REVSTAT – Statistical Journal
Volume 0, Number 0, Month 0000, 000–000
https://doi.org/00.00000/revstat.v00i0.000

An Estimator for the Hypervolume Under the ROC Manifold

Authors: Elena Nardi ✉
– IRCCS Azienda Ospedaliero-Universitaria di Bologna, Italy
elena.nardi@aosp.bo.it

Received: Month 0000 Revised: Month 0000 Accepted: Month 0000

Abstract:

- Measuring the diagnostic accuracy of biomedical tests is crucial in clinical research for both diagnosis and prognosis. For binary classification, ROC analysis and the area under the curve (AUC) are widely used. However, many real-world problems involve multiple classes, requiring extensions like the ROC manifold and the hypervolume under the manifold (HUM). This study develops a new estimator for assessing a continuous diagnostic marker's accuracy in multi-class settings, deriving its analytical form and variance. Simulation studies compare its performance with existing methods, and an empirical application is presented.

Keywords:

- *Receiver operating characteristic surface; hypervolume under ROC manifold; proportional hazards; classification accuracy, colorectal cancer; diagnostic medicine.*

AMS Subject Classification:

- 62P10, 62F10, 62H30

✉ Corresponding author

1. INTRODUCTION

In clinical studies, the accurate diagnosis of a patient's condition is crucial for appropriate treatment and evaluating prognosis. Thus, before implementing a new test, quantifying how well the medical test discriminates among different statuses is critical. Although the present work focuses mainly on clinical research, these problems are among the general classification issues that arise in almost all fields of scientific and social research. The procedures considered in this work could be generalized to any classification procedure that assigns a subject or an object to a class on the basis of the information observed.

To depict the quality of a diagnostic marker or a diagnostic test in a supervised classification problem, receiver operating characteristic (ROC) curve analysis plays a prominent role. This analysis was introduced in the second half of the last century in a two-class classification problem. It consists of a graphical representation of the relationship between the *sensitivity* and *specificity* of a test as the cut-off of the test varies. At each value of the decision-making threshold, the curve depicts the trade-off between the true positive rate (sensitivity) and false positive rate (1-specificity).

The area under the curve (AUC), is perhaps the most frequently used summary measure of the information reported in the ROC curve. It is a global measure of the performance of a diagnostic marker in discriminating between two statuses. An alternative interpretation refers to the measure of separability between the statistical distributions of the diagnostic test in the two populations Hand (2001).

The ROC curve and AUC remain major instruments used in the evaluation of a twofold classifier. However, many real situations in diagnostic decisions are not limited to a binary choice. Examples include staging the level of an illness or classifying an individual at low risk, moderate risk or high risk for a certain pathology Lusted (1960); Pepe (2003). To address this complicated m-class classification problem, in the past century, ROC analysis has been focused on deriving suitable generalizations of the curve. The ROC surface was introduced to address three-class issues, and the ROC manifold was introduced for when more than three classes are considered. Consequently, the notion of the AUC has been extended to the volume under the surface (VUS) and, in more complex situations with more than three classes, to the hypervolume under the manifold (HUM), Scurfield (1996, 1998); Mossman (1999). From a statistical perspective, theoretical inferential studies on generalized ROC analysis appeared only at the beginning of 2000. Furthermore, since those first works, few theoretical and empirical contributions have been reported, thus leaving the four-class classification issue as an area that still offers ample opportunities for further research Dreiseitl et al. (2000); Nakas and Yiannoutsos (2004); Kang and Tian (2013); Li and Fine (2008); Xiong et al. (2006); Gönen and Heller (2010); Nze Ossima et al. (2015); Jialiang Li and Pencina (2017); Waegeman et al. (2008); Feng et al. (2023). This study is dedicated to addressing this specific topic. It focuses on classifying subjects within an m-class framework. In such situations, a frequently used approach to ROC analysis consists of reducing the dimensionality of the problem by using a pairwise two-class ROC curve. Despite its simplicity, the standard approach has the drawback of considering a subset of the entire sample, thus potentially ignoring hidden patterns detectable only through a detailed analysis of the entire sample. The aim of this work is thus to propose a new method and to evaluate its performance in terms of computational

effort. The proposed HUM estimator takes into account the entire information of the sample and, at the same time, overcomes the computational burden; furthermore, being based on a well known estimation framework, it is very easy to implement. The article is organized as follows. The next Section is dedicated to the presentation of the statistical methodology proposed; specifically, we show how to derive an estimator of the hypervolume under the ROC surface in a m-class framework. Moreover, we derive the analytical form of the variance of the estimator. In the third section, simulation exercises in a four-class framework are presented under different data generating processes, in each scenario our method is compared with other three alternatives reported in the literature. In section four the methodology is evaluated through an empirical application. Finally, in the fifth section, some issues and potential future research are discussed.

2. Methods

Herein, a new HUM estimator, called HUM_{LM}, is proposed. It is a generalization of an approach proposed for the dichotomous framework Gönen and Heller (2010), generalized to estimate the volume under the ROC surface in a three-group classification framework Nze Ossima et al. (2015). Similarly, the analytical formula for the m-category HUM estimator is derived together with the calculation of its variance in the particular case of four classes. As discussed in the following sections, this represents a novelty in the literature. To derive our estimator, we assume the proportional hazard specification of the continuous marker results, and in the next section, we briefly revise these conditions. This assumption is also called the Lehmann assumption.

2.1. Lehmann assumption

Suppose we use a diagnostic test with continuous values to distinguish among a number m of classes or degree of disease. Let $X_1, X_2, X_j, \ldots, X_m$, be the continuous variables of the test result for subjects from class 1 to m with $j < m$. Moreover, let D be an ordinal categorical variable with values from 1 to m, and indicating the class for each subject. Suppose, further, that the test results for class 1, $X_{1,i}$, with $i = 1, 2, \ldots, n_1$ and n_1 as sample size of class 1 and are i.i.d., and the same for all the other classes. Moreover, let S_1, S_j, S_m indicate the corresponding survival functions. The survival distributions are assumed to have the family of Lehmann alternatives Lehmann (1953), i.e.:

$$S_2(x) = S_1(x)^{\theta_1}, \quad 0 < \theta_1 \leq 1 \tag{2.1}$$

$$S_j(x) = S_{j-1}(x)^{\theta_{j-1}}, \quad 0 < \theta_{j-1} \leq 1$$

$$S_m(x) = S_{m-1}(x)^{\theta_{m-1}}, \quad 0 < \theta_{m-1} \leq 1. \tag{2.2}$$

Of note, the condition on the parameters θs indicates that subjects from class m tend to have higher diagnostic test values than those from class j, that subjects from class j tend to have higher measurements than those from class $j-1$ and so on.

Using the log transformation allows us to rewrite the relationships among the survival

functions as

$$\log\big(S_{j+1}(x)\big) = \theta_j \log\big(S_j(x)\big), \quad j = \{1, 2, \ldots, m-1\}.$$

Moreover, according to the general definition of the hazard function

$$h(x) = \frac{-dS(x)}{dx}\frac{1}{S(x)} = \frac{-d[\log S(x)]}{dx}$$

and taking the first derivative with respect to x, we obtain

$$h_{j+1}(x) = h_j \theta_j.$$

Thus, the unknown parameters θ can be modeled with the Cox proportional hazards model, by assuming the marker value instead of the time index as the argument of the hazard function. The general formula of the Cox model is:

$$h(x|d) = h_1(x) \exp\{\boldsymbol{\beta}' \mathbf{d}\}$$

where x is the marker value, $\mathbf{d}$ is the vector of appropriate dummy variables to detect the group, $\boldsymbol{\beta}$ is the vector of parameters with $\theta_j = \exp\{\beta_j\}$, and h_1 is the hazard function of the baseline group.

2.2. The ROC manifold - generalization to m classes

Suppose $m-1$ assigned thresholds exist, with $c_1 < \cdots < c_j < \cdots < c_{m-1}$. The m probabilities of correct classification, in this case, are:

$$u_1 = P(X_1 < c_1); \;\; u_j = P(c_{j-1} \leq X_j < c_j); \;\; u_m = P(X_m \geq c_{m-1}).$$

In terms of survival functions, we can write the recursive expression for:

$$\begin{aligned} u_1 &= 1 - S_1(c_1) \qquad (2.3)\\ u_j &= S_j(c_{j-1}) - S_j(c_j) \\ u_m &= S_m(c_{m-1}). \qquad (2.4) \end{aligned}$$

and, as a consequence, the recursive expression for the survival function

$$\begin{aligned} S_1(c_1) &= 1 - u_1 \\ S_2(c_1) &= (S_1(c_1))_1^{\theta} = (1-u_1)^{\theta_1} \\ S_2(c_2) &= (S_2(c_1)) - u_2) = (1-u_1)^{\theta_1} - u_2 \\ S_j(c_{j-1}) &= (S_{j-1}(c_{j-1}))^{\theta_{j-1}} = (((1-u_1)^{\theta_1} - u_2)^{\theta_2} + \cdots - u_{j-1})^{\theta_{j-1}} \\ S_j(c_j) &= S_j(c_{j-1}) - u_j = (((1-u_1)^{\theta_1} - u_2)^{\theta_2} + \cdots - u_{j-1})^{\theta_{j-1}} - u_j \\ S_m(c_{m-1}) &= (S_{m-1}(c_{m-1}))^{\theta_{m-1}} = ((((1-u_1)^{\theta_1} - u_2)^{\theta_2} + \cdots - u_{m-2})^{\theta_{m-2}} - u_{m-1})^{\theta_{m-1}} \end{aligned} \qquad (2.5)$$

Now, the equation for the ROC surface can be obtained substituting (2.5) in (2.4) and deriving the correct-classification probability of class m as:

$$u_m = ((((1-u_1)^{\theta_1} - u_2)^{\theta_2} + \cdots - u_{m-2})^{\theta_{m-2}} - u_{m-1})^{\theta_{m-1}}.$$

The ROC hypersurface is thus an m-dimensional manifold with the following expression:

$$ROC(\mathbf{u}) = ((((1-u_1)^{\theta_1} - u_2)^{\theta_2} + \cdots - u_j)^{\theta_j} - u_{m-1})^{\theta_{m-1}} \tag{2.6}$$

where

$$\mathbf{u} = (u_1, \ldots, u_j, \ldots, u_m), \text{ with } u_j \in [0,1], \quad j = \{1, \ldots, m-1\},$$

and

$$0 \leq u_1 \leq 1; \;\; 0 \leq u_2 \leq S_2(c_1); \;\; 0 \leq u_j \leq S_j(c_{j-1}); \;\; 0 \leq u_m < S_m(c_{m-1}).$$

Moreover, from eqs (2.3)-(2.4), we can rewrite

$$0 \leq u_2 \leq (1-u_1)^{\theta_1}$$

and

$$0 \leq u_j < ((((1-u_1)^{\theta_1} - u_2)^{\theta_2} - u_3)^{\theta_3} + \cdots - u_{j-1})^{j-1}.$$

If the m distributions are identical, the discriminating power of the diagnostic test is null, and the ROC hypersurface satisfies the equation $u_1 + \cdots + u_j + \cdots + u_m = 1$.

2.3. The hypervolume under the manifold

For the simpler cases of two- and three-classification issues, the accuracy measure of the discriminating function $ROC(\mathbf{u})$ can be given by the AUC and VUS respectively. In the next theorem, an analytical formula for calculating the HUM is described. The analytical formula depends on only θ^s parameters.

Theorem 2.1. *Consider a m-class classification problem wherein the survival functions, under the Lehmann condition, are given as in eqs (2.1)-(2.2). Moreover, let the quantities u_1, u_j and u_m be defined as in eqs (2.3)-(2.4).*

If the discriminating function is given by the $ROC(\mathbf{u})$ function, as defined in eq. (2.6), then the HUM is given by

$$HUM(\theta_1, \theta_2, \ldots, \theta_{m-2}, \theta_{m-1}) = \prod_{n=1}^{m-1} \frac{1}{R(m-1,n)} \tag{2.7}$$

where:

$$R(m-1,n) = 1 + \sum_{i=m-n}^{m-1} \prod_{j=m-n}^{i} \theta_j.$$

for some parameter $0 < \theta_1 \leq 1$, $0 < \theta_2 \leq 1 \ldots 0 < \theta_{m-1} \leq 1$.

Proof: The hypervolume under the ROC manifold represents the accuracy measure of interest and is obtained by integrating the ROC surface defined in eq. (2.6) over its domain:

$$HUM(\theta_1,\ldots,\theta_{m-2},\theta_{m-1}) =$$

$$= \int_0^1 \int_0^{(1-u_1)^{\theta_1}} \cdots \int_0^{(((1-u_1)^{\theta_1}-u_2)^{\theta_2}\cdots-u_{m-2})^{\theta_{m-2}}} ((((1-u_1)^{\theta_1}-u_2)^{\theta_2}\cdots-u_{m-2})^{\theta_{m-2}}-u_{m-1})^{\theta_{m-1}}\, du_{m-1}\, du_{m-2}\ldots du_1$$

$$= \frac{1}{1+\theta_{m-1}} \int_0^{(((1-u_1)^{\theta_1}-u_2)^{\theta_2}\cdots-u_{m-3})^{\theta_{m-3}}} -\Big((1-u_1)^{\theta_1}-u_2)^{\theta_2}\cdots-u_{m-1}\Big)^{1+\theta_{m-1}}\Big|_0^{(((1-u_1)^{\theta_1}-u_2)^{\theta_2}\cdots-u_{m-2})^{\theta_{m-2}}} du_{m-2}\ldots du_1$$

$$= \frac{1}{1+\theta_{m-1}} \int_0^{(((1-u_1)^{\theta_1}-u_2)^{\theta_2}\cdots-u_{m-3})^{\theta_{m-3}}} ((1-u_1)^{\theta_1}-u_2)^{\theta_2}\cdots-u_{m-2})^{\theta_{m-2}(1+\theta_{m-1})}\, du_{m-2}\ldots du_1$$

$$= \frac{1}{1+\theta_{m-1}}\frac{1}{1+\theta_{m-2}(1+\theta_{m-1})} \int_0^{(((1-u_1)^{\theta_1}-u_2)^{\theta_2}\cdots-u_{m-4})^{\theta_{m-4}}} -\Big((1-u_1)^{\theta_1}-u_2)^{\theta_2}\cdots-u_{m-2})^{\theta_{m-2}(1+\theta_{m-1})}\Big|_0^{(((1-u_1)^{\theta_1}-u_2)^{\theta_2}-u_3)^{\theta_3}} du_{m-3}\ldots du_1$$

after some algebra a generalized expression for HUM can be obtained:

$$HUM(\theta_1,\ldots,\theta_{m-2},\theta_{m-1}) = \frac{1}{1+\theta_{m-1}}\cdot\frac{1}{1+\theta_{m-2}+\theta_{m-2}\theta_{m-1}}\cdots \frac{1}{1+\theta_{m-3}+\theta_{m-3}\theta_{m-2}+\theta_{m-3}\theta_{m-2}\theta_{m-1}}\cdot \frac{1}{1+\theta_1+\theta_1\theta_2+\cdots+\theta_1\theta_2\ldots\theta_{m-2}\theta_{m-1}}. \tag{2.8}$$

or, in a more compact form:

$$\mathrm{HUM}(\theta_1,\theta_2,\ldots,\theta_{m-1}) = \prod_{n=1}^{m-1}\frac{1}{R(m-1,n)},$$

with:

$$R(m-1,n) = 1+\sum_{i=m-n}^{m-1}\prod_{j=m-n}^{i}\theta_j.$$

□

Eq. (2.7) represents the entire HUM in an m-dimensional classification problem. As can be seen, the closed form depends only on the parameters of the Lehmann assumption. The following corollary to Theorem 2.1 provides an interesting result when the ROC function associated with the classification problem is unable to discriminate among the different classes.

Corollary 2.1. *Consider a m-class classification problem as the one defined in Theorem 2.1. If the diagnostic test is non-discriminatory, the HUM is equal to*

$$HUM(\theta_1,\ldots,\theta_{m-2},\theta_{m-1}) = 1/m!$$

which is its minimum possible value.

Proof of Corollary 2.1: In the case of a non-discriminatory test, $\theta_1 = \theta_2 = \cdots = \theta_{m-1} = 1$, thus indicating that no difference exists among the m survival functions. As a consequence, the HUM in eq. (2.7) simply becomes $HUM = 1/m!$, which represents the minimum value. □

Corollary 2.2. *If the problem reduces to three classes, that is when the parameters from θ_{m-1} to θ_3 equal zero, then $HUM = VUS = 1/((\theta_2+1)(\theta_1(\theta_2+1)+1))$, as in Nze Ossima et al. (2015). If θ_2 equals zero too, then $HUM = VUS = AUC = 1/(\theta_1+1)$; i.e., the volume under the surface collapses to the area under the Lehmann family ROC curve, as shown in Gönen and Heller (2010).*

Proof of Corollary 2.2: The result can be easily obtained from the proof of Theorem 2.1, through solving the integral by fixing the parameters from θ_{m-1} to $\theta_3 = 0$ first, and then $\theta_2 = 0$ too. □

The result in Corollary 2.2 simply states that the analytic formula for the HUM developed in Theorem 2.1 is a generalization, in the m-class classification framework, of the findings for the three- Nze Ossima et al. (2015) and two-class Gönen and Heller (2010) classification procedures.

2.4. Estimation

A semi-parametric approach is proposed to estimate the hypersurface and the hypervolume derived above. For the sake of simplicity, only the estimation procedure in a four class framework is shown. The generalization to m classes is straightforward.

Under the Lehmann condition, the four survival functions are related by the parameters $\theta_1, \ldots, \theta_{m-1}$. As a consequence, both the ROC surface and the hypervolume are functions of these unknown parameters, which represent the object of inferential analysis.

According to the intuition regarding the two-class classification issue Gönen and Heller (2010), we propose to estimate the parameters $\theta_1, \ldots$ and θ_{m-1} with the proportional hazards regression model already present in several statistical packages. As shown below, the problem in fact can be written in terms of a regression model. Let x be the generic realization of the continuous diagnostic test X. Under the Lehmann condition we have:

$$\theta_1 = \frac{h_2(x)}{h_1(x)} \tag{2.9}$$

$$\vdots$$

$$\theta_{m-1} = \frac{h_m(x)}{h_{m-1}(x)} \tag{2.10}$$

where $h_i(x)$ are the hazard functions of X in the i-th group, with $i = 1, \ldots, m$.

To estimate the parameters, we define the Cox model with the diagnostic test X in place of the usual "time" variable. To implement the model, the m-level categorical variable D can be replaced by a combination of $m-1$ *ad hoc* dummy variables $D_1, \ldots, D_{m-1}$.

The Cox proportional hazards model thus can be written as:

$$h(x|d_1, \ldots, d_{m-1}) = h_1(x) \exp\{\beta_1 d_1 + \cdots + \beta_{m-1} d_{m-1}\}$$

where $d_1, \ldots, d_{m-1}$ are the realizations of the dummy variables $D_1, \ldots, D_{m-1}$, respectively, and $h_1(x)$ is the baseline hazard function. Specifically, the hazard in group 1 is:

$$h(x|d_1 = 0, \ldots, d_m = 0) = h_1(x), \tag{2.11}$$

the hazard in group $m-2$ is:

$$h(x|d_1 = 1, \ldots, d_{m-3} = 1, d_{m-2} = 0) = h_{m-2} = h_1(x) \exp\{\beta_1 + \cdots + \beta_{m-3}\},$$

the one in group $m-1$ is:

$$h(x|d_1 = 1, d_{m-2} = 1, d_{m-1} = 0) = h_{m-1}(x) = h_1(x) \exp\{\beta_1 + \cdots + \beta_{m-2}\},$$

and that in group m is:

$$h(x|d_1 = 1, \ldots, d_{m-1} = 1, d_m = 0) = h_m(x) = h_1(x) \exp\{\beta_1 + \cdots + \beta_{m-1}\}, \tag{2.12}$$

where the scalars $\beta_1, \ldots$ and β_{m-1} are the parameters of the regression model to be estimated.

Now, by substituting the hazard functions in eqs (2.11)-(2.12) into the definition of the θ parameters in eqs (2.9)-(2.10), the latter equations can be rewritten as a function of the β parameters:

$$\begin{aligned}
\theta_1 &= \frac{h_1(x) \exp\{\beta_1\}}{h_1(x)} = \exp\{\beta_1\} \\
&\vdots \\
\theta_{m-2} &= \frac{h_1(x) \exp\{\beta_1 + \cdots + \beta_{m-2}\}}{h_1(x) \exp\{\beta_1 + \cdots + \beta_{m-3}\}} = \exp\{\beta_{m-2}\} \\
\theta_{m-1} &= \frac{h_1(x) \exp\{\beta_1 + \cdots + \beta_{m-1}\}}{h_1(x) \exp\{\beta_1 + \cdots + \beta_{m-2}\}} = \exp\{\beta_{m-1}\}.
\end{aligned}$$

Therefore, the vector of parameters $\boldsymbol{\theta}$ can be estimated by estimating the vector of parameters $\boldsymbol{\beta}$. For the latter estimation, the well known estimation techniques for the Cox proportional hazards model based on the maximization of the partial likelihood can be used. Moreover, because the properties of the ML estimators hold for the parameters βs, and the θs are obtained by applying a monotonic and continuous transformation, they maintain the same properties. Under the usual regularity conditions, the estimators $\hat{\theta}_1, \ldots, \hat{\theta}_{m-1}$ are thus consistent and asymptotically normally distributed.

Finally, by substituting the $\hat{\theta}$s in eq. (2.7) we obtain the partial maximum likelihood estimate of $\widehat{HUM}_{LN}$:

$$\widehat{HUM}_{LN} = \prod_{n=1}^{m-1} \frac{1}{R(m-1,n)}$$

where:

$$R(m-1,n) = 1 + \sum_{i=m-n}^{m-1} \prod_{j=m-n}^{i} \hat{\theta}_j.$$

2.5. An analytical formula for the variance of the estimator

The analytical formula for the asymptotic variance is derived only for the four-class framework, and it is obtained according to Nze Ossima et al. (2015) using the Delta method. The variance-covariance matrix for the vector of parameters $\hat{\boldsymbol{\theta}}$ can be decomposed as

$$\begin{aligned} \Sigma_{\hat{\theta}_1,\hat{\theta}_2,\hat{\theta}_3} &= \begin{bmatrix} \sigma^2_{\hat{\theta}_1} & \sigma_{\hat{\theta}_1,\hat{\theta}_2} & \sigma_{\hat{\theta}_1,\hat{\theta}_3} \\ \sigma_{\hat{\theta}_2,\hat{\theta}_1} & \sigma^2_{\hat{\theta}_2} & \sigma_{\hat{\theta}_2,\hat{\theta}_3} \\ \sigma_{\hat{\theta}_3,\hat{\theta}_1} & \sigma_{\hat{\theta}_3,\hat{\theta}_2} & \sigma^2_{\hat{\theta}_3} \end{bmatrix} \\ &= J^T V J \end{aligned} \tag{2.13}$$

where J is the Jacobian of $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$:

$$J = \begin{bmatrix} \exp\{\hat{\beta}_1\} & 0 & 0 \\ 0 & \exp\{\hat{\beta}_2\} & 0 \\ 0 & 0 & \exp\{\hat{\beta}_3\} \end{bmatrix} \tag{2.14}$$

and V is the variance-covariance matrix of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$:

$$V = \Sigma_{\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3} = \begin{bmatrix} \sigma^2_{\hat{\beta}_1} & \sigma_{\hat{\beta}_1,\hat{\beta}_2} & \sigma_{\hat{\beta}_1,\hat{\beta}_3} \\ \sigma_{\hat{\beta}_2,\hat{\beta}_1} & \sigma^2_{\hat{\beta}_2} & \sigma_{\hat{\beta}_2,\hat{\beta}_3} \\ \sigma_{\hat{\beta}_3,\hat{\beta}_1} & \sigma_{\hat{\beta}_3,\hat{\beta}_2} & \sigma^2_{\hat{\beta}_3} \end{bmatrix}. \tag{2.15}$$

By substituting (2.14) and (2.15) in (2.13), the variance-covariance matrix for $\hat{\boldsymbol{\theta}}$ becomes:

$$\begin{aligned} \Sigma_{\hat{\theta}_1,\hat{\theta}_2,\hat{\theta}_3} &= \begin{bmatrix} \exp\{2\hat{\beta}_1\}\sigma^2_{\hat{\beta}_1} & \exp\{\hat{\beta}_1\}\exp\{\hat{\beta}_2\}\sigma_{\hat{\beta}_1\hat{\beta}_2} & \exp\{\hat{\beta}_1\}\exp\{\hat{\beta}_3\}\sigma_{\hat{\beta}_1\hat{\beta}_3} \\ \exp\{\hat{\beta}_1\}\exp\{\hat{\beta}_2\}\sigma_{\hat{\beta}_1\hat{\beta}_2} & \exp\{2\hat{\beta}_2\}\sigma^2_{\hat{\beta}_2} & \exp\{\hat{\beta}_2\}\exp\{\hat{\beta}_3\}\sigma_{\hat{\beta}_2\hat{\beta}_3} \\ \exp\{\hat{\beta}_3\}\exp\{\hat{\beta}_1\}\sigma_{\hat{\beta}_3\hat{\beta}_1} & \exp\{\hat{\beta}_3\}\exp\{\hat{\beta}_2\}\sigma_{\hat{\beta}_3\hat{\beta}_2} & \exp\{2\hat{\beta}_3\}\sigma^2_{\hat{\beta}_3} \end{bmatrix} \\ &= \begin{bmatrix} \exp\{2\hat{\beta}_1\}\sigma^2_{\hat{\beta}_1} & \exp\{\hat{\beta}_1+\hat{\beta}_2\}\sigma_{\hat{\beta}_1\hat{\beta}_2} & \exp\{\hat{\beta}_1+\hat{\beta}_3\}\sigma_{\hat{\beta}_1\hat{\beta}_3} \\ \exp\{\hat{\beta}_1+\hat{\beta}_2\}\sigma_{\hat{\beta}_1\hat{\beta}_2} & \exp\{2\hat{\beta}_2\}\sigma^2_{\hat{\beta}_2} & \exp\{\hat{\beta}_2+\hat{\beta}_3\}\sigma_{\hat{\beta}_2\hat{\beta}_3} \\ \exp\{\hat{\beta}_3+\hat{\beta}_1\}\sigma_{\hat{\beta}_3\hat{\beta}_1} & \exp\{\hat{\beta}_3+\hat{\beta}_2\}\sigma_{\hat{\beta}_3\hat{\beta}_2} & \exp\{2\hat{\beta}_3\}\sigma^2_{\hat{\beta}_3} \end{bmatrix}. \end{aligned}$$

Using the Delta method, the variance for HUM_{LN} is:

$$\begin{aligned}\sigma^2_{\widehat{HUM}_{LN}} = &\sigma^2_{\hat{\theta}_1}\left(\frac{\partial\widehat{HUM}_{LN}}{\partial\hat{\theta}_1}\right)^2+\sigma^2_{\hat{\theta}_2}\left(\frac{\partial\widehat{HUM}_{LN}}{\partial\hat{\theta}_2}\right)^2+\sigma^2_{\hat{\theta}_3}\left(\frac{\partial\widehat{HUM}_{LN}}{\partial\hat{\theta}_3}\right)^2+\\ &+2\left(\frac{\partial\widehat{HUM}_{LN}}{\partial\hat{\theta}_1}\frac{\partial\widehat{HUM}_{LN}}{\partial\hat{\theta}_2}\right)\sigma_{\hat{\theta}_1\hat{\theta}_2}+2\left(\frac{\partial\widehat{HUM}_{LN}}{\partial\hat{\theta}_1}\frac{\partial\widehat{HUM}_{LN}}{\partial\hat{\theta}_3}\right)\sigma_{\hat{\theta}_1\hat{\theta}_3}+\\ &+2\left(\frac{\partial\widehat{HUM}_{LN}}{\partial\hat{\theta}_2}\frac{\partial\widehat{HUM}_{LN}}{\partial\hat{\theta}_3}\right)\sigma_{\hat{\theta}_2\hat{\theta}_3}\end{aligned} \tag{2.16}$$

where the single partial derivatives are given by:

$$\begin{aligned}\frac{\partial\widehat{HUM}_{LN}}{\partial\hat{\theta}_1} &= -\frac{1}{(\hat{\theta}_3+1)[\hat{\theta}_1(\hat{\theta}_2\hat{\theta}_3+\hat{\theta}_2+1)+1]^2}\\ \frac{\partial\widehat{HUM}_{LN}}{\partial\hat{\theta}_2} &= -\frac{2\hat{\theta}_1(\hat{\theta}_2\hat{\theta}_3+\hat{\theta}_2+1)-1}{(\hat{\theta}_2\hat{\theta}_3+\hat{\theta}_2+1)^2[\hat{\theta}_1(\hat{\theta}_2\hat{\theta}_3+\hat{\theta}_2+1)+1]^2}\\ \frac{\partial\widehat{HUM}_{LN}}{\partial\hat{\theta}_3} &= \frac{-\hat{\theta}_2}{(\hat{\theta}_3+1)[\hat{\theta}_2(\hat{\theta}_3+1)+1]^2[\hat{\theta}_1(\hat{\theta}_2(\hat{\theta}_3+1)+1)+1]}\\ &- \frac{\hat{\theta}_1\hat{\theta}_2}{(\hat{\theta}_3+1)[\hat{\theta}_2(\hat{\theta}_3+1)+1][\hat{\theta}_1[\hat{\theta}_2(\hat{\theta}_3+1)+1]+1]^2}\\ &- \frac{1}{(\hat{\theta}_3+1)^2[\hat{\theta}_2(\hat{\theta}_3+1)+1][\hat{\theta}_1(\hat{\theta}_2(\hat{\theta}_3+1)+1)+1]}.\end{aligned}$$

This result is extremely interesting because, to the best of our knowledge, it represents the first time in which a standard error for the HUM related to a four-class classification problem has been derived analytically. Since all the derivatives are derived analytically, the empirical calculation is extremely simple and notably faster than any other technique based on simulation. Examples are provided in the next section.

3. Simulation studies

This section compares three approaches for estimating the HUM value in the specific case of four-class classification. We report simulation studies with multiple scenarios to evaluate the effects of different data generating processes (DGP) on the estimator performance.

3.1. General setting

We hypothesize a setting in which a continuous diagnostic marker is evaluated in a sample with subjects belonging to four different groups, for example, four stages of disease. We consider a set of scenarios according to different sample sizes and different characteristics of the data generating process (DGP). As discussed in the previous sections, an important assumption characterizing this approach is the Lehmann condition, which is not necessary for the other estimators selected for comparison herein. An important distinction in the

following simulation exercises is whether the DGP satisfies such conditions. The Lehmann assumption states that

$$S_2 = S_1^{\theta_1}; \qquad S_3 = S_2^{\theta_2}; \qquad S_4 = S_3^{\theta_3}$$

where S_i, with $i = 1, \ldots, 4$, represents the survival function in each stage of disease, and θ_1, θ_2 and θ_3 are the parameters. The main challenge in the simulation exercises lies in the fact that the Cox model is defined in terms of hazard functions, while data generation requires starting from probability distributions. To address this issue, the Monte Carlo inversion method, as described in Nze Ossima et al. (2015), is employed.

3.2. Comparison with other HUM estimators

To the best of our knowledge, four studies have reported estimators of the hypervolume as a measure of the discrimination accuracy of a continuous biomarker. The four methods -our proposed approach and three alternatives- share the common principle of disregarding the marker's distribution. Notably, three of these methods are nonparametric, while ours is best described as semi-parametric. The first alternative, referred to as HUM_{EX}, adopts a nonparametric approach, with its theoretical and inferential framework built on the Mann-Whitney U statistic Nakas and Yiannoutsos (2004). This estimator is implemented in the R package "Biocomb"Novoselova et al. (2017). The second, HUM_{LF}, is the estimator presented by Li and Fine (2008). Like the HUM_{EX}, it does not require any assumptions regarding the functional form of the distribution of the biomarker in the population, and it is based on the multinomial logistic model. Furthermore, the condition for ordering the test results with respect to the class of disease is relaxed. This estimator has been implemented in a recent R package called "mcca"Gao and Li (2018). The third, recently proposed by Feng et al. (2023), is based on the HUM definition given by Nakas and Yiannoutsos (2004) and Feng et al. (2021). The authors present a graph-based method for efficiently computing the HUM and its asymptotic variance. Interestingly, the method is computationally efficient and applicable to continuous, discrete, and mixed-distribution biomarkers. Unfortunately, no packages are currently available to enable its implementation. Our proposal, HUM_{LN}, is a semi-parametric estimator in the sense that it is obtained through an approach that does not require a full parametric specification of the marker distribution for the four populations. It is based on the Lehmann assumption, which postulates the existence of a monotonic transformation producing marker values with an extreme value distribution without specifying and estimating the transformation. Thus, the only parameters to be estimated are those governing the relationships among the survival distributions. This approach, like that of HUM_{EX}, requires an assumption regarding the ordering of the disease categories. The variance of HUM_{LN} has an analytical form obtained with the Delta method shown in Section 2.5. It is computationally very fast because it relies on the proportional hazards framework, thus enabling inference with standard statistical softwares.

3.3. Data generation processes

In this section, a simulation study is reported, based on a continuous dataset of 1000 random samples generated from different DGPs. The first set of simulations (cases 1–3), summarized in Table 1 , is conducted under the assumption that the Lehmann condition holds. The data are generated by starting from different Weibull distributions. Specifically, the data are generated from a Cox proportional hazards model with different vectors of parameters $\boldsymbol{\beta} = (\beta_1; \beta_2; \beta_3)'$ and groups covariate vector $\boldsymbol{d}$. The hazard functions thus assume the form

$$h_i(t|d_1, d_2, d_3) = \begin{cases} h_1(t) & d_{1i} = 0, d_{2i} = 0, d_{3i} = 0 \\ h_1(t)\exp\{\beta_1 d_{1i}\} & d_{1i} = 1, d_{2i} = 0, d_{3i} = 0 \\ h_1(t)\exp\{\beta_1 d_{1i} + \beta_2 d_{2i}\} & d_{1i} = 1, d_{2i} = 1, d_{3i} = 0 \\ h_1(t)\exp\{\beta_1 d_{1i} + \beta_2 d_{2i} + \beta_3 d_{3i}\} & d_{1i} = 1, d_{2i} = 1, d_{3i} = 1. \end{cases}$$

Data generation under these conditions corresponds to the assumption that $X_i \sim Wei(\lambda_i, \nu)$ is the marker's distribution in the i-th class, with $i = 1, \ldots, 4$. In the second set of simulations (cases 4–6) in Table 2, the shape parameter of the Weibull distribution is group-dependent, meaning that the Lehmann condition does not hold. Finally, the last set of simulations refers to Gaussian distributions (cases 7–9) in Table 3.

- Case 1. is characterized by the vector of parameters $\boldsymbol{\beta} = (-1.4; -0.8; -0.6)'$. That is:

$$\begin{aligned} X_1 &\sim Wei(4, 2) \\ X_2 &\sim Wei(4 * \exp\{-1.4\}, 2) \\ X_3 &\sim Wei(4 * \exp\{-1.4 - 0.8\}, 2) \\ X_4 &\sim Wei(4 * \exp\{-1.4 - 0.8 - 0.6\}, 2). \end{aligned}$$

- Case 2: $\boldsymbol{\beta} = (-2.5; -1.2; -1.7)'$.
- Case 3: $\boldsymbol{\beta}$=$(-4.1; -3.5; -3.8)'$.
- Case 4: $\boldsymbol{\beta} = (-1.2, -0.5, -0.8)'$ and $\nu_1 = 5$, $\nu_2 = 4$, $\nu_3 = 2$, $\nu_4 = 5$.
- Case 5: $\boldsymbol{\beta} = (-1.4, -1.2, -3.2)'$ and $\nu_1 = 2$, $\nu_2 = 2$, $\nu_3 = 3$, $\nu_4 = 4$.
- Case 6: $\boldsymbol{\beta} = (-2.0, -3.0, -6.0)'$ and $\nu_1 = 2$, $\nu_2 = 2.5$, $\nu_3 = 4$, $\nu_4 = 7$.
- Case 7: $\mu_1 = 0.1$, $\mu_2 = 0.3$, $\mu_3 = 0.5$, $\mu_4 = 0.7$ and $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$.
- Case 8: $\mu_1 = 1$, $\mu_2 = 2$, $\mu_3 = 3$, $\mu_4 = 4$ and $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$.
- Case 9: $\mu_1 = 1$, $\mu_2 = 3$, $\mu_3 = 5$, $\mu_4 = 7$ and $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$.

The estimation results are shown in Table 1, Table 2 and Table 3. For each of the three estimators, the first three columns of the tables report the characteristics of the DGP, in terms of the parameters β, the true value of the HUM, and the sample size of each simulation. The following columns show the simulation results, the estimated HUM (hum), the standard error (se) and the bias, expressed in both absolute values and percentages, with respect to the true HUM. Regarding the calculation of the standard errors, for our estimator, we use the formula in eq. (2.16), whereas for the other two estimators, because analytical results do not exist, we apply the bootstrap technique.

	$\boldsymbol{\beta}$	true hum	N	est. hum	se	bias (abs value)	bias %
$\mathbf{HUM_{LN}}$							
case 1	$(-1.4; -0.8; -0.6)'$	0.264	120	0.270	0.042	0.006	2.133
			200	0.267	0.032	0.003	0.996
			320	0.266	0.026	0.002	0.638
case 2	$(-2.5; -1.2; -1.7)'$	0.561	120	0.565	0.053	0.004	0.712
			200	0.563	0.041	0.002	0.396
			320	0.561	0.033	<0.001	0.069
case 3	$(-4.1; -3.5; -3.8)'$	0.933	120	0.933	0.024	<0.001	0.005
			200	0.933	0.018	<0.001	0.034
			320	0.933	0.015	<0.001	0.036
$\mathbf{HUM_{EX}}$							
case 1	$(-1.4; -0.8; -0.6)'$	0.264	120	0.270	0.047	0.006	2.283
			200	0.266	0.037	0.003	0.952
			320	0.265	0.030	0.001	0.283
case 2	$(-2.5; -1.2; -1.7)'$	0.561	120	0.565	0.061	0.004	0.740
			200	0.563	0.048	0.002	0.360
			320	0.561	0.039	<0.001	0.083
case 3	$(-4.1; -3.5; -3.8)'$	0.933	120	0.934	0.032	0.001	0.130
			200	0.933	0.025	<0.001	0.012
			320	0.933	0.020	<0.001	0.006
$\mathbf{HUM_{LF}}$							
case 1	$(-1.4; -0.8; -0.6)'$	0.264	120	0.252	0.048	0.012	4.609
			200	0.249	0.038	0.015	5.563
			320	0.248	0.031	0.016	6.127
case 2	$(-2.5; -1.2; -1.7)'$	0.561	120	0.531	0.063	0.030	5.315
			200	0.529	0.049	0.032	5.680
			320	0.526	0.038	0.035	6.218
case 3	$(-4.1; -3.5; -3.8)'$	0.933	120	0.925	0.035	0.008	0.879
			200	0.922	0.027	0.011	1.175
			320	0.922	0.022	0.011	1.205

Table 1: Simulation results for the Weibull case under the Lehmann assumption and three different vectors of parameters $\boldsymbol{\beta}$ of the Cox proportional hazards regression model. Bias is expressed in absolute value, and the percentage is calculated with respect to the true HUM value.

	$\boldsymbol{\beta}$	true hum	N	est. hum	se	bias (abs value)	bias %
$\mathbf{HUM_{LN}}$							
case 4	$\boldsymbol{\beta} = (-1.2, -0.5, -0.8)'$ $\nu_1 = 5, \nu_2 = 4,$ $\nu_3 = 2, \nu_4 = 5$	0.234	120	0.160	0.024	0.075	31.972
			200	0.159	0.018	0.076	32.332
			320	0.158	0.015	0.076	32.447
case 5	$\boldsymbol{\beta} = (-1.4, -1.2, -3.2)'$ $\nu_1 = 2, \nu_2 = 2,$ $\nu_3 = 3, \nu_4 = 4$	0.553	120	0.481	0.055	0.073	13.150
			200	0.478	0.045	0.076	13.644
			320	0.475	0.035	0.079	14.267
case 6	$\boldsymbol{\beta} = (-2.0, -3.0, -6.0)'$ $\nu_1 = 2, \nu_2 = 2.5,$ $\nu_3 = 4, \nu_4 = 7$	0.836	120	0.721	0.049	0.115	13.753
			200	0.716	0.039	0.120	14.375
			320	0.713	0.032	0.123	14.683
$\mathbf{HUM_{EX}}$							
case 4	$\boldsymbol{\beta} = (-1.2, -0.5, -0.8)'$ $\nu_1 = 5, \nu_2 = 4,$ $\nu_3 = 2, \nu_4 = 5$	0.234	120	0.186	0.034	0.048	20.558
			200	0.182	0.028	0.053	22.484
			320	0.235	0.028	0.001	0.237
case 5	$\boldsymbol{\beta} = (-1.4, -1.2, -3.2)'$ $\nu_1 = 2, \nu_2 = 2,$ $\nu_3 = 3, \nu_4 = 4$	0.553	120	0.513	0.059	0.040	7.283
			200	0.514	0.048	0.040	7.195
			320	0.554	0.038	0.001	0.094
case 6	$\boldsymbol{\beta} = (-2.0, -3.0, -6.0)'$ $\nu_1 = 2, \nu_2 = 2.5,$ $\nu_3 = 4, \nu_4 = 7$	0.836	120	0.773	0.049	0.063	7.508
			200	0.771	0.038	0.064	7.715
			320	0.771	0.030	0.065	7.818
$\mathbf{HUM_{LF}}$							
case 4	$\boldsymbol{\beta} = (-1.2, -0.5, -0.8)'$ $\nu_1 = 5, \nu_2 = 4,$ $\nu_3 = 2, \nu_4 = 5$	0.234	120	0.117	0.042	0.118	50.175
			200	0.111	0.033	0.124	52.768
			320	0.104	0.023	0.130	55.513
case 5	$\boldsymbol{\beta} = (-1.4, -1.2, -3.2)'$ $\nu_1 = 2, \nu_2 = 2,$ $\nu_3 = 3, \nu_4 = 4$	0.553	120	0.473	0.058	0.080	14.540
			200	0.473	0.047	0.080	14.487
			320	0.470	0.036	0.083	15.049
case 6	$\boldsymbol{\beta} = (-2.0, -3.0, -6.0)'$ $\nu_1 = 2, \nu_2 = 2.5,$ $\nu_3 = 4, \nu_4 = 7$	0.836	120	0.714	0.056	0.122	14.547
			200	0.711	0.044	0.125	14.939
			320	0.710	0.035	0.126	15.073

Table 2: Simulation results for the Weibull case with group-specific shape parameters (Lehmann condition not satisfied) and three different vectors of parameters $\boldsymbol{\beta}$. Bias is expressed in absolute value, and the percentage is calculated with respect to the true HUM value.

When the Lehmann condition is satisfied, HUM_{LN} and HUM_{EX} perform very well and highly similarly. Both estimators present decreased bias. Case 3 has practically no bias, even in small samples, although our estimator has systematically smaller standard errors. The third estimator, HUM_{LF}, performs systematically more poorly, in terms of both bias and standard errors. In the Supplementary file, the computational times (in seconds) for obtaining the results are presented. The results clearly indicate that our estimator is computationally extremely efficient with respect to the others. When the Lehmann condition is not met, the HUM_{LN} continues to perform systematically better than the HUM_{LF}, in terms of both bias and precision (standard error). Regardless of the sample size and the

	parameters β	true hum	N	est. hum	se	bias (abs value)	bias %
HUM_{LN}							
case 7	$\mu_1 = 0.1$, $\mu_2 = 0.3$	0.077	120	0.069	0.016	0.009	11.218
	$\mu_3 = 0.5$, $\mu_4 = 0.7$		200	0.069	0.012	0.008	10.792
	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$		320	0.069	0.010	0.008	10.633
case 8	$\mu_1 = 1$, $\mu_2 = 2$	0.369	120	0.305	0.050	0.064	17.397
	$\mu_3 = 3$, $\mu_4 = 4$		200	0.300	0.039	0.069	18.633
	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$		320	0.299	0.031	0.071	19.148
case 9	$\mu_1 = 1$, $\mu_2 = 3$	0.771	120	0.672	0.060	0.099	12.818
	$\mu_3 = 5$, $\mu_4 = 7$		200	0.664	0.047	0.107	13.876
	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$		320	0.659	0.039	0.111	14.450
HUM_{EX}							
case 7	$\mu_1 = 0.1$, $\mu_2 = 0.3$	0.077	120	0.090	0.018	0.012	15.754
	$\mu_3 = 0.5$, $\mu_4 = 0.7$		200	0.084	0.014	0.007	8.552
	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$		320	0.082	0.012	0.004	5.712
case 8	$\mu_1 = 1$, $\mu_2 = 2$	0.369	120	0.365	0.053	0.004	1.206
	$\mu_3 = 3$, $\mu_4 = 4$		200	0.367	0.042	0.003	0.717
	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$		320	0.369	0.034	<0.001	0.037
case 9	$\mu_1 = 1$, $\mu_2 = 3$	0.771	120	0.768	0.049	0.002	0.322
	$\mu_3 = 5$, $\mu_4 = 7$		200	0.769	0.038	0.002	0.248
	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$		320	0.770	0.031	0.001	0.080
HUM_{LF}							
case 7	$\mu_1 = 0.1$, $\mu_2 = 0.3$	0.077	120	0.077	0.021	0.001	0.844
	$\mu_3 = 0.5$, $\mu_4 = 0.7$		200	0.074	0.016	0.003	4.354
	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$		320	0.073	0.014	0.005	5.928
case 8	$\mu_1 = 1$, $\mu_2 = 2$	0.369	120	0.334	0.052	0.035	9.558
	$\mu_3 = 3$, $\mu_4 = 4$		200	0.337	0.041	0.033	8.875
	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$		320	0.339	0.033	0.030	8.085
case 9	$\mu_1 = 1$, $\mu_2 = 3$	0.771	120	0.703	0.057	0.068	8.768
	$\mu_3 = 5$, $\mu_4 = 7$		200	0.702	0.045	0.068	8.872
	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$		320	0.703	0.036	0.068	8.808

Table 3: Simulation results for Normal distributions with group-specific expected values and equal variances (Lehmann condition not satisfied). Bias is expressed in absolute value, and the percentage is calculated with respect to the true HUM value.

magnitude of the HUM, the best estimator in terms of bias is the HUM_{EX}. The computational times reported the Supplementary file clearly indicate that our estimator is much faster than the others, regardless of the size of the sample. The computational time instead markedly increases with both HUM_{EX} and HUM_{LF}. In consideration of departures from the Lehmann conditions originating from Gaussian distributions, the HUM_{LN} estimators show increasing bias with larger sample sizes and more separate variables for the groups. Despite such unfavorable conditions, the bias remains relatively contained and in line with that of the HUM_{LF} estimator, which is, however, much more computationally demanding. The HUM_{EX} estimator in general has the lowest bias, although it becomes computationally time consuming as the sample size increases.

4. Blood markers for colorectal cancer

Cancer detection at early stages has been one of the main research topics undertaken by the scientific community over the recent decades. As the deterioration from pre malignant lesion to carcinoma and metastasis involves several molecular events, the idea of detecting solid tumours through simple blood tests has received growing interest. At the same time, medical and chemical research has greatly expanded the amount of testable components using human blood samples, including cell-free DNA (cfDNA) and RNA (cfRNA), as well as proteins and circulating vesicles, known as exosomes.

With the aim of early detection of CRC, over the recent years many countries have promoted a massive campaigns to resort to faecal immunochemical test (FIT) as a simple, non-invasive and acceptable test.

4.1. Data and descriptive statistics

The data we analyse come from a study on colorectal cancer conducted with the purpose of evaluating a panel of four messenger RNAs (mRNAs) as putative markers of the cancer (Rodia et al., 2018). Specifically, the authors tested four markers: carcinoembryonic antigen-related cell-adhesion molecule 6 (CEACAM6), lectin galactoside binding soluble 4 (LGALS4), tetraspanin 8 (TSPAN8), collagen type I alpha 2 chain (COL1A2), hereafter referred to with the acronym of CELTiC (CEACAM6, LGALS4, TSPAN8 and COL1A2), on subjects positive for the faecal immunochemical test (FIT) and undergoing colonoscopy. The researchers investigated 231 participants that can be classified into four distinct groups: 67 healthy subjects (N), 36 FIT positive with negative colonoscopy (NFIT), 36 low risk that is FIT positive with small polyps (LR), 92 FIT positive with advanced adenomas or a histologically confirmed diagnosis of colorectal cancer (HR/CCR).
Before proceeding with the discussion, a clarification about the study design is in order. In fact, as recognized by the authors, the study presents some limitations in the data. In particular, an in-depth analysis of the design would suggest to examining additional healthy subjects as well as FIT negative subjects and to increase the record of FIT positive and CRC subjects. Although we are aware of the preliminary nature of the study, the reason we decided to apply our methodology to the CELTiC dataset is twofold: firstly, the dataset presents the characteristics of classifying the subjects in four categories; secondly, a recent parallel paper investigating the same data offers a way to compare our results to those obtained through the traditional dichotomous-forced approach. Table 4 provides the descriptive statistics of the four biomarkers characterizing each group. It is important to note that, in line with standard practice in molecular genetics, the marker measures have been transformed such that they are inversely correlated to the amount of gene expression, thus high values indicate low levels of the relative gene. In Figure 2, the relative values for the four groups of healthy control subjects (N), negative colonoscopy (NFIT), low risk lesion (LR), high risk lesion or colorectal cancers (HR/CRC) are reported for each marker CEACAM6, LGALS4, TSPAN8 and COL1A2. A more detailed descriptive statistical analysis can be found in Rodia et al. (2018).

4.2. Statistical analysis using the HUM

In the paper by Rodia et al. (2018), the authors propose to use a multinomial logistic regression model in order to study the association between outcome and a linear combination of the proposed markers; two-tailed p-values less than 0.05 were considered statistically significant; the reference group is N (healthy subjects). However, the authors force the statistical methodology and use dichotomous ROC curve and AUC analysis to assess the accuracy of the model in discriminating among the four groups of subjects. In this section, instead, we implement the proposed methodology. Specifically, we aim to compute the ability of the four biomarkers in discriminating subjects among the four groups. The accuracy summary measure we adopt, thus, is given by the HUM. We estimated HUM_{LN} for each single marker and the analytical asymptotic standard errors of the HUM_{LN}, which provide a measure of the sample uncertainty associated with the accuracy summary indicator. The HUM_{LN}, moreover, is compared with the other two estimators already existing in the literature: the HUM_{LF} by Li and Fine (2008) and the HUM_{EX} by Nakas and Yiannoutsos (2004) already presented in Section 3.2. For the latter, however, since no analytical formulas are available for calculating the standard errors, bootstrap techniques have been employed. Finally, for the sake of completeness, bootstrap standard errors have also been calculated for the HUM_{LN} estimator. As suggested in Li and Fine (2008), bootstrap estimation of the standard errors for the three estimates are calculated with $B = 100$ bootstrap resamples. [1]

4.3. Results

Before presenting the results, we check for the Lehmann assumption through the graphical method and the statistical test proposed by Grambsch and Therneau (1994). The plots of the survival curves for all the markers are reported in Figure 1. Even though it is not easy to interpret the plots in a four-class framework, when considering TSPAN8 and COL1A2, it emerges that the curves are overall parallel, although three of them are practically indistinguishable for a wide range of marker values. For the two remaining markers, instead, the parallelism is questionable, especially for LGALS4. As can be seen in the Table 5, the null hypothesis of the proportional hazards assumption, cannot be rejected for TSPAN8 and COL1A2, while only at the 1% critical level for CEACAM6. It has to be rejected, instead, for the LGALS4 marker. Overall, however, the p-values remain relatively low, even when the null hypothesis cannot be rejected by the data. For each of the three estimators the point estimates and the bootstrap standard errors are shown in Table 6. Moreover, for the HUM_{LN}, in brackets, the asymptotic analytical standard errors is also retported. From Table 6, we can deduce some general results: for all the markers, despite the estimator used, the HUM is rather low; the HUM_{EX} estimator always produces the highest values of the HUM while the HUM_{LN}, on the contrary, is the one giving the lowest values of the hypervolume. The worst results for the HUM_{LN} are those associated to LGALS4 and CEACAM6, a possible explanation for the poor performance of the HUM_{LN} in terms of the magnitude of the point

[1] Specifically, the HUM_{EX} has been estimated by the R-function *Calculate HUM-EX* included in the R-package *Biocomb* , which internally computes the maximal HUM value between all the possible permutations of class labels. We recall that HUM_{LF}, instead, is not affected by the ordering of the categories.

estimate can be ascribed to the fact that the Lehmann condition is only marginally supported for the first two markers while must be rejected for the last two, at least at the 5% critical level (see Table 5). The HUM_{LN} is largely the most efficient estimator, both in terms of analytical and bootstrapped standard errors. Moreover, our approach, through the analysis of the p-values of the Cox regression coefficients, offers detailed information about the discriminatory power of the marker for each single class. In Table 7, for each marker, we show the estimated coefficients of the Cox model with the associated p-values. On the other side, regardless of the estimator used, the estimated HUM values confirm the LGALS4 biomarker as the most powerful blood marker discriminating among the four groups. This result reinforces the one obtained in Rodia et al. (2018) with the rough pairwise ROC analysis. This marker is able to correctly classifying four subjects randomly chosen from the four groups with a probability that ranges between 0.129 of the HUM_{LN} estimator and 0.219 of the HUM_{EX} estimator. If we recall that the null value of HUM for a four-category classification problem is $1/4! = 0.042$, we can argue that the accuracy of this marker is sufficiently better than a random guess.

	N		N FIT		LR		HR-CCR	
	mean	sd	mean	sd	mean	sd	mean	sd
TSPAN8	11.330	1.718	9.997	1.199	9.924	1.420	9.558	1.851
COL1A2	11.449	1.920	9.674	1.286	9.674	1.373	9.608	1.973
LGALS4	12.893	1.971	15.662	1.300	15.284	0.775	14.693	1.275
CEACAM6	12.343	1.893	14.249	1.096	13.589	1.206	13.346	1.247

Table 4: Means, standard deviations and relative effects of biomarkers by class.

marker	p-value
TSPAN8	0.185
COL1A2	0.139
LGALS4	< 0.01
CEACAM6	0.011

Table 5: Test for proportional hazards assumption.

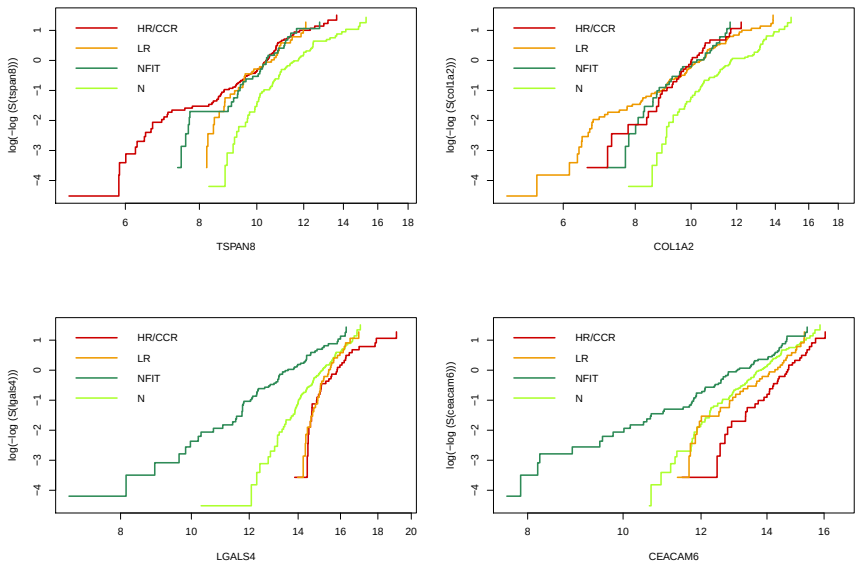

Figure 1: Log(-log(survival))curves as function of marker value for TSPAN8, COL1A2, LGALS4 and CEACAM6.

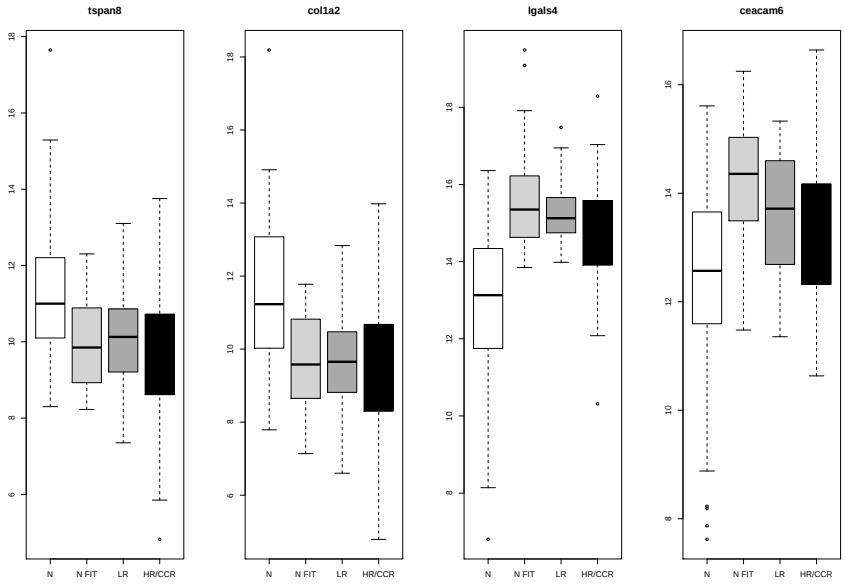

Figure 2: Box plot of the four blood markers for colorectal cancer detection in the four groups.

5. Discussion

In the present work, we sought to develop a method to evaluate the accuracy measure of a biomarker in discriminating a sample of patients divided into a number m of classes according to the severity of a disease. Because our problem was related to an m-class sample, we focused on ROC manifolds and relative HUM. We proposed a semi-parametric approach to

	HUM_{LN}		HUM_{EX}		HUM_{LF}	
	$\widehat{\text{hum}}$	se	$\widehat{\text{hum}}$	se	$\widehat{\text{hum}}$	se
TSPAN8	0.087	0.012 (0.013)	0.110	0.019	0.108	0.029
COL1A2[a]	0.102	0.015 (0.015)	0.111	0.020	0.111	0.035
LGALS4	0.129	0.021 (0.020)	0.219	0.034	0.151	0.041
CEACAM6	0.096	0.016 (0.015)	0.144	0.024	0.139	0.029

Table 6: Estimated HUMs by marker.
[a] The HUM_{EX} estimator suggests a different order of categories (HR/CCR, LR, NFIT, N); however, if we impose this order on the HUM_{LN} estimator, the estimated HUM decreases to 0.081.

TSPAN8					
	β	$\theta = \exp(\beta)$	se	z	p-value
LR	-0.065	0.937	0.197	-0.330	0.740
N FIT	0.014	1.014	0.236	0.060	0.950
N	-0.839	0.432	0.216	-3.890	< 0.001
COL1A2					
	β	$\theta = \exp(\beta)$	se	z	p-value
N FIT	-0.006	0.994	0.236	-0.030	0.980
HR/CCR	-0.207	0.813	0.199	-1.040	0.300
N	-0.836	0.433	0.168	-4.960	< 0.001
LGALS4					
	β	$\theta = \exp(\beta)$	se	z	p-value
HR/CCR	-0.918	0.399	0.164	-5.590	< 0.001
LR	-0.228	0.796	0.197	-1.160	0.250
N FIT	-0.356	0.700	0.243	-1.460	0.140
CEACAM6					
	β	$\theta = \exp(\beta)$	se	z	p-value
HR/CCR	-0.446	0.641	0.162	-2.750	0.006
LR	-0.099	0.906	0.199	-0.500	0.620
N FIT	-0.465	0.628	0.239	-1.950	0.052

Table 7: Estimated coefficients for the Cox proportional hazards regression model. Note that, for each marker, the ordering is the one established by the relative effects and the omitted class is the reference one.

derive the ROC manifold and the HUM. Our contribution relies on the Lehmann assumption and constitutes the generalization of the work by Gönen and Heller (2010) and Nze Ossima et al. (2015) to a m-class setting. We derived the analytical expression of the HUM estimator, referred to as HUM_{LN}, along with its variance, specifically restricted to the special case of four classes. Furthermore, an inferential solution for the estimator and for the variance is proposed.

To evaluate the performance of the suggested estimator, we carried out extensive simulation studies and applied the proposed approach to an empirical problem. In addition, both in the simulation study and in the empirical application, we compared our estimator with two other estimators reported in the literature. As expected, our estimator presented highly satisfactory performance under the Lehmann condition in both small and large samples. Furthermore, we observed that the behavior of HUM_{LN} did not depend on the distance among

the distributions. In fact, it performed well in estimating both small and large hypervolumes. Moreover, the coverage rate was always large and very close to the nominal level.

We acknowledge that the Lehmann assumption is a strong one, and we were mindful of this limitation in our analysis. For this reason, we investigated the behavior of the estimator when these conditions are not met. Departures from the Lehmann assumption were found to influence the bias of the estimator, with the bias increasing as the sample size grew. Nonetheless, our results demonstrated that the performance of our estimator was never substantially inferior to that of the two other competitors, even under very unfavorable conditions, such as significant departures from the Lehmann assumption.

Furthermore, we also evaluated the proposed method in terms of computational time. With our approach, calculating the HUM takes no more than a few seconds, even for large samples. This aspect indicates the advantage of our procedure compared to the two alternative estimators in the literature: in addition to requiring more than double in small samples, they present exploding computational times when the sample size becomes relatively large.

Moreover, our method is based on a well developed framework, and is easy to handle and implementable in all standard statistical packages. In addition, the regression framework on which our approach is based provides multiple advantages. First, it enables controlling the possible effects of covariates on the accuracy of the diagnostic test. Second, the estimates of the coefficients related to the class variables indirectly provide a way to test whether the classes significantly differ. If no significant difference is found, the analysis can be simplified by grouping classes with the same distributions. These are notable aspects of our method.

Finally, in medical research, interest in combining information from multiple markers is still increasing, see for example Hsu and Hsueh (2013); thus, future research might focus on the optimal combination of biomarkers, such as attempting to solve for an efficient algorithm in maximizing the HUM.

ACKNOWLEDGMENTS

This work takes the inspiration from my PhD thesis Nardi (2019), so I would like to express my deepest gratitude to Prof. Rossella Miglio for her steady guidance during my PhD experience. I would also like to thank Eng. Stefano Battista, a dear friend, for the fruitful discussions on the concept of Hypervolume.

REFERENCES

Dreiseitl, S., Ohno-Machado, L., and Binder, M. (2000). Comparing three-class diagnostic tests by three-way roc analysis. *Medical Decision Making*, 20(3):323–331.

Feng, Q., Li, J., Ping, X., and Calster, B. V. (2021). Hypervolume under roc manifold for discrete biomarkers with ties. *Journal of Statistical Computation and Simulation*, 91(18):3864–3879.

Feng, Q., Liu, P., Kuan, P.-F., Zou, F., Chen, J., and Li, J. (2023). A network approach to compute hypervolume under receiver operating characteristic manifold for multi-class biomarkers. *Statistics in Medicine*, 42(6):834–859.

Gao, M. and Li, J. (2018). *mcca: Multi-Category Classification Accuracy*. R package version 0.2.0.

Gönen, M. and Heller, G. (2010). Lehmann family of roc curves. *Medical Decision Making*, 30(4):509–517.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526.

Hand, David J.and Till, R. J. (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning*, 45(2):171–186.

Hsu, M. and Hsueh, H. (2013). The linear combinations of biomarkers which maximize the partial area under the roc curves. *Computational Statistics*, 28:647–666.

Jialiang Li, J. P. F. and Pencina, M. J. (2017). Multi-category diagnostic accuracy based on logistic regression. *Statistical Theory and Related Fields*, 1(2):143–158.

Kang, L. and Tian, L. (2013). Estimation of the volume under the roc surface with three ordinal diagnostic categories. *Computational Statistics & Data Analysis*, 62:39–51.

Lehmann, E. L. (1953). The power of rank tests. *The Annals of Mathematical Statistics*, pages 23–43.

Li, J. and Fine, J. P. (2008). Roc analysis with multiple classes and multiple tests: methodology and its application in microarray studies. *Biostatistics*, 9(3):566–576.

Lusted, L. B. (1960). Logical analysis in roentgen diagnosis. *Radiology*, 74(2):178–193. PMID: 14419034.

Mossman, D. (1999). Three-way rocs. *Medical Decision Making*, 19(1):78–89. PMID: 9917023.

Nakas, C. T. and Yiannoutsos, C. T. (2004). Ordered multiple-class roc analysis with continuous measurements. *Statistics in medicine*, 23(22):3437–3449.

Nardi, E. (2019). *ROC curves and the generalization to multiple classes*. PhD thesis, University of Bologna, Bologna (IT).

Novoselova, N., Wang, J., and Pessler, F. K. (2017). *Biocomb: Feature Selection and Classification with the Embedded Validation Procedures for Biomedical Data Analysis*. R package version 0.3.

Nze Ossima, A. D., Daurès, J.-P., Bessaoud, F., and Trétarre, B. (2015). The generalized lehmann roc curves: Lehmann family of roc surfaces. *Journal of Statistical Computation and Simulation*, 85(3):596–607.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

Rodia, M. T., Solmi, R., Pasini, F., Nardi, E., Mattei, G., Ugolini, G., Ricciardiello, L., Strippoli, P., Miglio, R., and Lauriola, M. (2018). Lgals4, ceacam6, tspan8, col1a2: blood markers for colorectal cancer. validation in a cohort of subjects positive to faecal immunochemical test. *Clinical colorectal cancer*, 17(2):217–228.

Scurfield, B. K. (1996). Multiple-event forced-choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology*, 40(3):253–269. PMID:8979976.

Scurfield, B. K. (1998). Generalization of the theory of signal detectability ton-eventm-dimensional forced-choice tasks. *Journal of Mathematical Psychology*, 42(1):5–31.

Waegeman, W., Baets, B. D., and Boullart, L. (2008). On the scalability of ordered multi-class roc analysis. *Computational Statistics & Data Analysis*, 52(7):3371–3388.

Xiong, C., van Belle, G., Miller, J. P., and Morris, J. C. (2006). Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. *Statistics in medicine*, 25(7):1251–1273.