Reconciling Binary Replicates: Beyond the Average

Hadrien Lorenzo, Pierre Pudlo, Manuela Royer-Carenzi

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Reconciling Binary Replicates: Beyond the Average

H. Lorenzo | **P. Pudlo** | **M. Royer-Carenzi**

[1]UMR 7373, CNRS, Centrale Marseille, I2M, Aix Marseille Univ, Marseille, France

Correspondence
Corresponding author P. Pudlo.
Email: pierre.pudlo@univ-amu.fr

Abstract

To improve data quality, observations are frequently repeated on the same individual, producing technical replicates, which we assume here to be binary (0 or 1). These replicates are prone to natural variability and measurement errors. In medical contexts, state 1 often indicates the presence of a pathology, while state 0 indicates its absence. From these replicates, we aim to address two critical questions: (1) how to infer the latent state of individuals, represented as 0, 1, or ¡ in cases of indecision, and (2) how to estimate the probability that an individual is affected by the pathology. Although the average of replicates is widely used for such tasks, it has limitations, particularly in terms of bias and reliability. In this paper, we propose alternative approaches: the median of replicates, maximum likelihood estimation, and a Bayesian algorithm. We provide a comprehensive theoretical analysis demonstrating the superiority of these methods over the average-based approach. The Bayesian method, in particular, offers significant advantages by incorporating uncertainty and yielding credible intervals for parameter estimates. We support our findings with extensive simulations and apply the proposed methods to real-world medical datasets. These include a periodontal dataset and a mammogram screening dataset, where we highlight the practical implications of our methodology for improving diagnostic accuracy and estimating disease prevalence. Our work demonstrates the importance of using robust statistical methods to process binary replicates, offering valuable tools for reliable decision-making in noisy measurement contexts.

KEYWORDS
Technical replicates, prevalence estimation, medical diagnosis, median, Expectation-Maximization algorithm, Bayesian algorithm

1 | INTRODUCTION

In Medicine, Biology, and more generally, applied sciences, conclusions are drawn from carefully analyzing noisy data. Noise may come from the measurement protocol, the instruments, or the intrinsic variability of the phenomenon under study. In this context, using technical replicates is a common practice to improve the quality of the conclusions. As Palmer said in [7], "Measurement errors are unavoidable and repetitions should be the rule to quantify its magnitude". However, dealing with technical replicates is not always straightforward: replicates of the same individuals are more likely to be similar than replicates of different individuals. This phenomenon is known as overdispersion, and it has been widely studied in the literature [11, 5, 4, 3]. Jaeger [5] showed that not taking into account the correlation of the responses for the same subject leads to biased estimates and artificially increases our trust in the estimators, which can lead to invalid statistical analyses. A common practice to deal with this correlation is to summarize the replicates of the same individual by a single score, which is usually the empirical average of the replicates.

In this paper, we focus on binary data: each replicate is a binary variable and the true state of the individual is also a binary variable. The major difficulty is that a binary variable carries very little information, and noise can change entirely the value to its opposite. In medical diagnostics, the true state is equal to 1 when the subject carries the disease and 0 otherwise. The replicates are binary values that aim to infer the true state of the individual.

To deal with the binary replicates, we introduce and compare three statistical methods: an average-based, a median-based, a maximum-likelihood-based and a Bayesian methods. Each method provides a different way of scoring the individuals. We

also design a classifier for each method that makes decisions for each individual and returns a trustworthy diagnostic. Because of binary variables' relative lack of information, we have introduced an indecision response, which our classifiers can return in cases of significant doubt. Finally, we propose different methods to estimate the prevalence of the disease in the sampled population.

We organized the paper as follows. Section 2 describes the statistical model, the four competing approaches, and their rationale. After that, in Section 3, we give mathematical results to compare their efficiency. In Section 4 simulations are used to compare the methods and numerical illustrations on medical datasets are provided: a periodontal dataset [3] and a synthetic mammogram screening dataset [1, 6]. The methods described in the paper are implemented in an R-package, available on GitHub at the following address: https://github.com/pierrepudlo/BinaryReplicates.

2 | METHODS

In Section 2.1 we introduce a statistical model for binary replicates that intends to capture a binary state of an individual. To infer the prevalence of each state in the population and to classify the individuals, we introduce four statistical methods: an average-based, a median-based, a maximum a posteriori, and a Bayesian methods. All these methods rely on scoring the individuals given their binary replicates; see Sections 2.3, 2.4 and 2.5. Finally, for each method, we introduce classifiers to recover the binary state of each individual based on the scoring and estimators of the prevalence in Section 2.7.

2.1 | The statistical model

We consider a dataset comprising binary observations $X_{ij} \in 0, 1$, where $i = 1, \dots, N$ indexes individuals and $j = 1, \dots, n_i$ indexes technical replicates. Each replicate X_{ij} provides an approximation of the true status $T_i \in 0, 1$, used to mitigate measurement imprecision. For instance, in medical diagnostics, $T_i = 1$ denotes disease presence, while $T_i = 0$ indicates health. We assume that the replicates X_{ij} are independent noisy measurements of T_i. The marginal distribution of T_i is Bernoulli with parameter θ_T, representing disease prevalence. For each replicate j,

$$\begin{aligned} p &= \mathbb{P}(X_{ij} = 1 | T_i = 0) \quad \text{is the false-positivity rate,} \\ q &= \mathbb{P}(X_{ij} = 0 | T_i = 1) \quad \text{is the false-negativity rate.} \end{aligned}$$

The measurement error $T_i - X_{ij}$ takes values in $-1, 0, 1$, with 0 indicating no error, 1 a false positive, and -1 a false negative.

By the law of total probability, $\mathbb{E}(T_i - X_{ij}) = \theta_T q + (\theta_T - 1)p$, which is non-zero if $\theta_T \neq p/(p + q)$. Thus, the diagnostic test is biased under these conditions. If either p or q exceeds 1/2, the replicate is more likely wrong than correct, resulting in unreliable data. We assume $p, q \in (0, 1/2)$. In what follows, we assume that $p, q \in (0, 1/2)$.

2.2 | Sufficient statistics

We summarize the entire dataset of X_{ij}, for $i = 1, \dots, N$ and $j = 1, \dots, n_i$, by introducing the sum $S_i = X_{i1} + \cdots + X_{in_i}$ for each individual i. This sum represents the number of positive replicates for individual i. The vector $(S_1, \dots, S_N)$ serves as a sufficient statistic for the dataset. Indeed, the only information lost in this transformation is the ordering of the replicates for each individual. However, since the technical replicates for a given individual are exchangeable, their ordering carries no relevance to the statistical analysis. Therefore, no statistical information is lost when we replace the dataset with the vector $(S_1, \dots, S_N)$.

Given $T_i = 1$, resp. $T_i = 0$, the technical replicates X_{ij} are independent Bernoulli random variables with parameter $1 - q$, resp. p. Thus, S_i given the true value T_i is

$$\left[S_i \middle| T_i\right] \sim \mathscr{B}\text{in}\big(n_i, T_i(1 - q) + (1 - T_i)p\big) = \mathbb{1}_{T_i=1}\mathscr{B}\text{in}(n_i, 1 - q) + \mathbb{1}_{T_i=0}\mathscr{B}\text{in}(n_i, p) \tag{1}$$

with independence between the S_i's given $(T_1, \dots, T_N)$.

2.3 | Scoring technical replicates with the mean and the median

The average-based score $Y_{A,i}$ is defined as

$$Y_{A,i} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} = \frac{S_i}{n_i}.$$

This score is heavily used in practice, as it is a simple scaling of the sufficient statistic S_i.
The median-based score $Y_{M,i}$ is defined as

$$Y_{M,i} = \text{median}\left(X_{i1}, \ldots, X_{in_i}\right) = \mathbb{1}_{S_i > n_i/2} + \frac{1}{2}\mathbb{1}_{S_i = n_i/2}. \tag{2}$$

In the specific case where n_i is even and when there is a tie in the frequencies of 0 and 1 within the replicates of individual i, we set $Y_{M,i} = 1/2$.
Note that with the above definitions, the mean and median scores are equal when $n_i \in \{1, 2\}$. However, their effectiveness is not guaranteed, and we will compare them with other scoring methods that are non-linear functions of S_i.

2.4 | Maximum-A-Posteriori scoring

The most likely value of T_i given the observed values of S_i depends on the value of $\mathbb{P}(T_i = 1|S_i = s_i)$, whether below or above 1/2. Using (1), $T_i \sim \mathscr{B}\text{er}(\theta_T)$ and the Bayes formula, the distribution of T_i given S_i is

$$\left[T_i \middle| S_i = s_i\right] \sim \mathscr{B}\text{er}\left(\frac{\theta_T q^{n_i - s_i}(1-q)^{s_i}}{\theta_T q^{n_i - s_i}(1-q)^{s_i} + (1-\theta_T)p^{s_i}(1-p)^{n_i - s_i}}\right). \tag{3}$$

Thus, we introduce the likelihood-based score $Y_{L,i}(\theta_T, p, q)$ as

$$Y_{L,i}(\theta_T, p, q) = \mathbb{P}(T_i = 1|S_i = s_i) = \frac{\theta_T q^{n_i - s_i}(1-q)^{s_i}}{\theta_T q^{n_i - s_i}(1-q)^{s_i} + (1-\theta_T)p^{s_i}(1-p)^{n_i - s_i}}. \tag{4}$$

This score is an increasing function of s_i because of Lemma 2 in A.3. This score $Y_{L,i}(\theta_T, p, q)$ depends on the values of p, q, and θ_T. It is possible to estimate the fixed parameters θ_T, p, and q by maximum likelihood. The likelihood of the fixed parameters θ_T, p, and q given the data $(s_1, \ldots, s_N)$ is, up to a multiplicative constant,

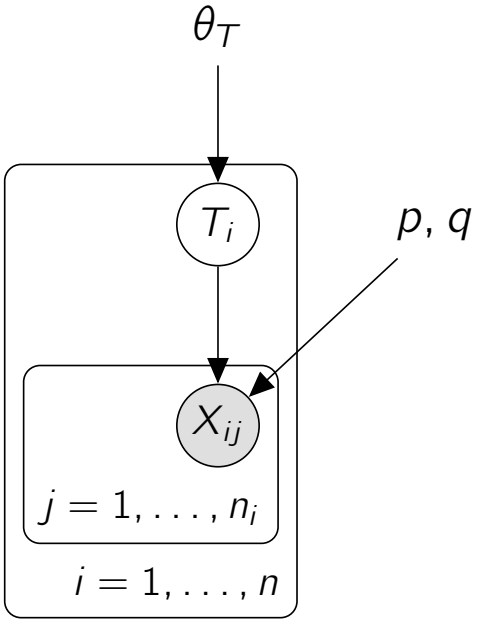
$$\mathcal{L}(\theta_T, p, q) = \prod_{i=1}^{n}\left\{\theta_T(1-q)^{s_i}q^{n_i - s_i} + (1-\theta_T)p^{s_i}(1-p)^{n_i - s_i}\right\}. \tag{5}$$

Even if the above likelihood is an explicit function of the fixed parameters, we cannot maximize it to obtain an explicit formula of the maximum-likelihood estimator. Besides, this maximum might suffer from unrealistic maxima, corresponding to $p = 0$ or $q = 0$ for example. For this reason we consider the posterior distribution of the parameter considering $\mathscr{B}\text{eta}(2, 2)$ prior on both p and q while the prior on θ_T is kept uniform. This will have to effect to penalize negatively those extreme solutions putting 0 weight on them. The maximum of the corresponding functional, the Maximum-A-Posteriori (MAP), is not explicit. As for maximum likelihood, we need to resort to numerical optimization. The standard algorithm to fulfill this task in mixture models is the Expectation-Maximization (EM) algorithm, see, e.g, Chapters 1, 2 and 9 of [2]. For further details, we refer to A.1.

With the result of the EM algorithm we get an approximation of the maximum a posterior estimator, we define the MAP score $Y_{\text{MAP},i}$ as

$$Y_{MAP,i} = Y_{L,i}(\widehat{\theta}_{T,MAP}, \widehat{p}_{MAP}, \widehat{q}_{MAP}), \tag{6}$$

where $(\widehat{\theta}_{T,MAP}, \widehat{p}_{MAP}, \widehat{q}_{MAP})$ is the approximation of the Maximum-A-Posteriori estimator such as described in Algorithm 1 in A.1.



$$
\begin{aligned}
\theta_T &\sim \mathscr{B}\text{eta}(a_T, b_T),\\
p &\sim \mathscr{B}\text{eta}(a_{\text{FP}}, b_{\text{FP}})_{|p\in(0,1/2)},\\
q &\sim \mathscr{B}\text{eta}(a_{\text{FN}}, b_{\text{FN}})_{|q\in(0,1/2)},\\
\big[T_i\big|\theta_T\big] &\sim \mathscr{B}\text{er}(\theta_T),\\
\big[X_{ij}\big|T_i, p, q\big] &\sim \mathscr{B}\text{er}\big(T_i(1-q) + (1-T_i)p\big),\\
\big[S_i\big|T_i, p, q\big] &\sim \mathscr{B}\text{in}\big(n_i, T_i(1-q) + (1-T_i)p\big),\\
\text{where } S_i &= \sum_{j=1}^{n_i} X_{ij}.
\end{aligned}
$$

FIGURE 1 The Bayesian model: the Directed Acyclic Graph (left) and the generative model (right). Variables θ_T, p, and q are fixed parameters (with a prior distribution), the T_i's are latent variables, and the X_{ij}'s are the observed data. The hyperparameters $a_T, b_T, a_{\text{FP}}, b_{\text{FP}}, a_{\text{FN}}, b_{\text{FN}}$ set the prior distribution and should be chosen by the user. The prior distribution on p and q are Beta distributions truncated so that p and q are both in $(0, 1/2)$. The loop over the replicates $j = 1, \ldots, n_i$ can be replaced by a single node S_i, counting the number of ones among the X_{ij}'s since it is a sufficient statistic.

2.5 | Bayesian scoring

We encompass the stochastic model described above in a Bayesian framework. We consider a Bayesian model with a prior distribution on θ_T, p, and q whose density is denoted $\pi(\theta_T, p, q)$, see Figure 1. In the Bayesian framework, $\mathbb{P}_\pi$ and $\mathbb{E}_\pi$ denote the probability measure and the expected value, as detailed in A.2

Using the Bayesian model defined above and its posterior distribution, the Bayesian score $Y_{B,i}$ is defined as the following posterior expected value

$$Y_{B,i} = \mathbb{E}_\pi\big(T_i\big|S_1, \ldots, S_N\big) = \mathbb{P}_\pi\big(T_i = 1\big|S_1, \ldots, S_N\big),$$

that integrates T_i over the posterior values of θ_T, p, and q and thus over the uncertainty on the fixed parameters. Alternatively, assuming $\pi(\theta_T, p, q|S_1, \ldots, S_N)$ is the posterior distribution of the fixed parameters given the data, it can be viewed as the posterior expected value of the likelihood-based score $Y_{L,i}(\theta_T, p, q) = \mathbb{P}(T_i = 1|S_i)$, namely

$$Y_{B,i} = \int Y_{L,i}(\theta_T, p, q)\pi(\theta_T, p, q|S_1, \ldots, S_N)\mathrm{d}\theta_T\mathrm{d}p\mathrm{d}q. \tag{7}$$

Unlike scores $Y_{A,i}$ and $Y_{M,i}$, values of $Y_{MAP,i}$ and $Y_{B,i}$ depend not only on the replicates observed on the i-th individual (i.e., S_i) but also on the whole dataset through the posterior distribution on the fixed parameters.

To decide how to set the hyperparameters for the prior distribution of the fixed parameters θ_T, p, and q, see Figure 1, remember that each of these parameters represents the probability of getting a 1 in a binary (0/1) trial. For these probabilities, we use the Beta distribution $\mathscr{B}\text{eta}(a, b)$, where $a, b > 0$. This distribution represents information from a sample of size $a + b$ consisting of a observations equal to 1 and b equal to 0. A non-informative prior is the Beta distribution with $a = b = 1/2$, representing minimal prior knowledge. This can be thought of as a fictive sample of size 1 with equal likelihoods for 0 and 1. We recommend using this non-informative prior for the prevalence θ_T. The uniform distribution on $(0, 1)$ is a Beta distribution

with $a = b = 1$, providing a weakly informative prior based on a balanced sample of size 2. For the false-positive rate p and the false-negative rate q, setting $a_{\text{FP}} = a_{\text{FN}} = b_{\text{FP}} = b_{\text{FN}} = 1/2$ is not ideal: this distribution places too much weight near 0, where the likelihood-based scoring performs poorly. Since replicates imply noisy measurements, meaning p and q should not be too close to 0, we recommend using $a_{\text{FP}} = a_{\text{FN}} = b_{\text{FP}} = b_{\text{FN}} = 2$, which puts less weight near 0. In summary, we propose using $a_T = b_T = 1/2$ and $a_{\text{FP}} = a_{\text{FN}} = b_{\text{FP}} = b_{\text{FN}} = 2$ as default values for the hyperparameters, adjusting them as needed if more information is available. See Figure 4, Section 4.2, for an example.

Finally, we can think of the median-based score as a special case of the Bayesian score. Specifically, when the prior is highly concentrated around $(\theta_T, p, q) = (1/2, 0, 0)$, the posterior distribution of the fixed parameters also stays concentrated around $(1/2, 0, 0)$. In this scenario, the Bayesian score $Y_{B,i}$ defined in (7) becomes approximately equal to $Y_{L,i}(1/2, 0, 0)$, which corresponds to the median-based score $Y_{M,i}$. Thus, the median-based score can be seen as an approximation of the Bayesian score when using a strongly informative prior. This strongly informative prior assumes that the false-positive and false-negative rates are negligible and that the prevalence is roughly 1/2. Such a situation occurs with the prior defined in Figure 1, when the hyperparameters $b_{\text{FN}} = b_{\text{FP}} \to \infty$ and $a_T = b_T \to \infty$ simultaneously.

2.6 | From scorings to prevalence and error rates estimators

To infer the prevalence θ_T, we consider the four following estimates:

$$\forall K \in \{A, M, MAP\},\ \widehat{\theta}_{T,K} = \frac{1}{N}\sum_{i=1}^{N} Y_{K,i} \quad \text{and} \quad \widehat{\theta}_{T,B} = \mathbb{E}_{\pi}\left(\theta_T \middle| S_1, \ldots, S_N\right). \tag{8}$$

For average-, median- and maximum a posteriori-based methods, scorings can be used to estimate the false-positivity and false-negativity rates p and q:

$$\forall K \in \{A, M, MAP\},\ \widehat{p}_K = \frac{\sum_{i=1}^{N} S_i(1 - Y_{K,i})}{\sum_{i=1}^{N} n_i(1 - Y_{K,i})} \text{ and } \widehat{q}_K = \frac{\sum_{i=1}^{N}(n_i - S_i)Y_{K,i}}{\sum_{i=1}^{N} n_i Y_{K,i}}. \tag{9}$$

Bayesian statistics has its way of estimating parameters using their posterior expected values. We can approximate them easily by computing the average of a sample of θ_T, p or q drawn from the posterior distribution with the Hamiltonian Monte Carlo algorithm of `rstan`.

2.7 | From scorings to classifiers

We compute our predictions of the latent T_i values by thresholding the scores. The above scoring statistics take values in $[0, 1]$ and summarize information from a given dataset, the s_is, into a single value. We should interpret them as a tool to infer the latent T_i value: when $T_i = 1$, we expect high scores; when $T_i = 0$, we expect low scores. We introduce an indecision response (1/2) when we would not trust a decision based on the observed replicates. Since binary data carry little information, this could happen. It is an invitation to add more replicates related to this individual before deciding.

To classify the individuals, we introduce two thresholds $0 < v_L \leq 1/2 \leq v_U < 1$ and set the classifiers as

$$\widehat{T}_{K,i} = \Phi(Y_{K,i}), \quad \text{where } \Phi(y) = \begin{cases} 0 & \text{if } y < v_L, \\ 1/2 & \text{if } v_L \leq y \leq v_U, \\ 1 & \text{if } y > v_U, \end{cases} \tag{10}$$

for all methods $K \in \{A, M, MAP, B\}$ and all individuals $i = 1, \ldots, n$. Since the median-based score $Y_{M,i}$ is always in $\{0, 1/2, 1\}$, we always have $\widehat{T}_{M,i} = Y_{M,i}$. This double thresholding method is related to risk theory. To compute the risk of a classifier, we introduce a loss function $\ell(t, \widehat{t})$ that quantifies the cost of predicting $\widehat{t} \in \{0, 1/2, 1\}$ when the truth is $t \in \{0, 1\}$. See Section 3.1 for more details.

2.8 | Predictions

Let assume that a new individual $n+1$ is given to the agent through n_{n+1} and s_{n+1}. It is possible to give the posterior prediction of its score $\widehat{Y}_{n+1}$ based on the dataset composed by the n previous individuals. For $K \in \{A, M, MAP\}$, we estimate the parameters $\widehat{\theta}_{T,K}, \widehat{p}_K$ and $\widehat{q}_K$, from the individuals $\{1, \cdots, n\}$, and we compute

$$\widehat{Y}_{K,n+1} = Y_{L,n+1}(\widehat{\theta}_{T,K}, \widehat{p}_K, \widehat{q}_K), \tag{11}$$

where $Y_{L,n+1}(\theta_T, p, q) = \mathbb{P}(T_{n+1} = 1 | S_{n+1} = s_{n+1})$ is computed as in Equation (4). It is also possible to build the prediction score for Bayesian approach, for which the form is

$$\widehat{Y}_{B,n+1} = \int Y_{L,n+1}(\theta_T, p, q)\, \pi(\theta_T, p, q | S_1, \ldots, S_n)\, \mathrm{d}\theta_T\, \mathrm{d}p\, \mathrm{d}q. \tag{12}$$

A Monte-Carlo estimator is chosen in order to approximate the previous integral, such as:

$$\widehat{Y}_{B,n+1} = \frac{1}{H} \sum_{h=1}^{H} Y_{L,n+1}(\theta_{T,h}, p_h, q_h),$$

where each parameter $(\theta_{T,h}, p_h, q_h)$ is sampled from the posterior distribution $\pi(\theta_T, p, q | S_1, \ldots, S_n)$ such as described in Section 2.5.

Note that for $K \in \{A, M, MAP\}$, K-prediction scores do not include any variability over the parameters, while Bayesian-prediction scores do. Thus these K-prediction scores might drive to over-confident decisions.

3 | THEORETICAL RESULTS

In this Section, we provide efficiency results that compare the various methods introduced above. We start with the efficiency of the classifiers in terms of sensitivity and specificity, and of specific loss functions, as introduced in Section 3.1, that deal with indecision responses on the replicates. Section 3.3 gives the results on the classifiers. Section 3.4 gives the results on the prevalence estimators. To state the results, we may need the following hypotheses.

(H1) The false-positivity and false-negativity rates p and q are in $(0, 1/2)$.
(H2) There exists at least one individual for which the number of replicates $n_i \geq 3$.
(H3) The v_L and v_U that define the classifiers in Section 2.7 are such that $0 < v_L \leq 1/2 \leq v_U < 1$.
(H4) The loss function $\ell(t, \widehat{t})$ satisfies the conditions (13) of Lemma 1 in Section 3.1.

In the following, we only look at the following problem. Assume we want to predict a binary random $T \in \{0, 1\}$, with three possible decisions: 0, 1/2 (inconclusive) and 1 based on the known value of $\vartheta = \mathbb{P}(T = 1)$. (In the following section, ϑ can be θ_T or $\mathbb{P}(T_i = 1 | S_i)$ if we reason given the data.)

3.1 | Loss functions for classifiers with indecision response

Consider the loss function $\ell(t, \widehat{t})$, defined on $\{0, 1\} \times \{0, 1/2, 1\}$ such as

$\ell(t, \widehat{t})$	0	1/2	1
0	0	a	b
1	c	d	0

where a, b, c and d are positive constants. When $a = b = 1$ and $c = d = 0$, the loss function is related to the specificity. When $c = d = 1$ and $a = b = 0$, the loss function is related to the sensitivity. And, when $b = c = 1$ and $a = d = 0$, the loss function is the misclassification error. The general loss function $\ell(t, \widehat{t})$ can be interpreted as follows. If a and d are small enough compared to b and c, the indecision response 1/2 may be the best choice in case of a strong uncertainty between 0 and 1. In medical applications, asking for further tests may cost less than making a doubtful decision.

In the following we would consider ℓ in symmetrical form, taking $b = c = 1$ and $a = d$, denoted ℓ_a, where $0 < a < 1$ is the indecision cost. Thus the cost of a false positive, $\ell_a(0,1)$, is equal to the cost of a false negative, $\ell_a(1,0)$. This is unrealistic in the context of medical diagnosis. Indeed, we often wish to avoid false negatives, so as not to leave a diseased patient untreated. In this case, an absence of decision with $\widehat{t} = 1/2$ is better than the false negative. The resulting cost of indecision, a, is thus lower than that of a false negative, set to 1. An indecision response is always an error, whether the truth is 0 or 1. Therefore, decision methods propose $\widehat{t} = 1/2$ only if the cost of indecision is sufficiently low. The constraints are given in Equation (13) of Lemma 1. This gives $a < 1/2$. However, this cost of indecision a must remain high (i.e. not too close to 0) for decisions to be made in most cases.

3.2 | Minimal risk classifiers

For a given loss function ℓ, we are interested in the best classifiers $\widehat{T}^\star$ that minimize the risk $r(\widehat{t}) = \mathbb{E}\big(\ell(T,\widehat{t})\big)$, defined such as:

$$\widehat{T}^\star = \underset{\widehat{t}\in\{0,1/2,1\}}{\arg\min}\ r(\widehat{t}),$$

with an advantage for the indecision response, i.e., 1/2, if there is a tie in the risks. Lemma 1 (a proof is given in A.4) gives the conditions on a, b, c and d under which the indecision response can appear. It also gives the best decision in this case.

Lemma 1. *Assume $\vartheta = \mathbb{P}(T = 1)$ is known. The best classifier $\widehat{T}^\star$ is*

$$\widehat{T}^\star = \Phi(\vartheta),$$

where Φ is defined in Equation (10) *with $v_L = \frac{a}{c-(d-a)}$ and $v_U = \frac{b-a}{(d-a)+b}$*

$$\text{if and only if} \quad \frac{bc}{b+c} > a + (d-a)\frac{b}{b+c} \quad \text{and} \quad -b < (d-a) < c. \tag{13}$$

If we apply Lemma 1 to the loss function ℓ_a, to obtain the best classifier, we must choose $v_L = a$ and $v_U = 1 - v_L$.

3.3 | Accuracy of the classifiers as a diagnostic tool

Whatever the scoring method, we rely on the same two thresholds $v_L \le v_U$ to transform the scores into diagnostics. Let us denote $n_0 = \max_i n_i$ and $\delta_0 = \frac{1}{2n_0}$ if n_0 is odd, $\delta_0 = \frac{1}{2(n_0-1)}$ otherwise. Note that in the limit case where $v_L \in \left[\frac{1}{2} - \delta_0; \frac{1}{2}\right]$ and $v_U \in \left[\frac{1}{2}; \frac{1}{2} + \delta_0\right]$, both average- and median-based classifications are identical. Otherwise, we can compare the sensitivity and specificity of the average-based and median-based classifiers.

Theorem 1. *Assume (H1) and (H3). We have*

$$\text{sensitivity}(\widehat{T}_{A,i}) \le \text{sensitivity}(\widehat{T}_{M,i}).$$

The above inequality is strict if and only if $v_U > 1/2$, becoming $v_U > 1/2 + \delta_0$. Moreover, we have

$$\text{specificity}(\widehat{T}_{A,i}) \le \text{specificity}(\widehat{T}_{M,i})$$

The above inequality is strict if and only if $v_L < 1/2$, becoming $v_L < 1/2 - \delta_0$.

As a consequence, the median-based classifier is also better in terms of the misclassification rate and informedness.

The proof is given in B.1. The theorem states that the median-based classifier is better than the average-based classifier regarding sensitivity and specificity when we introduce the inclusive response, i.e., as soon as $v_L < 0.5 < v_U$. Both classifiers reflect the properties of their respective scoring. Hence, Theorem 1 yields a first conclusion on the efficiency of the scoring methods in favor of the median-based scoring.

We also obtained efficiency results on the Bayesian classifier. To state it, we must refer to the loss function $\ell(t,\widehat{t})$ defined in Section 3.1. Bayesian statistics, which is well grounded in decision theory, is known to be efficient in the sense that it provides

the best possible estimators and classifiers given the data if the statistics are computed wisely with the posterior distribution, see, e.g., [9]. Here, we can prove the following results.

Theorem 2. *Assume (H1) and (H4). Consider any classifier $\widehat{T}_i$ of the i-th individual that returns a decision in $\{0, 1/2, 1\}$, based on the data $S_1, \ldots, S_n$.*

(i) (Optimality of the likelihood-based classifier) Whatever the values of (θ_T, p, q), we have

$$\mathbb{E}(\ell(T_i, \widehat{T}_{L,i}(\theta_T, p, q))) \leq \mathbb{E}(\ell(T_i, \widehat{T}_i)).$$

(ii) (Bayesian optimality of the Bayesian classifier) We have

$$\mathbb{E}_\pi(\ell(T_i, \widehat{T}_{B,i})) \leq \mathbb{E}_\pi(\ell(T_i, \widehat{T}_i)).$$

(iii) (Admissibility of $\widehat{T}_{B,i}$) If on a set of values of (θ_T, p, q) with positive prior probability we have

$$\mathbb{E}\Big[\ell\big(T_i, \widehat{T}_i\big)\Big] < \mathbb{E}\Big[\ell\big(T_i, \widehat{T}_{B,i}\big)\Big],$$

then there exists another set of values of (θ_T, p, q) with positive prior probability for which the inequality is reversed (strictly).

The proof is given in B.2. Note that $\mathbb{E}$ is the expected value given the fixed parameters θ_T, p, and q, whereas $\mathbb{E}_\pi$ integrates them according to the prior distribution of density π. Item (i) states that the likelihood-based classifier is optimal in terms of risk, i.e., of expected loss at fixed values of θ_T, p, and q. Item (ii) considers the Bayesian risk, which is the expected loss integrated over the prior distribution of the fixed parameters. The Bayesian classifier is optimal in this sense. As stated in item (iii), the Bayesian classifier is admissible: no other classifier can outperform it uniformly over the entire set of fixed parameter values. Because of these results, we recommend using the Bayesian classifier in practice.

3.4 | Accuracy of the prevalence estimators

As defined in Equation (8), we have four prevalence estimators at our disposal: the average-based, the median-based, the Maximum-A-Posteriori and the Bayesian prevalence estimators. The latter $\widehat{\theta}_{T,B}$ is the expected value of the posterior distribution of the prevalence θ_T given the data. In contrast, the former three $\widehat{\theta}_{T,K}$, with $K = \{A, M, MAP\}$, are the empirical means of the associated scores $Y_{K,i}$ for $i = 1, \cdots N$.

We first consider the two empirical means $\widehat{\theta}_{T,A}$ and $\widehat{\theta}_{T,M}$. They are heavily biased, with a bias that does not tend to 0 as the number of individuals increases. On the other hand, their variances are proportional to $1/n$, see D.3. Thus, asymptotically, their squared biases dominate their mean squared errors. Moreover, we can compare their bias as follows.

Theorem 3. *(Bias of $\widehat{\theta}_{T,A}$ and $\widehat{\theta}_{T,M}$) Assume (H1) and set $n_0 = \min\{n_i, i = 1, \ldots, N\}$.*

For any values of p and q in $(0, 1/2)$, there exists an interval J that contains $p/(p+q)$ such that

$$|\mathbb{E}\big(\widehat{\theta}_{T,M}\big) - \theta_T| \leq |\mathbb{E}\big(\widehat{\theta}_{T,A}\big) - \theta_T|,$$

except if $\theta_T \in J$.

The bias of $\widehat{\theta}_{T,A}$ is not influenced by the numbers of replicates n_i's. Whereas, as $n_0 \to \infty$, then the bias of $\widehat{\theta}_{T,M}$ tends to 0 and the length of the interval J tends to 0.

The proof is given in D.2. The theorem states that the median-based prevalence estimator $\widehat{\theta}_{T,M}$ is better than the average-based prevalence estimator $\widehat{\theta}_{T,A}$ in terms of bias, except when θ_T is in an interval J. And the length of J is small when the number of replicates is always large. In the latter case, we can rely on the median-based $\widehat{\theta}_{T,M}$ to estimate the prevalence. Otherwise, both estimators are heavily biased and should be used with caution.

As always, Bayesian statistics come with its efficiency. In terms of mean squared error, we can prove the following results on the Bayesian prevalence estimator $\widehat{\theta}_{T,B}$.

Theorem 4. *Assume (H1), and consider an estimator $\widehat{\theta}_T$ of θ_T, that is to say any function of the data $S_1, \ldots, S_N$.*

(i) (Bayesian optimality of $\widehat{\theta}_{T,B}$) We have

$$\mathbb{E}_\pi\left[(\widehat{\theta}_{T,B}-\theta_T)^2\right] \leq \mathbb{E}_\pi\left[(\widehat{\theta}_T-\theta_T)^2\right].$$

(ii) (Admissibility of $\widehat{\theta}_{T,B}$) If on a set of values of (θ_T, p, q) with positive prior probability we have

$$\mathbb{E}\left[\left(\widehat{\theta}_T-\theta_T\right)^2\right] < \mathbb{E}\left[\left(\widehat{\theta}_{T,B}-\theta_T\right)^2\right],$$

then there exists another set of values of (θ_T, p, q) with positive prior probability for which the inequality is reversed (strictly).

The proof is given in D.4. The above Theorem states that $\widehat{\theta}_{T,B}$ is optimal in terms of the Bayesian L^2-risk, which is the mean squared error integrated over the prior distribution. Moreover, the Bayesian estimator is admissible, which means that there is no other statistic whose mean squared error is always smaller than the one of $\widehat{\theta}_{T,B}$ whatever the values of the fixed parameters. Additionally, the Bayesian methodology evaluates the uncertainty of the estimated value with credible intervals. These are strong arguments in favor of the Bayesian prevalence estimator, and we recommend its use in practice.

4 | NUMERICAL RESULTS

To evaluate the performance of a method on a specific dataset, we use the empirical $\overline{\ell}_a$-risk, defined as

$$\overline{\ell}_a\text{-risk} = \frac{1}{N}\sum_{i=1}^{N} \ell_a(t_i, \widehat{t}_i),$$

i.e., the average of the losses we commit with all decisions taken for each observation i, where the loss function ℓ_a was defined in Section 3.1.

4.1 | Some simulations

Simulations are based on the Bayesian model described in Figure 1 where $p = 0.1, q = 0.05$ and n_is are sampled through a following uniform distribution $n_i \sim \mathscr{U}_{[\![2,6]\!]}$ and the Table 1 details how priors are built.

	a_{FP}	b_{FP}	a_{FN}	b_{FN}	a_{T}	b_{T}
Default Bayesian	2	2	2	2	0.5	0.5
Misguided Bayesian	50	50	50	50	0.5	0.5

TABLE 1 Chosen priors for simulations for $\theta_T \in [0.1, 0.5]$.

First we compare the quality of estimating θ_T through the different approaches. Parameter θ_T is evenly sampled between 0.01 and 0.5. For each θ_T, datasets of size $N = 200$, have been sampled. Figure 2 gives the results of estimating the parameter θ_T, where $\hat{\theta}_{T,K}$ is one of estimators produced by any of the 6 considered approaches. Medians (thick lines) and quartiles (shaded areas) are represented. The closer each curve is to the straight black line, the better the approach is. As expected from Theorem 3, the sample mean of Median-scorings produce better prevalence-estimates than the sample mean of Average-scorings. Moreover, for Average and Median approaches, the bias obtained on simulations follows the linear bias expected from D.2, plotted as thin straight lines. Furthermore Maximum-A-Posteriori outperforms other approaches. Default Bayesian keeps very close of these optimal methods, followed by Median. On the contrary, the prevalence value inferred by Average is the worst-performing.

Next Figure 3 gives the $\overline{\ell}_a$-risks, for $\theta_T = 0.4$, through their median and their $[0.4, 0.6]$-quantile area in shaded. As expected from B.3, the sample mean of Median-risks is linear and increasing, whereas the sample mean of Average-risks is linear and increasing, piecewise, where jumps occur at all the observed values of $\frac{s_i}{n_i} < 0.5$, when n_i takes values in $[\![2, 6]\!]$. In our simulations, the Average approach gives the poorest results. The Bayesian and MAP approaches are always better than both the Average and the Median solutions. The Default Bayesian performs better than Misguided.

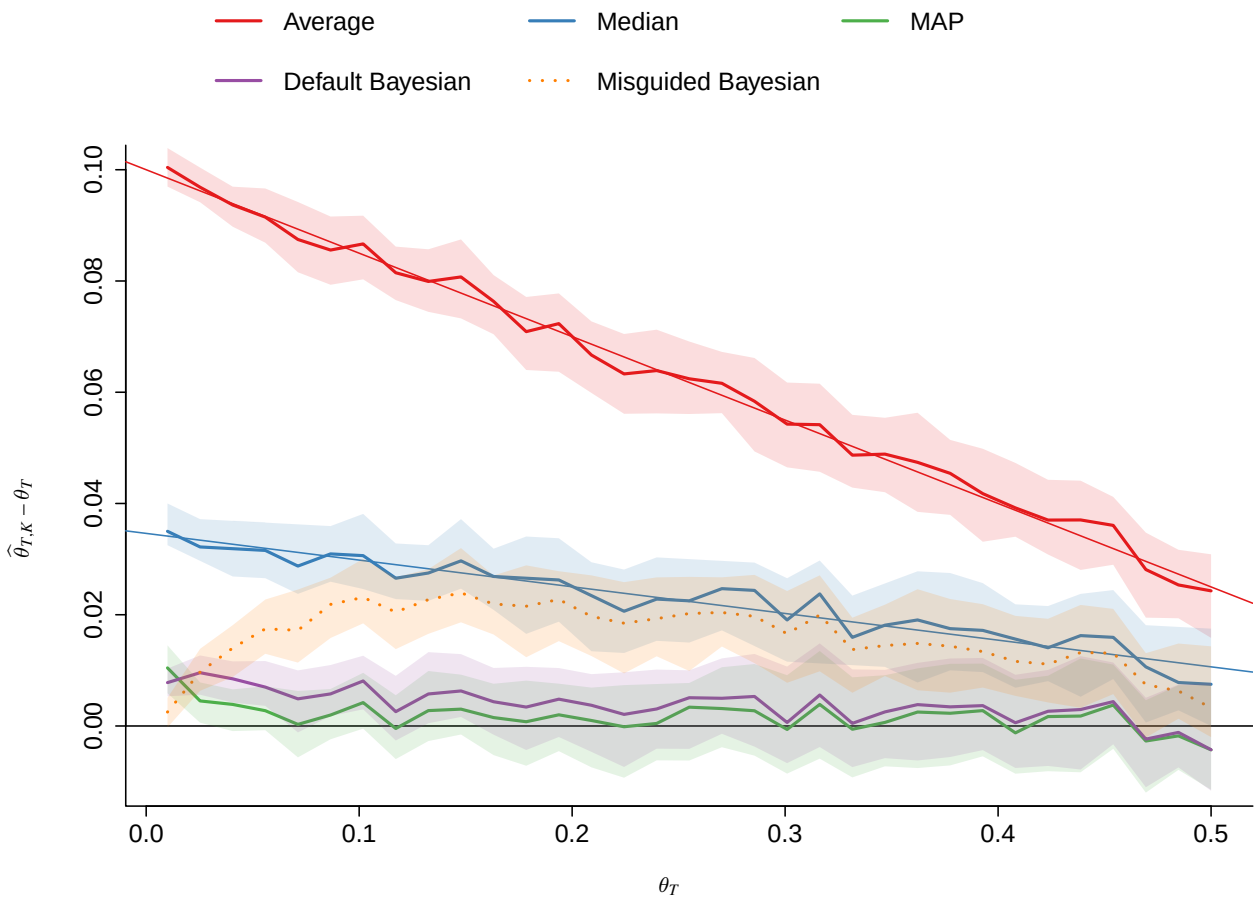


FIGURE 2 Bias in prevalence estimation for simulated datasets, versus $\theta_T \in [0.01, 0.5]$. Medians are plotted in thick lines and $[0.4, 0.6]$-quantile areas are filled with shaded colors, 4 datasets have been sampled for each value given to θ_T. For Average and Median approaches, the theoretical bias, computed in D.2, are plotted as thin straight lines. Horizontal black line correspond with the objective: null error on estimating θ_T.

In Section 3.4, we recommended using the Bayesian prevalence estimator. The numerical results are in agreement with this recommendation.

4.2 | A periodontal dataset

We tried the proposed methods on the periodontal dataset of [3], which includes $N = 50$ individuals. The number of replicates per individual, n_i, varies from 1 to 6. The status variable of the dataset provides a trustworthy diagnostic (*healthy* or *infected*). We used it as the true value of T_i. We ran 1) the average-based method 2) the median-based method 3) the Maximum-A-Posteriori estimator approximation such as described in the Algorithm 1 4) the default Bayesian method with the default prior given in Section 2.5, which is weakly informative and 5) a misguided Bayesian method. The a priori and the posterior distributions of Bayesian methods on the fixed parameters are displayed in Figure 4.

Since we are in the exceptional cases where the latent status of each patient T_i is known, it is possible to infer the theoretical parameters θ_T, p and q, with the following estimates:

$$\widehat{\theta}_{T,\text{latent}} = \frac{1}{N}\sum_{i=1}^{N} T_i, \quad \widehat{p}_{\text{latent}} = \frac{\sum_{i=1}^{N} S_i(1-T_i)}{\sum_{i=1}^{N} n_i(1-T_i)} \text{ and } \widehat{q}_{\text{latent}} = \frac{\sum_{i=1}^{N}(n_i - S_i)T_i}{\sum_{i=1}^{N} n_i T_i}. \tag{14}$$

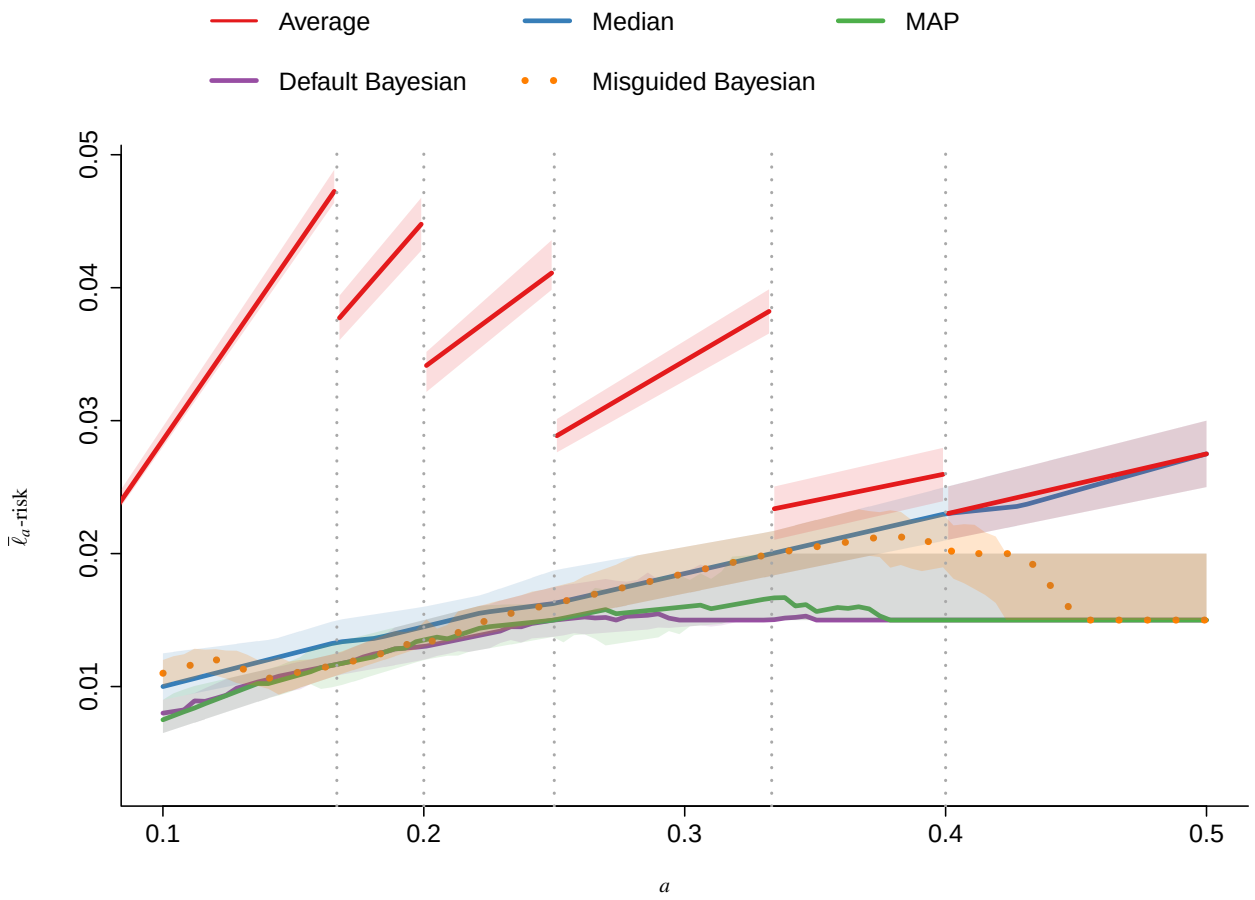


FIGURE 3 Comparison of all methods when the decision cost a varies and $\theta_T = 0.4$. Their performance is measured in terms of empirical $\overline{\ell}_a$-risk. We run the methods on 300 simulated datasets for $a \in [0.1, 0.5]$. The plain line is the median of the 300 empirical risks and the band represents the $[0.4, 0.6]$-quantile interval. The Average-risk jumps when a is equal to an observed value $\frac{s_i}{n_i}$, where $n_i \in [\![2, 6]\!]$. In other words, jumps occur when a crosses $\{\frac{1}{6}; \frac{1}{5}; \frac{1}{4}; \frac{1}{3}; \frac{2}{5}\}$, represented as vertical dotted lines.

Then we can infer the prevalence in the data set as $\widehat{\theta}_{T,\text{latent}} = 29/50 = 0.58$, the false positivity rate as $\widehat{p}_{\text{latent}} = 9/48 = 0.187$ and the false negavity rate as $\widehat{q}_{\text{latent}} = 48/142 = 0.338$. Globally we see that Default and Misguided Bayesian methods provides very different distributions, illustrating the influence of priors in Bayesian methods. Here Default Bayesian is quite well centered on the parameter values estimated from the latent status T_i. Indeed, the modes of the the Default Bayesian posterior density are 0.67, 0.12 and 0.29 for θ_T, p and q respectively. On the contrary, Misguided Bayesian provides a peaked posterior distribution for p and q, that is completely shifted towards 0.5 (respective modes are 0.50 and 0.46), whereas it provides an uniform posterior distribution for θ_T estimation.

Furthermore, the estimation of parameters p, q and θ_T provided by Average, Median and Maximum-A-Posteriori methods, computed from Equations (8) and (9), are represented in Figure 4 as points below the associated graphics. Contrary to Bayesian methods that provide a posterior distribution, these methods just provide an estimated value.

Next, we set the thresholds as $v_L = 0.45$ and $v_U = 1 - v_L$ to compute the classifiers. This means that an indecision response is given when the scorings are too close to 1/2. We compared classifications to the diagnostics given by the status variable in Table 2, which is a kind of confusion matrix. Since the thresholds v_L and v_U are close to 0.5, the average-based and the median-based classifiers give the same classification. Indeed, in this example we have $n_0 = 6$, then $a = 0.45 > 1/2 - \delta_0 = 0.4$, as defined in Section 3.3.

According to Lemma 1, classification-accuracy is fitted to the loss function $\ell_{0.45}(t, \widehat{t})$. According to the empirical $\ell_{0.45}$-risk function, given in Table 2, the Default Bayesian methods is the best, and the misguided Bayesian method is the worst. The

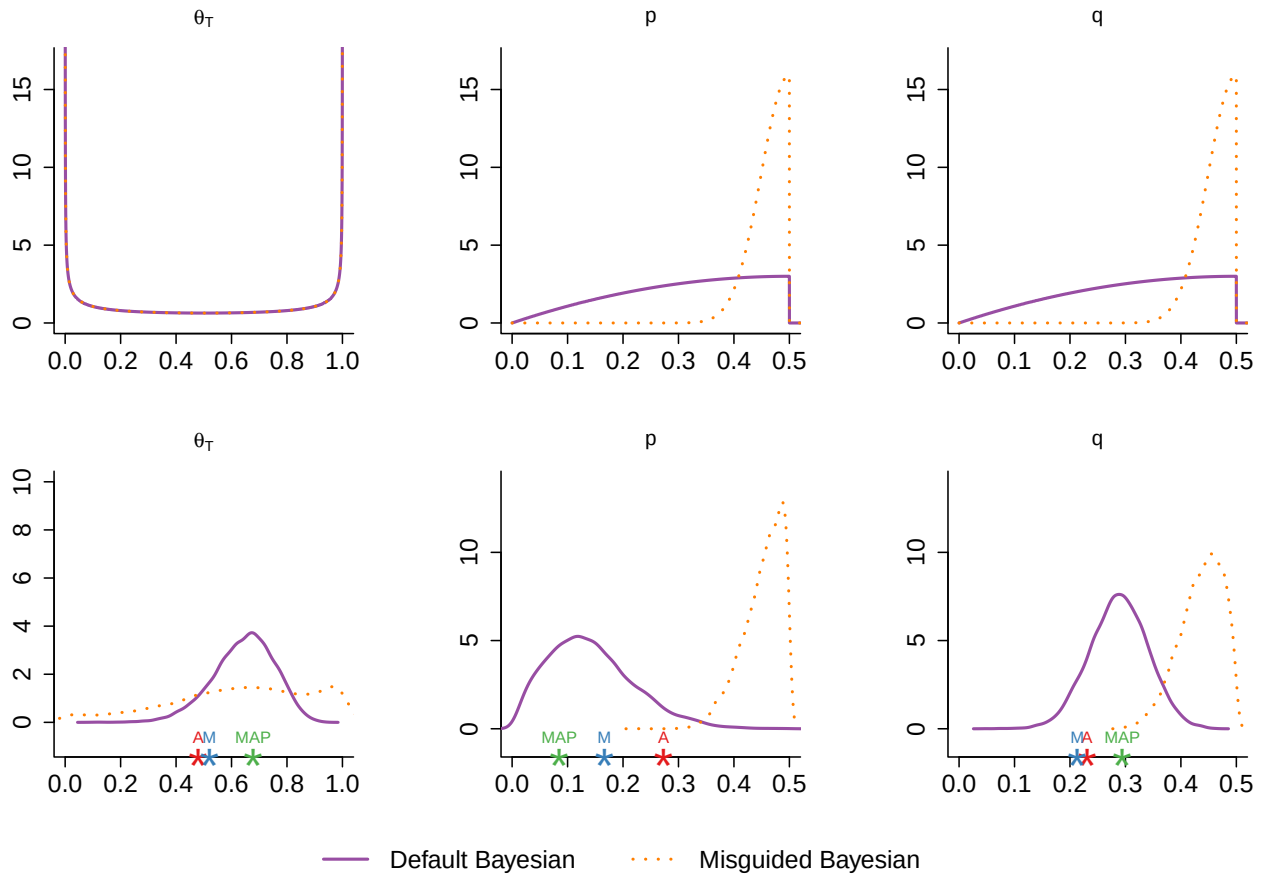


FIGURE 4 Comparison of the prior distributions (first row) and the associated posterior distributions (second row). From the latent status of each patient, we infer the parameters $\widehat{\theta}_{T,\text{latent}} = 0.58$, $\widehat{p}_{\text{latent}} = 0.187$ and $\widehat{q}_{\text{latent}} = 0.338$. The parameters $\widehat{\theta}_{T,K}$, $\widehat{p}_K$ and $\widehat{q}_K$, estimated by non-Bayesian methods (K=A, M, MAP) are displayed below the respective graphs with stars. .

TABLE 2 Classification of the periodontal dataset. The table on the left counts how many times a decision has occurred given the method and the status of the individual. The thresholds were set to $v_L = a = 0.45$ and $v_U = 0.55$. The table on the right gives the empirical $\ell_{0.45}$-risk of each method, corresponding to $a = 0.45$.

Status	Method	$\widehat{T} = 0$	$\widehat{T} = 1/2$	$\widehat{T} = 1$
Healthy	Average	16	1	4
Healthy	Median	16	1	4
Healthy	MAP	13	0	8
Healthy	Default	13	3	5
Healthy	Misguided	1	7	13
Infected	Average	5	5	19
Infected	Median	5	5	19
Infected	MAP	3	2	24
Infected	Default	3	2	24
Infected	Misguided	1	4	24

Method	Emp. $\ell_{0.45}$-risk
Average	11.7
Median	11.7
MAP	11.9
Default	10.2
Misguided	18.9

Misguided has a bad $\ell_{0.45}$-risk because it diagnoses all patients as infected, whether they are actually infected or not. Note that Default Bayesian and Maximum-A-Posteriori provide the same decisions, except for 3 healthy individuals. On those two individuals, the Default method is in favor of 1/2 (inconclusive), whereas the Maximum-A-Posteriori method is in favor of 1. Since predicting 1 for an healthy patient is a larger error than predicting 1/2, the empirical $\ell_{0.45}$-risk of the default Bayesian method (10.2) is smaller than that of the Maximum-A-Posteriori method (11.9).

4.3 | A mammogram screening dataset

In [1, 6], the authors consider $N = 148$ women that may suffer or not from a breast cancer. The infection-status of the patient is given by a gold-standard diagnostic test. 64 patients are confirmed cancer cases, and 84 are not affected. $n_{rad} = 110$ radiologists read the n mammography films. Further information on this dataset are given in E.1. Since mammography data are sensitive, they are often proprietary and confidential, generally maintained by the American College of Radiology. An agreement with the ACR is necessary to use complete datasets. With only partial data available, we had to run simulations to synthesize complete data.

Moreover, in this dataset, error rates are non homogeneous, which is out of our model. We have simulated datasets corresponding to this non-homogeneity of the error rates. The complete description of the simulation procedure is given in E.2. In the context of mammograms, authorizing a diagnosis of uncertainty is crucial, as it may correspond to situations where radiologists disagree or mammogram images are difficult to interpret. These cases correspond to situations where clinical follow-up (additional tests such as biopsies or a second reading by specialists) is preferable to a categorical decision.

Here we simulate 100 datasets when the number of radiologists are restricted to $n_i = 4$. For each simulation, we use 90% of the 110 mammograms to estimate $(\theta_{T,K}, p_K, q_K)$ for $K = A, M, MAP$ and B. Next we compute the predictions $\widehat{Y}_{K,n+1}$, as described in Equations (11) and (12), for the 15 remaining mammograms. Figure 5 displays the predictions scores versus the observed number of positive replicates $s_i = 0, 1 \cdots, 4$. For each simulation, we additionally compute $\sum_a \overline{\ell}_a$-risk, for a varying from 0.15 to 0.45. The Default Bayesian and the Maximum-A-Posteriori methods outperform the others, whereas Average provides the worst empirical risk.

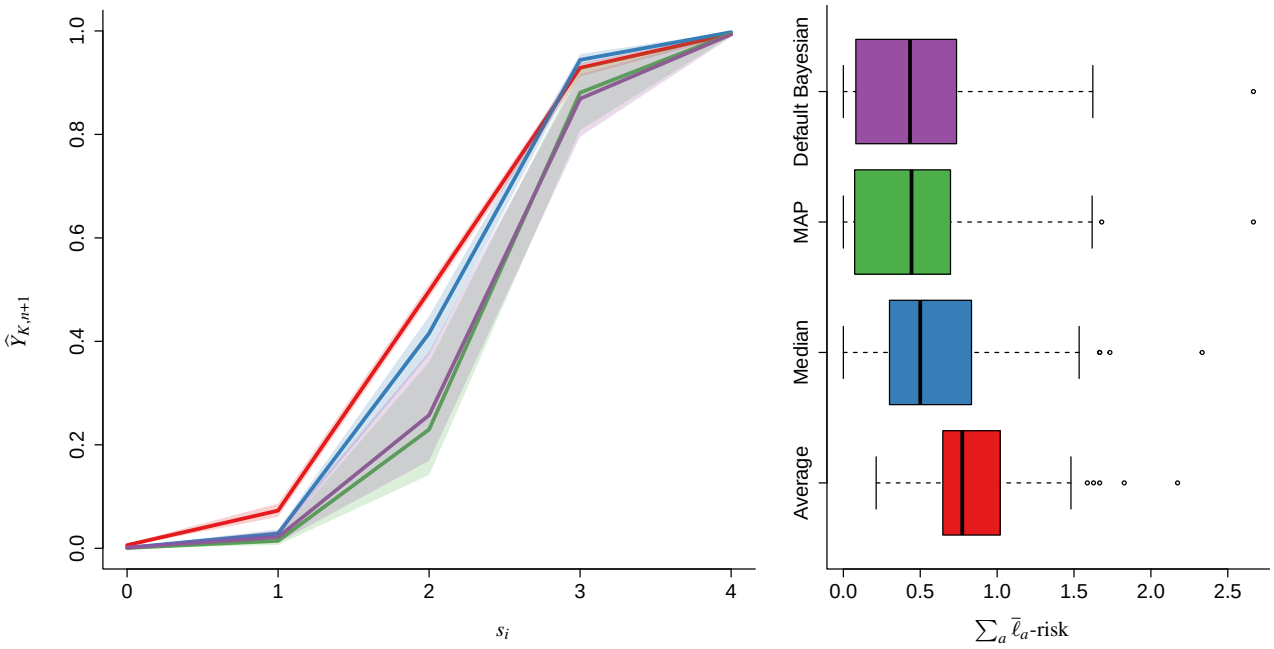


FIGURE 5 Comparison of prediction scores $\widehat{Y}_{K,n+1}$. In the left subfigure, medians (lines) and quartiles (shaded areas) are represented. In the right subfigure, predisctions performance is measured in terms of empirical risk $\sum_a \overline{\ell}_a$-risk, for a varying from 0.15 to 0.45. We run the methods on 100 simulated mammography datasets subsampled with 4 remaining radiologists.

5 | CONCLUSION AND PERSPECTIVES

To compensate for the imprecision of certain measurements, it is common practice to collect several measurements on the same individual. These non-independent measurements are called technical replicates, and must then be summarized by a value,

called scoring. Generally, the scoring used to summarize technical replicates is the average of these values. We additionally consider scorings defined by the median of the values, by the maximum a posteriori values, where parameters p, q and θ_T are estimated with an EM algorithm. We also introduce a Bayesian algorithm, to propose a new, so-called Bayesian scoring. Note that methods based on average and median only require information from the individual concerned, whereas Bayesian and Maximum-A-Posteriori approaches use the entire data set. All these scoring methods generally do not provide the same scorings.

We incorporate an indecision response (score = 1/2). This introduces flexibility into classification methods, which is essential in contexts where rapid but potentially erroneous decisions can have serious consequences, as in diagnostic medicine. For example, in the diagnosis of serious pathologies, false negatives (patients misdiagnosed as healthy) present high risks, including delayed treatment, which can significantly affect prognosis. The introduction of an indecision threshold ($v_L \leq Y \leq v_U$) makes it possible to mark certain patients as requiring further investigation. This mechanism avoids the hasty classification of an ambiguous case as healthy, thus reducing the risk of false negatives. In this way, indecision offers a pragmatic approach to managing uncertainty. It's also important to consider the cost of false positives (incorrect diagnoses of cancer, for example, leading to unnecessary stress and invasive examinations). Although false positives are generally less critical than false negatives, a judiciously placed indecision threshold can reduce misdiagnoses by filtering out ambiguous cases.

Scorings can be used to deduce two types of information. The first is to use scoring to infer the status of every patient, either infected or not. We set a threshold v_U (resp. v_L) for scoring above (resp. below) which the patient is considered infected (resp. non-infected). We obtain a diagnosis for patients with a score outside the range $[v_L; v_U]$ and indecision response for the others. Let us remark that when choosing the simple thresholds $v_L = v_U = 0.5$, both average and median-based classifications always provide the same inferred values, and so the same diagnostic. For less trivial thresholds, classifications differ.

Secondly, we can infer the prevalence θ_T, such as the false-positivity and false-negativity rates p, q. In reality, we obtain an estimate of the prevalence of the pathology among the patients on whom we have collected data, which is not necessarily the prevalence of the global population. For Average, Median and Maximum-A-Posteriori approaches, we infer the prevalence by averaging the associated scorings, whereas Bayesian approach uses the posterior chains. Although MAP and Bayesiang performances are very close, the superiority of Bayesian methods lies in the fact that they provide a posterior distribution and not just an estimated value.

We show that $\widehat{\theta}_{T,M}$ provides a better overall estimate of the prevalence of θ_T than $\widehat{\theta}_{T,A}$, in terms of bias and MSE, when the number of individuals N is sufficiently large. Because of the linearity of the bias, the bias will be maximal for extreme values of θ_T. In other words, prevalence estimates based on mean or median scorings may not be accurate if we want to quantify the prevalence of a pathology that we know to be very rare or, on the contrary, very frequent.

It has also been suggested that parameters estimated by MAP or Bayesian can be used to quickly predict a diagnosis on a new patient, without having to rerun calculations on the entire cohort.

In practice, average- and median-based methods are easier to calculate, whereas Bayesian method shows several drawbacks: (1) it is computationally expensive, (2) if the prior is set poorly, the results can be heavily biased, and (3) the scoring $Y_{B,i}$ depends on the whole dataset through the posterior distribution of the fixed parameters θ_T, p, and q. The fact that the score of the i-th individual depends on the whole dataset may seem odd at first sight. However many scoring methods in machine learning, including the scores of the logistic regression, depend on parameters of the model that are fitted on the whole dataset. In practice though, if data collection is done sequentially, the Bayesian methodology can be applied only at the end, which is a major advantage. Moreover, Bayesian method is efficient, especially if the prior distribution is weakly informative, or if the information included in the prior is correct.

In conclusion, calculating the average of the technical replicates seems natural, but it can result in poor scorings, and therefore a poor classification or a significant bias in the estimation of θ_T. If we want to keep the calculation method just as simple, it's better to calculate the median of the technical replicates. But of course, the optimal method in most cases is to use our Bayesian algorithm and to return its scorings as well as the parameter estimates, accompanied by their credibility intervals, provided by this algorithm.

Our methodology is applicable to several medical contexts where diagnoses rely on noisy data, such as genetic testing or microbiological diagnostics. It highlights the importance of incorporating uncertainty into decisions, an aspect often neglected in conventional methods.

AUTHOR CONTRIBUTIONS

All the authors contributed equally.

FINANCIAL DISCLOSURE
None reported.

CONFLICT OF INTEREST
The authors declare no potential conflict of interests.

REFERENCES

1. Beam, C., Conant, E., Sickles, E.: Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. JNCI 95, 282–290 (2003)
2. Fruhwirth-Schnatter, S., Celeux, G., Robert, C.P.: Handbook of mixture analysis. CRC press (2019)
3. Hujoel, P., Moulton, L., Loesche, W.: Estimation of sensitivity and specificity of site-specific diagnostic tests (with erratum). J. Periodont. Res. 25, 193–196 (1990). Erratum in *J. Periodont. Res.* (1990) 25(6):377
4. Im, S.: Performance of the beta-binomial model for clustered binary responses: Comparison with generalized estimating equations. J. Mod. Appl. Stat. Methods 19(1), 1–25 (2021)
5. Jaeger, T.: Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. J. Mem. Lang. 59(4), 434–446 (2008)
6. Kim, J., Lee, J.H.: The Validation of a Beta-Binomial Model for Overdispersed Binomial Data. Commun Stat Simul Comput 46(2), 807–814 (2017)
7. Palmer, R., Strobeck, C.: Fluctuating asymmetry: Measurement, Analysis, Patterns. Annual Review of Ecology and Systematics 17, 391–421 (1986)
8. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2023). URL https://www.R-project.org/
9. Robert, C.P.: The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer (2007)
10. Stan Development Team: RStan: the R interface to Stan (2024). URL https://mc-stan.org/. R package version 2.32.6
11. Williams, D.: Extra-binomial variation in logistic linear models. Appl. Statist. 31(2), 144–148 (1982)

How to cite this article: Lorenzo H., Pudlo P., and Royer-Carenzi M.. Reconciling Binary Replicates: Beyond the Average. *Stat Med.* 2021;00(00):1–18.

APPENDIX

A TECHNICAL RESULTS

A.1 Complete description of the maximum a posteriori EM estimation algorithm

Recall the likelihood according to Equation 5:

$$\mathcal{L}(\theta_T, p, q | s_1, \dots, s_n) = \prod_{i=1}^{n} \left(\theta_T (1-q)^{s_i} q^{n_i - s_i} + (1-\theta_T) p^{s_i} (1-p)^{n_i - s_i} \right).$$

The maximization of this functional might stuck to border points, corresponding to $\hat{p} = 0$ or $\hat{q} = 0$ for example. Those points are not expected and in order to avoid them we have chosen to use priors on both p and q such that

$$\pi(p, q) \propto p(1-p)q(1-q),$$

corresponding to $p, q \sim \mathscr{B}\text{eta}(2, 2)$, corresponding to weakly informative priors. The posterior distribution of the parameters is thus proportional to

$$\pi(\theta_T, p, q, s_1, \dots, s_n) = p(1-p)q(1-q) \prod_{i=1}^{n} \left(\theta_T (1-q)^{s_i} q^{n_i - s_i} + (1-\theta_T) p^{s_i} (1-p)^{n_i - s_i} \right). \tag{A1}$$

The following details the EM algorithm in order to maximize the likelihood of the problem. Rather than maximizing the likelihood, the objective is to maximize the posterior distribution of the parameter, but the EM algorithm can be used the same way.

The EM algorithm is an iterative procedure that alternates between the Estimation (E) and Maximization (M) steps. The E-step computes the expected value of the latent variables T_i's given the observed data s_i's and the current values of the fixed parameters. The M-step maximizes the completed posterior distribution of the fixed parameters given the expected values of the latent variables. The completed log-posterior distribution of the fixed parameters θ_T, p, and q given the observed data $(s_1, \ldots, s_n)$ and the latent $(y_1, \ldots, y_n)$ can be extracted from (A5) and is, up to an additive constant,

$$\begin{aligned}\pi_c(\theta_T, p, q, s_1, \ldots, s_n, y_1, \ldots, y_n) = \sum_{i=1}^{n} \Big\{ & y_i \Big(s_i \log(1-q) + (n_i - s_i)\log(q) + \theta_T \Big) \\ & + (1-y_i)\Big(s_i \log(p) + (n_i - s_i)\log(1-p) + (1-\theta_T) \Big) \Big\}, \\ & + \log(p(1-p)) + \log(q(1-q)).\end{aligned}$$

We can easily optimize the above completed log-posterior distribution in θ_T, p and q given the values of $(s_1, \ldots, s_N)$ and $(y_1, \ldots, y_N)$. This means that the M-steps of the EM algorithm are explicit:

$$\theta_{\text{T,M-step}} = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad p_{\text{M-step}} = \frac{1 + \sum_{i=1}^{N} s_i(1-y_i)}{2 + \sum_{i=1}^{N} n_i(1-y_i)}, \quad q_{\text{M-step}} = \frac{1 + \sum_{i=1}^{N}(n_i - s_i)y_i}{2 + \sum_{i=1}^{N} n_i y_i}. \tag{A2}$$

The value of y_i that is plugged in the M-step is the expected value of the latent variable T_i given the observed data s_i and the current values of the fixed parameters. (Note that those values of y_i are now real numbers between 0 and 1 exactly as our different scores.) This is the result of the E-step in the EM algorithm. Here, it is also explicit: since T_i is a binary variable, this conditional expected value is equal to the probability of $T_i = 1$ given the observed data s_i and the current values of the fixed parameters. We can recognize the definition of the likelihood-based score $Y_{L,i}(\theta_T, p, q)$ given in (4). Thus, the E-step of the EM algorithm is

$$y_{i\,\text{E-step}} = Y_{L,i}(\theta_T, p, q), \tag{A3}$$

where we use the last values of the fixed parameters θ_T, p, and q available during the iterative procedure.

As with many numerical optimization algorithms, the EM algorithm is sensitive to the initial values. We thus have to run the EM algorithm several times. Each time, we initialize the latent y_i's by drawing them at random from

$$y_i \sim \mathscr{B}\text{eta}(s_i + 1/2, n_i - s_i + 1/2). \tag{A4}$$

We then run the EM algorithm starting with an M-step until convergence. We compare the results obtained from the several runs by using the posterior distribution given in (A1) and the best model is chosen.

The posterior distribution of our mixture model in (A1) suffers from the label-switching problem, see Chapter 1 of [2]. Indeed, there is a symmetry in the posterior : we can exchange the 0 and 1 labels in the latent space and get the same posterior value with $\theta_T^* = 1 - \theta_T$, $p^* = 1 - q$ and $q^* = 1 - p$ as the new fixed parameters. However, we have assumed that the replicates are noisy measurements of the true value T_i and, for this reason, we have set the constraint that the false-positivity rate p and false-negativity rate q are both smaller than 1/2. This constraint gives us a relabelling of the latent y_i's that removes entirely the label-switching problem when applying the EM algorithm to our mixture model: each time the constraint on p is violated, we use the symmetry described above, switch the 0 and 1 and change the current values of the fixed parameters to $1-\theta_T$, $1-q$ and $1-p$ and the current values of all y_i's to $1-y_i$. Algorithm 1 synthesizes the complete estimation procedure.

A.2 Bayesian posterior calculation

Using the observed value $(s_1, \ldots, s_N)$ of the sufficient statistic $(S_1, \ldots, S_N)$, the joint density of the Bayesian model is thus, see Figure 1,

$$\begin{aligned}\pi(\theta_T, p, q, y_1, \ldots, t_N, t_1, \ldots, s_N) &= \pi(\theta_T, p, q) \prod_{i=1}^{N} \mathbb{P}_\pi\big(T_i = t_i \big| \theta_T\big)\, \mathbb{P}_\pi\big(S_i = s_i \big| T_i = t_i, p, q\big) \\ &= \pi(\theta_T, p, q) \prod_{i=1}^{N} \theta_T^{t_i}(1-\theta_T)^{1-t_i} \binom{n_i}{s_i} \Big\{ t_i (1-q)^{s_i} q^{n_i - s_i} + (1-t_i) p^{s_i} (1-p)^{n_i - s_i} \Big\}. \end{aligned} \tag{A5}$$

Algorithm 1 Maximum-A-Posteriori estimator approximation of the fixed parameters

for $r \in [\![1, R]\!]$ **do**
  $\forall i \in [\![1, N]\!]$, initialize $y_{\mathtt{r},\mathtt{i}}$: $y_{\mathtt{r},\mathtt{i}} \sim \mathscr{B}\mathtt{eta}(s_i + 1/2, n_i - s_i + 1/2)$
  while Convergence not reached **do**
    Perform M-step:

$$\theta_{\mathtt{T},\mathtt{EM},\mathtt{r}} \leftarrow \frac{1}{N}\sum_{i=1}^{N} y_{\mathtt{r},\mathtt{i}}, \quad p_{\mathtt{EM},\mathtt{r}} \leftarrow \frac{1+\sum_{i=1}^{N} s_i(1-y_{\mathtt{r},\mathtt{i}})}{2+\sum_{i=1}^{N} n_i(1-y_{\mathtt{r},\mathtt{i}})}, \quad q_{\mathtt{EM},\mathtt{r}} \leftarrow \frac{1+\sum_{i=1}^{N}(n_i-s_i)y_{\mathtt{r},\mathtt{i}}}{2+\sum_{i=1}^{N} n_i y_{\mathtt{r},\mathtt{i}}}$$

    if $q > 1/2$ **then**
      $(\theta_{T,\mathtt{EM},\mathtt{r}}, p_{\mathtt{EM},\mathtt{r}}, q_{\mathtt{EM},\mathtt{r}}) \leftarrow (1-\theta_{T,\mathtt{EM},\mathtt{r}}, 1-p_{\mathtt{EM},\mathtt{r}}, 1-q_{\mathtt{EM},\mathtt{r}})$
    end if
    $\forall i \in [\![1, n]\!]$, perform E-step: $y_{\mathtt{r},\mathtt{i}} \leftarrow Y_{L,i}(\theta_{T,\mathtt{EM},r}, p_{\mathtt{EM},r}, q_{\mathtt{EM},r})$
  end while
end for
The best model $(\widehat{\theta}_{T,\mathtt{MAP}}, \widehat{p}_{\mathtt{MAP}}, \widehat{q}_{\mathtt{MAP}})$ verifies

$$(\widehat{\theta}_{T,\mathtt{MAP}}, \widehat{p}_{\mathtt{MAP}}, \widehat{q}_{\mathtt{MAP}}) = \underset{r\in[\![1,R]\!]}{\arg\max}\, \pi(\theta_{T,\mathtt{EM},\mathtt{r}}, p_{\mathtt{EM},\mathtt{r}}, q_{\mathtt{EM},\mathtt{r}}, s_1, \ldots, s_n)$$

If we integrate over the latent T_i's, we have the joint density of the observed data and the fixed parameters as

$$\pi(\theta_T, p, q, s_1, \ldots, s_N) = \pi(\theta_T, p, q) \prod_{i=1}^{N} \binom{n_i}{s_i} \left\{ \theta_T (1-q)^{s_i} q^{n_i - s_i} + (1-\theta_T) p^{s_i} (1-p)^{n_i - s_i} \right\}. \tag{A6}$$

Despite this explicit expression of the joint density, the posterior distribution of the fixed parameters θ_T, p, and q given the data $(s_1, \ldots, s_N)$ is difficult to compute explicitly. We use the R-package `rstan` [10], which implements an Hamiltonian Monte Carlo algorithm to sample from the posterior distribution of the fixed parameters θ_T, p, and q given the data. Latent values of the T_i's are drawn at each iteration of the MCMC algorithm from the predictive distribution of the T_i's given the data X_{ij}'s and the current values of the fixed parameters θ_T, p, and q. The predictive distribution of the T_i's given the data S_i's and the current values of the parameters θ_T, p, and q is a product of Bernoulli distributions given by (3).

A.3 On the likelihood-based scores

Lemma 2. *Assume $\theta_T \in (0,1)$, $p \in (0, 1/2)$ and $q \in (0, 1/2)$. The likelihood-based score, defined as in* (4)*, is an increasing function of s.*

Proof. It is enough to prove that the following function is increasing in s:

$$Y_L(s) = \theta_T \left(\theta_T + (1-\theta_T) \left(\frac{1-p}{q} \right)^{n-s} \left(\frac{p}{1-q} \right)^{s} \right)^{-1}.$$

Actually $s \mapsto \left(\frac{1-p}{q} \right)^{n-s} \left(\frac{p}{1-q} \right)^{s}$ is a decreasing function of s because

$$\left(\frac{q}{1-p} \right)^{n-s+1} \left(\frac{1-q}{p} \right)^{s-1} < \left(\frac{q}{1-p} \right)^{n-s} \left(\frac{1-q}{p} \right)^{s} \iff \frac{q}{1-p} < \frac{1-q}{p}$$
$$\iff 0 < 1 - (p+q).$$

And the last inequality is true because $p < 1/2$ and $q < 1/2$. □

A.4 Proof of Lemma 1

Since $r(\widehat{t}) = \mathbb{E}\big[\ell(T,\widehat{t})\big]$, we have

$$\begin{aligned} r(0) &= \ell(0,0) + \vartheta(\ell(1,0) - \ell(0,0)) = c\vartheta, \\ r(1) &= \ell(0,1) + \vartheta(\ell(1,1) - \ell(0,1)) = b - b\vartheta, \\ r(1/2) &= \ell(0,1/2) + \vartheta(\ell(1,1/2) - \ell(0,1/2)) = a + (d-a)\vartheta. \end{aligned}$$

The three risks are thus affine functions of ϑ, see Figure A1.

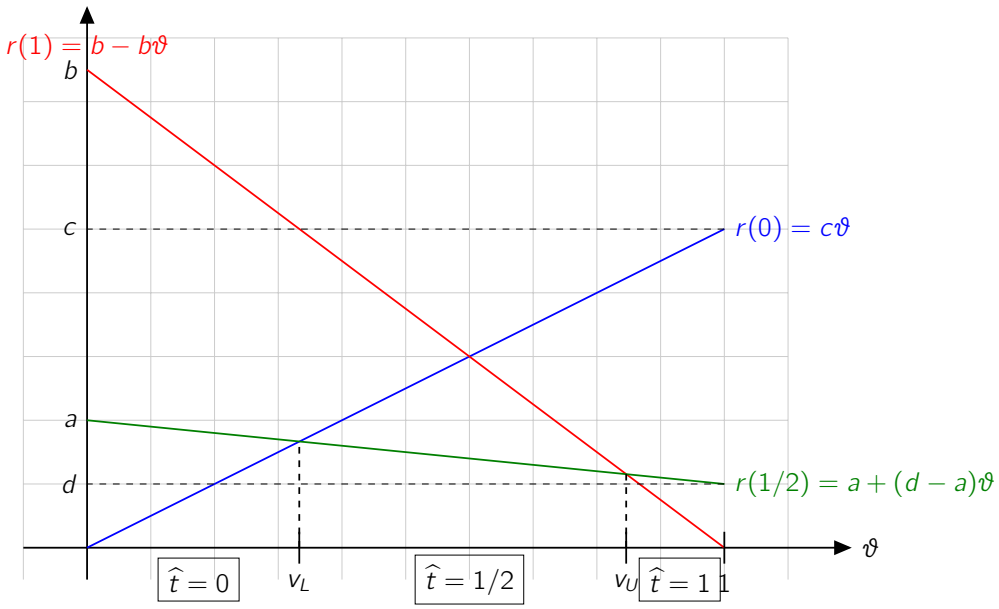


FIGURE A1 Risks $r(\widehat{t})$ of deciding $\widehat{t} = 0, 1/2$ or 1 as a function of ϑ. When $\vartheta \in (0, v_U)$, the lowest risk is $r(0)$, thus the best decision is $\widehat{t} = 0$. When $\vartheta \in (v_L, v_U)$, the lowest risk is $r(1/2)$, thus the best decision is $\widehat{t} = 1/2$. When $\vartheta \in (v_U, 1)$, the lowest risk is $r(1)$, thus the best decision is $\widehat{t} = 1$.

The best decision is of the desired form if and only if, as functions of ϑ, the three risks $r(0)$, $r(1/2)$ and $r(1)$ intersect each other as in Figure A1. This happens if and only if the two following conditions are satisfied:

(*i*) The point where the risks $r(0)$ and $r(1)$ intersect is above the line of $r(1/2)$.
(*ii*) The slope of $r(1/2)$ is between those of $r(0)$ and $r(1)$, that is to say in $(-b, c)$.

Condition (*ii*) is clearly equivalent to $-b < (d-a) < c$. Moreover, the point at which $r(0)$ and $r(1)$ intersect has coordinates

$$\left(\frac{b}{b+c}, \frac{bc}{b+c}\right).$$

And at $\vartheta = b/(b+c)$, the risk $r(1/2)$ is equal to $a + (d-a)b/(b+c)$. Hence condition (*i*) is equivalent to the first inequality in the Proposition. Finally, the values of v_L and v_U are obtained by solving the equations $r(0) = r(1/2)$ and $r(1) = r(1/2)$, respectively. □

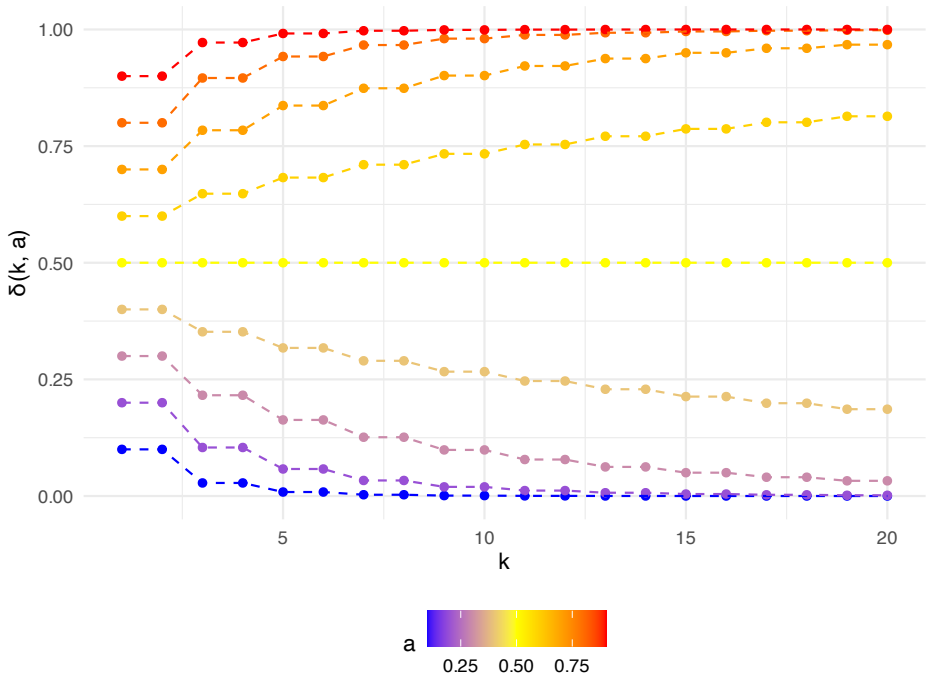


FIGURE A2 The values of $\delta(k,a)$ for $a = 0.1, 0.2, \ldots, 0.9$, as defined in Equation (A7).

A.5 Properties of the Binomial distribution

To understand the probabilistic properties of the median-based scoring, we need to introduce some coefficients related to the Binomial distribution. For any integer k and any real $a \in (0,1)$, we set

$$u(k,a) = \mathbb{P}\Big(\mathscr{B}\text{in}(k,a) > k/2\Big), \quad v(k,a) = \mathbb{P}\Big(\mathscr{B}\text{in}(k,a) = k/2\Big) \text{ and}$$
$$\delta(k,a) = u(k,a) + v(k,a)/2. \tag{A7}$$

Note that, If k is an odd number, then $v(k,a) = 0$. Values of $\delta(k,a)$ are given in Figure A2.

Proposition 1. *The δ-coefficients have the following properties.*

(i) $\delta(1,a) = a$.
(ii) $1 - \delta(k,a) = \delta(k, 1-a)$.
(iii) If k is an even number, then $\delta(k,a) = \delta(k-1,a)$.
(iv) If $a < 1/2$, then $\delta(k,a)$ is a non-increasing function of k.
(v) If $a < 1/2$, then $\delta(k,a)$ tends to 0 when $k \to \infty$.

Proof. We introduce B_i, $i = 1,2,\ldots$, an iid sequence of random variables distributed according to $\mathscr{B}\text{er}(a)$, and set $Z_k = B_1 + \cdots + B_k \sim \mathscr{B}\text{in}(k,a)$ for all $k \geq 1$.

A.5.0.1 *Proof of* (i)

We have $\delta(1,a) = \mathbb{P}(Z_1 > 1/2) + \mathbb{P}(Z_1 = 1/2)/2 = \mathbb{P}(Z_1 = 1) = a$.

A.5.0.2 *Proof of* (ii)

Consider first the case where k is odd. We have $1 - u(k,a) = 1 - \mathbb{P}(Z_k > k/2) = \mathbb{P}(Z_k < k/2)$ since $k/2$ is not an integer. And $\mathbb{P}(Z_k < k/2) = \mathbb{P}(k - Z_k > k/2)$. Now $k - Z_k = (1-B_1) + \cdots + (1-B_k) \sim \mathscr{B}\text{in}(k, 1-a)$. Thus, $1 - u(k,a) = u(k, 1-a)$. Moreover, when k is odd, $v(k,a) = 0$ and $\delta(a,k) = u(a,k)$. Hence, (ii) is proven when k is odd.

Consider now the case where k is even. We have $1 - u(k,a) = 1 - \mathbb{P}(Z_k > k/2) = \mathbb{P}(Z_k \leq k/2) = \mathbb{P}(Z_k < k/2) + \mathbb{P}(Z_k = k/2)$. Like above, $\mathbb{P}(Z_k < k/2) = u(k, 1-a)$. And $v(k,a) = \mathbb{P}(Z_k = k/2) = \mathbb{P}(k - Z_k = k/2) = v(k, 1-a)$. So $1 - u(k,a) = \big[u(k,1-a) + v(k,1-a)\big]$.

Thus,

$$1-\delta(k,a) = 1-u(k,a)-\frac{1}{2}v(k,a) = \left[u(k,1-a)+v(k,1-a)\right]-\frac{1}{2}v(k,1-a) = \delta(k,1-a).$$

A.5.0.3 *Proof of* (iii)
Set $k = 2m$, where m is an integer. We have $\delta(2m-1,a) = u(2m-1,a) = \mathbb{P}(Z_{2m-1} > m-1/2) = \mathbb{P}(Z_{2m-1} \geq m)$. And $Z_{2m-1} = Z_{2m} - B_{2m}$. Since B_{2m} is either 0 or 1, we have

$$\left\{Z_{2m-1} \geq m\right\} = \left\{Z_{2m} \geq m+1\right\} \cup \left\{Z_{2m} = m, B_{2m} = 0\right\}$$

and the two events are disjoint. Thus $u(2m-1,a) = u(2m,a) + \mathbb{P}(Z_{2m} = m, B_{2m} = 0)$. To compute the last probability, recall that $Z_{2m} = B_1 + \cdots + B_{2m}$, and so, $Z_{2m} = m$ means that exactly half of the B_i's are equal to 0. Thus, given $Z_{2m} = m$, the probability that $B_{2m} = 0$ is $m/(2m) = 1/2$. Hence $\mathbb{P}(Z_{2m} = m, B_{2m} = 0) = \mathbb{P}(Z_{2m} = m)/2 = v(2m,a)/2$ and finally, $u(2m-1,a) = u(2m,a) + v(2m,a)/2$.

A.5.0.4 *Proof of* (iv)
Set $a < 1/2$. It is enough to prove that, for all m, $\delta(2m-1,a) = \delta(2m,a) > \delta(2m+1,a)$. The equality is the result of (iii). Thus, it is enough to prove that $\delta(2m,a) > \delta(2m+1,a)$ for all m.

Fix any m. We have $\delta(2m+1,a) = \mathbb{P}(Z_{2m+1} \geq m+1)$. And $Z_{2m+1} = Z_{2m} + B_{2m+1}$ with independence between these last two variables. Thus,

$$\delta(2m+1,a) = (1-a)\mathbb{P}(Z_{2m} \geq m+1) + a\mathbb{P}(Z_{2m} \geq m).$$

Using $\mathbb{P}(Z_{2m} \geq m) = \mathbb{P}(Z_{2m} \geq m+1) + \mathbb{P}(Z_{2m} = m)$ in the above, we get

$$\delta(2m+1,a) = \mathbb{P}(Z_{2m} \geq m+1) + a\mathbb{P}(Z_{2m} = m) = u(2m,a) + a\,v(2m,a).$$

Since $a < 1/2$, we have $a\,v(2m,a) < v(2m,a)/2$ and thus $\delta(2m+1,a) < \delta(2m,a)$.

A.5.0.5 *Proof of* (v)
Since $a < 1/2$, Hoefding's inequality gives

$$\mathbb{P}(Z_k \geq k/2) \leq \exp\left(-2k(1/2-a)^2\right).$$

Thus, $\delta(k,a) = \mathbb{P}(Z_k > k/2) + \mathbb{P}(Z_k = k/2)/2 \leq \mathbb{P}(Z_k \geq k/2)$ tends to 0 when $k \to \infty$. □

B THEORETICAL RESULTS ON THE CLASSIFIERS

B.1 Proof of Theorem 1

The sensitivity of a classifier $\widehat{T}$ is $\mathbb{P}(\widehat{T} = 1|T = 1)$. The average-based classifier is one if and only if $S_i > n_i v_U$. And the median-based classifier is one if and only if $S_i > n_i/2$. Because of (H3), $\{S_i > n_i v_U\} \subset \{S_i > n_i/2\}$, and thus the median-based classifier is more sensitive than the average-based classifier.

There is equality if and only if both events are equal. This happens if and only if $v_U = 1/2$. In the particular case where all n_i are equal, this happens if and only if $v_U < 1/2 + \delta_0$, where δ_0 is defined in Section 3.3. The same kind of reasoning yields to the inequality on the specificities of the classifiers. □

B.2 Proof of Theorem 2

B.2.0.1 *Proof of* (*i*)
The likelihood-based scoring is $\mathbb{P}(T_i = 1)$. Because of (H4) and Lemma 1, the thresholding of this scoring with v_L and v_U gives the best estimator in terms of ℓ-risk. Hence the desired inequality. □

B.2.0.2 *Proof of* (*ii*)
We can also apply Lemma 1, using the expected value $\mathbb{E}_\pi$ given the data $S = (S_1, \ldots, S_n)$, with $\vartheta = \mathbb{P}_\pi(T_i = 1|S)$ which is the definition of $Y_{B,i}$. Because of (H4), Lemma 1 gives that $\widehat{T}_{B,i}$ is the best estimator in terms of ℓ-posterior risk, i.e.,

$$\mathbb{E}_\pi\left(\ell(T_i, \widehat{T}_{B,i})\middle|S\right) \leq \mathbb{E}_\pi\left(\ell(T_i, \widehat{T}_i)\middle|S\right).$$

Integrating over the marginal distribution of S in the Bayesian model yields the desired inequality. □

B.2.0.3 *Proof of* (*iii*)
If (*iii*) is wrong, then the inequality has a prior probability of 1. I.e., with prior probability equal to 1,

$$\mathbb{E}_\pi\Big[\ell\big(T_i,\widehat{T}_i\big)\Big|\theta_T,p,q\Big] < \mathbb{E}_\pi\Big[\ell\big(T_i,\widehat{T}_{B,i}\big)\Big|\theta_T,p,q\Big].$$

If we integrate this over the prior distribution, we get a classifier $\widehat{T}_i$ strictly better than the Bayesian classifier $\widehat{T}_{B,i}$, which is impossible because of (*ii*). Hence (*iii*) is true. □

B.3 Average and median-based empirical risks

Let us denote by H the set of the healthy individuals, and by I the set of the infected ones. Let us consider a decision cost $a \in]0, \frac{1}{2}]$. From, the definition of the loss function given in subsection 3.1, we have

$$\ell_a^M = a\sum_{i=1}^N \mathbb{1}_{\frac{s_i}{n_i}=\frac{1}{2}} + \sum_{i=1}^N \mathbb{1}_{\frac{s_i}{n_i}>\frac{1}{2}}\mathbb{1}_{i\in H} + \sum_{i=1}^N \mathbb{1}_{\frac{s_i}{n_i}<\frac{1}{2}}\mathbb{1}_{i\in I}.$$

Then ℓ_a^M is a linear and increasing function of the decision cost a. In a similar way,

$$\ell_a^A = a\sum_{i=1}^N \mathbb{1}_{\frac{s_i}{n_i}\in[a,1-a]} + \sum_{i=1}^N \mathbb{1}_{\frac{s_i}{n_i}\in]1-a,1]}\mathbb{1}_{i\in H} + \sum_{i=1}^N \mathbb{1}_{\frac{s_i}{n_i}\in[0,a[}\mathbb{1}_{i\in I}.$$

It is obvious that as long as a (or $1-a$) does not cross an observable value of *Si*/*ni*, the terms involved in the sums are constant, and so the average-risk ℓ_a^A is linear and increasing with a. Then jumps occur only when a crosses an observed value of $\frac{s_i}{n_i}$. In this case, the average-risk increases from $1-a$ times the number of infected patients with $\frac{s_i}{n_i}=a$, but decreases from a times the number of healthy patients with $\frac{s_i}{n_i}=a$. At the same time, the average-risk increases from $1-a$ times the number of healthy patients with $\frac{s_i}{n_i}=1-a$, but decreases from a times the number of infected patients with $\frac{s_i}{n_i}=1-a$. Note that infected patients are mainly expected to have $\frac{s_i}{n_i}>\frac{1}{2}$ whereas healthy patients are mainly expected to have $\frac{s_i}{n_i}<\frac{1}{2}$.

C THEORETICAL RESULTS ON THE SCORES

C.1 Bias of the scores

For an individual i, the two frequentist scores $Y_{A,i}$ and $Y_{M,i}$ are biased when we consider recovering the latent T_i variable or the prevalence θ_T. Their expected values and conditional expected values given T_i are explicit, see Proposition 2 below.

Proposition 2 (Bias)**.** *Fix an individual i for which we observe n_i replicates. The average-based scoring $Y_{A,i}$ is such that*

$$\mathbb{E}\big(Y_{A,i}\big|T_i\big) = T_i(1-q) + (1-T_i)p,$$
$$\mathbb{E}\big(Y_{A,i}\big) = \theta_T(1-q) + (1-\theta_T)p.$$

The median-based scoring $Y_{M,i}$ is such that

$$\mathbb{E}\big(Y_{M,i}\big|T_i\big) = T_i\delta(n_i,1-q) + (1-T_i)\delta(n_i,p),$$
$$\mathbb{E}\big(Y_{M,i}\big) = \theta_T\delta(n_i,1-q) + (1-\theta_T)\delta(n_i,p),$$

with the coefficients defined in Equation (A7)*.*

Proof. For all $K \in \{A,M\}$, $\mathbb{E}(T_{K,i}|T_i)$ is a linear function of T_i. Thus, computing $\mathbb{E}(T_{K,i})$ from $\mathbb{E}(T_{K,i}|T_i)$ is straightforward: we replace T_i in the expression by its expected value θ_T.

The two scores are explicit functions of S_i, given by $Y_{A,i} = S_i/n_i$ and Equation (2) above. Integrating those formulae over the Binomial distribution of $[S_i|T_i]$ in (1) yields the desired results. □

C.2 Variance of the scores

The random variability of each score around its expected value can be measured by the variance. We can compute them explicitly for the average- and median-based scores as follows. To state the results, we need to introduce another coefficient as

in Equation (A7):

$$\gamma(k,a) = u(k,a)(1-u(k,a)) + \frac{1}{4}v(k,a)(1-v(k,a)) - u(k,a)v(k,a). \tag{C8}$$

Proposition 3 (Variance)**.** *Fix an individual i for which we observe n_i replicates. The average-based scoring $Y_{A,i}$ is such that*

$$\begin{aligned}\mathrm{var}\big(Y_{A,i}\big|T_i\big) &= T_i\frac{q(1-q)}{n_i} + (1-T_i)\frac{p(1-p)}{n_i},\\ \mathrm{var}\big(Y_{A,i}\big) &= \theta_T(1-\theta_T)(1-q-p)^2 + \theta_T\frac{q(1-q)}{n_i} + (1-\theta_T)\frac{p(1-p)}{n_i}.\end{aligned}$$

The median-based scoring $Y_{M,i}$ is such that

$$\begin{aligned}\mathrm{var}\big(Y_{M,i}\big|T_i\big) &= T_i\gamma(n_i,1-q) + (1-T_i)\gamma(n_i,p),\\ \mathrm{var}\big(Y_{M,i}\big) &= \theta_T(1-\theta_T)\big(1-\delta(n_i,q)-\delta(n_i,p)\big)^2 + \theta_T\gamma(n_i,1-q) + (1-\theta_T)\gamma(n_i,p).\end{aligned}$$

Proof. For the average-based scoring $Y_{A,i}$, we used $Y_{A,i} = S_i/n_i$ and the Binomial distribution of $[S_i|T_i]$ in (1) to compute the conditional variance given T_i. The unconditional variance is obtained by using

$$\mathrm{var}(Y_{A,i}) = \mathbb{E}(\mathrm{var}(Y_{A,i}|T_i)) + \mathrm{var}(\mathbb{E}(Y_{A,i}|T_i)).$$

Likewise for the variance of the median-based scoring $Y_{M,i}$, using Equation (2). □

D THEORETICAL RESULTS ON THE PREVALENCE ESTIMATES

D.1 Prevalence estimates bias

There is a systematic bias in the average-based and median-based prevalence estimates, which are empirical means of their respective scores. Indeed, using Proposition 2, we can compute their bias. We need the following coefficients to state the results: for all $a \in (0,1)$, we set

$$\overline{\delta}(a) = \frac{1}{n}\sum_{i=1}^{n}\delta(n_i,a),$$

with the coefficients introduced in Equation (A7).

Proposition 4. *The average-based prevalence estimate $\widehat{\theta}_{T,A}$ is such that*

$$\mathbb{E}(\widehat{\theta}_{T,A}) = \theta_T(1-q) + (1-\theta_T)p.$$

The median-based prevalence estimate $\widehat{\theta}_{T,M}$ is such that

$$\mathbb{E}(\widehat{\theta}_{T,M}) = \theta_T(1-\overline{\delta}(q)) + (1-\theta_T)\overline{\delta}(p),$$

Moreover, if (H2) is true, then

$$\overline{\delta}(p) < p \quad \textit{and} \quad \overline{\delta}(q) < q.$$

D.2 Proof of Theorem 3

Fix p and q in $(0, 1/2)$ and assume (H2). The bias of the prevalence estimates are the linear functions of θ_T because of Proposition 4:

$$\begin{aligned}\mathrm{bias}(\widehat{\theta}_{T,A}) &= \mathbb{E}(\widehat{\theta}_{T,A}) - \theta_T = p - \theta_T(p+q),\\ \mathrm{bias}(\widehat{\theta}_{T,M}) &= \mathbb{E}(\widehat{\theta}_{T,M}) - \theta_T = \overline{\delta}(p) - \theta_T(\overline{\delta}(p) + \overline{\delta}(q)).\end{aligned}$$

The first one is null at $\theta_T = p/(p+q)$, and the second one is null at $\theta_T = \overline{\delta}(p)/(\overline{\delta}(p) + \overline{\delta}(q))$. The absolute values of these biases are thus as in Figure D3. And, because of (H2), $\overline{\delta}(p) < p$ and $\overline{\delta}(q) < q$. Thus,

$$\Big|\mathrm{bias}(\widehat{\theta}_{T,M})\Big| < \Big|\mathrm{bias}(\widehat{\theta}_{T,A})\Big|$$

expect when θ_T is in the interval, we denote J, where the blue curve is below the red one in Figure D3. And, since the blue curve is null at $p/(p+q)$, this value is in J.

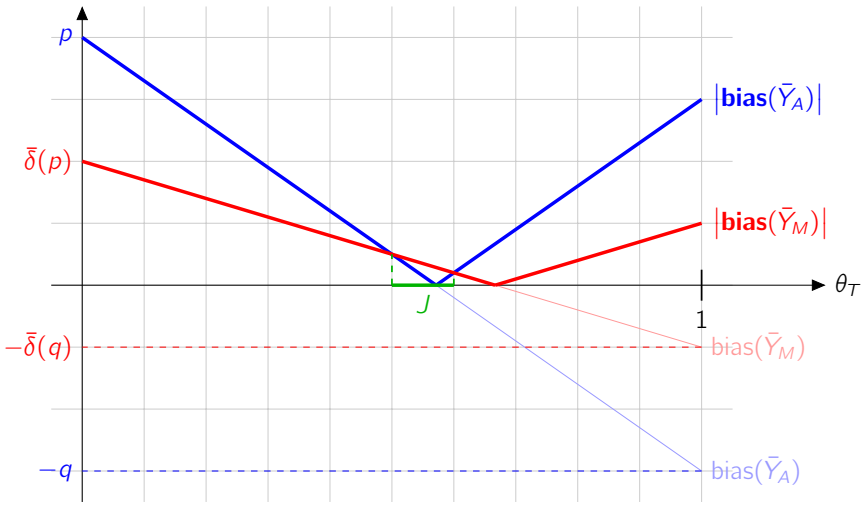


FIGURE D3 Typical behavior of the Biases of the average- and median-based prevalence estimates $\widehat{\theta}_{T,A}$ and $\widehat{\theta}_{T,M}$ as a function of θ_T, and their absolute values. In absolute values, the bias of the average-based estimate in thick blue is smaller than the bias of the median-based one if and only if $\theta_T \in J$.

Because of Proposition 1, $\delta(k,p)$ is a non-increasing function of k, and thus $\overline{\delta}(p) \leq \delta(n_0,p)$. And likewise, $\overline{\delta}(q) \leq \delta(n_0,q)$. Moreover, the two bounds $\delta(n_0,p)$ and $\delta(n_0,q)$ tends to 0 when $n_0 \to \infty$. That is to say that the red curve in Figure D3 tends to zero when $n_0 \to \infty$, whereas the blue curve does not change. Hence, the length of J tends to 0 when $n_0 \to \infty$. □

D.3 Variance of the prevalence estimates

The variance of the average- or median-based estimate is the variance of the empirical means of the scores. As for the biases, we can compute them explicitly. We need the following coefficients to state the results: for all $a \in (0,1)$, we set

$$\overline{\gamma}(a) = \frac{1}{N}\sum_{i=1}^{N} \gamma(n_i, a) \quad \text{and} \quad \tilde{n} = \left(\frac{1}{N}\sum_{i=1}^{N} \frac{1}{n_i}\right)^{-1}.$$

Note that $\tilde{n}$ is the harmonic mean of the n_i's.

Proposition 5. *The average-based prevalence estimate $\widehat{\theta}_{T,A}$ is such that*

$$\mathrm{var}(\widehat{\theta}_{T,A}) = \frac{1}{N}\left\{\theta_T(1-\theta_T)(1-q+p) + \theta_T\frac{q(1-q)}{\tilde{n}} + (1-\theta_T)\frac{p(1-p)}{\tilde{n}}\right\}.$$

The median-based prevalence estimate $\widehat{\theta}_{T,M}$ is such that

$$\mathrm{var}(\widehat{\theta}_{T,M}) = \frac{1}{N}\left\{\theta_T(1-\theta_T)(1-\overline{\delta}(q)+\overline{\delta}(p)) + \theta_T\overline{\gamma}(1-q) + (1-\theta_T)\overline{\gamma}(p)\right\}.$$

Proof. Both prevalence estimates are empirical means of their respective scores, which are independent. Thus, the proposed formulae are straightforward consequences of Proposition 3 that gives the variances of the scores. □

D.4 Proof of Theorem 4

The proof is similar to the proof of Theorem 2 given in Section B.2. We have to replace the loss function ℓ by the squared error loss function $L^2(\theta_T, \widehat{\theta}) = (\theta_T - \widehat{\theta})^2$. And, to replace Lemma 1, we use the fact that the best estimator in terms of the posterior L^2-risk is the posterior mean. Likewise, it achieves the minimum Bayesian L^2-risk and is admissible. □

E REAL CASE DATA INFORMATION

E.1 Details of the mammogram dataset

Three radiologists did not consider all the exams: radiologist code `1201` to whom 63 positive cases are presented, corresponding to 1 missing positive case exam. Radiologist `7714` to whom 61 positive cases and 79 negative cases are presented, corresponding to 3 missing positive case exams and 5 missing negative case exams. Radiologist `9007` to whom 83 negative cases are presented, corresponding to 1 missing negative case exam. Thus in total, there are 4 missing positive case exams and 6 missing negative case exams among the n_{rad} radiologists.

For any radiologist j, it is possible to estimate a radiologist false-positivity rate p_j, respectively a radiologist false-negativity q_j, such as

$$\widehat{p}_j = \frac{\sum_{i=1}^{N} \mathbb{1}_{T_i=0}\mathbb{1}_{X_{ij}=1}\mathbb{1}_{m_{ij}=1}}{\sum_{i=1}^{N} \mathbb{1}_{T_i=0}\mathbb{1}_{m_{ij}=1}} \quad \text{and} \quad \widehat{q}_j = \frac{\sum_{i=1}^{N} \mathbb{1}_{T_i=1}\mathbb{1}_{X_{ij}=0}\mathbb{1}_{m_{ij}=1}}{\sum_{i=1}^{N} \mathbb{1}_{T_i=1}\mathbb{1}_{m_{ij}=1}}, \tag{E9}$$

where m_{ij} is equal to 1 if radiologist j observes mammography i and 0 otherwise.

E.2 A simulation procedure for synthetic mammography datasets

This paragraph details a simulation procedure in order to generate a dataset $\boldsymbol{X} = \left\{ \left\{x_{ij}\right\}_j, t_i, n_i \right\}_i$. x_{ij}s are filled with 0 and 1 corresponding to realizations of the variables X_{ij} for each of the n patients and each of the n_{rad} radiologists. This dataset must verify different points:

- t_i is set equal to 0 for $i \in [\![1, 84]\!]$ and to 1 for $i \in [\![85, 148]\!]$
- according to the equations (E9) and for any radiologist j, the $\left(x_{ij}\right)_i$s are such as
  - there are exactly $84\widehat{p}_j$ 1s and $84(1-\widehat{p}_j)$ 0s in the 84 first coordinates of $\left(x_{ij}\right)_i$
  - there are exactly $64\widehat{q}_j$ 0s and $64(1-\widehat{q}_j)$ 1s in the 64 last coordinates of $\left(x_{ij}\right)_i$
- n_i is set equal to n_{rad} except for radiologists of ids 1201, 7714 and 9007, as discussed in the following point.
- concerning the missing values discussed earlier:
  - there is 1 missing positive case for radiologist of id 1201 and n_i is set to 109
  - there are 3 missing positive cases and 5 missing negative cases for radiologist of id 7714 and n_i is set to 102
  - there is 1 missing negative case for radiologist of id 9007 and n_i is set to 109.
- among the patients, some must be hard to be properly diagnosed.

For the last point we have chosen the following procedure for any radiologist j: treat $x_j^{(0)} := \left\{x_{ij}\right\}_{i\in[\![1,84]\!]}$ and $x_j^{(1)} := \left\{x_{ij}\right\}_{i=[\![85,148]\!]}$ in parallel. In $x_j^{(0)}$, there are exactly $84\widehat{p}_j$ values equal to 1 and $84(1-\widehat{p}_j)$ values equal to 0. In $x_j^{(1)}$, there are exactly $64\widehat{q}_j$ values equal to 0 and $64(1-\widehat{q}_j)$ values equal to 1. In both cases we throw the desired positions of $84\widehat{p}_j$ values among the vector $[\![1, 84]\!]$, respectively the desired positions of $64\widehat{q}_j$ values among the vector $[\![85, 148]\!]$. Each patient is weighted, based on weights w_i, corresponding to the difficulty of correctly diagnosing patient i (say $x_{ij} = 1$ *while* $t_i = 0$ and $x_{ij} = 0$ *while* $t_i = 1$) meaning that if $w_k > w_i$ then patient i is harder to diagnose correctly than patient k. Note, there are many possible datasets:

$$\prod_{j=1}^{n_{rad}} C_{64}^{64\widehat{q}_j} C_{84}^{84\widehat{p}_j},$$

in the mean case $\widehat{p}_j$ are all equal to $1/N\sum_{j=1}^{N}\widehat{p}_j \approx 0.22$ and $\widehat{q}_j$ are all equal to $1/N\sum_{j=1}^{N}\widehat{q}_j \approx 0.13$ and the formula simplifies in $\left(C_{64}^{32}C_{84}^{42}\right)^{110}$, which is not reachable by common computers. So, testing all the possible datasets might be inappropriate. Although it is possible to derive the law of $x_j^{(0)}$ and $x_j^{(1)}$. Let us denote

$$\mathcal{U}([\![\alpha, \beta]\!], n_0) = \{\text{Draw of size } n_0 \text{ in } [\![\alpha, \beta]\!] \text{ without replacement}\}.$$

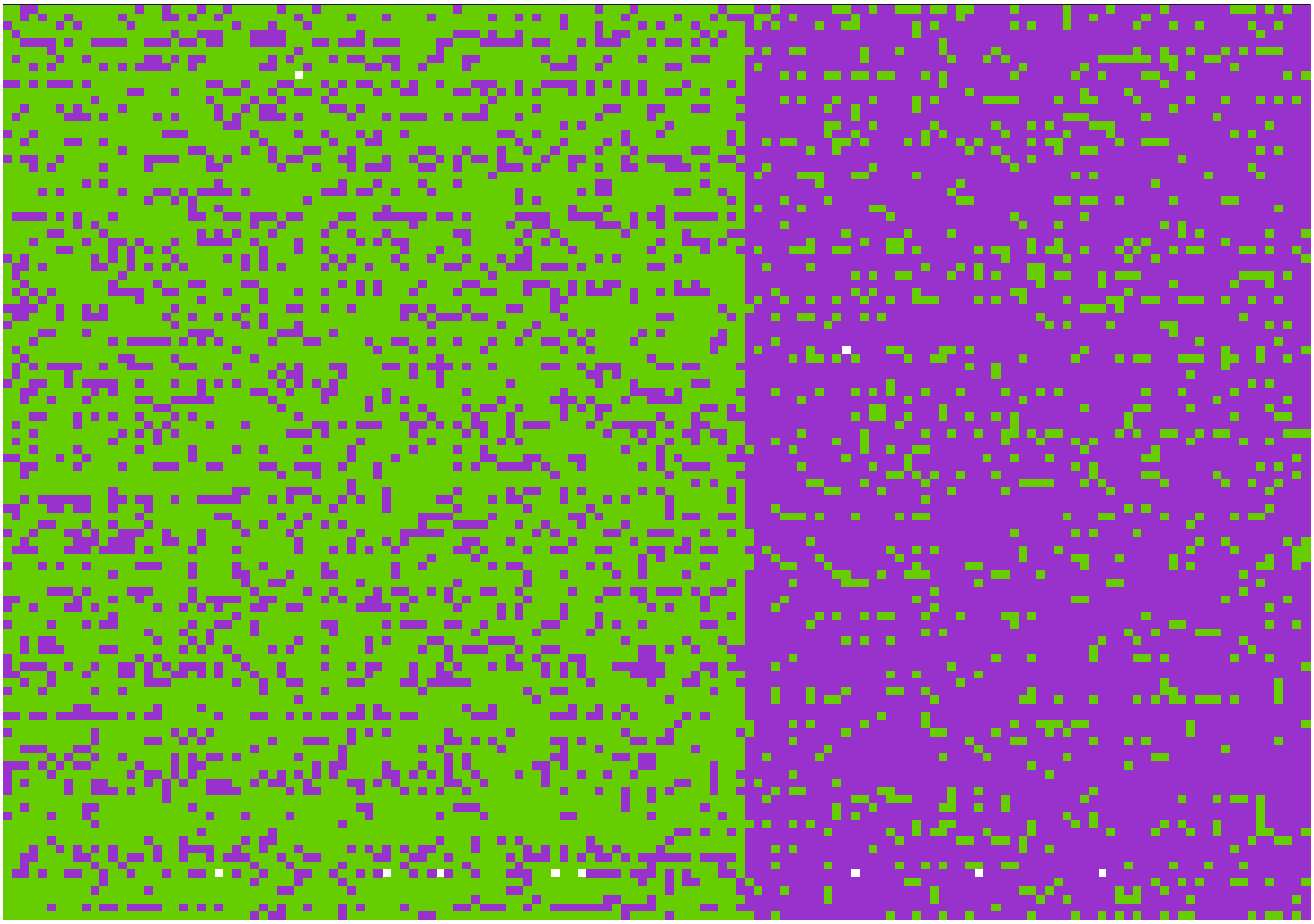


FIGURE E4 Example of a simulated dataset for the mammography analysis where green pixels correspond with 0s, purple pixels correspond with 1s and white correspond with NAs. Each row is a radiologist and each column is a mammogram.

Then $u \in \mathcal{U}([\![1, 84]\!], 84\widehat{p}_j)$ gives the indices of the non-diseased women who are diagnosed as diseased by the radiologist. Denoting $u = \{u_i\}_{i \in [\![1, 84\widehat{p}_j]\!]}$:

$$\mathbb{P}(\{x_{ij}^{(0)}\}_{i \in u} = 1, \{x_{ij}^{(0)}\}_{i \in u^c} = 0 | \widehat{p}_j) = \frac{\prod_{i \in [\![1, 84\widehat{p}_j]\!]} w_{u_i}}{\prod_{i \in [\![1, 84\widehat{p}_j]\!]} \left(\sum_{k \in u_{[\![i, 84\widehat{p}_j]\!]}} w_k + \sum_{k \in u^c} w_k \right)},$$

where u^c is the complementary of u in $[\![1, 84]\!]$. Equivalently, for any $v \in \mathcal{U}([\![85, 148]\!], 64\widehat{q}_j)$:

$$\mathbb{P}(\{x_{ij}^{(1)}\}_{i \in v} = 0, \{x_{ij}^{(1)}\}_{i \in v^c} = 1 | \widehat{q}_j) = \frac{\prod_{i \in [\![85, 64\widehat{q}_j]\!]} w_{v_i}}{\prod_{i \in [\![85, 64\widehat{q}_j]\!]} \left(\sum_{k \in v_{[\![i, 64\widehat{q}_j]\!]}} w_k + \sum_{k \in v^c} w_k \right)},$$

where v^c is the complementary of v in $[\![85, 148]\!]$. These probabilities were calculated by considering the order which is the solution chosen by the function `sample` of the software **R** [8].