Published online: 2019-11-19

Overrating Classifier Performance in ROC Analysis in the Absence of a Test Set: Evidence from Simulation and Italian CARATkids Validation

Giovanna Cilluffo[1,2,§] Salvatore Fasola[1,2,§] Giuliana Ferrante[3] Laura Montalbano[1] Ilaria Baiardini[4] Luciana Indinnimeo[5] Giovanni Viegi[1,6] Joao A. Fonseca[7] Stefania La Grutta[1]

[1] Institute for Biomedical Research and Innovation, National Research Council of Italy, Palermo, Italy
[2] Department of Economical, Business and Statistical Science, University of Palermo, Palermo, Italy
[3] Department of Health Promotion Sciences, Maternal and Infant Care, Internal Medicine and Medical Specialities, University of Palermo, Italy
[4] Department of Biomedical Sciences, Humanitas University, Milan, Italy
[5] Department of Pediatrics and NPI, University of Roma Sapienza, Rome, Italy
[6] Institute of Clinical Physiology, Pulmonary Environmental Epidemiology Unit, National Research Council of Italy, Pisa, Italy
[7] Department of Immunoallergy, CUF Porto Hospital and Institute, Porto, Portugal

Address for correspondence Salvatore Fasola, PhD, Institute for Biomedical Research and Innovation, National Research Council of Italy, Palermo, Italy (e-mail: salvatore.fasola@irib.cnr.it).

Methods Inf Med 2019;58:e27–e42.

Abstract

Background The use of receiver operating characteristic curves, or "ROC analysis," has become quite common in biomedical research to support decisions. However, sensitivity, specificity, and misclassification rates are still often estimated using the training sample, overlooking the risk of overrating the test performance.

Methods A simulation study was performed to highlight the inferential implications of splitting (or not) the dataset into training and test set. The normality assumption was made for the classifier given the disease status, and the Youden's criterion considered for the detection of the optimal cutoff. Then, an ROC analysis with sample split was applied to assess the discriminant validity of the Italian version of the Control of Allergic Rhinitis and Asthma Test (CARATkids) questionnaire for children with asthma and rhinitis, for which recent studies may have reported liberal performance estimates.

Results The simulation study showed that both single split and cross-validation (CV) provided unbiased estimators of sensitivity, specificity, and misclassification rate, therefore allowing computation of confidence intervals. For the Italian CARATkids questionnaire, the misclassification rate estimated by fivefold CV was 0.22, with 95% confidence interval 0.14 to 0.30, indicating an acceptable discriminant validity.

Conclusions Splitting into training and test set avoids overrating the test performance in ROC analysis. Validated through this method, the Italian CARATkids is valid for assessing disease control in children with asthma and rhinitis.

Keywords
- asthma control test
- sample split
- performance estimators
- optimal cutoff
- simulation study
- true predictive performance

§ *These two authors contributed equally.*

received
October 4, 2018
accepted after revision
May 21, 2019

DOI https://doi.org/10.1055/s-0039-1693732.
ISSN 0026-1270.

© 2019 Georg Thieme Verlag KG Stuttgart · New York

License terms

Introduction

The use of receiver operating characteristic curves, or "ROC analysis," has become quite common in biomedical research to support decisions.[1–3] In fact, continuous developments in clinical, biological, and psychometric methods provide a wide range of measurements that can be evaluated as potential diagnostic or prognostic tools. Several advanced nonparametric, semiparametric, and parametric methods have been developed for estimating and comparing ROC curves derived from continuous classifiers.[4] However, the most widespread approach to ROC analysis, routinely used in a clinical setting, is still the simplest one: several values of some numerical (continuous or discrete) classifier are evaluated as possible "optimal" cutoff for labeling individuals as "diseased" or "nondiseased."[5–8] The goal is to set up a simple screening test, therefore avoiding performing a more invasive, expensive, or time-consuming "gold standard" test.

To derive a ROC curve, sensitivity is plotted against one minus specificity derived from cross-tabulations (CVs) of the true binary status and several binary classifiers obtained through different cutoffs. Different criteria have been proposed for establishing the "optimal" cutoff, mainly based on a trade-off between sensitivity and specificity. However, there is no general criterion that guarantees optimality in all situations, since optimality may depend on different test characteristics and implications (costs, psychological consequences) of false positivities and false negativities. The most widely used criteria are minimization of the distance from (0,1)[9] and maximization of the Youden's index (sensitivity + specificity-1),[10] the latter being somewhat more appropriate.[11] Sensitivity, specificity, and misclassification rates, obtained with the optimal cutoff, together with the area under the ROC curve, are commonly used to report the predictive performance of a classifier.[12,13]

The need to assess the predictive performance of a classifier on an independent test sample has been well demonstrated, for example, in the context of machine learning,[14] decision trees,[15] and penalized least square discriminant analysis.[16] By contrast, this topic appears to have been overlooked in medical literature about ROC analysis, with the result that the aforementioned performance indicators are still quite often estimated using the same sample of data where the test was developed.

Although the issue of deriving appropriate estimators for the performance error rates could be bypassed using parametric[17–19] or Bayesian approaches,[15] these methods may be unfamiliar to medical researchers. In addition, the main issue of the training-test set approach is the choice of the training set proportion (usually 1:2 or 2:3), especially when the sample size is small.[20] An alternative approach is CV.[21,22] CV leaves out one or more observations in turn to be used as the test sample; all the test samples form a partition of the whole sample, so that all the observations are involved in estimation of the classification error. The dilemma, however, is about choosing the classifier to retain, since different classifiers may be obtained from different training subsets. In general, one may then return to the full dataset.[23]

The motivation for writing this article concerns the increasing acknowledgment of the prognostic value of patient-reported outcomes in patients with asthma,[24] rhinitis,[25,26] or both.[27,28] In fact, recent studies have provided simple screening tests for assessing the disease control and therefore monitoring its course. However, out of the five studies referenced above, only one[24] appears to have randomly divided the total sample into a "development" (or "training") sample (75%) and a "confirmatory" (or " test") sample (25%). In particular, for pediatric patients with asthma and rhinitis, one of the previous validation studies of the "Control of Allergic Rhinitis and Asthma Test" (Control of Allergic Rhinitis and Asthma Test (CARAT) CARATkids questionnaire) in Brazilian children[27] reported an estimated probability of 1 for the CARATkids score being larger than 3 with uncontrolled asthma (sensitivity), and an estimated probability of 0.93 for the CARATkids score being lower than 7 with controlled asthma (specificity). Since they report sensitivity and specificity from the same sample where they were maximized, such estimates may be affected by positive bias, that is, they probably overestimate the true sensitivity and specificity in the general population.

The aim of this study was to highlight the positive inferential implications of splitting the study sample into a training sample (where the optimal test is derived) and a test sample (where performance or error rates are estimated) in the setting of ROC analysis. This was accomplished by using a well-known data generating mechanisms and a simple simulation study, as a possible reference for medical researchers dealing with such data.

Methods

Statistical Characterization

Let Y_i be a dichotomous random variable for which

$$Y_i = \begin{cases} 0 & \text{if individual } i \text{ is not diseased} \\ 1 & \text{if individual } i \text{ is diseased} \end{cases}$$

$i = 1,2,\ldots,n$, where n is the size of a given sample of individuals from some target population. It is possible to define

$$p = prob(Y_i = 1) \quad (1)$$

as the prevalence of the disease in the target population. Now consider a quantitative random variable X_i, and suppose that, on average, the X values are greater in diseased individuals. Given this property, X_i may be considered as a potential classifier for Y_i. For the illustrative purpose of this article, the distribution of X_i conditional to the disease status is supposed to be Normal (or Gaussian), so that

$$\begin{cases} X_i \sim \mathcal{N}(\mu_1, \sigma_1^2) & \text{if } Y_i = 0 \text{ (individual } i \text{ is not diseased)} \\ X_i \sim \mathcal{N}(\mu_2, \sigma_2^2) & \text{if } Y_i = 1 \text{ (individual } i \text{ is diseased)} \end{cases}$$

Here μ_1 and σ_1 are, respectively, the true mean and standard deviation of the classifier among nondiseased individuals, while μ_2 and σ_2 are their counterparts among diseased

individuals (with $\mu_2 > \mu_1$). On this ground, the rationale of ROC analyses is that the "working variable"

$$\hat{Y}_i(c) = \begin{cases} 0 \text{ (say that individual } i \text{ is nondiseased)} \\ \qquad\qquad \text{if } x_i \le c \text{ (negative test)} \\ 1 \text{ (say that individual } i \text{ is diseased)} \\ \qquad\qquad \text{if } x_i > c \text{ (positive test)} \end{cases}$$

can be used as a simple classification rule in the target population for some given cutoff c. The accuracy of the test depends on its ability to correctly detect diseased and nondiseased individuals. In particular, the performance indicators of interest are usually sensitivity (probability that the test is positive in diseased individuals), specificity (probability that the test is negative in nondiseased individuals), and the misclassification rate (probability of incorrectly classifying an individual). The true performance has to be evaluated in the target population, and of course, it depends on the cutoff c. The true sensitivity is defined as:

$$Se(c) = prob(X > c|Y = 1) = 1 - \Phi(c,\mu_2,\sigma_2^2), \quad (2)$$

where $\Phi(\cdot)$ represents the Gaussian distribution function, with parameters of the diseased population in this case. Similarly, the true specificity is:

$$Sp(c) = prob(X \le c|Y = 0) = \Phi(c,\mu_1,\sigma_1^2), \quad (3)$$

where now the distribution of X in nondiseased individuals is involved. Finally, the true misclassification rate is:

$$\begin{aligned} \epsilon(c) &= \text{prob}\{(X > c \cap Y = 0) \cup (X \le c \cap Y = 1)\} \\ &= \text{prob}(X > c \cap Y = 0) + \text{prob}(X \le c \cap Y = 1) \\ &= \text{prob}(Y = 0)\,\text{prob}(X > c|Y = 0) + \text{prob}(Y = 1)\,\text{prob}(X \le c|Y = 1) \\ &= (1 - p)[1 - Sp(c)] + p[1 - Se(c)]. \quad (4) \end{aligned}$$

In ROC analyses, pairs (x_i,y_i) are collected on n individuals; in particular, the disease status y_i is assessed through some validated gold standard test. To set up the classification rule, the cutoff to use is selected among several candidates on a grid of x values, as the value that optimizes a given criterion. The sample of individuals on which this optimization is performed is called the "training set." The size of the training set will be denoted by n_δ, where $\delta = n_\delta/n$ (e.g., $\delta = 50\%$, $\delta = 67\%$ or $\delta = 100\%$) indicates the training percentage. According to Youden's criterion,[10] the optimal cutoff is estimated as:

$$\begin{aligned} \hat{c}_\delta &= arg\ max_{1\le j\le J}\left\{\frac{\sum_{i=1}^{n_\delta}[\hat{Y}_i(c_j)]y_i}{\sum_{i=1}^{n_\delta} y_i} + \frac{\sum_{i=1}^{n_\delta}[1-\hat{Y}_i(c_j)][1-y_i]}{\sum_{i=1}^{n_\delta}[1-y_i]} - 1\right\} \\ &= arg\ max_{1\le j\le J}\{\widehat{Se}_\delta(c_j) + \widehat{Sp}_\delta(c_j) - 1\}, \quad (5) \end{aligned}$$

where c_j is the j-th candidate cutoff on a J-dimensional discrete grid of x values, $\widehat{Se}_\delta$ is the test sensitivity in the training sample, and $\widehat{Sp}_\delta$ is the test specificity in the training sample.

Once the optimal cutoff $\hat{c}_\delta$ has been identified using the training set, the next step is to estimate the true predictive performance of the optimized test, that is, to estimate (2), (3), and (4) for $c = \hat{c}_\delta$. To accomplish this, two simple approaches are commonly used. The first, liberal approach consists in estimating the performance in the training set:

$$\widehat{Se}_\delta(\hat{c}) = \frac{\sum_{i=1}^{n_\delta}[\hat{Y}_i(\hat{c}_\delta)]y_i}{\sum_{i=1}^{n_\delta} y_i} \quad (6)$$

$$\widehat{Sp}_\delta(\hat{c}_\delta) = \frac{\sum_{i=1}^{n_\delta}[1-\hat{Y}_i(\hat{c}_\delta)][1-y_i]}{\sum_{i=1}^{n_\delta}[1-y_i]} \quad (7)$$

$$\hat{\epsilon}_\delta(\hat{c}_\delta) = (1 - \hat{p}_\delta)[1 - \widehat{Sp}_\delta(\hat{c}_\delta)] + \hat{p}_\delta[1 - \widehat{Se}_\delta(\hat{c}_\delta)]. \quad (8)$$

If the true disease prevalence (i.e., the prevalence in the general population) is simply estimated by $\hat{p}_\delta = \frac{\sum_{i=1}^{n_\delta} y_i}{n_\delta}$ (i.e., the prevalence in the study sample), Eq. (8) reduces to the following, more familiar expression:

$$\hat{\epsilon}_\delta(\hat{c}_\delta) = \frac{\sum_{i=1}^{n_\delta} I[y_i \ne \hat{Y}_i(\hat{c}_\delta)]}{n_\delta}, \quad (9)$$

where $I(\cdot)$ is an indicator function. However, if the number of diseased and nondiseased individuals is fixed a priori in the study design, using Eq. (9) would be definitely wrong for obvious reasons; in this case, the true prevalence should be inferred from previous studies or just hypothesized.

The second, more conservative approach consists in randomly leaving out a given proportion of the sample, say n_γ individuals (e.g., $\gamma = 50\%$ or $\gamma = 33\%$), to use as the test sample, that is, to estimate (2), (3), and (4) for $c = \hat{c}_\delta$, where $\hat{c}_\delta$ comes from the training sample ($\delta + \gamma = 100\%$). The performance estimators will therefore be denoted with $\widehat{Se}_\gamma(\hat{c}_\delta)$, $\widehat{Sp}_\gamma(\hat{c}_\delta)$, and $\hat{\epsilon}_\gamma(\hat{c}_\delta)$.

Sometimes, separation into a training and a test set is difficult due to the small sample size. In this case, k-fold CV makes better use of the data. With this approach, the whole sample is randomly partitioned into k subgroups to be used as the test set in different steps. At each step, $k-1$ groups (training set) are used to develop a classifier, and the outcome predictions are derived in the test set. This procedure is repeated until all the k subgroups have been used as the test set, and the overall classifier performance is therefore evaluated. The main issue with this approach is the choice of the classifier to retain, since different classifiers may be obtained at each step; in this case, one may return to the full dataset using $\hat{c}_{100\%}$.[23] The k-fold CV performance estimators will be denoted with $\widehat{Se}_k$, $\widehat{Sp}_k$ and $\hat{\epsilon}_k$; when $k = n$, CV is referred to as leave-one-out CV (LOOCV).

The next section is intended to show, empirically, that the first approach (100% training) leads to an overestimation of the true sensitivity and specificity (and consequently an underestimation of the misclassification rate in the target population), while the second approach (independent test sample) provides unbiased estimates. The following true performance indicators, averaged over $\hat{c}_\delta$, will be considered for the different procedures ($\delta = 100\%$, $\delta = 67\%$ and $\delta = 50\%$):

$$\begin{aligned} Se_\delta &= \sum_{j=1}^{J} Se(c_j)\text{prob}(\hat{c}_\delta = c_j), \quad (10) \\ Sp_\delta &= \sum_{j=1}^{J} Sp(c_j)\text{prob}(\hat{c}_\delta = c_j), \quad (11) \\ \epsilon_\delta &= \sum_{j=1}^{J} \epsilon(c_j)\text{prob}(\hat{c}_\delta = c_j), \quad (12) \end{aligned}$$

where the terms $\text{prob}(\hat{c}_\delta = c_j)$ are estimated through simulation.

Simulation Study

In the simulation study showed in the next section, the n pairs (y_i,x_i) were generated as follows: first, y_i was generated from a Bernoulli random variable with probability of success equal to p, then, x_i was generated from a Gaussian distribution, using parameters μ_1 and σ_1^2 if $y_i=0$, μ_2 and σ_2^2 if $y_i=1$. Simulations were performed to assess the true sensitivity, the specificity and the misclassification rate, and the properties of the estimators presented in the previous section.

Different configurations of target populations were considered by varying the mean difference (distances between μ_1 and μ_2, $\delta_\mu=2, 4, 6, 8$), the variances $[(\sigma_1,\sigma_2)=(1,4),(2,3),(2,2),(3,2),(4,1)]$ and the disease prevalence ($p=0.6, 0.4$). ►**Figure 1** illustrates the hypothesized populations of diseased and nondiseased individuals. From each population, 1,000 random data samples were generated using different sample sizes ($n=50, 100, 200$). For each simulated sample, a ROC analysis was performed to detect the optimal cutoff using Youden's criterion (on a discrete grid of $J=30$ equally spaced candidates c_j), and the performance of the obtained test was estimated using different training percentages ($\delta=100\%, 67\%, 50\%$), fivefold CV and LOOCV.

Clinical Data

The data analyzed in the article come from a cross-sectional study performed at the Pediatric Pulmonology-Allergology outpatient clinic of the CNR Institute for Biomedical Research and Innovation of Palermo, and at the Department of Pediatrics of the Sapienza University of Rome, Italy. Children aged 6 to 11 years, with a medical diagnosis of allergic rhinitis and asthma, were consecutively enrolled from March 2015 to December 2016. Children with other respiratory or chronic diseases that might interfere with the study measurements, as well as children with psychiatric disorders and/or cognitive impairment, were excluded. The $n=112$ patients were assessed at baseline (T0) and after a mean period of 3 months (T1). All children attended both visits and completed an Italian version of the CARATkids questionnaire[27–29] and the Childhood Asthma Control Test (C-ACT).[24]

Some psychometric characteristics of the Italian CARATkids questionnaire were assessed. In particular, the *discriminant validity* of CARATkids was evaluated in previous studies[27,28] as its ability to detect children with uncontrolled asthma, defined as C-ACT score $\leq$19.[24] Moreover, the more general *cross-sectional and longitudinal validity* was assessed through the correlation between the total score of the CARATkids and the total score of C-ACT. A ROC analysis was performed and the optimal cutoff value for CARATkids selected according to Youden's method. The area under the curve (AUC) was estimated and its significance ($AUC>0.5$) tested using the method described by DeLong et al.[30]

The study was approved by the local ethic committee (N 11/2014 Azienda ospedaliera Universitaria Policlinico Paolo Giaccone) and conducted in accordance with the Declaration of Helsinki and Good Clinical Practice guidelines. All parents provided written informed consent. The study was registered on the central registration system ClinicalTrials.gov (NCT 02409550).

Results

Simulation Study

►**Tables 1** to **6** show the means and standard deviations of the different estimates obtained in the simulated data samples. Scenarios with $\delta_\mu=2$, 8 were reported in ►**Supplementary Tables S1** to **S6** (online only). In general, small differences are observed in the expected value of the cutoff estimator ($\hat{c}$), which locates approximately at the intersection point between the distributions in ►**Fig. 1**, that is, the optimal cutoff in the population. As expected,

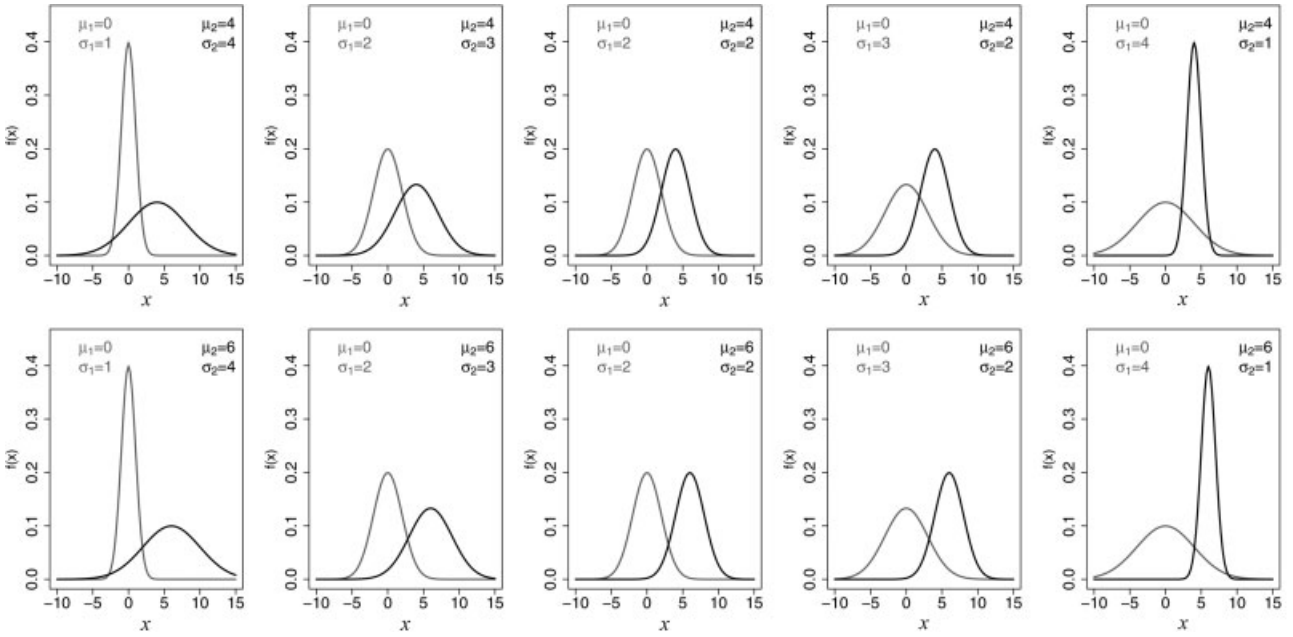

Fig. 1 Theoretical scenarios of populations considered in the simulation study. Gray curves indicate nondiseased individuals, and black curves indicate diseased individuals.

Table 1 Simulated means and standard deviations (σ) of the detected cutoff ($\hat{c}$), and of estimated sensitivity $\widehat{Se}$, specificity $\widehat{Sp}$, and misclassification rate ($\hat{\epsilon}$) with $n = 50$ and $p = 0.60$. *Se*, *Sp*, and ϵ indicate the true performances

Δ_μ	σ_1, σ_2	δ	$\hat{c}$	$\sigma_{\hat{c}}$	*Se*	$\widehat{Se}$	$\sigma_{\widehat{Se}}$	*Sp*	$\widehat{Sp}$	$\sigma_{\widehat{Sp}}$	ϵ	$\hat{\epsilon}$	$\sigma_{\hat{\epsilon}}$
4	1, 4	100%	1.660	0.512	0.719	0.738	0.085	0.930	0.974	0.043	0.196	0.168	0.052
		67%	1.560	0.539	0.727	0.729	0.151	0.916	0.921	0.131	0.197	0.195	0.098
		50%	1.473	0.576	0.734	0.739	0.123	0.899	0.900	0.133	0.200	0.197	0.080
		Fivefold CV	1.660	0.512	0.719	0.727	0.084	0.930	0.926	0.058	0.196	0.193	0.060
		LOOCV	1.660	0.512	0.719	0.722	0.087	0.930	0.935	0.056	0.196	0.193	0.062
	2, 3	100%	2.019	0.894	0.737	0.768	0.106	0.822	0.883	0.095	0.229	0.187	0.056
		67%	2.021	0.986	0.734	0.731	0.169	0.818	0.824	0.192	0.232	0.233	0.106
		50%	1.880	1.080	0.747	0.748	0.151	0.796	0.804	0.181	0.233	0.231	0.089
		Fivefold CV	2.019	0.894	0.737	0.739	0.099	0.822	0.815	0.100	0.229	0.230	0.070
		LOOCV	2.019	0.894	0.737	0.735	0.110	0.822	0.824	0.105	0.229	0.229	0.078
	2, 2	100%	1.906	0.700	0.838	0.866	0.086	0.816	0.874	0.089	0.171	0.131	0.048
		67%	1.881	0.813	0.837	0.836	0.153	0.808	0.813	0.190	0.175	0.174	0.098
		50%	1.823	0.899	0.840	0.838	0.138	0.797	0.797	0.175	0.177	0.178	0.084
		Fivefold CV	1.906	0.700	0.838	0.838	0.082	0.816	0.814	0.092	0.171	0.171	0.063
		LOOCV	1.906	0.700	0.838	0.836	0.093	0.816	0.819	0.096	0.171	0.170	0.072
	3, 2	100%	1.729	0.903	0.850	0.877	0.094	0.710	0.771	0.113	0.206	0.166	0.053
		67%	1.668	1.067	0.848	0.848	0.156	0.700	0.701	0.213	0.211	0.211	0.104
		50%	1.614	1.151	0.850	0.849	0.145	0.693	0.698	0.195	0.213	0.212	0.086
		Fivefold CV	1.729	0.903	0.850	0.845	0.092	0.710	0.707	0.114	0.206	0.210	0.070
		LOOCV	1.729	0.903	0.850	0.843	0.105	0.710	0.712	0.119	0.206	0.209	0.078
	4, 1	100%	2.153	0.657	0.942	0.963	0.046	0.703	0.749	0.102	0.154	0.124	0.045
		67%	2.023	0.866	0.942	0.940	0.100	0.690	0.695	0.204	0.159	0.157	0.090
		50%	1.916	1.063	0.936	0.938	0.103	0.679	0.677	0.186	0.166	0.168	0.083
		Fivefold CV	2.153	0.657	0.942	0.945	0.053	0.703	0.696	0.107	0.154	0.155	0.056
		LOOCV	2.153	0.657	0.942	0.946	0.057	0.703	0.703	0.105	0.154	0.151	0.058
6	1, 4	100%	1.835	0.506	0.849	0.860	0.067	0.950	0.984	0.031	0.111	0.091	0.040
		67%	1.721	0.584	0.855	0.853	0.120	0.932	0.931	0.121	0.114	0.117	0.081
		50%	1.632	0.618	0.860	0.863	0.096	0.918	0.920	0.119	0.117	0.114	0.068
		Fivefold CV	1.835	0.506	0.849	0.853	0.068	0.950	0.938	0.051	0.111	0.113	0.049
		LOOCV	1.835	0.506	0.849	0.849	0.069	0.950	0.944	0.046	0.111	0.112	0.050
	2, 3	100%	2.614	0.828	0.862	0.882	0.071	0.886	0.935	0.065	0.128	0.097	0.041
		67%	2.567	0.937	0.863	0.859	0.125	0.878	0.883	0.150	0.131	0.132	0.081
		50%	2.465	1.030	0.867	0.862	0.115	0.863	0.859	0.150	0.134	0.139	0.072
		Fivefold CV	2.614	0.828	0.862	0.862	0.071	0.886	0.879	0.073	0.128	0.131	0.055
		LOOCV	2.614	0.828	0.862	0.857	0.077	0.886	0.888	0.072	0.128	0.130	0.058
	2, 2	100%	2.877	0.696	0.930	0.947	0.049	0.913	0.955	0.049	0.077	0.051	0.032
		67%	2.734	0.802	0.935	0.935	0.093	0.898	0.895	0.142	0.080	0.082	0.070
		50%	2.605	0.947	0.938	0.937	0.087	0.880	0.879	0.135	0.085	0.086	0.063
		Fivefold CV	2.877	0.696	0.930	0.932	0.055	0.913	0.903	0.061	0.077	0.079	0.045
		LOOCV	2.877	0.696	0.930	0.930	0.057	0.913	0.911	0.057	0.077	0.077	0.047
	3, 2	100%	3.095	0.825	0.910	0.931	0.063	0.840	0.886	0.076	0.118	0.087	0.040
		67%	3.005	0.999	0.910	0.906	0.120	0.829	0.823	0.170	0.122	0.127	0.086
		50%	2.892	1.134	0.912	0.912	0.108	0.816	0.807	0.162	0.126	0.129	0.069
		Fivefold CV	3.095	0.825	0.910	0.909	0.065	0.840	0.828	0.083	0.118	0.123	0.054
		LOOCV	3.095	0.825	0.910	0.907	0.071	0.840	0.833	0.084	0.118	0.122	0.060
	4, 1	100%	3.818	0.731	0.965	0.978	0.036	0.826	0.867	0.080	0.091	0.068	0.037
		67%	3.679	0.929	0.964	0.963	0.079	0.815	0.812	0.171	0.096	0.097	0.075
		50%	3.426	1.216	0.963	0.964	0.076	0.794	0.796	0.160	0.104	0.104	0.071
		Fivefold CV	3.818	0.731	0.965	0.963	0.046	0.826	0.817	0.087	0.091	0.096	0.049
		LOOCV	3.818	0.731	0.965	0.964	0.046	0.826	0.823	0.087	0.091	0.092	0.049

Abbreviation: LOOCV, leave-one-out cross-validation.

Table 2 Simulated means and standard deviations (σ) of the detected cutoff ($\hat{c}$), and of estimated sensitivity $\widehat{Se}$, specificity $\widehat{Sp}$, and misclassification rate ($\hat{\epsilon}$) with $n = 50$ and $p = 0.4$. *Se*, *Sp*, and ϵ indicate the true performances

Δ_μ	σ_1, σ_2	δ	$\hat{c}$	$\sigma_{\hat{c}}$	*Se*	$\widehat{Se}$	$\sigma_{\widehat{Se}}$	*Sp*	$\widehat{Sp}$	$\sigma_{\widehat{Sp}}$	ϵ	$\hat{\epsilon}$	$\sigma_{\hat{\epsilon}}$
4	1, 4	100%	1.699	0.492	0.716	0.743	0.103	0.937	0.963	0.049	0.152	0.125	0.046
		67%	1.635	0.532	0.721	0.715	0.192	0.926	0.926	0.107	0.156	0.157	0.091
		50%	1.544	0.597	0.728	0.729	0.152	0.908	0.908	0.114	0.164	0.164	0.082
		Fivefold CV	1.699	0.492	0.716	0.719	0.104	0.937	0.927	0.055	0.152	0.156	0.058
		LOOCV	1.699	0.492	0.716	0.715	0.107	0.937	0.935	0.055	0.152	0.152	0.059
	2, 3	100%	2.129	0.836	0.726	0.776	0.114	0.837	0.873	0.093	0.207	0.166	0.051
		67%	2.087	0.965	0.728	0.730	0.203	0.826	0.827	0.162	0.213	0.211	0.103
		50%	1.986	1.022	0.737	0.738	0.175	0.812	0.811	0.162	0.218	0.219	0.094
		Fivefold CV	2.129	0.836	0.726	0.730	0.111	0.837	0.826	0.091	0.207	0.212	0.069
		LOOCV	2.129	0.836	0.726	0.725	0.122	0.837	0.834	0.099	0.207	0.209	0.077
	2, 2	100%	1.966	0.689	0.832	0.879	0.088	0.824	0.865	0.085	0.173	0.130	0.048
		67%	1.947	0.799	0.830	0.827	0.172	0.817	0.818	0.155	0.178	0.179	0.099
		50%	1.919	0.877	0.830	0.832	0.156	0.810	0.818	0.148	0.182	0.177	0.086
		Fivefold CV	1.966	0.689	0.832	0.833	0.097	0.824	0.823	0.085	0.173	0.173	0.065
		LOOCV	1.966	0.689	0.832	0.834	0.100	0.824	0.831	0.088	0.173	0.168	0.069
	3, 2	100%	1.833	0.901	0.838	0.891	0.091	0.721	0.760	0.107	0.232	0.188	0.055
		67%	1.798	1.064	0.835	0.840	0.183	0.714	0.719	0.176	0.238	0.234	0.106
		50%	1.785	1.102	0.834	0.835	0.172	0.712	0.715	0.163	0.239	0.237	0.093
		Fivefold CV	1.833	0.901	0.838	0.844	0.094	0.721	0.712	0.101	0.232	0.234	0.071
		LOOCV	1.833	0.901	0.838	0.836	0.105	0.721	0.716	0.109	0.232	0.235	0.081
	4, 1	100%	2.179	0.585	0.943	0.973	0.040	0.705	0.737	0.085	0.200	0.167	0.051
		67%	2.108	0.720	0.940	0.943	0.114	0.698	0.703	0.152	0.205	0.200	0.096
		50%	2.050	0.867	0.934	0.934	0.113	0.692	0.697	0.139	0.211	0.209	0.087
		Fivefold CV	2.179	0.585	0.943	0.942	0.057	0.705	0.705	0.087	0.200	0.199	0.059
		LOOCV	2.179	0.585	0.943	0.941	0.059	0.705	0.708	0.088	0.200	0.197	0.063
6	1, 4	100%	1.884	0.502	0.846	0.870	0.076	0.955	0.979	0.032	0.089	0.065	0.034
		67%	1.783	0.542	0.852	0.858	0.141	0.942	0.942	0.096	0.094	0.092	0.076
		50%	1.737	0.574	0.854	0.857	0.119	0.935	0.936	0.088	0.097	0.095	0.063
		Fivefold CV	1.884	0.502	0.846	0.853	0.080	0.955	0.950	0.042	0.089	0.089	0.045
		LOOCV	1.884	0.502	0.846	0.850	0.080	0.955	0.954	0.040	0.089	0.087	0.046
	2, 3	100%	2.726	0.777	0.855	0.889	0.074	0.898	0.928	0.059	0.119	0.088	0.037
		67%	2.671	0.877	0.857	0.853	0.150	0.889	0.888	0.122	0.124	0.128	0.083
		50%	2.551	0.981	0.863	0.861	0.133	0.874	0.870	0.130	0.131	0.134	0.078
		Fivefold CV	2.726	0.777	0.855	0.853	0.080	0.898	0.890	0.063	0.119	0.124	0.053
		LOOCV	2.726	0.777	0.855	0.851	0.082	0.898	0.894	0.066	0.119	0.123	0.057
	2, 2	100%	2.928	0.687	0.927	0.956	0.049	0.917	0.947	0.047	0.079	0.050	0.030
		67%	2.824	0.767	0.931	0.940	0.105	0.906	0.911	0.112	0.084	0.079	0.071
		50%	2.759	0.851	0.932	0.932	0.100	0.898	0.896	0.110	0.088	0.090	0.066
		Fivefold CV	2.928	0.687	0.927	0.929	0.059	0.917	0.912	0.053	0.079	0.080	0.044
		LOOCV	2.928	0.687	0.927	0.928	0.061	0.917	0.918	0.052	0.079	0.078	0.045
	3, 2	100%	3.175	0.831	0.904	0.942	0.064	0.846	0.876	0.071	0.131	0.098	0.042
		67%	3.126	0.952	0.903	0.898	0.142	0.840	0.833	0.139	0.135	0.139	0.087
		50%	3.088	1.059	0.901	0.901	0.128	0.834	0.828	0.125	0.139	0.143	0.075
		Fivefold CV	3.175	0.831	0.904	0.903	0.074	0.846	0.839	0.073	0.131	0.135	0.057
		LOOCV	3.175	0.831	0.904	0.902	0.077	0.846	0.842	0.077	0.131	0.134	0.061
	4, 1	100%	3.930	0.643	0.962	0.985	0.030	0.834	0.858	0.067	0.115	0.091	0.041
		67%	3.831	0.827	0.957	0.954	0.109	0.826	0.822	0.133	0.121	0.125	0.083
		50%	3.716	1.000	0.956	0.956	0.091	0.816	0.818	0.122	0.128	0.127	0.072
		Fivefold CV	3.930	0.643	0.962	0.961	0.047	0.834	0.825	0.072	0.115	0.119	0.051
		LOOCV	3.930	0.643	0.962	0.962	0.046	0.834	0.830	0.070	0.115	0.117	0.051

Abbreviation: LOOCV, leave-one-out cross-validation.

Table 3 Simulated means and standard deviations (σ) of the detected cutoff ($\hat{c}$), and of estimated sensitivity $\widehat{Se}$, specificity $\widehat{Sp}$ and misclassification rate ($\hat{\epsilon}$) with $n = 100$ and $p = 0.6$. *Se*, *Sp* and ϵ indicate the true performances

Δ_μ	σ_1, σ_2	δ	$\hat{c}$	$\sigma_{\hat{c}}$	*Se*	$\widehat{Se}$	$\sigma_{\widehat{Se}}$	*Sp*	$\widehat{Sp}$	$\sigma_{\widehat{Sp}}$	ϵ	$\hat{\epsilon}$	$\sigma_{\hat{\epsilon}}$
4	1, 4	100%	1.754	0.432	0.712	0.725	0.067	0.947	0.967	0.035	0.194	0.178	0.039
		67%	1.721	0.480	0.714	0.707	0.116	0.940	0.938	0.085	0.196	0.200	0.073
		50%	1.669	0.496	0.719	0.718	0.094	0.933	0.931	0.084	0.196	0.197	0.056
		Fivefold CV	1.754	0.432	0.712	0.714	0.064	0.947	0.940	0.042	0.194	0.195	0.044
		LOOCV	1.754	0.432	0.712	0.712	0.068	0.947	0.944	0.043	0.194	0.195	0.047
	2, 3	100%	2.107	0.700	0.731	0.755	0.084	0.840	0.874	0.075	0.226	0.198	0.039
		67%	2.102	0.784	0.730	0.730	0.128	0.836	0.834	0.136	0.228	0.228	0.072
		50%	2.082	0.878	0.730	0.732	0.123	0.830	0.830	0.134	0.230	0.230	0.064
		Fivefold CV	2.107	0.700	0.731	0.734	0.076	0.840	0.836	0.072	0.226	0.225	0.048
		LOOCV	2.107	0.700	0.731	0.732	0.087	0.840	0.840	0.080	0.226	0.225	0.057
	2, 2	100%	1.967	0.575	0.836	0.856	0.073	0.828	0.861	0.073	0.168	0.142	0.037
		67%	1.972	0.641	0.833	0.832	0.115	0.826	0.825	0.135	0.170	0.170	0.069
		50%	1.980	0.691	0.830	0.829	0.107	0.825	0.829	0.123	0.172	0.171	0.059
		Fivefold CV	1.967	0.575	0.836	0.835	0.066	0.828	0.826	0.072	0.168	0.169	0.046
		LOOCV	1.967	0.575	0.836	0.836	0.077	0.828	0.829	0.077	0.168	0.167	0.054
	3, 2	100%	1.769	0.688	0.854	0.876	0.076	0.717	0.750	0.091	0.201	0.174	0.039
		67%	1.780	0.789	0.849	0.848	0.120	0.717	0.712	0.154	0.204	0.206	0.077
		50%	1.692	0.877	0.855	0.858	0.108	0.706	0.703	0.144	0.205	0.204	0.059
		Fivefold CV	1.769	0.688	0.854	0.853	0.071	0.717	0.715	0.088	0.201	0.202	0.050
		LOOCV	1.769	0.688	0.854	0.853	0.082	0.717	0.716	0.095	0.201	0.202	0.058
	4, 1	100%	2.202	0.434	0.951	0.964	0.036	0.708	0.730	0.073	0.146	0.130	0.032
		67%	2.185	0.528	0.947	0.949	0.070	0.706	0.706	0.132	0.149	0.149	0.063
		50%	2.115	0.665	0.945	0.945	0.074	0.699	0.698	0.117	0.153	0.155	0.054
		Fivefold CV	2.202	0.434	0.951	0.950	0.039	0.708	0.705	0.075	0.146	0.149	0.039
		LOOCV	2.202	0.434	0.951	0.952	0.041	0.708	0.707	0.076	0.146	0.146	0.041
6	1, 4	100%	1.921	0.423	0.845	0.852	0.051	0.962	0.980	0.026	0.108	0.097	0.031
		67%	1.859	0.470	0.848	0.847	0.086	0.954	0.954	0.073	0.109	0.111	0.057
		50%	1.840	0.514	0.849	0.848	0.074	0.950	0.952	0.069	0.111	0.111	0.048
		Fivefold CV	1.921	0.423	0.845	0.844	0.050	0.962	0.959	0.033	0.108	0.110	0.035
		LOOCV	1.921	0.423	0.845	0.843	0.052	0.962	0.962	0.034	0.108	0.110	0.037
	2, 3	100%	2.730	0.650	0.857	0.874	0.057	0.903	0.930	0.049	0.125	0.103	0.031
		67%	2.675	0.780	0.858	0.856	0.100	0.894	0.896	0.109	0.128	0.127	0.061
		50%	2.646	0.824	0.860	0.864	0.085	0.889	0.893	0.098	0.129	0.125	0.050
		Fivefold CV	2.730	0.650	0.857	0.859	0.054	0.903	0.899	0.052	0.125	0.125	0.040
		LOOCV	2.730	0.650	0.857	0.858	0.060	0.903	0.903	0.054	0.125	0.124	0.044
	2, 2	100%	2.939	0.537	0.930	0.943	0.038	0.922	0.946	0.042	0.073	0.056	0.024
		67%	2.887	0.622	0.931	0.931	0.072	0.916	0.916	0.091	0.075	0.075	0.049
		50%	2.830	0.709	0.933	0.933	0.068	0.909	0.908	0.088	0.077	0.077	0.043
		Fivefold CV	2.939	0.537	0.930	0.930	0.040	0.922	0.922	0.045	0.073	0.073	0.032
		LOOCV	2.939	0.537	0.930	0.930	0.042	0.922	0.922	0.047	0.073	0.073	0.035
	3, 2	100%	3.168	0.645	0.911	0.929	0.048	0.849	0.876	0.059	0.114	0.092	0.029
		67%	3.136	0.736	0.910	0.913	0.086	0.845	0.844	0.115	0.116	0.114	0.057
		50%	3.115	0.856	0.908	0.909	0.084	0.841	0.839	0.106	0.119	0.119	0.051
		Fivefold CV	3.168	0.645	0.911	0.911	0.048	0.849	0.848	0.060	0.114	0.114	0.038
		LOOCV	3.168	0.645	0.911	0.912	0.052	0.849	0.849	0.064	0.114	0.113	0.041
	4, 1	100%	3.949	0.500	0.968	0.978	0.025	0.836	0.856	0.059	0.085	0.071	0.026
		67%	3.861	0.634	0.968	0.969	0.051	0.830	0.828	0.112	0.087	0.088	0.050
		50%	3.801	0.748	0.966	0.968	0.053	0.825	0.821	0.100	0.090	0.091	0.045
		Fivefold CV	3.949	0.500	0.968	0.968	0.028	0.836	0.832	0.062	0.085	0.087	0.032
		LOOCV	3.949	0.500	0.968	0.969	0.029	0.836	0.835	0.061	0.085	0.085	0.032

Abbreviation: LOOCV, leave-one-out cross-validation.

Table 4 Simulated means and standard deviations (σ) of the detected cutoff ($\hat{c}$), and of estimated sensitivity $\widehat{Se}$, specificity $\widehat{Sp}$, and misclassification rate ($\hat{\epsilon}$) with $n = 100$ and $p = 0.4$. *Se*, *Sp*, and ϵ indicate the true performances

Δ_μ	σ_1, σ_2	δ	$\hat{c}$	$\sigma_{\hat{c}}$	*Se*	$\widehat{Se}$	$\sigma_{\widehat{Se}}$	*Sp*	$\widehat{Sp}$	$\sigma_{\widehat{Sp}}$	ϵ	$\hat{\epsilon}$	$\sigma_{\hat{\epsilon}}$
4	1, 4	100%	1.759	0.421	0.711	0.728	0.075	0.948	0.962	0.037	0.147	0.131	0.032
		67%	1.721	0.466	0.714	0.712	0.134	0.941	0.940	0.071	0.150	0.151	0.064
		50%	1.678	0.502	0.718	0.713	0.110	0.934	0.934	0.074	0.152	0.154	0.056
		Fivefold CV	1.759	0.421	0.711	0.709	0.075	0.948	0.944	0.037	0.147	0.149	0.039
		LOOCV	1.759	0.421	0.711	0.709	0.076	0.948	0.947	0.040	0.147	0.147	0.041
	2, 3	100%	2.153	0.674	0.726	0.752	0.089	0.846	0.868	0.076	0.202	0.178	0.039
		67%	2.174	0.778	0.722	0.712	0.151	0.844	0.842	0.120	0.204	0.210	0.074
		50%	2.138	0.811	0.725	0.717	0.132	0.839	0.838	0.116	0.206	0.210	0.065
		Fivefold CV	2.153	0.674	0.726	0.720	0.085	0.846	0.844	0.071	0.202	0.206	0.049
		LOOCV	2.153	0.674	0.726	0.720	0.093	0.846	0.846	0.081	0.202	0.204	0.057
	2, 2	100%	2.034	0.550	0.828	0.859	0.072	0.837	0.856	0.069	0.167	0.143	0.036
		67%	2.006	0.644	0.829	0.827	0.129	0.830	0.829	0.114	0.170	0.171	0.068
		50%	2.008	0.695	0.827	0.827	0.124	0.829	0.829	0.107	0.172	0.172	0.057
		Fivefold CV	2.034	0.550	0.828	0.828	0.070	0.837	0.830	0.066	0.167	0.170	0.046
		LOOCV	2.034	0.550	0.828	0.829	0.077	0.837	0.836	0.074	0.167	0.167	0.051
	3, 2	100%	1.862	0.699	0.843	0.876	0.076	0.727	0.752	0.085	0.226	0.198	0.042
		67%	1.875	0.760	0.840	0.841	0.134	0.728	0.730	0.128	0.228	0.226	0.073
		50%	1.878	0.870	0.835	0.833	0.132	0.726	0.726	0.121	0.230	0.231	0.062
		Fivefold CV	1.862	0.699	0.843	0.843	0.072	0.727	0.729	0.077	0.226	0.225	0.049
		LOOCV	1.862	0.699	0.843	0.844	0.080	0.727	0.732	0.085	0.226	0.223	0.055
	4, 1	100%	2.235	0.427	0.948	0.966	0.035	0.711	0.725	0.064	0.194	0.178	0.038
		67%	2.225	0.504	0.944	0.938	0.089	0.709	0.702	0.112	0.197	0.202	0.071
		50%	2.194	0.586	0.942	0.939	0.082	0.706	0.705	0.096	0.199	0.201	0.057
		Fivefold CV	2.235	0.427	0.948	0.945	0.040	0.711	0.707	0.064	0.194	0.197	0.043
		LOOCV	2.235	0.427	0.948	0.945	0.043	0.711	0.708	0.067	0.194	0.197	0.045
6	1, 4	100%	1.950	0.409	0.843	0.856	0.058	0.965	0.978	0.025	0.084	0.071	0.026
		67%	1.908	0.454	0.845	0.841	0.112	0.960	0.960	0.056	0.086	0.087	0.050
		50%	1.882	0.485	0.847	0.846	0.091	0.955	0.957	0.060	0.088	0.088	0.045
		Fivefold CV	1.950	0.409	0.843	0.843	0.059	0.965	0.961	0.029	0.084	0.086	0.032
		LOOCV	1.950	0.409	0.843	0.843	0.061	0.965	0.964	0.031	0.084	0.084	0.034
	2, 3	100%	2.792	0.623	0.852	0.877	0.056	0.909	0.926	0.047	0.114	0.093	0.029
		67%	2.790	0.710	0.851	0.848	0.111	0.906	0.907	0.088	0.116	0.116	0.060
		50%	2.741	0.781	0.853	0.855	0.096	0.899	0.899	0.085	0.119	0.119	0.050
		Fivefold CV	2.792	0.623	0.852	0.853	0.058	0.909	0.906	0.048	0.114	0.115	0.038
		LOOCV	2.792	0.623	0.852	0.852	0.062	0.909	0.908	0.053	0.114	0.114	0.043
	2, 2	100%	2.988	0.543	0.927	0.947	0.039	0.925	0.941	0.038	0.074	0.057	0.023
		67%	2.955	0.631	0.927	0.924	0.089	0.921	0.919	0.077	0.077	0.079	0.049
		50%	2.904	0.678	0.929	0.930	0.071	0.916	0.916	0.068	0.079	0.078	0.041
		Fivefold CV	2.988	0.543	0.927	0.926	0.043	0.925	0.923	0.038	0.074	0.075	0.031
		LOOCV	2.988	0.543	0.927	0.926	0.044	0.925	0.925	0.040	0.074	0.074	0.033
	3, 2	100%	3.210	0.634	0.908	0.933	0.047	0.852	0.871	0.055	0.125	0.104	0.031
		67%	3.182	0.716	0.908	0.911	0.098	0.849	0.849	0.098	0.128	0.125	0.061
		50%	3.133	0.820	0.908	0.910	0.090	0.843	0.843	0.091	0.131	0.129	0.051
		Fivefold CV	3.210	0.634	0.908	0.906	0.051	0.852	0.851	0.053	0.125	0.127	0.039
		LOOCV	3.210	0.634	0.908	0.908	0.054	0.852	0.853	0.058	0.125	0.124	0.043
	4, 1	100%	3.989	0.451	0.967	0.982	0.023	0.839	0.855	0.048	0.110	0.094	0.029
		67%	3.944	0.564	0.965	0.967	0.062	0.836	0.838	0.091	0.113	0.111	0.056
		50%	3.903	0.630	0.964	0.965	0.059	0.832	0.838	0.081	0.115	0.111	0.048
		Fivefold CV	3.989	0.451	0.967	0.964	0.032	0.839	0.842	0.049	0.110	0.109	0.034
		LOOCV	3.989	0.451	0.967	0.965	0.033	0.839	0.843	0.051	0.110	0.108	0.036

Abbreviation: LOOCV, leave-one-out cross-validation.

Table 5 Simulated means and standard deviations (σ) of the detected cutoff ($\hat{c}$), and of estimated sensitivity $\widehat{Se}$, specificity $\widehat{Sp}$, and misclassification rate ($\hat{\epsilon}$) with $n = 200$ and $p = 0.6$. *Se*, *Sp*, and ϵ indicate the true performances

Δ_μ	σ_1, σ_2	δ	$\hat{c}$	$\sigma_{\hat{c}}$	*Se*	$\widehat{Se}$	$\sigma_{\widehat{Se}}$	*Sp*	$\widehat{Sp}$	$\sigma_{\widehat{Sp}}$	ϵ	$\hat{\epsilon}$	$\sigma_{\hat{\epsilon}}$
4	1, 4	100%	1.740	0.342	0.713	0.723	0.050	0.950	0.963	0.028	0.192	0.181	0.029
		67%	1.721	0.384	0.715	0.715	0.081	0.946	0.946	0.057	0.193	0.193	0.051
		50%	1.710	0.427	0.715	0.716	0.071	0.942	0.942	0.058	0.194	0.193	0.042
		Fivefold CV	1.740	0.342	0.713	0.716	0.048	0.950	0.949	0.029	0.192	0.191	0.032
		LOOCV	1.740	0.342	0.713	0.716	0.051	0.950	0.950	0.032	0.192	0.190	0.034
	2, 3	100%	2.212	0.559	0.721	0.736	0.069	0.857	0.877	0.060	0.225	0.208	0.030
		67%	2.175	0.645	0.724	0.724	0.099	0.850	0.853	0.102	0.226	0.225	0.053
		50%	2.167	0.676	0.724	0.726	0.095	0.848	0.846	0.096	0.226	0.226	0.045
		Fivefold CV	2.212	0.559	0.721	0.727	0.060	0.857	0.851	0.057	0.225	0.224	0.035
		LOOCV	2.212	0.559	0.721	0.723	0.070	0.857	0.855	0.064	0.225	0.224	0.042
	2, 2	100%	1.960	0.436	0.840	0.851	0.053	0.831	0.852	0.057	0.163	0.149	0.025
		67%	1.958	0.501	0.839	0.836	0.083	0.829	0.829	0.095	0.165	0.167	0.048
		50%	1.952	0.567	0.838	0.836	0.082	0.826	0.829	0.093	0.167	0.167	0.041
		Fivefold CV	1.960	0.436	0.840	0.838	0.048	0.831	0.833	0.054	0.163	0.164	0.031
		LOOCV	1.960	0.436	0.840	0.838	0.054	0.831	0.834	0.060	0.163	0.163	0.035
	3, 2	100%	1.824	0.555	0.853	0.867	0.063	0.725	0.746	0.072	0.198	0.181	0.028
		67%	1.807	0.627	0.852	0.857	0.088	0.722	0.725	0.117	0.200	0.195	0.052
		50%	1.826	0.676	0.849	0.850	0.087	0.724	0.725	0.103	0.201	0.200	0.043
		Fivefold CV	1.824	0.555	0.853	0.853	0.058	0.725	0.726	0.067	0.198	0.198	0.034
		LOOCV	1.824	0.555	0.853	0.855	0.066	0.725	0.726	0.073	0.198	0.197	0.040
	4, 1	100%	2.217	0.350	0.954	0.963	0.028	0.710	0.723	0.053	0.144	0.133	0.024
		67%	2.195	0.410	0.953	0.955	0.047	0.707	0.712	0.094	0.145	0.142	0.045
		50%	2.186	0.462	0.951	0.952	0.050	0.706	0.706	0.083	0.147	0.146	0.037
		Fivefold CV	2.217	0.350	0.954	0.955	0.027	0.710	0.709	0.053	0.144	0.143	0.027
		LOOCV	2.217	0.350	0.954	0.954	0.032	0.710	0.712	0.055	0.144	0.142	0.030
6	1, 4	100%	1.964	0.358	0.843	0.848	0.036	0.968	0.978	0.021	0.107	0.100	0.021
		67%	1.947	0.389	0.843	0.839	0.063	0.965	0.966	0.047	0.108	0.110	0.039
		50%	1.939	0.438	0.844	0.843	0.054	0.963	0.964	0.042	0.109	0.109	0.032
		Fivefold CV	1.964	0.358	0.843	0.842	0.036	0.968	0.966	0.023	0.107	0.108	0.025
		LOOCV	1.964	0.358	0.843	0.842	0.038	0.968	0.967	0.024	0.107	0.108	0.026
	2, 3	100%	2.783	0.506	0.855	0.865	0.044	0.911	0.926	0.039	0.123	0.111	0.022
		67%	2.757	0.588	0.856	0.853	0.070	0.907	0.906	0.076	0.124	0.126	0.043
		50%	2.751	0.633	0.855	0.855	0.064	0.905	0.905	0.068	0.125	0.125	0.034
		Fivefold CV	2.783	0.506	0.855	0.855	0.041	0.911	0.910	0.039	0.123	0.123	0.026
		LOOCV	2.783	0.506	0.855	0.855	0.045	0.911	0.912	0.042	0.123	0.122	0.029
	2, 2	100%	2.971	0.450	0.930	0.939	0.031	0.926	0.940	0.032	0.071	0.061	0.018
		67%	2.949	0.497	0.931	0.931	0.052	0.924	0.925	0.061	0.072	0.072	0.033
		50%	2.958	0.541	0.929	0.930	0.049	0.923	0.925	0.057	0.073	0.072	0.028
		Fivefold CV	2.971	0.450	0.930	0.929	0.030	0.926	0.926	0.032	0.071	0.072	0.022
		LOOCV	2.971	0.450	0.930	0.930	0.033	0.926	0.927	0.035	0.071	0.071	0.025
	3, 2	100%	3.202	0.515	0.912	0.923	0.039	0.854	0.870	0.046	0.111	0.098	0.022
		67%	3.182	0.594	0.912	0.914	0.062	0.851	0.853	0.085	0.113	0.110	0.040
		50%	3.192	0.638	0.910	0.911	0.060	0.851	0.853	0.072	0.114	0.112	0.035
		Fivefold CV	3.202	0.515	0.912	0.913	0.036	0.854	0.853	0.046	0.111	0.111	0.027
		LOOCV	3.202	0.515	0.912	0.914	0.041	0.854	0.854	0.049	0.111	0.110	0.031
	4, 1	100%	3.999	0.362	0.970	0.977	0.020	0.840	0.852	0.043	0.082	0.073	0.019
		67%	3.988	0.425	0.969	0.969	0.040	0.839	0.839	0.079	0.083	0.084	0.037
		50%	3.977	0.455	0.968	0.967	0.038	0.838	0.839	0.064	0.084	0.084	0.030
		Fivefold CV	3.999	0.362	0.970	0.970	0.020	0.840	0.840	0.043	0.082	0.082	0.022
		LOOCV	3.999	0.362	0.970	0.971	0.022	0.840	0.841	0.044	0.082	0.081	0.023

Abbreviation: LOOCV, leave-one-out cross-validation.

Table 6 Simulated means and standard deviations (σ) of the detected cutoff ($\hat{c}$), and of estimated sensitivity $\widehat{Se}$, specificity $\widehat{Sp}$ and misclassification rate ($\hat{\epsilon}$) with $n = 200$ and $p = 0.4$. *Se*, *Sp*, and ϵ indicate the true performances

Δ_μ	σ_1, σ_2	δ	$\hat{c}$	$\sigma_{\hat{c}}$	*Se*	$\widehat{Se}$	$\sigma_{\widehat{Se}}$	*Sp*	$\widehat{Sp}$	$\sigma_{\widehat{Sp}}$	ϵ	$\hat{\epsilon}$	$\sigma_{\hat{\epsilon}}$
4	1, 4	100%	1.783	0.353	0.710	0.721	0.055	0.954	0.962	0.030	0.144	0.134	0.024
		67%	1.774	0.386	0.710	0.710	0.094	0.951	0.952	0.049	0.145	0.146	0.044
		50%	1.763	0.424	0.711	0.709	0.079	0.948	0.946	0.052	0.147	0.149	0.039
		Fivefold CV	1.783	0.353	0.710	0.709	0.053	0.954	0.953	0.028	0.144	0.144	0.027
		LOOCV	1.783	0.353	0.710	0.708	0.056	0.954	0.955	0.031	0.144	0.144	0.030
	2, 3	100%	2.147	0.562	0.728	0.748	0.071	0.849	0.863	0.062	0.199	0.183	0.028
		67%	2.145	0.633	0.727	0.726	0.110	0.847	0.846	0.092	0.201	0.203	0.052
		50%	2.151	0.689	0.726	0.724	0.104	0.845	0.846	0.092	0.202	0.203	0.046
		Fivefold CV	2.147	0.562	0.728	0.726	0.066	0.849	0.849	0.055	0.199	0.200	0.034
		LOOCV	2.147	0.562	0.728	0.726	0.073	0.849	0.848	0.063	0.199	0.201	0.040
	2, 2	100%	2.014	0.463	0.833	0.851	0.059	0.837	0.851	0.057	0.165	0.149	0.026
		67%	2.043	0.504	0.829	0.828	0.097	0.839	0.838	0.085	0.165	0.166	0.048
		50%	2.017	0.578	0.830	0.830	0.095	0.834	0.834	0.085	0.168	0.167	0.041
		Fivefold CV	2.014	0.463	0.833	0.831	0.054	0.837	0.837	0.050	0.165	0.165	0.031
		LOOCV	2.014	0.463	0.833	0.831	0.061	0.837	0.840	0.058	0.165	0.163	0.037
	3, 2	100%	1.838	0.545	0.852	0.873	0.059	0.727	0.740	0.066	0.223	0.207	0.030
		67%	1.860	0.599	0.847	0.848	0.097	0.728	0.728	0.095	0.224	0.225	0.053
		50%	1.837	0.663	0.848	0.848	0.094	0.725	0.726	0.091	0.226	0.225	0.045
		Fivefold CV	1.838	0.545	0.852	0.849	0.055	0.727	0.727	0.060	0.223	0.224	0.035
		LOOCV	1.838	0.545	0.852	0.851	0.062	0.727	0.726	0.068	0.223	0.224	0.042
	4, 1	100%	2.242	0.336	0.952	0.964	0.029	0.712	0.722	0.048	0.192	0.181	0.027
		67%	2.229	0.397	0.950	0.950	0.060	0.710	0.712	0.079	0.194	0.193	0.047
		50%	2.225	0.424	0.949	0.949	0.051	0.710	0.715	0.068	0.194	0.192	0.040
		Fivefold CV	2.242	0.336	0.952	0.951	0.030	0.712	0.713	0.046	0.192	0.192	0.030
		LOOCV	2.242	0.336	0.952	0.952	0.032	0.712	0.713	0.048	0.192	0.191	0.031
6	1, 4	100%	1.977	0.345	0.842	0.849	0.043	0.969	0.975	0.021	0.082	0.075	0.019
		67%	1.975	0.394	0.842	0.837	0.074	0.967	0.966	0.039	0.083	0.086	0.034
		50%	1.943	0.419	0.843	0.843	0.065	0.964	0.963	0.040	0.084	0.085	0.030
		Fivefold CV	1.977	0.345	0.842	0.840	0.043	0.969	0.967	0.022	0.082	0.084	0.022
		LOOCV	1.977	0.345	0.842	0.840	0.044	0.969	0.969	0.024	0.082	0.083	0.023
	2, 3	100%	2.809	0.511	0.853	0.869	0.047	0.913	0.923	0.038	0.111	0.099	0.021
		67%	2.820	0.591	0.851	0.853	0.084	0.912	0.913	0.064	0.113	0.112	0.042
		50%	2.763	0.651	0.854	0.857	0.076	0.905	0.905	0.066	0.115	0.114	0.037
		Fivefold CV	2.809	0.511	0.853	0.853	0.045	0.913	0.913	0.035	0.111	0.111	0.026
		LOOCV	2.809	0.511	0.853	0.853	0.049	0.913	0.913	0.041	0.111	0.111	0.031
	2, 2	100%	3.013	0.419	0.928	0.940	0.031	0.930	0.938	0.030	0.071	0.061	0.017
		67%	3.008	0.475	0.927	0.926	0.063	0.928	0.927	0.054	0.072	0.074	0.035
		50%	2.967	0.542	0.928	0.929	0.055	0.924	0.923	0.054	0.074	0.075	0.029
		Fivefold CV	3.013	0.419	0.928	0.926	0.033	0.930	0.929	0.029	0.071	0.072	0.022
		LOOCV	3.013	0.419	0.928	0.927	0.034	0.930	0.930	0.032	0.071	0.071	0.024
	3, 2	100%	3.245	0.506	0.909	0.924	0.040	0.857	0.865	0.044	0.122	0.112	0.023
		67%	3.235	0.585	0.908	0.909	0.073	0.855	0.852	0.071	0.124	0.125	0.042
		50%	3.244	0.658	0.905	0.905	0.072	0.855	0.854	0.067	0.125	0.126	0.036
		Fivefold CV	3.245	0.506	0.909	0.908	0.039	0.857	0.853	0.042	0.122	0.125	0.028
		LOOCV	3.245	0.506	0.909	0.908	0.044	0.857	0.854	0.046	0.122	0.124	0.032
	4, 1	100%	4.043	0.360	0.967	0.978	0.020	0.843	0.849	0.036	0.107	0.099	0.021
		67%	4.035	0.427	0.965	0.966	0.047	0.842	0.842	0.063	0.109	0.109	0.039
		50%	4.019	0.468	0.965	0.965	0.042	0.841	0.839	0.056	0.110	0.111	0.032
		Fivefold CV	4.043	0.360	0.967	0.967	0.022	0.843	0.842	0.036	0.107	0.108	0.024
		LOOCV	4.043	0.360	0.967	0.968	0.024	0.843	0.843	0.037	0.107	0.107	0.025

Abbreviation: LOOCV, leave-one-out cross-validation.

the variance of $\hat{c}$ increases as the training percentage decreases; as a consequence, the true performance indicators appear to get a little worse as δ decreases, since there is greater probability that the estimated cutoff assumes values far from the aforementioned intersection point. As expected, the true performances improve as the sample size and the distance between the distributions of diseased and nondiseased individuals (δ_μ) increase. Similarly, the true sensitivity increases as the variability of the classifier decreases among diseased individuals (σ_2), just as the true specificity increases as σ_1 decreases.

Quite evidently, assessing the test performance on the training set (δ = 100%) leads to an overestimation of the true sensitivity and specificity, and consequently an underestimation of the true misclassification rate in all scenarios. In particular, the bias is higher when estimating sensitivity with a low prevalence (on average less fewer diseased individuals are sampled) and when estimating specificity with a high prevalence (on average less fewer nondiseased individuals are sampled). The bias is substantially not affected by increasing δ_μ or changing σ_1 and σ_2, while it decreases as the sample size increases.

When a single split or CV is performed, $\widehat{Se}$, $\widehat{Sp}$, and $\hat{\epsilon}$ are unbiased estimators of the true performance indicators in all scenarios. However, in the case of a single split, $\widehat{Se}$, $\widehat{Sp}$, and $\hat{\epsilon}$ are affected by higher variability, especially with smaller sample sizes, higher population variances of the classifier, and a higher training proportion (δ = 67%). Conversely, CV performs the best in all scenarios.

For each scenario, ►Figs. 2, 3, and 4 represent the true misclassification rates for each candidate cutoffs c_j (black curve) and the empirical (simulated) means of the error rates estimated as the different c_j are selected as the optimal cutoff using the different training percentages (δ = 100%, 67%, 50%) and CV. For all scenarios, the lines in green (δ = 67%), blue (δ = 50%), light blue (fivefold CV), and magenta (LOOCV) lie on the black curve [the true $\epsilon(c_j)$], meaning that, once a given $\hat{c}_\delta$ is obtained, the true classification error is correctly estimated. By contrast, the red line (δ = 100%) lies below the black line, indicating an underestimation of the true classification error. Since the red and the black curves appear to be parallel, the bias is approximately the same regardless of the optimal cutoff estimated. ►Figures 3 and 4 show that the bias reduces as the sample size n increases from 50 to 100 or 200 individuals.

It is worth noting that the means of the simulated error estimates get more wiggly for the most external cutoffs; this is due to the low probability of detecting such cutoffs as optimal over the 1,000 replicates. Due to the lower variance of $\hat{c}_{100\%}$ highlighted in ►Tables 1–5 to 6, the lines in red, light blue, and magenta are shorter than the others. Since the most external cutoffs are associated with higher true misclassification rates, this would explain the somewhat better true overall performances highlighted in ►Tables 1–5 to 6 when δ = 100%.

Validation of the Italian CARATkids Questionnaire

Due to the moderate sample size ($n = 112$), the low estimated prevalence (17.86%, i.e., C-ACT total score ≤19 in 20/112 children) and the sample estimates $\hat{\Delta}_\mu = 5.4$, $\hat{\sigma}_1 = 2.61$ and $\hat{\sigma}_2 = 3$, δ = 50% were used as it should provide lower uncertainty of $\hat{\epsilon}_\gamma$ with a negligible loss of true performance (►Table 3); CV was also performed. Due to the random group assignment, no significant differences were found between the training set and the test set (►Table 7).

At T0 and T1 CARATkids score (increasing for decreasing disease control) showed significant intra-visit correlation (Spearman's rho in the training set) with C-ACT (increasing for increasing disease control): $\rho = -0.65$ (p-value < 0.001) at T0 and $\rho = -0.61$ (p-value < 0.001) at T1. The inter-visit correlation was weaker but statistically significant: $\rho = -0.52$ (p-value < 0.001).

►Figure 5 depicts the ROC curve evaluated on the training set. The overall CARATkids score showed a good ability to predict a C-ACT score ≤19: the AUC was 0.91 (95% confidence interval [CI]: 0.82–0.99), and the optimal threshold was $\hat{c}_{50\%} = 5.5$, associated with a 91% sensitivity, a 76% specificity, and a 21% misclassification rate on the same training sample. The rule obtained (CARATkids > 5.5 for identifying uncontrolled asthma) was therefore tested on the test set. It yielded an estimated sensitivity of 78% and an estimated specificity of 77%. The estimated misclassification rate was 0.23, with 95% CI: 0.11 to 0.34 (usual CI for a proportion), indicating an acceptable discriminant validity. The optimal threshold was the same when estimated on the whole sample ($\hat{c}_{100\%} = 5.5$). In this case, fivefold CV yielded an estimated sensitivity of 75%, an estimated specificity of 78%, and a misclassification rate of 0.22, with 95% CI: 0.14 to 0.30. LOOCV yielded an estimated sensitivity of 75%, an estimated specificity of 76%, and a misclassification rate of 0.24, with 95% CI: 0.16 to 0.32.

Discussion

In this article, two common approaches for estimating and validating simple classification rules have been described in the context of ROC analysis: the focus has been on the inferential implications of splitting (once or repeatedly) or not the dataset into training and test sets. In fact, though well addressed in other areas, this topic still appears to be overlooked among medical researchers dealing with clinical data.

A simulation study showed that splitting the sample into training and test sets allows unbiased estimation of sensitivity, specificity, and misclassification rate. A single split of the sample produces more fluctuating estimates (higher variance) for both the cutoff and the performance indicators. The problem of higher variance is of some importance, and has raised questions about the ideal splitting proportion, especially when the sample size n is small (<100). Moreover, this approach slightly reduces the true performance of the classification method. This aspect should not discourage the use of a test set, even when the total sample size is $n \approx 50$, but rather, it may suggest reducing the training proportion. Indeed, while using a smaller training set does not appear to affect true performances considerably (►Tables 1 and 2), it may help to reduce the variance of their estimates by increasing the number of individuals in the test set. In general, very small sample sizes ($n < 50$) together with

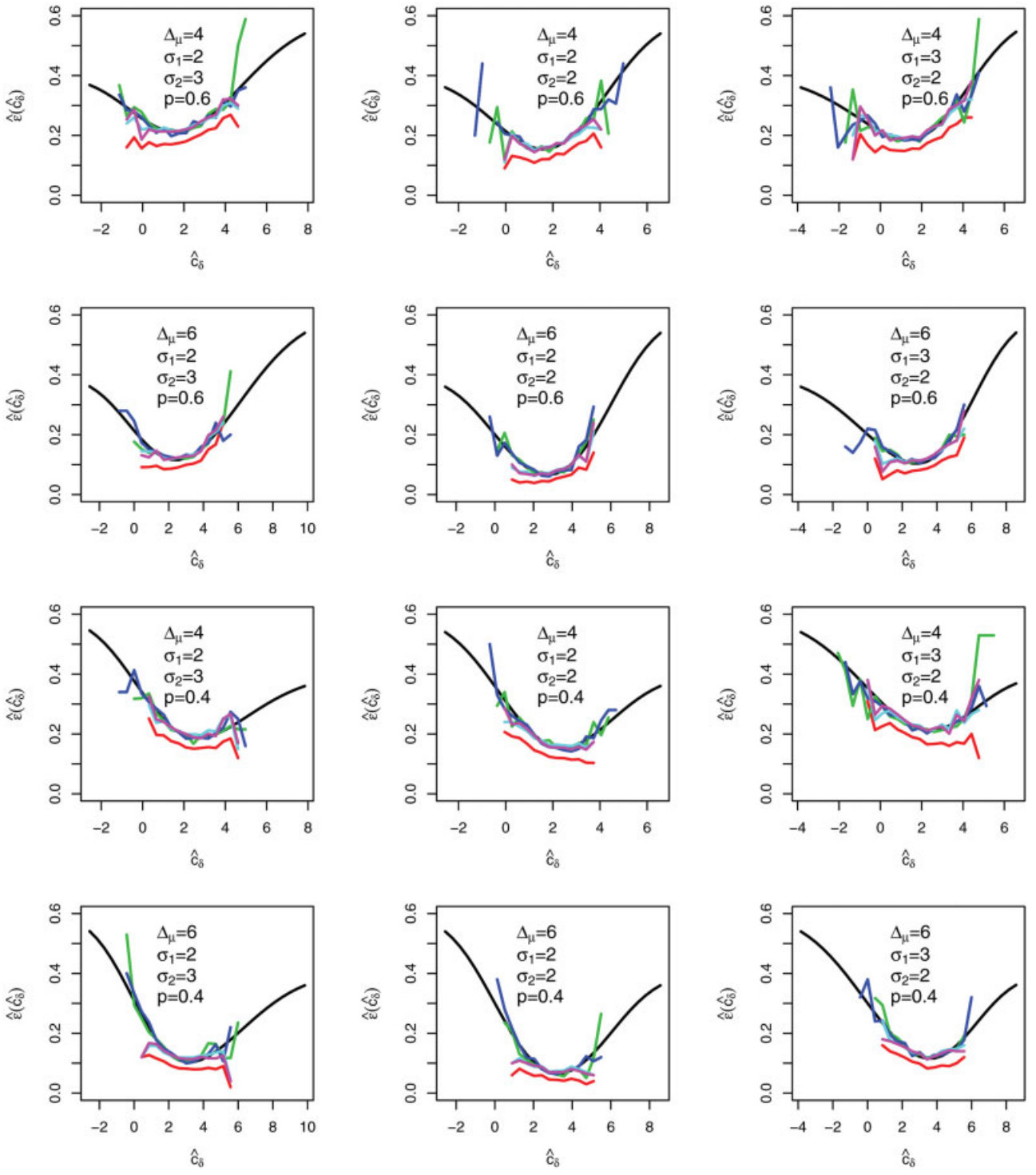

Fig. 2 Misclassification rates given the optimal cutoff, for $n = 50$ (►**Tables 1** and **2**). Black line: true error. Red line: mean error estimated using the whole sample. Green line: mean error estimated with 67% training and 33% test. Blue line: mean error estimated with 50% training and 50% test. Light blue line: mean error estimated with fivefold cross-validation. Magenta line: mean error estimated with leave one out cross-validation.

very small (or very high) expected prevalences ($p < 0.20$) would discourage the use of ROC analysis in simple random samples (due to the small proportion of diseased individuals in the sample). In such situations, it would be preferable to use CV, or alternatively, stratified sampling (to fix the number of diseased and nondiseased individuals), or more advanced methods like SMOTE.[31,32] The issue of the higher variance after a single sample split can be quantified by deriving standard CIs for proportions; conversely, the bias of 6, 7, and 8 does not allow computation of valid CIs.

For the sake of brevity, the simulation study was limited to normal distribution only, and moreover the properties of a

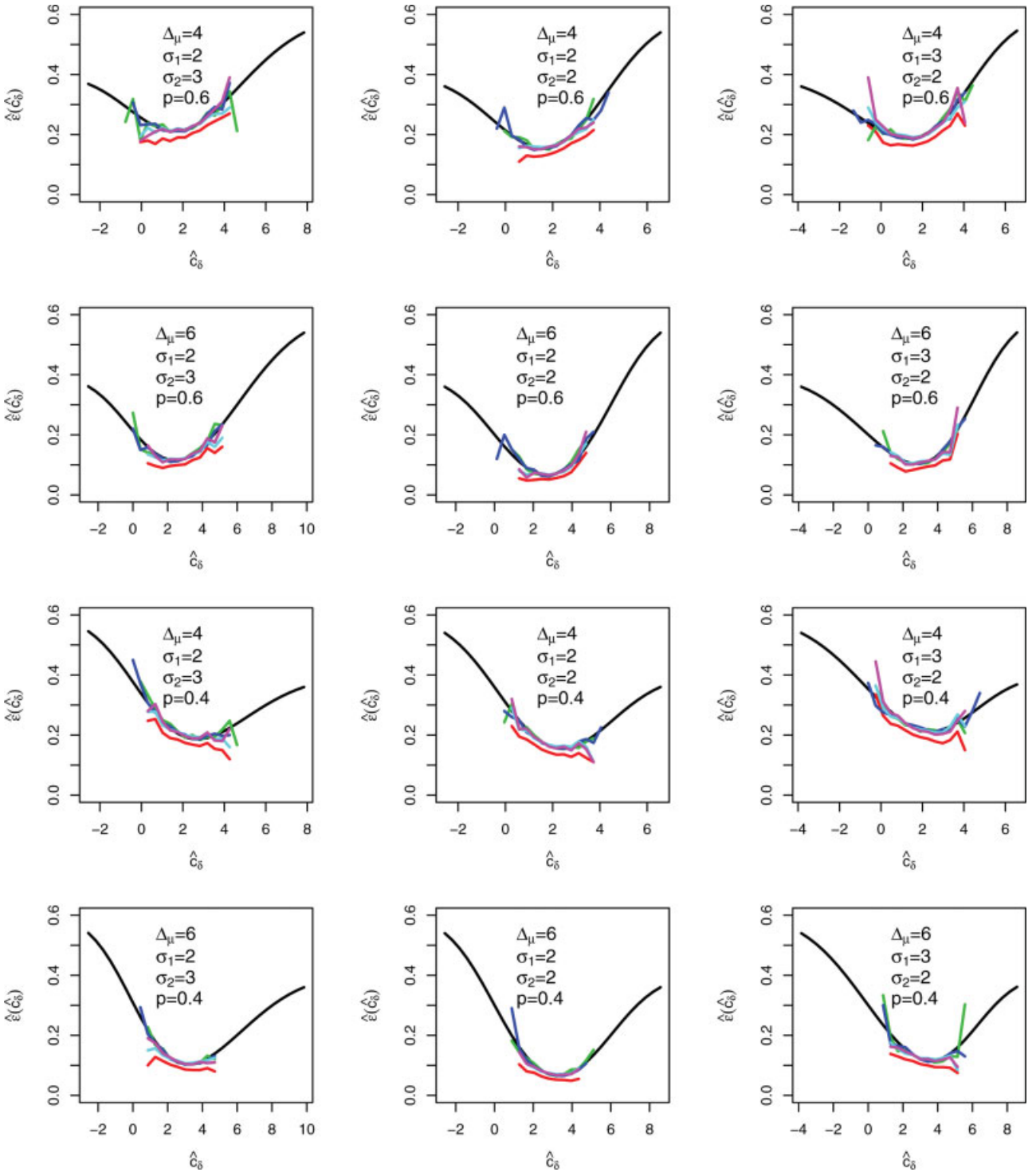

Fig. 3 Misclassification rates given the optimal cutoff, for $n = 100$ (►**Tables 3** and **4**). Black line: true error. Red line: mean error estimated using the whole sample. Green line: mean error estimated with 67% training and 33% test. Blue line: mean error estimated with 50% training and 50% test. Light blue line: mean error estimated with fivefold cross-validation. Magenta line: mean error estimated with leave one out cross-validation.

single utility function (Youden's index) were assessed. Despite representing a possible limitation of the present study, it might be speculated that the main findings should not be much influenced by the choice of simulation setting. Indeed, the main goal of the article was simply to show, empirically, the usefulness of using independent test samples, a topic that has been well studied in other contexts, but overlooked in medical literature about ROC analysis.

The motivating dataset of 112 Italian outpatient children represented a case in point. By now the CARATkids questionnaire for assessing disease control in children with asthma and rhinitis has been validated in three different

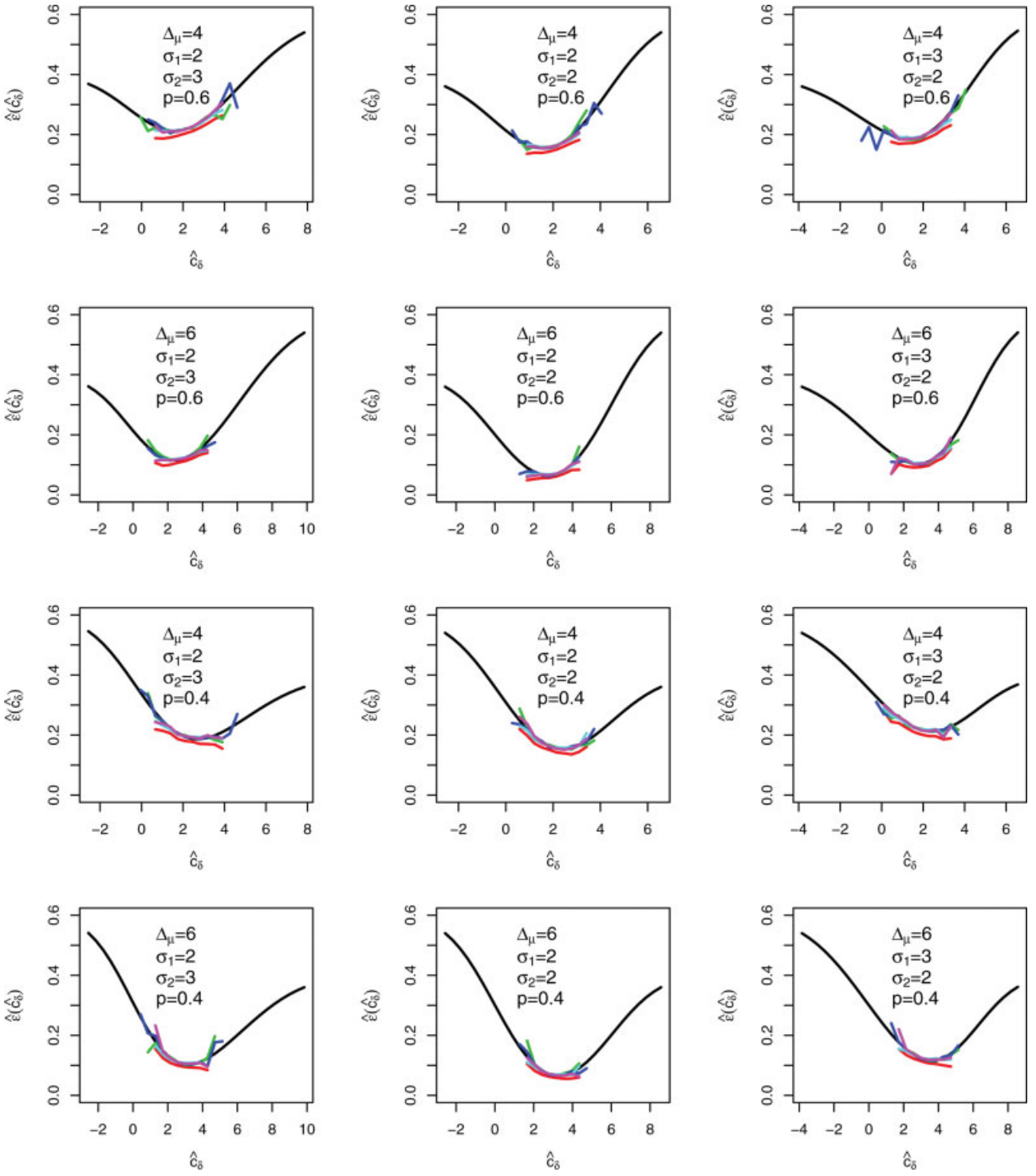

Fig. 4 Misclassification rates given the optimal cutoff, for $n = 200$ (►**Tables 5** and **6**). Black line: true error. Red line: mean error estimated using the whole sample. Green line: mean error estimated with 67% training and 33% test. Blue line: mean error estimated with 50% training and 50% test. Light blue line: mean error estimated with fivefold cross-validation. Magenta line: mean error estimated with leave one out cross-validation.

languages.[27–29] However, none of these studies performed a ROC analysis with a sample split.

The intravisit (cross-sectional validity) and intervisit (longitudinal validity) correlations with C-ACT were similar to those found in all the previous aforementioned studies.[27–29] Concerning prediction of asthma control (discriminant validity), it is worth noting that the ROC analysis using a 50% sample split showed high sensitivity in the training set as in the previous studies,[27,28] while it was of more moderate intensity in the independent test set. This might suggest that the CARATkids misclassification rate may have been underestimated in previous assessments. Overall, the results

Table 7 Patient characteristics

	Training set $n = 56$	Test set $n = 56$	All $n = 112$	*p*-Value
Gender, *n* (%)				0.69
Male	36 (64.29%)	39 (69.64%)	75 (66.96%)	
Female	20 (35.71%)	17 (30.36%)	37 (33.04%)	
Age, mean (SD)	8.29 (1.69)	8.32 (1.57)	8.3 (1.63)	0.91
Height, mean (SD)	131.84 (11.87)	132.60 (9.45)	132.72 (10.68)	0.91
Asthma severity, *n* (%)				0.33
Intermittent	13 (23.21%)	8 (14.29%)	21 (18.75%)	
Persistent	43 (76.79%)	48 (85.71%)	91 (81.25%)	
Rhinitis severity, *n* (%)				0.93
Intermittent	25 (44.64%)	24 (42.86%)	49 (43.75%)	
Persistent	31 (55.36%)	32 (57.14%)	63 (56.25%)	
CARATkids, mean (SD)	4.8 (3.35)	4.54 (3.34)	4.67 (3.33)	0.71
C-ACT, mean (SD)	22.96 (3.45)	22.88 (3.58)	22.92 (3.5)	0.89
C-ACT$\leq$19, *n* (%)	11 (19.64%)	9 (16.07%)	20 (17.86%)	

Abbreviations: C-ACT, the Childhood Asthma Control Test; SD, standard deviation.

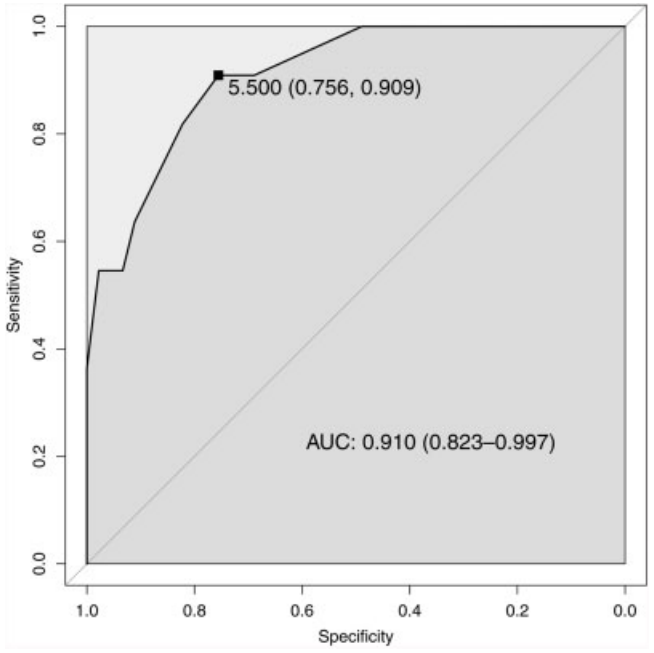

Fig. 5 Receiver operating characteristic curve for the ability of CARATkids to predict a the Childhood Asthma Control Test $\leq$19. AUC, area under the curve.

highlight the clinical validity of the Italian version of the CARATkids questionnaire for assessment of disease control in children with asthma and rhinitis.

Conclusions

Medical researchers dealing with clinical data should carefully consider the usefulness of splitting, when possible, the study sample into a training sample (where the optimal test is derived) and a test sample (where performance or error rates are estimated) when performing a ROC analysis, or alternatively using CV. The results of the present study support the use of CARATkids as a valid questionnaire to assess disease control in Italian children with asthma and rhinitis; its use helps to optimize simultaneous evaluation of allergic rhinitis and asthma, contributing to more comprehensive health care in children.

Note

All data and materials are available upon request.

Funding
None.

Conflict of Interest
None declared.

References

1 Hajian-Tilaki K. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. Caspian J Intern Med 2013;4(02):627–635
2 Fawcett T. An introduction to roc analysis. Pattern Recognit Lett 2006;27(08):861–874
3 Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39(04):561–577
4 Zhou X-H, McClish DK, Obuchowski NA. Statistical Methods in Diagnostic Medicine, volume 569. New York: John Wiley & Sons; 2009
5 Chinellato I, Piazza M, Sandri M, et al. Evaluation of association between exercise-induced bronchoconstriction and childhood asthma control test questionnaire scores in children. Pediatr Pulmonol 2012;47(03):226–232
6 Voorend-van Bergen S, Vaessen-Verberne AA, Landstra AM, et al. Monitoring childhood asthma: web-based diaries and the asthma control test. J Allergy Clin Immunol 2014;133(06):1599–605.e2
7 Behan L, Dimitrov BD, Kuehni CE, et al. PICADAR: a diagnostic predictive tool for primary ciliary dyskinesia. Eur Respir J 2016;47 (04):1103–1112
8 Takemura M, Nishio M, Fukumitsu K, et al. Optimal cut-off value and clinical usefulness of the Adherence Starts with Knowledge-12 in patients with asthma taking inhaled corticosteroids. J Thorac Dis 2017;9(08):2350–2359
9 Coffin M, Sukhatme S. Receiver operating characteristic studies and measurement errors. Biometrics 1997;53(03):823–837
10 Youden WJ. Index for rating diagnostic tests. Cancer 1950;3(01): 32–35
11 Perkins NJ, Schisterman EF. The inconsistency of “optimal” cut-points obtained using two criteria based on the receiver operating characteristic curve. Am J Epidemiol 2006;163(07):670–675
12 Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform 2005;38(05):404–415
13 McNeil BJ, Adelstein SJ. Determining the value of diagnostic and screening tests. J Nucl Med 1976;17(06):439–448
14 Fawcett T. Roc graphs: notes and practical considerations for researchers. Mach Learn 2004;31(01):1–38
15 Zhong M. An Analysis of Misclassification Rates for Decision Trees. Orlando: University of Central Florida; 2007
16 Westerhuis JA, Hoefsloot HCJ, Smit S, et al. Assessment of PLSDA cross validation. Metabolomics 2008;4(01):81–89
17 Zou KH, Warfield SK, Fielding JR, et al. Statistical validation based on parametric receiver operating characteristic analysis of continuous classification data. Acad Radiol 2003;10(12):1359–1368
18 Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Stat Med 1998;17(09):1033–1053
19 Hui SL, Walter SD. Estimating the error rates of diagnostic tests. Biometrics 1980;36(01):167–171
20 Dobbin KK, Simon RM. Optimally splitting cases for training and testing high dimensional classifiers. BMC Med Genomics 2011;4 (01):31
21 James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning, volume 112. New York: Springer; 2013
22 Ounpraseuth S, Lensing SY, Spencer HJ, Kodell RL. Estimating misclassification error: a closer look at cross-validation based methods. BMC Res Notes 2012;5(01):656
23 Brereton RG. Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data. TrAC Trends in Analytical Chemistry 2006;25(11):1103–1111
24 Liu AH, Zeiger R, Sorkness C, et al. Development and cross-sectional validation of the Childhood Asthma Control Test. J Allergy Clin Immunol 2007;119(04):817–825
25 Fernandes PH, Matsumoto F, Solé D, Wandalsen GF. Translation into Portuguese and validation of the rhinitis control assessment test (RCAT) questionnaire. Rev Bras Otorrinolaringol (Engl Ed) 2016;82(06):674–679
26 Meltzer EO, Schatz M, Nathan R, Garris C, Stanford RH, Kosinski M. Reliability, validity, and responsiveness of the Rhinitis Control Assessment Test in patients with rhinitis. J Allergy Clin Immunol 2013;131(02):379–386
27 Amaral R, Carneiro AC, Wandalsen G, Fonseca JA, Sole D. Control of allergic rhinitis and asthma test for children (CARATKids): validation in brazil and cutoff values. Ann Allergy Asthma Immunol 2017;118(05):551–556.e2
28 Linhares DV, da Fonseca JA, Borrego LM, et al; CARATKids study group. Validation of control of allergic rhinitis and asthma test for children (CARATKids)–a prospective multicenter study. Pediatr Allergy Immunol 2014;25(02):173–179
29 Emons JA, Flokstra-de Blok BM, Jong C, et al. Use of the control of allergic rhinitis and asthma test (CARATKids) in children and adolescents; validation in Dutch. Pediatr Allergy Immunol 2017; 28(02):185–190
30 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(03):837–845
31 Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. J Artif Intell Res 2002; 16:321–357
32 Adler W, Gefeller O, Gul A, Horn FK, Khan Z, Lausen B. Ensemble pruning for glaucoma detection in an unbalanced data set. Methods Inf Med 2016;55(06):557–563