0.097
$\Delta(k, 0.5)$ for Δm	0.076	0.099	0.081	0.097	0.083	0.096	0.085	0.095	0.086	0.095	0.091
$\Delta(k, 0.7)$ for $\Delta\beta$	0.036	0.030	0.053	0.045	0.040	0.055	0.049	0.044	0.041	0.051	0.053
$\Delta(k, 0.7)$ for Δm	0.034	0.028	0.050	0.043	0.037	0.052	0.046	0.042	0.038	0.048	0.049
$\Delta(k, 0.8)$ for $\Delta\beta$	0.037	0.030	0.025	0.022	0.019	0.035	0.032	0.029	0.026	0.024	0.034
$\Delta(k, 0.8)$ for Δm	0.034	0.028	0.023	0.020	0.018	0.033	0.030	0.027	0.025	0.023	0.032

those tables we state that, with respect to both reliability measures, IRP(k) and IAP(k), $\Delta\beta$ appears to be less reliable than Δm.

Finally, we address the reliability of $\Delta\beta$ and Δm by means of the refined measure $\Delta(k, s)$ as in Sect. 3. In particular, Table 5 and Fig. 6 present the obtained estimates for $s \geq \frac{1}{2}$ where as Fig. 7 shows the results for $s < \frac{1}{2}$. We observe that also in case of the refined measure $\Delta(k, s)$ $\Delta\beta$ still appears to be less reliable compared to Δm, even if such a difference in reliability is not as substantial as in case of the IRP(k) and IAP(k) measures.

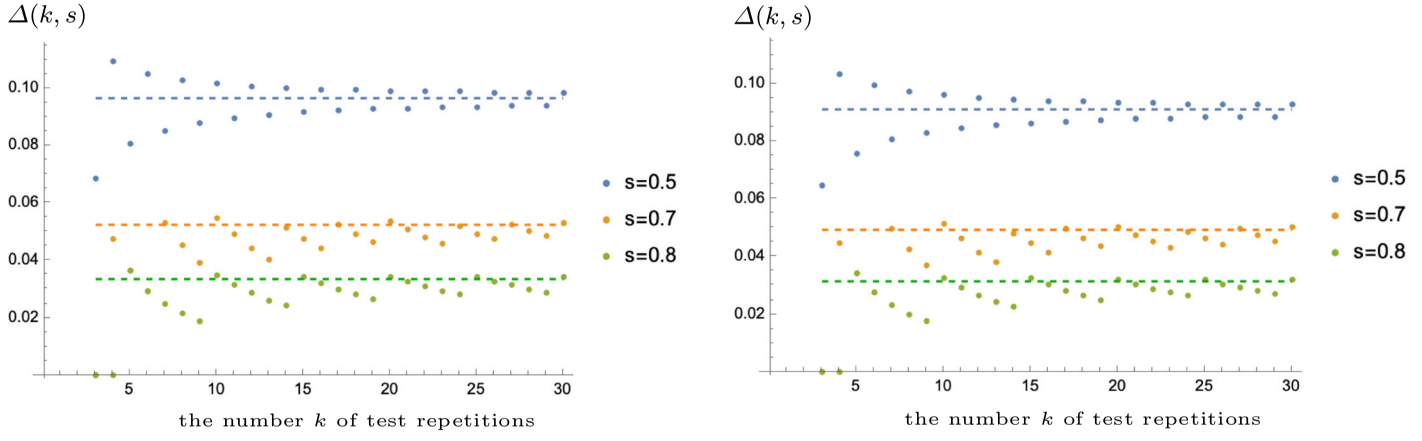

Fig. 6 Limiting behavior of the reliability measure $\Delta(k, s)$ for $\Delta\beta$ (left-hand panel) and Δm (right-hand panel), different values of $s \geq \frac{1}{2}$ and increasing values of k. Horizontal lines represent the corresponding limiting values

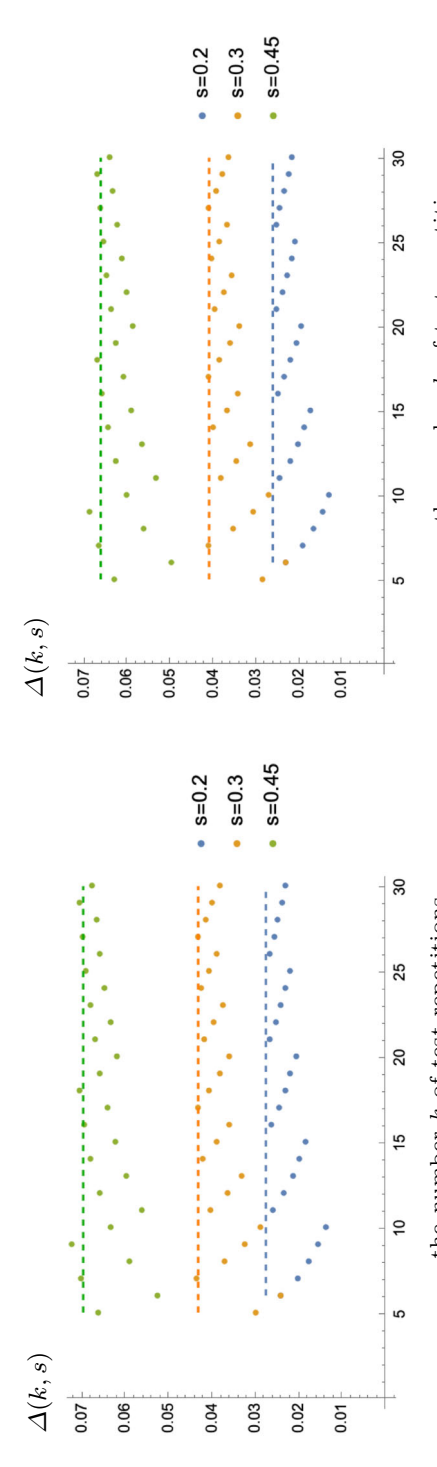

Fig. 7 Limiting behavior of the reliability measure $\Delta(k, s)$ for $\Delta\beta$ (left-hand panel) and Δm (right-hand panel), different values of $s < \frac{1}{2}$ and increasing values of k. Horizontal lines represent the corresponding limiting values

6 Conclusion

One of the most important characteristics of a new test is its reliability, interpreted as the consistency of the test outcomes obtained, e.g., as a result of a number of test repetitions. To address test reliability, the knowledge of the true disease status of the patient is desired. In practice, that knowledge might become crucial, in particular, when there is no gold standard available. In such cases, the true disease status of the patient can be estimated by applying one or several imperfect reference tests as proposed in Alonzo and Pepe (1999), Hui and Zhou (1988), Hui and Walter (1980), Pepe and Janes (2007).

In this paper we suggest a method for approaching the reliability of a new test that does not require any additional reference test. In the context of that method, the reliability is evaluated based on the results of test repetitions. In particular, we first analyze the reliability measures IRP and IAP as proposed in Akkerhuis et al. (2019). Further, we introduce a one-parameter family of refined reliability measures $\Delta(k, s)$ and investigate their limiting behavior as the number k of test repetitions increases.

By allowing for a variable threshold for the positive test result, the proposed reliability measures allow to overcome several limitations of IRP and IAP which are mode-based reliability measures. Also, using $\Delta(k, s)$, we can avoid some issues that may arise when applying one or several reference tests for a new test's reliability evaluation. Among those issues are, e.g., a possible reference bias caused by the choice of the reference measure(s) or conditional dependence among the reference tests as well as between any reference test and a new test Thibodeau (1981), Vacek (1985).

Acknowledgements The author is grateful to Jeroen de Mast and Stefan Steiner for introducing her to the concepts of IRP and IAP and to Akkerhuis et al. (2019).

Funding Not applicable.

Data availability (data transparency) Upon request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Code availability (software application or custom code) Upon request.

References

Akkerhuis TS, De Mast J, Erdmann TP (2019) Estimation of the random error of binary test using adaptive polynomials. J Qual Technol 51(1):81–93

Albert PS (2007) Random effects modeling approaches for estimating ROC curves from repeated ordinal tests without a gold standard. Biometrics 63(2):593–602

Alonzo TA, Pepe MS (1999) Using a combination of reference tests to assess the accuracy of a new diagnostic test. Stat Med 18:2987–3003

Chiang CL (1951) On the design of mass medical surveys. Hum Biol 23(3):242–71

Field SF, Beraldi D, Bachman M, Stewart SK, Beck S et al (2015) Accurate measurement of 5-methylcytosine and 5-hydroxymethylcytosine in human cerebellum DNA by oxidative bisulfite on an array (OxBS-array). PLoS ONE 10(2):e0118202

Gelaye B, Tadesse MG, Williams MA, Fann JR, Vander SA, Zhou X (2014) Assessing validity of a depression screening instrument in the absence of a gold standard. Ann Epidemiol 24(7):527–531

Groeneveld RA, Meeden G (1977) The mode, median, and mean inequality. Am Stat 31(3):120–121

Guglielmi G (2020) Fast coronavirus tests: what they can and can't do. Nature 585:496–498

Hui SL, Walter SD (1980) Estimating the error rates of diagnostic tests. Biometrics 36(1):167–171

Hui SL, Zhou XH (1998) Evaluation of diagnostic tests without gold standards. Stat Methods Med Res 7:354–370

Lachenbruch P (1988) Multiple reading procedures: the performance of diagnostic tests. Stat Med 7:549–557

Nissen-Meyer S (1964) Evaluation of screening tests in medical diagnosis. Biometrics 20(4):730–755

Pepe MS, Janes H (2007) Insights into latent class analysis of diagnostic test performance. Biostatistics 8(2):474–484

Politser P (1982) Reliability, decision rules, and the value of repeated tests. Med Decis Making 2(1):1

Slynko A, Benner A (2019) Statistical methods for classification of 5hmC levels based on the Illumina Inifinium HumanMethylation450 (450k) array data, under the paired bisulfite (BS) and oxidative bisulfite (oxBS) treatment. PLoS ONE 14(6):e0218103

Thibodeau LA (1981) Evaluating diagnostic tests. Biometrics 37(4):801–804

Vacek PM (1985) The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics 41(4):959–968

Wang C, Hanson TE (2019) Estimation of sensitivity and specificity of multiple repeated binary tests without a gold standard. Stat Med 38(13):2381

Wang Z, Dendukuri N, Zar HJ, Joseph L (2017) Modeling conditional dependence among multiple diagnostic tests. Stat Med 36:1–17

Yerushalmy J (1947) Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. Public Health Rep 62(40):1432–1449