Covariate-Specific Estimation of the Sensitivity to the Early Disease Stage in Diagnostic Tests

Covariate-specific estimation of the sensitivity to the early disease stage in diagnostic tests

Duc-Khanh To[1,2], Gianfranco Adimari[3], Monica Chiogna[4], and Gloria Remoli[4]

[1] Faculty of Mathematics and Computer Science, University of Science, Ho Chi Minh City, Vietnam,
[2] Vietnam National University, Ho Chi Minh City, Vietnam
[3] Unit of Biostatistics, Epidemiology and Public Health - Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, Italy,
[4] Department of Statistical Sciences "Paolo Fortunati", University of Bologna, Italy
gloria.remoli@unibo.it

Abstract. Within the three-class ROC analysis framework, we address the problem of making inferences about a true class fraction (TCF), given the remaining two. More precisely, we propose a procedure to estimate the covariate-specific TCF_2, i.e., the covariate-specific probability of correct classification at the so-called early stage, when the values for the true class fractions at first and third classes are fixed. An application to a real dataset is also presented.

Keywords: ROC analysis, quantile regression, logistic regression, biomarker

1 Introduction

Three-class receiver operating characteristic (ROC) analysis is a widely utilized statistical framework for assessing the discriminatory capability of a diagnostic test (or biomarker) across three distinct disease states, as detailed in [9]. Let Y denote the diagnostic test outcome, typically measured on a continuous scale, and let Y_1, Y_2, Y_3 represent the test outcomes for individuals in classes 1, 2, and 3, respectively. Without loss of generality, we assume that higher test values correspond to greater disease severity. Given thresholds t_1 and t_2, where $t_1 < t_2$, the true class fractions (TCFs) are defined as follows [10]:

$$\begin{aligned} \theta_1 &\equiv \mathrm{TCF}_1(t_1) = \Pr(Y_1 \le t_1) = F_1(t_1) \\ \theta_2 &\equiv \mathrm{TCF}_2(t_1, t_2) = \Pr(t_1 < Y_2 \le t_2) = F_2(t_2) - F_2(t_1) \\ \theta_3 &\equiv \mathrm{TCF}_3(t_2) = \Pr(Y_3 > t_2) = 1 - F_3(t_2) \end{aligned} \tag{1}$$

where $F_j(\cdot)$ denotes the cumulative distribution function of Y_j for $j = 1, 2, 3$.

Within the context of three-class ROC analysis, a key focus lies in making statistical inferences about one TCF given fixed values for the other two. Of particular interest to medical practitioners is the evaluation of the probability θ_2, which quantifies the diagnostic test's ability to correctly classify individuals

in the second disease class (commonly referred to as the early stage), while ensuring predefined values for θ_1 and θ_3, corresponding to the first and third disease classes, respectively. Over the past decade, various methodologies have been proposed to estimate or construct confidence intervals for θ_2, conditional on fixed values of θ_1 and θ_3. Notable contributions include [1], [2], and [5].

However, in biomedical research, it is common for additional covariate information (e.g., age, sex, comorbidities) to be collected alongside potential diagnostic test results. These covariates may influence the performance of the diagnostic test and associated accuracy metrics. To date, only [8] and [12] have proposed methods that incorporate covariates into three-class ROC analysis.

In this study, we introduce a novel methodology for estimating the covariate-specific TCF_2, i.e., the covariate-specific probability θ_2 of correct classification at the early stage of disease, under fixed values of θ_1 and θ_3. Our approach is grounded in quantile regression techniques [6] and logistic regression modeling applied to a constructed "working" binary sample. Specifically, we utilize quantile regressions to estimate covariate-specific thresholds corresponding to fixed values of θ_1 and θ_3. These thresholds are then used to generate a "working" binary sample at specified covariate values. Subsequently, logistic regression is employed to estimate the covariate-specific TCF_2. Additionally, we extend our methodology to estimate covariate-adjusted TCF_2, maintaining fixed values for θ_1 and θ_3.

We demonstrate that the proposed estimators for both covariate-specific and covariate-adjusted TCF_2 are consistent and asymptotically normal. While theoretical derivations are omitted here for brevity, we focus on illustrating our method through an application to a real-world dataset.

2 Model setting and estimation

Let $\mathbf{X} = (X_1, \ldots, X_p)^\top$ be a p-dimensional random variable describing covariates measured together with test values, and let $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ denote the covariates associated with class 1, 2 and 3, respectively. Given a set of values $\mathbf{x}$ of interest, let $F_j(y;\mathbf{x}) = \Pr(Y_j \le y|\mathbf{X}_j = \mathbf{x})$ be the conditional cumulative distribution function of Y_j, and let $F_j^{-1}(\cdot;\mathbf{x})$ be the corresponding conditional quantile function. The covariate-specific true class fractions are defined as follows:

$$\begin{aligned}\theta_1(t_1;\mathbf{x}) &= F_1(t_1;\mathbf{x}),\\ \theta_2(t_1,t_2;\mathbf{x}) &= F_2(t_2;\mathbf{x}) - F_2(t_1;\mathbf{x}),\\ \theta_3(t_2;\mathbf{x}) &= 1 - F_3(t_2;\mathbf{x}),\end{aligned}$$

for $t_1 < t_2$, at $\mathbf{x}$.

Now, by controlling the levels of TCF_1 and TCF_3 at θ_{10} and θ_{30}, respectively, we can obtain covariate-specific thresholds $t_{10}(\mathbf{x}) = F_1^{-1}(\theta_{10};\mathbf{x})$ and $t_{20}(\mathbf{x}) = F_3^{-1}(1-\theta_{30};\mathbf{x})$. As a consequence, the covariate-specific TCF_2, at fixed values of θ_{10} and θ_{30}, is given by:

$$\theta_{20}(\theta_{10},\theta_{30};\mathbf{x}) = \Pr\left(F_1^{-1}(\theta_{10};\mathbf{x}) \le Y_2 \le F_3^{-1}(1-\theta_{30};\mathbf{x})\middle|\mathbf{X}_2 = \mathbf{x}\right). \quad (2)$$

The covariate-adjusted TCF_2, at fixed values of θ_{10} and θ_{30}, is then defined as

$$\theta_{20}(\theta_{10},\theta_{30}) = \Pr\left(F_1^{-1}(\theta_{10};\mathbf{X}_2) \le Y_2 \le F_3^{-1}(1-\theta_{30};\mathbf{X}_2)\right). \tag{3}$$

Let $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ be sets of covariate values from class 1, 2 and 3, respectively. We use the linear quantile regression [6] to model covariate-specific thresholds, i.e.,

$$F_1^{-1}(\theta_{10};\mathbf{x}_1) = \mathbf{z}_{1,\mathbf{x}_1}^\top \boldsymbol{\alpha}(\theta_{10}), \tag{4}$$

$$F_3^{-1}(1-\theta_{30};\mathbf{x}_3) = \mathbf{z}_{3,\mathbf{x}_3}^\top \boldsymbol{\beta}(\theta_{30}), \tag{5}$$

where $\mathbf{z}_{1,\mathbf{x}_1}$ and $\mathbf{z}_{3,\mathbf{x}_3}$ are rows of design matrices, and $\boldsymbol{\alpha}(\cdot)$, $\boldsymbol{\beta}(\cdot)$ are parameter vectors depending on θ_{10} and θ_{30}, respectively. The covariate-specific TCF_2, $\theta_{20}(\theta_{10},\theta_{30};\mathbf{x}_2)$, can be modeled by using a generalized linear model for a proportion, for instance, we adopt a logistic model:

$$\Pr\left(\mathbf{z}_{1,\mathbf{x}_2}^\top \boldsymbol{\alpha}(\theta_{10}) \le Y_2 \le \mathbf{z}_{3,\mathbf{x}_2}^\top \boldsymbol{\beta}(\theta_{30}) \middle| \mathbf{X}_2 = \mathbf{x}_2\right) = \frac{\exp\left(\mathbf{z}_{2,\mathbf{x}_2}^\top \boldsymbol{\gamma}\right)}{1+\exp\left(\mathbf{z}_{2,\mathbf{x}_2}^\top \boldsymbol{\gamma}\right)}. \tag{6}$$

For $j = 1, 2, 3$ and $i = 1, \ldots, n_j$, let y_{ji} be the observed test values, and $\mathbf{x}_{ji}$ be the associated observed covariate values. The estimates $\widehat{\boldsymbol{\alpha}}(\theta_{10})$ and $\widehat{\boldsymbol{\beta}}(\theta_{30})$ are obtained by the standard quantile regression method by `R` package `quantreg` [7]. Given the estimates $\widehat{\boldsymbol{\alpha}}(\theta_{10})$, $\widehat{\boldsymbol{\beta}}(\theta_{30})$ and the covariate values $\mathbf{x}_{2i}$, we can construct "working" thresholds $\widehat{t}_{10}(\mathbf{x}_{2i}) = \boldsymbol{z}_{1,\mathbf{x}_{2i}}^\top \widehat{\boldsymbol{\alpha}}(\theta_{10})$ and $\widehat{t}_{20}(\mathbf{x}_{2i}) = \boldsymbol{z}_{3,\mathbf{x}_{2i}}^\top \widehat{\boldsymbol{\beta}}(\theta_{30})$, and obtain binary "working" diagnostic responses for the second class as $\widetilde{y}_{2i} = \mathrm{I}\left(\boldsymbol{z}_{1,\mathbf{x}_{2i}}^\top \widehat{\boldsymbol{\alpha}}(\theta_{10}) \le y_{2i} \le \boldsymbol{z}_{3,\mathbf{x}_{2i}}^\top \widehat{\boldsymbol{\beta}}(\theta_{30})\right)$, for $i = 1, \ldots, n_2$. Then an estimator $\widehat{\boldsymbol{\gamma}}$ is obtained by implementing a standard GLM estimation process with the working sample as response and the covariate value $\mathbf{x}$ of interest. Consequently, an estimate $\widehat{\theta}_{20}(\theta_{10},\theta_{30};\mathbf{x})$ is obtained as $\frac{\exp\left(\mathbf{z}_{2,\mathbf{x}}^\top \widehat{\boldsymbol{\gamma}}\right)}{1+\exp\left(\mathbf{z}_{2,\mathbf{x}}^\top \widehat{\boldsymbol{\gamma}}\right)}$. Finally, as

$$\theta_2(\theta_{10},\theta_{30}) = \mathbb{E}\left[\mathrm{I}\left(\boldsymbol{z}_{1,\mathbf{X}_2}^\top \boldsymbol{\alpha}(\theta_{10}) \le Y_2 \le \boldsymbol{z}_{3,\mathbf{X}_2}^\top \boldsymbol{\beta}(\theta_{30})\right)\right]$$

from (3), in order to estimate the covariate-adjusted TCF_2 at fixed values of θ_{10} and θ_{30} we use the plug-in principle, i.e.

$$\widehat{\theta}_2(\theta_{10},\theta_{30}) = \frac{1}{n_2}\sum_{i=1}^{n_2} \mathrm{I}\left(\boldsymbol{z}_{1,\mathbf{x}_{2i}}^\top \widehat{\boldsymbol{\alpha}}(\theta_{10}) \le y_{2i} \le \boldsymbol{z}_{3,\mathbf{x}_{2i}}^\top \widehat{\boldsymbol{\beta}}(\theta_{30})\right).$$

Standard errors for the proposed estimators can be obtained by using bootstrap procedures.

3 Real data analysis

In this section, we apply our proposed method to study the effect of some covariates on the ability of Bio-adrenomedullin (bio-ADM) to distinguish between

levels of congestion in patients with new-onset or worsening heart failure [11]. Data were obtained from 'A systems BIOlogy Study to TAilored Treatment in Chronic Heart Failure' (BIOSTAT-CHF [13]), a multicentre, multinational, prospective observational study. Given clinical congestion score (CCS) values obtained as the sum of peripheral oedema (0 to 3), jugular venous pressure (0, 1) and orthopnea (0, 1), "true" levels of congestion are defined as no congestion (CCS = 0), mild congestion (CCS=1,2) and severe congestion (CCS=3,4,5). Of the 2516 patients enrolled, we keep those with complete information on both congestion score and bio-ADM. Among 1344 subjects, 427 (31.8%) are not congested, 546 (40.6%) are mildly congested and 372 (27.6%) are severely congested. See Figure 1 for a plot of the non-parametrically estimated marginal densities by congestion level.

In our analysis, we estimate covariate-specific TCF_2 as well as covariate-adjusted TCF_2, including as covariates age, presence of diabetes and renal disease, and their interaction. The true class fractions are fixed as $\theta_{10} = 0.2$ and $\theta_{30} = 0.6$. The choice of these true class fractions, specifically the relatively low value of 0.2 for the first category, is grounded in clinical considerations rather than economic priorities. This decision reflects the understanding that misclassifying a healthy individual into a category associated with mild or severe congestion does not lead to significant clinical consequences. Such misclassifications would typically prompt further diagnostic evaluations, which would ultimately confirm the individual's healthy status.

In the first quantile regression, only renal disease is significant, whereas in the second, all clinical variables, that is, diabetes, renal disease and their interaction, are significant. Table 1 presents the estimated thresholds $\widehat{t}_{10}(\mathbf{x})$ and $\widehat{t}_{30}(\mathbf{x})$ as a function of age across four health conditions: no co-morbidities, renal disease, diabetes, and both co-morbidities combined. Across all ages, patients with no comorbidities consistently exhibit the lowest values for both thresholds. At the same time, those with both renal disease and diabetes generally have the highest thresholds, highlighting the combined effect of the two conditions in the diagnosis of mild congestion. Specifically, renal disease alone results in slightly higher thresholds than diabetes alone. Furthermore, as age increases, threshold $\widehat{t}_{10}(\mathbf{x})$ for all groups shows a gradual rise, while threshold $\widehat{t}_{10}(\mathbf{x})$ for all groups shows a gradual decrease, although the rate of change appears different in different groups. Taking also into account that if no covariates are considered, the estimated thresholds result to be $\widehat{t}_{10} = 16.55$, $\widehat{t}_{30} = 47.03$, these findings suggest that the presence of these co-morbidities, particularly when combined, along with consideration of age, has a measurable impact on the diagnosis of mild congestion, potentially reflecting a greater clinical burden.

As for the logistic regression on the "working" response, all covariates have a significant effect. Figure 2 shows the covariate-specific estimates of TCF_2. Finally, the estimated covariate-adjusted TCF_2 is 0.66, with a standard error of 0.0314, based on the bootstrap process with 200 replications. The corresponding 95% confidence interval is $(0.59,\ 0.73)$.

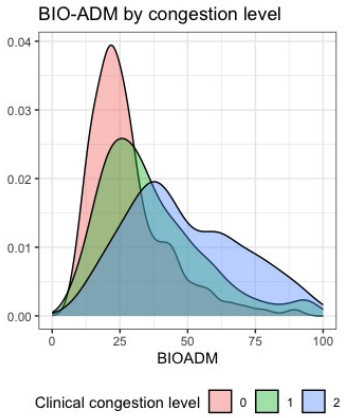

Fig. 1: Bio-ADM marginal density by congestion level

Age	None		Renal Dis.	
	$\widehat{t}_{10}(\mathbf{x})$	$\widehat{t}_{30}(\mathbf{x})$	$\widehat{t}_{10}(\mathbf{x})$	$\widehat{t}_{30}(\mathbf{x})$
50	15.04	44.73	18.95	75.22
60	15.28	42.01	19.18	72.50
70	15.51	39.30	19.42	69.79
80	15.75	36.58	19.65	67.07
90	15.98	33.87	19.89	64.36

Age	Diabetes		Both	
	$\widehat{t}_{10}(\mathbf{x})$	$\widehat{t}_{30}(\mathbf{x})$	$\widehat{t}_{10}(\mathbf{x})$	$\widehat{t}_{30}(\mathbf{x})$
50	15.70	64.45	19.32	62.90
60	15.94	61.73	19.56	60.19
70	16.17	59.02	19.79	57.47
80	16.41	56.30	20.03	54.76
90	16.64	53.59	20.26	52.04

Table 1: Estimated thresholds $\widehat{t}_{10}(\mathbf{x})$ and $\widehat{t}_{30}(\mathbf{x})$, as a function of age when no co-morbidities are recorded; in presence of renal diseases; of diabetes; of both co-morbidities.

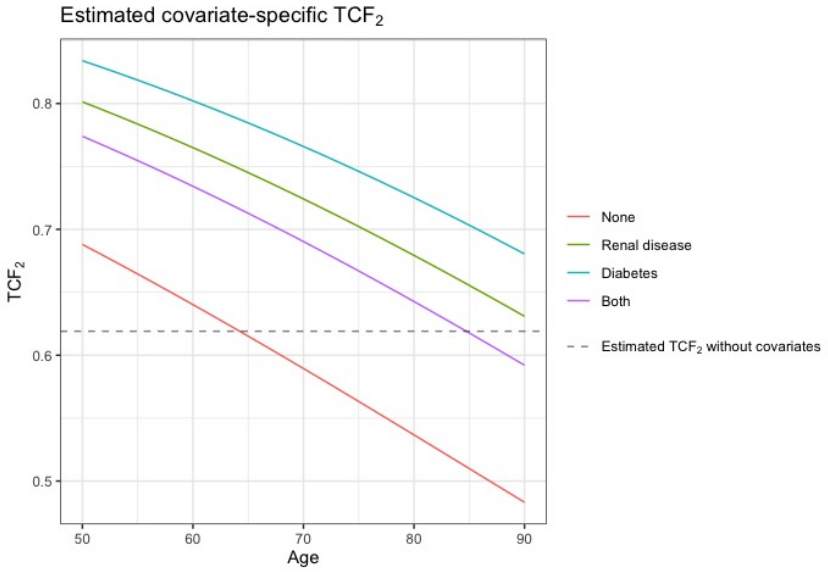

Fig. 2: Estimated covariate-specific TCF_2

Acknowledgements

The European Union - Next Generation EU - PNRR ex D.M. 117/2023, jointly with Istituto Superiore di Sanità (ISS) -CUP J33C23001470009- finances Gloria Remoli (GR). We thank HORIZON-HLTH-2022-TOOL-11-01 project: "BIOmarker based diagnostic TOOLkit to personalize pharmacological approaches in congestive heart failure" (project number: 101095653) for providing the data. Moreover, we thank L.Potena (IRCSS Ospedaliero-Universitaria Ospedale S. Orsola, Bologna), E.Toschi (ISS) for facilitating and G.Coutance (INSERM, Paris) for hosting GR during her visit at INSERM.

References

1. Dong, T., Tian, L.: Confidence interval estimation for sensitivity to the early diseased stage based on empirical likelihood. Journal of Biopharmaceutical Statistics **25**(6), 1215–1233 (2015)
2. Dong, T., Tian, L., Hutson, A., Xiong, C.: Parametric and non-parametric confidence intervals of the probability of identifying early disease stage given sensitivity to full disease and specificity with three ordinal diagnostic groups. Statistics in Medicine **30**(30), 3532–3545 (2011)
3. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. CRC press (1993)
4. Frumento, P., Bottai, M.: Parametric modeling of quantile regression coefficient functions. Biometrics **72**(1), 74–84 (2016)
5. Hai, Y., Shi, S., Qin, G.: Bayesian and influence function-based empirical likelihoods for inference of sensitivity to the early diseased stage in diagnostic tests. Biometrical Journal **65**(3), 2200,021 (2023). DOI https://doi.org/10.1002/bimj.202200021
6. Koenker, R.: Quantile regression, vol. 38. Cambridge University Press (2005)
7. Koenker, R.: quantreg: Quantile Regression (2023). URL https://CRAN.R-project.org/package=quantreg. R package version 5.95
8. Li, J., Zhou, X., Fine, J.P.: A regression approach to ROC surface, with applications to Alzheimer's disease. Sci. China Math. **55**(8), 1583–1595 (2012)
9. Nakas, C.T., Bantis, L.E., Gatsonis, C.A.: ROC Analysis for Classification and Prediction in Practice. CRC Press (2023)
10. Nakas, C.T., Yiannoutsos, C.T.: Ordered multiple-class ROC analysis with continuous measurements. Stat. Med **23**(22), 3437–3449 (2004)
11. Ter Maaten, J.M., Kremer, D., Demissei, B.G., Struck, J., Bergmann, A., Anker, S.D., Ng, L.L., Dickstein, K., Metra, M., Samani, N.J., et al.: Bio-adrenomedullin as a marker of congestion in patients with new-onset and worsening heart failure. European journal of heart failure **21**(6), 732–743 (2019)
12. To, D.K., Adimari, G., Chiogna, M.: Estimation of the volume under a ROC surface in presence of covariates. Computational Statistics & Data Analysis **174**, 107,434–107,449 (2022)
13. Voors, A.A., Anker, S.D., Cleland, J.G., Dickstein, K., Filippatos, G., van der Harst, P., Hillege, H.L., Lang, C.C., Ter Maaten, J.M., Ng, L., et al.: A systems biology study to tailored treatment in chronic heart failure: rationale, design, and baseline characteristics of biostat-chf. European journal of heart failure **18**(6), 716–726 (2016)