Receiver Operating Characteristic Analysis Under Extended Tree or Umbrella Ordering

Yingdong Feng[1], Li Yan[2], and Lili Tian[1]

Abstract

Let $Y_1, Y_2, \cdots, Y_K$ denote the marker values for independent disease classes $1, 2, \cdots, K$. Extended tree or umbrella ordering is defined as $(Y_1, \cdots, Y_{K_1}) \preceq (Y_{K_1+1}, \cdots, Y_K)$ or $(Y_1, \cdots, Y_{K_1}) \succeq (Y_{K_1+1}, \cdots, Y_K)$ where $K_1 \geq 1$, $K \geq 2$. In this paper, we consider the problem of measuring diagnostic ability of a biomarker under extended tree or umbrella ordering. Traditionally, researchers often first pool $Y_1, \cdots, Y_{K_1}$ as one major class and $Y_{K_1+1}, \cdots, Y_K$ as another, and then estimate area under ROC curve (AUC) in binary classification. The purpose of this paper is two-fold: 1) to investigate the inappropriateness of AUC obtained by such polling strategy as a diagnostic measure; 2) to propose a ROC framework for extended tree ordering (ETROC) and area under the curve ($ETAUC$) as a diagnostic measure. The generalized inference (GI) and nonparametric bootstrap (NB) methods are studied for confidence interval estimation of $ETAUC$. Simulation studies are carried out to assess the performance of the proposed methods. An ovarian cancer data set is analyzed using the proposed new measure $ETAUC$.

Keywords

Tree ordering; ROC curve; area under ROC (AUC); ovarian cancer

[1]Department of Biostatistics, University at Buffalo, Buffalo, NY 14214, USA.

[2]Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, NY 14263.

Corresponding author:

Lili Tian, Department of Biostatistics, 717 Kimball Tower, 3435 Main Street, Buffalo, NY 14214

Email: ltian@buffalo.edu

1 Introduction

The receiver operating characteristic (ROC) curve, i.e.true positive rate (sensitivity) against true negative rate (1-specificity), and its associated statistics such as area under ROC curve (AUC) serve as the most popular tools in the field of diagnostic studies for the purpose of marker evaluation under binary classification. The area under the ROC curve, AUC, is the most widely used diagnostic measure. There exist many comprehensive reviews regarding ROC curve and AUC (Pepe, 2003; Shapiro, 1999; Zhou *et al.*, 2011; Zou *et al.*, 2011).

In practice, many diseases can be classified into multiple K ($K > 2$) classes. Let $Y_1, Y_2, \cdots, Y_K$ be the marker values for independent disease classes $1, 2, \cdots, K$, respectively. Without loss of generality, assume higher value indicates worse condition. There exist many possible orderings for $Y_1, Y_2, \cdots, Y_K$ among which simple ordering ($Y_1 \preceq Y_2 \cdots \preceq Y_K$) and tree or umbrella ordering ($Y_1 \preceq (Y_2, Y_3, \cdots, Y_K)$ or $Y_1 \succeq (Y_2, Y_3, \cdots, Y_K)$) are most widely studied in statistical literature. Note that "$\preceq$" means "stochastically smaller" and there do not exists clearly defined orderings among the variables placed inside a parenthesis. Simple ordering often occurs in cancer and Alzheimer's disease diagnosis, e.g. the diagnosis result could be healthy, early diseased and fully diseased. (Cramer *et al.*, 2011; Leichtle *et al.*, 2013; Morris *et al.*, 2001; Partheen *et al.*, 2011; Scinto and Daffner, 2000). There exist many statistical literatures on the inference about the diagnostic measures under simple ordering (Alonzo *et al.*, 2009; Dong *et al.*, 2017; Li and Fine, 2008; Li *et al.*, 2012; Mossman, 1999; Nakas *et al.*, 2010, 2013; Nakas and Yiannoutsos, 2004; Sampat *et al.*, 2009; Scurfield, 1996, 1998; Tian *et al.*, 2011; Xiong *et al.*, 2006, 2007; Youden, 1950; Zhang and Li, 2011).

Besides simple ordering, tree and umbrella ordering also have important clinical applications, especially in molecular diagnostics of cancer subtypes. Take lymphomas diagnosis as an example. Lymphomas are a group of hematological malignancies that are derived from lymphocytes and occur predominantly in lymph nodes or other lymphoid structures. Non-Hodgkin lymphomas (NHL) is one of the major categories of lymphomas

and B-cell lymphoma accounts for about 90% of NHL. Subtypes of B-cell lymphoma include diffuse large B-cell lymphoma (DLBCL), follicular lymphoma, primary mediastinal large B-cell lymphoma (PMBL), Burkitt lymphoma and etc. One of the immunohistochemical biomarkers that are of diagnostic value for B-cell lymphomas is BCL6 (Sun *et al.*, 2016). For evaluating the diagnostic accuracy of BCL6, the classification problem of distinguishing B-cell lymphoma vs. healthy class falls in the framework of tree or umbrella ordering. As another example, lung cancer is classified as non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) based on immunohistological morphology and tumor cell histological, and NSCLC consists of three subtypes (adenocarcinoma, squamous-cell carcinoma, and large cell carcinoma) among which there are no clearly defined orderings. As a result, discriminating NSCLC from healthy class or from SCLC involves tree or umbrella ordering. There exist some research about diagnostic measures under tree or umbrella ordering, e.g. the umbrella volume (UV) proposed by Nakas and Alonzo (Nakas and Alonzo, 2007). Most recently, Wang et al. (Wang *et al.*, 2016) proposed a TROC (ROC for tree-ordering) framework for K-class tree ordering and the area under TROC curve ($TAUC$) as an appropriate diagnostic measure.

Besides simple ordering and tree (or umbrella) ordering, some scenarios often encountered in practice are $(Y_1, Y_2, \cdots, Y_{K_1}) \preceq (Y_{K_1+1}, \cdots, Y_K)$ or $(Y_1, Y_2, \cdots, Y_{K_1}) \succeq (Y_{K_1+1}, \cdots, Y_K)$ where $K_1 > 1$ and $K - K_1 > 1$. For example, a recent ovarian cancer study (Cramer *et al.*, 2011) collected data on healthy controls, benign disease cases, early stage, and late stage ovarian cancer patients. To evaluate the performance of biomarkers for distinguishing between non-cancer and ovarian cancer, Cramer et al. (Cramer *et al.*, 2011) estimated AUCs based on the pooled control group (healthy controls and benign disease cases) and the pooled diseased group (early and late ovarian cancer cases). More details about ovarian cancer data set will be presented in Section 2. As another example, in a recent study of prostate cancer, Chadha et al. (Chadha *et al.*, 2014) estimated AUCs of biomarkers by pooling all healthy and benign subjects as the control group, and all primary and metastatic subjects as the diseased group. In both examples,

the control group consists of two classes (healthy control and benign disease cases) as well as the diseased group (early stage and late stage), and researchers were interested in evaluating the diagnostic accuracy of biomarkers for separating the control group (non-cancer) from the diseased group (cancer). Both examples target the classification under $(Y_1, Y_2) \preceq (Y_3, Y_4)$ where Y_1, Y_2, Y_3, and Y_4 denote the marker values for healthy control, benign disease group, early stage and late stage, respectively. Such pooling strategy used by both of aforementioned studies (Chadha *et al.*, 2014; Cramer *et al.*, 2011) is a common practice for the purpose of evaluating the diagnostic ability of biomarkers when both healthy group and diseased group involved multiple subtypes. However, little has been done on checking the appropriateness and implication of such pooling approach for the intended purpose.

Henceforth in this paper, we refer ordering $(Y_1, Y_2, \cdots, Y_{K_1}) \preceq (Y_{K_1+1}, \cdots, Y_K)$ or $(Y_1, Y_2, \cdots, Y_{K_1}) \succeq (Y_{K_1+1}, \cdots, Y_K)$ where $K_1 \geq 1$ and $K - K_1 \geq 1$. Note that as $K_1 = 1$ and $K = K_1 + 1$, extended tree or umbrella ordering becomes tree or umbrella ordering, respectively. As $K_1 = 1$ and $K = 2$, extended tree or umbrella ordering becomes the traditional binary classification. We will focus on extended tree ordering as extended umbrella ordering can be handled similarly.

The purpose of this paper is two-fold: 1) to investigate the consequences of pooling strategy on biomarker evaluation; 2) to present a new diagnostic accuracy measure under extended tree or umbrella ordering. The rest of paper is organized as follows. In Section 2, using the example of ovarian cancer (Cramer *et al.*, 2011), we demonstrate AUC based on pooling strategy is not an appropriate measure for extended tree ordering. In Section 3, an ROC framework for extended tree ordering, namely ETROC, is introduced and the area under the ETROC curve, namely $ETAUC$, is proposed. Section 4 presents both parametric and nonparametric approaches for the confidence interval estimations of $ETAUC$. In Section 5, a simulation study is conducted to assess and compare the performance of the proposed methods. Section 6 analyzes an ovarian cancer data set. Section 7 gives summary and discussion.

2 Motivation

Ovarian cancer is the 5th leading cause of cancer death among women in developed countries (Chudecka-Głaz, 2015). It generally presents in advanced stages with high case fatality ratio (CFR) but has favorable survival if diagnosed earlier. However, clinical symptoms are not well manifested in early stages of the disease, resulting in late diagnosis and poor prognosis. The Prostate, Lung, Colorectal and Ovarian (PLCO) (Cramer *et al.*, 2011) cancer study is a randomized trial evaluating the effect of screening on cancer mortality. Women assigned to the ovarian screening arm received annual transvaginal ultrasound and CA125 testing. Serum samples were also collected and banked for scientific study. Four sites initiated a phase II study under the early detection research network using the pre-saved samples which included 480 healthy controls, 160 benign cases, 75 early stage cases, and 82 late stage cases. This data set can be downloaded from public portal at https://edrn.nci.nih.gov, and the results have been published by many researchers (Cramer *et al.*, 2011; Zhu *et al.*, 2011). To evaluate the performance of selected ovarian cancer biomarkers for distinguishng between non-cancer cases and cancer cases, Cramer et al. (Cramer *et al.*, 2011) pooled all the benign disease cases and general population controls into a pooled control group, and all the early and late ovarian cancer cases into a pooled diseased group; and estimated AUCs based on binary classification; i.e. control group vs. cancer group. The descriptive statistics for 13 biomarkers including two well studied biomarkers for ovarian cancer, i.e. cancer antigen 125 (CA125) and human epididymis protein 4 (HE4), are presented in Table 1 along with estimated AUCs and ranks.

While such pooling strategies are often used in research, the validity and implication of pooling in biomarker evaluation have never been carefully inspected. In the following, we will examine these aspects closely. We will use the setting of the ovarian cancer data set as an example. Let Y_1, Y_2, Y_3 and Y_4 denote the variables for marker values for healthy, benign, early stage, and late stage groups respectively. Let Y_{12} stand for the variable for the pooled control group, i.e. $Y_{12} \sim f_{Y_{12}}(y) = w_{1|12} f_{Y_1}(y) + (1 - w_{1|12}) f_{Y_2}(y)$

Table 1. Summary statistics and *AUC*s for 13 biomarkers for ovarian cancer data set. Estimated *AUC* is obtained under binary classification: pooled noncancer group (healthy control and benign cases) vs pooled cancer (early and late stage cancer cases) (Cramer *et al.*, 2011).

Biomarkers	Mean				Standard Deviation				AUC
	Healthy controls	Benign	Early	Late	Healthy controls	Benign	Early	Late	(Rank)
CA125	29.71	39.65	565.63	1556.41	136.29	73.57	2549.78	3740.10	0.911 (1)
CA153	17.77	21.10	38.47	147.35	11.13	25.45	56.04	299.66	0.732 (5)
CA199	16.96	31.92	161.82	37.82	31.50	139.69	553.99	106.78	0.603 (12)
KLK6	3.73	3.66	4.21	9.99	1.45	1.22	2.62	10.09	0.710 (8)
CA724	2.29	2.05	37.40	39.74	3.70	2.59	148.76	84.35	0.752 (4)
DD.O110	129.11	137.86	481.12	1033.12	127.93	283.46	979.62	1267.45	0.680 (9)
DD.C248	0.62	0.66	0.60	0.75	0.76	0.83	0.48	0.54	0.588 (13)
DD.P108	38.23	39.68	47.24	69.03	23.78	25.11	23.57	38.45	0.720 (6)
DD.X065	0.63	0.81	1.09	1.64	0.76	1.24	1.81	1.95	0.616 (11)
HE4	61.04	72.07	185.24	496.20	69.56	105.47	253.23	347.50	0.857 (2)
SMRP	0.86	0.94	0.97	4.46	0.81	1.38	0.76	7.41	0.679 (10)
YKL40	82.87	103.13	121.25	205.86	65.22	71.77	100.14	117.97	0.720 (6)
IGF2	1919.53	1852.88	1642.18	1208.97	441.74	488.05	417.98	456.66	0.785 (3)

Table 2. Scenarios of random sampling based on the ovarian cancer data set.

Scenario	$w_{1\|12}$	$w_{3\|34}$	(n_1, n_2, n_3, n_4)
1	0.5	0.5	(100,100,50,50)
2	0.5	0.25	(100,100,25,75)
3	0.5	0.75	(100,100,75,25)
4	0.25	0.5	(50,150,50,50)
5	0.25	0.25	(50,150,25,75)
6	0.25	0.75	(50,150,75,25)
7	0.75	0.5	(150,50,50,50)
8	0.75	0.25	(150,50,25,75)
9	0.75	0.75	(150,50,75,25)

where $w_{1|12}$ is the relative weight for healthy controls out of the pooled control group. Similarly, Y_{34} stands for the variable for the pooled cancer group, i.e. $Y_{34} \sim f_{Y_{34}}(y) = w_{3|34} f_{Y_3}(y) + (1 - w_{3|34}) f_{Y_4}(y)$ where $w_{3|34}$ is the relative weight for early stage ones out of the pooled cancer group. For the purpose of evaluating the diagnostic ability of distinguishing cancer from non-cancer cases, Cramer et al. (Cramer *et al.*, 2011) estimated AUCs for biomarkers as binary classification under $Y_{12} \preceq Y_{34}$, as presented in Table 1. In this paper we refer to such obtained AUCs (based on pooling) as naive AUCs ($NAUC$s) for the reasons which will be given in the following.

It is obvious that $NAUC$ depends on the relative weight $w_{1|12}$ and $w_{3|34}$ since the distributions of Y_{12} and Y_{34} depend on $w_{1|12}$ and $w_{3|34}$ respectively. To further demonstrate this point clearly, we assume Y_k's ($k = 1, 2, 3, 4$) follow exponential distributions with mean value of $1/\lambda_k$. Given $w_{1|12}$ and $w_{3|34}$, the $NAUC$ for $Y_{12} \preceq Y_{34}$ can be obtained as

$$NAUC = 1 - w_{1|12}[\frac{\lambda_3 w_{3|34}}{\lambda_1 + \lambda_3} + \frac{\lambda_4(1 - w_{3|34})}{\lambda_1 + \lambda_4}] - (1 - w_{1|12})[\frac{\lambda_3 w_{3|34}}{\lambda_2 + \lambda_3} + \frac{\lambda_4(1 - w_{3|34})}{\lambda_2 + \lambda_4}]. \quad (1)$$

Given $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, $NAUC$ is directly related to $w_{1|12}$ and $w_{3|34}$. Hence $NAUC$ is not an appropriate diagnostic measure as $NAUC$ should only be reported accompanied by both $w_{1|12}$ and $w_{3|34}$, and two $NAUC$s are only comparable when both $w_{1|12}$ and $w_{3|34}$ are the same for two biomarkers under comparison.

To further demonstrate the effect of relative prevalences $w_{1|12}$ and $w_{3|34}$ on $NAUC$, we

will conduct a random sampling study based on the original ovarian data set presented in Table 1. Let n_1, n_2, n_3 and n_4 stand for the sample sizes for healthy, benign, early and late cases respectively. Subjects are randomly selected from the original data set in each group. The sampling strategies are listed in Table 2. There are three settings for each of $w_{1|12}$ and $w_{3|34}$ yielding a total of nine scenarios for $(w_{1|12}, w_{3|34})$. The estimated $NAUC$ for each biomarker based on 1000 rounds of repetitions along with their corresponding ranks under each scenario are presented in Table 3. It is shown that the estimated $NAUC$s as well as the rankings for some biomakers could change dramatically among different settings of $w_{1|12}$ and $w_{3|34}$. For examples, in terms of estimated $NAUC$, CA72.4 ranges from 0.689 to 0.806, and HE4 ranges from 0.771 to 0.909; in terms of the ranking, CA19.9 ranges from 7 to 13, and YKL40 from 6 to 10 under different settings of $w_{1|12}$ and $w_{3|34}$. This random sampling study clearly demonstrates that $NAUC$ is not an appropriate measure for biomarker evaluation.

3 The ROC curve for extended tree orderings ($ETROC$) and the area under ETROC ($ETAUC$)

This section aims to present an appropriate diagnostic measure for extended tree ordering defined as $(Y_1, Y_2, \cdots, Y_{K_1}) \preceq (Y_{K_1+1}, \cdots, Y_K)$. The classification under extended tree ordering generally involve two major classes each of which contains several subtypes. Without loss of generality, assume higher response indicate worse condition and marker values are measured on a continuous scale. Denote K_1 ($K_1 \geq 1$) as the number of subtypes for the first class, and let $Y_1, Y_2, \cdots, Y_{K_1}$ denote the marker values for independent subtypes $1, 2, \cdots, K_1$. Denote K_2 ($K_2 \geq 1$) as the number of subtypes for the second class and let $Y_{K_1+1}, \cdots, Y_K$ denote the marker values for independent subtypes $K_1 + 1, \cdots, K$. Note $K_1 + K_2 = K$. Note that the case $K_1 = 1$ and $K_2 = 1$ corresponds to binary classification. When $K_1 = 1$ and $K_2 \geq 2$, extended tree ordering becomes tree ordering which was studied by Wang et al. (Wang *et al.*, 2016).

Table 3. Estimated $NAUC$ (Rank) for scenarios in Table 2 using random sampling from the ovarian cancer data set (1000 simulation runs).

Scenario	1	2	3	4	5	6	7	8	9
CA125	0.895 (1)	0.917 (1)	0.872 (1)	0.881 (1)	0.907 (1)	0.855 (1)	0.909 (1)	0.928 (1)	0.890 (1)
CA15.3	0.720 (5)	0.771 (5)	0.669 (6)	0.713 (5)	0.764 (6)	0.661 (6)	0.728 (5)	0.779 (5)	0.677 (5)
CA19.9	0.600 (12)	0.537 (13)	0.662 (7)	0.594 (12)	0.530 (13)	0.653 (7)	0.608 (12)	0.547 (13)	0.670 (7)
KLK6	0.708 (7)	0.767 (6)	0.647 (8)	0.710 (6)	0.769 (5)	0.650 (8)	0.705 (8)	0.765 (7)	0.644 (9)
CA72.4	0.750 (4)	0.804 (4)	0.692 (4)	0.750 (4)	0.806 (4)	0.694 (4)	0.745 (4)	0.801 (4)	0.689 (4)
DD.O110	0.683 (9)	0.718 (10)	0.645 (9)	0.688 (8)	0.722 (10)	0.650 (8)	0.678 (9)	0.715 (10)	0.639 (10)
DD.C248	0.583 (13)	0.612 (12)	0.553 (13)	0.580 (13)	0.611 (12)	0.552 (13)	0.586 (13)	0.614 (12)	0.555 (13)
DD.P108	0.712 (6)	0.753 (7)	0.671 (5)	0.707 (7)	0.749 (7)	0.665 (5)	0.717 (6)	0.757 (8)	0.676 (6)
DD.X065	0.608 (11)	0.647 (11)	0.566 (12)	0.600 (11)	0.642 (11)	0.562 (12)	0.614 (11)	0.652 (11)	0.572 (12)
HE4	0.844 (2)	0.903 (2)	0.783 (2)	0.834 (2)	0.897 (2)	0.771 (2)	0.851 (2)	0.909 (2)	0.794 (2)
SMRP	0.670 (10)	0.745 (9)	0.598 (11)	0.671 (10)	0.743 (8)	0.597 (11)	0.672 (10)	0.746 (9)	0.598 (11)
YKL40	0.695 (8)	0.753 (7)	0.639 (10)	0.676 (9)	0.736 (9)	0.617 (10)	0.715 (7)	0.770 (6)	0.661 (8)
IGF2	0.770 (3)	0.819 (3)	0.720 (3)	0.759 (3)	0.809 (3)	0.708 (3)	0.780 (3)	0.829 (3)	0.731 (3)

Define two random variables as $X = max(Y_1, Y_2, \cdots, Y_{K_1})$ and $Z = min(Y_{K_1+1}, Y_{K_1+2}, \cdots, Y_K)$. The cumulative distribution function of X and Z can be written as

$$F_X(x) = P(X < x) = \prod_{i=1}^{K_1} P(Y_i < x) = \prod_{i=1}^{K_1} F_i(x), \tag{2}$$

$$F_Z(z) = P(Z < z) = 1 - \prod_{j=K_1+1}^{K} P(Y_j \geq z) = 1 - \prod_{j=K_1+1}^{K} (1 - F_j(z)). \tag{3}$$

At a given cut-point c, define extended tree sensitivity ($ETSe$) as

$$ETSe(c) = P(Z > c) = P(min(Y_{K_1+1}, \cdots, Y_K) > c) = \prod_{j=K_1+1}^{K} (1 - F_j(c)), \tag{4}$$

and extended tree specificity ($ETSp$) as

$$ETSp(c) = P(X \leq c) = P(max(Y_1, Y_2, \cdots, Y_{K_1}) \leq c) = \prod_{i=1}^{K_1} F_i(c). \tag{5}$$

For $c \in (-\infty, +\infty)$, the ROC curve under extended tree ordering (ETROC) can be defined as:

$$ETROC(F_1, \cdots, F_{K_1}, \cdots, F_K) = (1 - ETSp(c), ETSe(c)). \tag{6}$$

It can be shown that the area under $ETROC$ (denoted as $ETAUC$) equals to the probability that any randomly chosen subject from the first major class $(1, \cdots, K_1)$ has lower marker value than that of any of the randomly chosen subject from the second

major class $(K_1+1, \cdots, K)$. The details are as follows.

$$\begin{aligned}
ETAUC &= \int_0^1 ETSe \; d(1-ETSp) \\
&= \int_{-\infty}^{+\infty} \prod_{j=K_1+1}^{K} (1-F_j(c)) d \prod_{i=1}^{K_1} F_i(c) \\
&= \sum_{p=1}^{K_1} \int_{-\infty}^{+\infty} \prod_{j=K_1+1}^{K} (1-F_j(c)) \prod_{i=1, i\neq p}^{K_1} F_i(c) f_p(c) dc \\
&= \sum_{p=1}^{K_1} E_{Y_p}[\prod_{j=K_1+1}^{K} P(Y_j > c | Y_p = c) \prod_{i=1, i\neq p}^{K_1} P(Y_i < c | Y_p = c)] \\
&= \sum_{p=1}^{K_1} E_{Y_p}[P(Z > c | Y_p = c) P(X \leq c | Y_p = c)] \\
&= P(X < Z) \\
&= P(max(Y_1, \cdots, Y_{K_1}) < min(Y_{K_1+1}, \cdots, Y_K)).
\end{aligned} \tag{7}$$

It is well known that for binary classification, the chance line for ROC curve is the diagonal line. For the extended tree ordering under consideration, a chance curve corresponds to the scenario that the marker under consideration has no discriminatory ability at all, i.e. the distributions of $Y_1, \cdots, Y_{K_1}, Y_{K_1+1}, \cdots, Y_K$ completely overlap. It is easy to show that the chance curve is

$$Y = (1 - (1-X)^{(1/K_1)})^{K-K_1}, \tag{8}$$

where Y stands for $ETSe$ and X stands for $1 - ETSp$. Note that when $K_1 = 1$ and $K = 2$ (i.e. binary classfication), the chance curve becomes the diagonal line.

The minimum of $ETAUC$, i.e. the area under the chance curve, can be obtained as:

$$ETAUC_{min} = \frac{\Gamma(K_1+1)\Gamma(K-K_1+1)}{\Gamma(K+1)}, \tag{9}$$

where $\Gamma(.)$ is the gamma function. Note that when $K_1 = 1$ and $K = 2$ (i.e. binary classfication), $ETAUC_{min} = 1/2$. If the marker under consideration can perfectly distinguish two major classes, $ETAUC$ reaches its maximum 1.

As an example, let $Y_1 \sim N(0,1), Y_2 \sim N(0.5,1), Y_3 \sim N(1,1), Y_4 \sim N(3,1), Y_5 \sim N(3.5,1), Y_6 \sim N(4,1)$. Figure 1 displays several ETROC curves along with their corresponding chance curves for the scenarios: $(Y_1, Y_2) \preceq (Y_5, Y_6)$(red), $(Y_1, Y_2) \preceq (Y_4, Y_5, Y_6)$(green), $(Y_1, Y_2, Y_3) \preceq (Y_5, Y_6)$(blue), and $(Y_1, Y_2, Y_3) \preceq (Y_4, Y_5, Y_6)$(orange).

Compared to the Naive AUC ($NAUC$), one of the major advantages of the proposed $ETAUC$ is that $ETAUC$ is independent of the relative frequency of subtypes in each major group. Using the same settings of the exponential example in Section 2, i.e. $(Y_1, Y_2) \preceq (Y_3, Y_4)$ and Y_k's $(k = 1, 2, 3, 4)$ follow exponential distributions with mean value of $1/\lambda_k$, $ETAUC$ can be easily obtained as

$$ETAUC = \frac{\lambda_1}{\lambda_1 + \lambda_3 + \lambda_4} + \frac{\lambda_2}{\lambda_2 + \lambda_3 + \lambda_4} - \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4}. \tag{10}$$

Clearly, $ETAUC$ is a function of only λ_1, λ_2 and λ_3, and is independent of both of $w_{1|12}$ and $w_{3|34}$. As stated above, $ETAUC$ is scaleless and has a well-defined probability interpretation. Therefore, $ETAUC$ can be a very useful diagnostic measure for biomarkers under extended tree ordering.

4 Confidence interval estimation of $ETAUC$

In this section, a parametric approach, i.e. generalized inference based on the concept of generalized pivotal quantity, as well as a nonparametric approach are proposed to construct the confidence interval estimation of $ETAUC$. Section 4.2 and 4.2 presents generalized inference methods with and without normality, respectively. Section 4.3 presents the nonparametric method.

4.1 The generalized inference (GI) method: Under Normality

Since the concepts of generalized variables and generalized pivots were introduced (Tsui and Weerahandi, 1989; Weerahandi, 1995), the generalized inference (GI)method has proved to be fruitful for providing finite sample solutions to a variety of problems for which

traditional exact statistical methods do not exist (Dong *et al.*, 2011; Krishnamoorthy *et al.*, 2009; Krishnamoorthy and Mathew, 2003; Lai *et al.*, 2012; Li *et al.*, 2008; Lin *et al.*, 2007; Tian and Cappelleri, 2004; Weerahandi, 2004; Yin and Tian, 2014a,b). The advantage of GI is that it typically has good performance even at small sample sizes as it is an exact test. The detailed discussion could be found in the book by Weerahandi (Weerahandi, 2013).

Suppose K independent groups follow the extended tree ordering, i.e. $(Y_1, \cdots, Y_{K_1}) \preceq (Y_{K_1+1}, \cdots, Y_K)$ and $Y_k \sim N(\mu_k, \sigma_k^2)$ where $k = 1, \cdots, K_1, \cdots, K$. Under normality, $ETAUC$ in (7) can be calculated as the following:

$$ETAUC = \sum_{p=1}^{K_1} \int_{-\infty}^{+\infty} \prod_{j=K_1+1}^{K} \Phi(-a_j c + b_j) \prod_{i=1, i\neq p}^{K_1} \Phi(m_i c - n_i)\phi(c)dc, \quad (11)$$

where $a_j = \frac{\sigma_1}{\sigma_j}$, $m_i = \frac{\sigma_1}{\sigma_i}$, $b_j = \frac{\mu_j - \mu_1}{\sigma_j}$, and $n_i = \frac{\mu_i - \mu_1}{\sigma_i}$.

The generalized pivotal quantities for normal means and variances are well-known as

$$R_{\sigma_k^2} = \frac{(n-1)S_k^2}{V_k}, \quad (12)$$

$$R_{\mu_k} = \bar{y}_k - Z_k\sqrt{\frac{R_{\sigma_k^2}}{n_k}}, \quad (13)$$

where $V_k = \frac{(n_k-1)S_k^2}{\sigma_k^2} \sim \chi^2_{n-1}$ and $Z_k = \frac{\sqrt{n}(\bar{Y}_k - \mu_k)}{\sigma_k} \sim N(0,1)$ for $k = 1, 2, \ldots, K$. The generalized pivotal quantity for $ETAUC$ can be easily written out as

$$R_{ETAUC} = \sum_{p=1}^{K_1} \int_{-\infty}^{+\infty} \prod_{j=K_1+1}^{K} \Phi(-R_{a_j} c + R_{b_j}) \prod_{i=1, i\neq p}^{K_1} \Phi(R_{m_i} c - R_{n_i})\phi(c)dc, \quad (14)$$

where $R_{a_j} = \frac{R_{\sigma_1}}{R_{\sigma_j}}$, $R_{m_i} = \frac{R_{\sigma_1}}{R_{\sigma_i}}$, $R_{b_j} = \frac{R_{\mu_j} - R_{\mu_1}}{R_{\sigma_j}}$, and $R_{n_i} = \frac{R_{\mu_i} - R_{\mu_1}}{R_{\sigma_i}}$. One can easily show that R_{ETAUC} satisfies the two following conditions to be a bona fide generalized pivotal quantity: 1) the distribution of R_{ETAUC} is independent of any unknown parameters; and 2) the observed value of R_{ETAUC} equals to $ETAUC$ for given $\bar{y}_k$ and S_k^2 for $(k = 1, \cdots, K_1, \cdots, K)$.

Given a specific data set Y_{kj}'s where $k = 1, \cdots, K_1, \cdots, K$, and $j = 1, 2, \ldots, n_k$, the generalized confidence interval for $ETAUC$ can be obtained via the following steps:

i) Calculate $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ for $k = 1, \cdots, K_1, \cdots, K$; *ii*) Generate V_k from $\chi^2_{n_k-1}$ and Z_k from standard normal distribution $N(0,1)$, then obtain $R_{\sigma_k^2}$ and R_{μ_k} following (12) and (13); *iii*) Compute R_{ETAUC} following (14); *iv*) Repeat first three steps for a total of $R = 2000$ times to obtain a set of R_{ETAUC}'s values; *v*) Arrange the set of R_{ETAUC}'s from small to large values. Denote $R_{ETAUC}(\alpha)$ as the 100αth percentile of R_{ETAUC}'s. Therefore $(R_{ETAUC}(\alpha/2), R_{ETAUC}(1-\alpha/2))$ is a two-sided $100(1-\alpha)\%$ confidence interval of $ETAUC$.

4.2 Generalized inference (GI): Without normality

In practice, it is common that the normality of the data is not satisfied. For such scenarios, Box-Cox transformation (Box and Cox, 1964) could be applied to achieve normality due to the fact that $ETAUC$ is invariant under monotonic transformation. This type of approach is widely used in ROC analysis for a variety of problems (Faraggi and Reiser, 2002; Fluss *et al.*, 2005; Molodianovitch *et al.*, 2006; Schisterman *et al.*, 2004; Zou and Hall, 2000).

For the i^{th} $(i = 1, \ldots, n_k)$ subject in the k^{th} group $(k = 1, \cdots, K_1, K_1 + 1 \cdots, K)$, a power transformation of the Box-Cox type is suggested as:

$$Y_{ki}^{(\lambda)} = \begin{cases} \frac{Y_{ki}^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ log(Y_{ki}) & \lambda = 0 \end{cases},$$

where it is assumed that $Y_{ki}^{\lambda} \sim N(\mu_k, \sigma_k^2)$. The likelihood function based on all the observations from K groups is:

$$\sum_{k}^{K} \sum_{i}^{n_k} [-\frac{1}{2} log(2\pi\sigma_k^2) - \frac{(Y_{ki}^{\lambda} - \mu_k)^2}{2\sigma_k^2} + (\lambda - 1) log Y_{ki})]. \tag{15}$$

We can obtain the MLE of λ, μ_k and σ_k by maximizing the function in (15). The proposed generalized inference approach in Section 4.1 can be implemented using the transformed data.

4.3 Nonparametric bootstrap (NB) method

The parametric approaches are only appropriate when the normality assumption of either original data or the transformed data is not violated. However, the normality is not always achievable. Therefore, it is necessary to provide a non-parametric approach to estimate the confidence interval of $ETAUC$.

Given an observed data set Y_{kj}'s where $k = 1, \cdots, K_1, \cdots, K$, and $j = 1, 2, \ldots, n_k$, the empirical estimate of $ETAUC$ can be calculated as following

$$\widehat{ETAUC}_{NB} = \frac{1}{\prod_{k=1}^{K} n_k} \sum_{v_1=1}^{n_1} \cdots \sum_{v_K=1}^{n_K} I(max(Y_{1v_1}, \cdots, Y_{K_1 v_{K_1}}) < min(Y_{K_1+1 v_{K_1+1}}, \cdots, Y_{K v_K})). \quad (16)$$

The nonparametric bootstrap confidence interval for $ETAUC$ can be obtained via the following steps: 1) For kth group ($k = 1, 2, \ldots, K$), resample with replacements n_k observations from $y_{k1}, y_{k2}, \ldots, y_{kn_k}$; 2) Obtain $\widehat{ETAUC}$ using the bootstrap samples following (16); 3) Repeat first two steps for a total of R=1000 times to get a set of $\widehat{ETAUC}_{NB}$; 4) Array the set $\widehat{ETAUC}_{NB}$ from small to large values. Define $\widehat{ETAUC}_{NB}(\alpha)$ as the 100αth percentile of nonparametric bootstrap samples of $ETAUC$'s. A two-sided $100(1-\alpha)\%$ nonparametric bootstrap confidence interval estimate of $ETAUC$ is $(\widehat{ETAUC}_{NB}(\alpha/2)$, $\widehat{ETAUC}_{NB}(1-\alpha/2))$.

5 Simulation studies

Simulation studies were conducted to assess and compare the performance of proposed confidence interval estimation by generalized inference (GI) and nonparametric bootstrap (NB) methods for $ETAUC$. A variety of scenarios are considered under a several sample sizes settings. As $(K_1, K) = (2, 4)$, the sample size settings are: $(n_1, n_2, n_3, n_4) =$ $(15, 15, 15, 15), (30, 30, 30, 30), (50, 50, 50, 50), (75, 75, 75, 75), (100, 100, 100, 100), (20, 30, 20, 20)$; and as $(K_1, K) = (2, 5)$, the sample size settings are: $(n_1, n_2, n_3, n_4, n_5) = (15, 15, 15, 15, 15), (30, 30, 30, 30, 3$ For each setting, 2000 independent random samples are generated and the coverage probability (CP) is calculated by the proportion of cases that 95% estimated confidence

intervals contained the true value of $ETAUC$. LT and UT in Table 4-7 stand for the proportion that the true value of $ETAUC$ falls below the lower confidence bound and above the upper confidence bound, respectively. The average of estimated interval length is denoted as MIL in the tables.

Tables 4 and 5 presented simulation results under normal distributions for $(K_1, K) = (2, 4)$ and $(K_1, K) = (2, 5)$ respectively, and Tables 6 and 7 under gamma distributions for $(K_1, K) = (2, 4)$ and $(K_1, K) = (2, 5)$ respectively. All four tables show that when $ETAUC$ gets larger, especially when it is greater than 0.8, the NB method does not perform as well as the GI method at small sample sizes; however, when sample sizes get larger, NB method produces satisfactory coverage probabilities for most scenarios. The GI method generally maintains the nominal level for all the parameter settings, despite the sample sizes. In terms of LT and UT, GI usually has heavier proportion of true value falling above the upper confidence bound whereas NB method has more balanced LT and UT when $ETAUC$ is small. Furthermore, NB has wider confidence intervals compared to GI method when $ETAUC$ is small and the observation is reversed when $ETAUC$ is large.

In summary, when the normality is satisfied, the GI method would be a good choice for confidence interval estimation. When the normality assumptions could not be achieved even with Box-Cox transformation and the sample sizes are not small, the NB method could be applied.

6 Real data analyses

In this section, we reanalyzed the ovarian cancer data set presented in Section 2 (Cramer *et al.*, 2011).

This data set contains four classes: 480 healthy controls, 160 benign cases, 75 early stage cases, and 82 late stage cases. One of the research aims was to evaluate biomarkers' diagnostic accuracy for distinguishing between cancer and non-cancer cancers and rank

these biomarkers accordingly. As reviewed in Section 2, Cramer et al. (Cramer *et al.*, 2011) estimated traditional AUCs based on pooled data. For the research interest in this paper, this data set falls into the framework of extended tree ordering with $K_1 = 2$, and $K = 4$. In this section, we will evaluate the biomarkers using the newly proposed diagnostic measure ($ETAUC$) for extended tree ordering.

Table 8. Estimated $ETAUC$s, corresponding ranking and 95% confidence intervals of the biomarlers in ovarian cancer data set.

Biomarkers	Estimated $ETAUC$	Rank	95% CI	
			LB	UB
CA125	0.705	1	0.634	0.774
CA153	0.401	4	0.325	0.482
CA199	0.243	12	0.187	0.304
KLK6	0.363	6	0.300	0.438
CA724	0.283	10	0.193	0.379
DD.O110	0.325	8	0.256	0.398
DD.C248	0.233	13	0.182	0.291
DD.P108	0.380	5	0.314	0.451
DD.X065	0.257	11	0.196	0.317
HE4	0.586	2	0.502	0.668
SMRP	0.308	9	0.240	0.381
YKL40	0.350	7	0.283	0.423
IGF2	0.458	3	0.385	0.525

LB, lower bound of the 95% confidence interval; UB, upper bound of the 95% confidence interval.

The Shapiro-Wilk test was applied to check normality for the biomarkers under each group, and it was found that none of 13 biomarkers satisfies normality before or after Box-Cox transformation. Table 8 presents the estimated $ETAUC$ with the corresponding ranking and 95% nonparametric confidence intervals. Compared with Table 1, for majority of biomarkers, discrepancies between rankings of based on estimated $ETAUC$ and $NAUC$ are observed. For example, based on estimated $ETAUC$, we observe KLK6 > YKL40, and CA15.3 > CA72.4 where ">" means "better", while these relationships are reversed by estimated $NAUC$ in Table 1.

Furthermore, to investigate the magnitude of sampling error for $ETAUC$, we performed

Table 9. Estimated $ETAUC$ (Rank) under each scenario from Table 2 using random sampling from the ovarian cancer data set (1000 simulation runs).

Scenario	1	2	3	4	5	6	7	8	9
CA125	0.704 (1)	0.705 (1)	0.704 (1)	0.705 (1)	0.703 (1)	0.705 (1)	0.704 (1)	0.706 (1)	0.705 (1)
CA15.3	0.402 (4)	0.405 (4)	0.402 (4)	0.401 (4)	0.400 (4)	0.402 (4)	0.402 (4)	0.403 (4)	0.401 (4)
CA19.9	0.242 (12)	0.242 (12)	0.242 (12)	0.243 (12)	0.243 (12)	0.241 (12)	0.241 (12)	0.243 (12)	0.243 (12)
KLK6	0.364 (6)	0.363 (6)	0.363 (6)	0.364 (6)	0.362 (6)	0.363 (6)	0.364 (6)	0.362 (6)	0.363 (6)
CA72.4	0.283 (10)	0.285 (10)	0.283 (10)	0.281 (10)	0.284 (10)	0.284 (10)	0.281 (10)	0.280 (10)	0.282 (10)
DD.O110	0.326 (8)	0.325 (8)	0.324 (8)	0.326 (8)	0.323 (8)	0.325 (8)	0.326 (8)	0.324 (8)	0.325 (8)
DD.C248	0.234 (13)	0.235 (13)	0.233 (13)	0.232 (13)	0.235 (13)	0.233 (13)	0.234 (13)	0.234 (13)	0.232 (13)
DD.P108	0.381 (5)	0.382 (5)	0.380 (5)	0.380 (5)	0.382 (5)	0.390 (5)	0.381 (5)	0.379 (5)	0.381 (5)
DD.X065	0.257 (11)	0.257 (11)	0.256 (11)	0.255 (11)	0.259 (11)	0.258 (11)	0.258 (11)	0.256 (11)	0.257 (11)
HE4	0.587 (2)	0.587 (2)	0.587 (2)	0.586 (2)	0.586 (2)	0.586 (2)	0.585 (2)	0.586 (2)	0.586 (2)
SMRP	0.306 (9)	0.308 (9)	0.308 (9)	0.308 (9)	0.306 (9)	0.307 (9)	0.308 (9)	0.308 (9)	0.308 (9)
YKL40	0.349 (7)	0.348 (7)	0.350 (7)	0.349 (7)	0.348 (7)	0.350 (7)	0.350 (7)	0.349 (7)	0.351 (7)
IGF2	0.458 (3)	0.457 (3)	0.458 (3)	0.457 (3)	0.457 (3)	0.458 (3)	0.458 (3)	0.457 (3)	0.458 (3)

the similar random sampling study for $ETAUC$ using the settings listed in Table 2. Table 9 presents the estimated $ETAUC$ based on 1000 rounds of repetitions along with the corresponding ranking. Unlike the results of $NAUC$ in Table 3, the estimated $ETAUC$s for all biomarkers barely change across sampling scenarios and they are very close to the estimates in Table 8. It is notable that all ranks remain the same across sampling scenarios, as expected. These observations demonstrate that $ETAUC$ under extended tree ordering could be used as an appropriate diagnostic measure while $NAUC$ could lead to misleading results.

7 Conclusion

Extended tree or umbrella ordering often occurs in the field of cancer research when both healthy group and diseased group involve multiple sub-classes. As a standard practice, researchers often treat the problem of biomarker evaluation as binary classification by pooling all the sub-classes within each group to create two major classes and estimated the traditional area under ROC curve as a diagnostic measure.

This paper investigates the consequences and implication of such pooling approach and demonstrated that such derived AUC based on pooling depends on the relative sampling weights of sub-classes and hence is not convenient to use as a meaningful diagnostic measure.

Therefore, we present an appropriate diagnostic measure, i.e. area under ROC curve for extended tree ordering ($ETAUC$). $ETAUC$ is independent of the relative frequency of sub-classes in each group, and also has a well-defined probability interpretation. Note that when each group only contains one class, $ETAUC$ just becomes the traditional AUC for binary classification.

An R program is available from Dr. Lili Tian at ltian@buffalo.edu.

References

Alonzo, T.A., Nakas, C.T., Yiannoutsos, C.T., and Bucher, S., 2009. A comparison of tests for restricted orderings in the three-class case, *Statistics in Medicine*, 28 (7), 1144–1158.

Box, G.E. and Cox, D.R., 1964. An analysis of transformations, *Journal of the Royal Statistical Society. Series B (Methodological)*, 211–252.

Chadha, K.C., Miller, A., Nair, B.B., Schwartz, S.A., Trump, D.L., and Underwood, W., 2014. New serum biomarkers for prostate cancer diagnosis, *Clinical Cancer Investigation Journal*, 3 (1), 72.

Chudecka-Głaz, A.M., 2015. Roma, an algorithm for ovarian cancer, *Clinica Chimica Acta*, 440, 143–151.

Cramer, D.W., Bast, R.C., Berg, C.D., Diamandis, E.P., Godwin, A.K., Hartge, P., Lokshin, A.E., Lu, K.H., McIntosh, M.W., Mor, G., *et al.*, 2011. Ovarian cancer biomarker performance in prostate, lung, colorectal, and ovarian cancer screening trial specimens, *Cancer Prevention Research*, 4 (3), 365–374.

Dong, T., Attwood, K., Hutson, A., Liu, S., and Tian, L., 2017. A new diagnostic accuracy measure and cut-point selection criterion, *Statistical Methods in Medical Research*, 26 (6), 2832–2852.

Dong, T., Tian, L., Hutson, A., and Xiong, C., 2011. 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.

Faraggi, D. and Reiser, B., 2002. Estimation of the area under the roc curve, *Statistics in Medicine*, 21 (20), 3093–3106.

Fluss, R., Faraggi, D., and Reiser, B., 2005. Estimation of the youden index and its associated cutoff point, *Biometrical Journal*, 47 (4), 458–472.

Krishnamoorthy, K., Lin, Y., and Xia, Y., 2009. Confidence limits and prediction limits for a weibull distribution based on the generalized variable approach, *Journal of Statistical Planning and Inference*, 139 (8), 2675–2684.

Krishnamoorthy, K. and Mathew, T., 2003. Inferences on the means of lognormal distributions using generalized p-values and generalized confidence intervals, *Journal of Statistical Planning and Inference*, 115 (1), 103–121.

Lai, C.Y., Tian, L., and Schisterman, E.F., 2012. Exact confidence interval estimation for the youden index and its corresponding optimal cut-point, *Computational Statistics & Data Analysis*, 56 (5), 1103–1114.

Leichtle, A.B., Ceglarek, U., Weinert, P., Nakas, C.T., Nuoffer, J.M., Kase, J., Conrad, T., Witzigmann, H., Thiery, J., and Fiedler, G.M., 2013. Pancreatic carcinoma, pancreatitis, and healthy controls: metabolite models in a three-class diagnostic dilemma, *Metabolomics*, 9 (3), 677–687.

Li, C.R., Liao, C.T., and Liu, J.P., 2008. A non-inferiority test for diagnostic accuracy based on the paired partial areas under roc curves, *Statistics in Medicine*, 27 (10), 1762–1776.

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.

Li, J., Jiang, B., and Fine, J.P., 2012. Multicategory reclassification statistics for assessing improvements in diagnostic accuracy, *Biostatistics*, 14 (2), 382–394.

Lin, S., Lee, J.C., and Wang, R., 2007. Generalized inferences on the common mean vector of several multivariate normal populations, *Journal of Statistical Planning and Inference*, 137 (7), 2240–2249.

Molodianovitch, K., Faraggi, D., and Reiser, B., 2006. Comparing the areas under two correlated roc curves: parametric and non-parametric approaches, *Biometrical Journal*, 48 (5), 745–757.

Morris, J.C., Storandt, M., Miller, J.P., McKeel, D.W., Price, J.L., Rubin, E.H., and Berg, L., 2001. Mild cognitive impairment represents early-stage alzheimer disease, *Archives of Neurology*, 58 (3), 397–405.

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

Nakas, C.T. and Alonzo, T.A., 2007. Roc graphs for assessing the ability of a diagnostic marker to detect three disease classes with an umbrella ordering, *Biometrics*, 63 (2), 603–609.

Nakas, C.T., Alonzo, T.A., and Yiannoutsos, C.T., 2010. Accuracy and cut-off point selection in three-class classification problems using a generalization of the youden index, *Statistics in Medicine*, 29 (28), 2946–2955.

Nakas, C.T., Dalrymple-Alford, J.C., Anderson, T.J., and Alonzo, T.A., 2013. Generalization of youden index for multiple-class classification problems applied to the assessment of externally validated cognition in parkinson disease screening, *Statistics in Medicine*, 32 (6), 995–1003.

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

Partheen, K., Kristjansdottir, B., and Sundfeldt, K., 2011. Evaluation of ovarian cancer biomarkers he4 and ca-125 in women presenting with a suspicious cystic ovarian mass, *Journal of Gynecologic Oncology*, 22 (4), 244–252.

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

Sampat, M.P., Patel, A.C., Wang, Y., Gupta, S., Kan, C.W., Bovik, A.C., and Markey, M.K., 2009. Indexes for three-class classification performance assessmentan empirical comparison, *IEEE Transactions on Information Technology in Biomedicine*, 13 (3), 300–312.

Schisterman, E.F., Faraggi, D., and Reiser, B., 2004. Adjusting the generalized roc curve for covariates, *Statistics in Medicine*, 23 (21), 3319–3331.

Scinto, L.F. and Daffner, K.R., 2000. *Early Diagnosis of Alzheimers Disease*, Springer Science & Business Media.

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

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

Shapiro, D.E., 1999. The interpretation of diagnostic tests, *Statistical Methods in Medical Research*, 8 (2), 113–134.

Sun, R., Medeiros, L.J., and Young, K.H., 2016. Diagnostic and predictive biomarkers for lymphoma diagnosis and treatment in the era of precision medicine, *Modern Pathology*, 29 (10), 1118.

Tian, L. and Cappelleri, J.C., 2004. A new approach for interval estimation and hypothesis testing of a certain intraclass correlation coefficient: the generalized variable method, *Statistics in Medicine*, 23 (13), 2125–2135.

Tian, L., Xiong, C., Lai, C.Y., and Vexler, A., 2011. Exact confidence interval estimation for the difference in diagnostic accuracy with three ordinal diagnostic groups, *Journal of Statistical Planning and Inference*, 141 (1), 549–558.

Tsui, K.W. and Weerahandi, S., 1989. Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters, *Journal of the American Statistical Association*, 84 (406), 602–607.

Wang, D., Attwood, K., and Tian, L., 2016. Receiver operating characteristic analysis under tree orderings of disease classes, *Statistics in Medicine*, 35 (11), 1907–1926.

Weerahandi, S., 1995. Generalized confidence intervals, *in: Exact Statistical Methods for Data Analysis*, Springer, 143–168.

Weerahandi, S., 2004. *Generalized Inference in Repeated Measures: Exact Methods in MANOVA and Mixed Models*, vol. 500, John Wiley & Sons.

Weerahandi, S., 2013. *Exact Statistical Methods for Data Analysis*, Springer Science & Business Media.

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.

Xiong, C., van Belle, G., Miller, J.P., Yan, Y., Gao, F., Yu, K., and Morris, J.C., 2007. A parametric comparison of diagnostic accuracy with three ordinal diagnostic groups, *Biometrical Journal*, 49 (5), 682–693.

Yin, J. and Tian, L., 2014a. Joint confidence region estimation for area under roc curve and youden index, *Statistics in Medicine*, 33 (6), 985–1000.

Yin, J. and Tian, L., 2014b. Joint inference about sensitivity and specificity at the optimal cut-off point associated with youden index, *Computational Statistics & Data Analysis*, 77, 1–13.

Youden, W.J., 1950. Index for rating diagnostic tests, *Cancer*, 3 (1), 32–35.

Zhang, Y. and Li, J., 2011. Combining multiple markers for multi-category classification: An roc surface approach, *Australian & New Zealand Journal of Statistics*, 53 (1), 63–78.

Zhou, X.H., McClish, D.K., and Obuchowski, N.A., 2011. *Statistical Methods in Diagnostic Medicine*, John Wiley & Sons.

Zhu, C.S., Pinsky, P.F., Cramer, D.W., Ransohoff, D.F., Hartge, P., Pfeiffer, R.M., Urban, N., Mor, G., Bast, R.C., Moore, L.E., *et al.*, 2011. A framework for evaluating biomarkers for early detection: validation of biomarker panels for ovarian cancer, *Cancer Prevention Research*, 4 (3), 375–383.

Zou, K.H. and Hall, W., 2000. Two transformation models for estimating an roc curve derived from continuous data, *Journal of Applied Statistics*, 27 (5), 621–631.

Zou, K.H., Liu, A., Bandos, A.I., Ohno-Machado, L., and Rockette, H.E., 2011. *Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis*, Chapman and Hall/CRC.

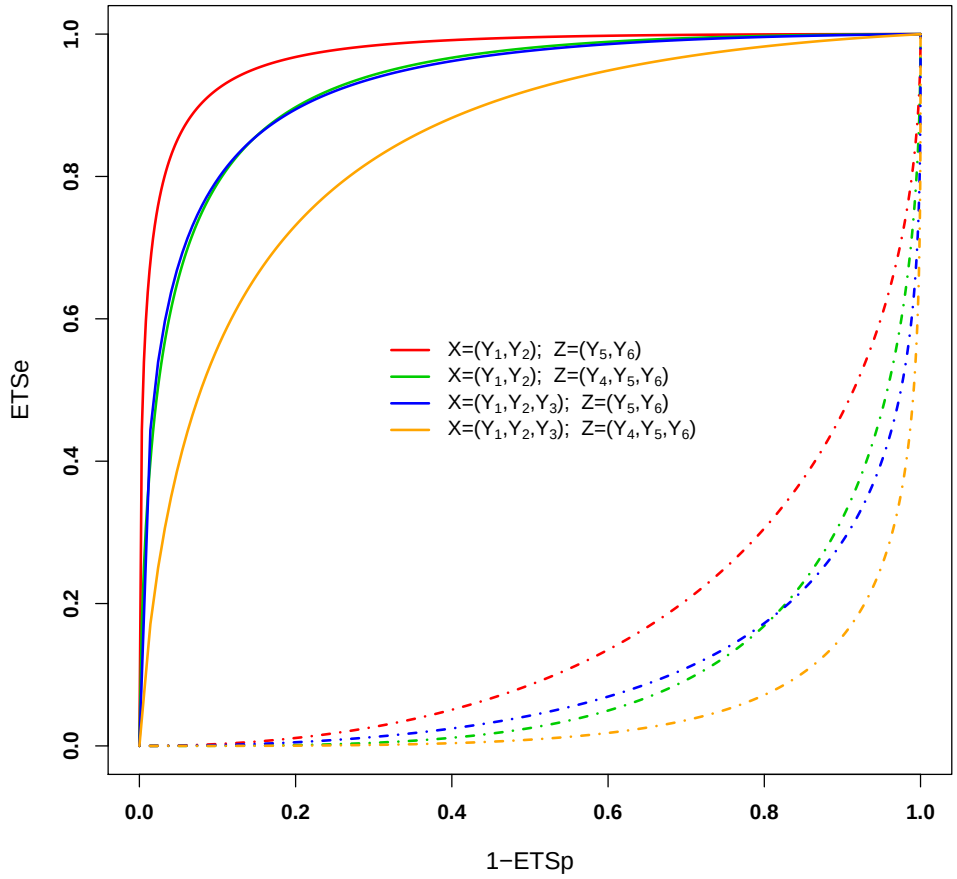

Figure 1: ETROC curves (solid lines) and the corresponding chance curves (dash-dot lines, same color) of $X \preceq Z$ for a variety of cases where $Y_1 \sim N(0,1), Y_2 \sim N(0.5,1), Y_3 \sim N(1,1), Y_4 \sim N(3,1), Y_5 \sim N(3.5,1), Y_6 \sim N(4,1)$.

Table 4. Estimated coverage probabilities for 95% confidence intervals for $ETAUC$ under $(Y_1, Y_2) \preceq (Y_3, Y_4)$ with normal distribution settings based on 2000 simulations.

$(Y_1, Y_2) \preceq (Y_3, Y_4)$								
	GI	NB	GI	NB	GI	NB	GI	NB
Sample size	CP		LT		UT		MIL	
$Y_1 \sim N(0,1); Y_2 \sim N(0.3,1); Y_3 \sim N(1,1); Y_4 \sim N(1.3,1);\ ETAUC = 0.4545$								
(15,15,15,15)	0.961	0.944	0.008	0.028	0.031	0.029	0.370	0.381
(30,30,30,30)	0.959	0.955	0.013	0.024	0.029	0.022	0.260	0.268
(50,50,50,50)	0.956	0.948	0.015	0.028	0.029	0.025	0.201	0.208
(75,75,75,75)	0.952	0.955	0.013	0.022	0.035	0.024	0.164	0.169
(100,100,100,100)	0.957	0.950	0.011	0.019	0.033	0.032	0.142	0.146
(20,30,20,20)	0.951	0.942	0.010	0.030	0.040	0.029	0.291	0.299
$Y_1 \sim N(0,1); Y_2 \sim N(0.5,1.1); Y_3 \sim N(2,1.2); Y_4 \sim N(3,1.3);\ ETAUC = 0.7258$								
(15,15,15,15)	0.959	0.938	0.005	0.050	0.036	0.013	0.366	0.347
(30,30,30,30)	0.959	0.940	0.006	0.041	0.036	0.020	0.249	0.246
(50,50,50,50)	0.957	0.955	0.011	0.027	0.033	0.018	0.190	0.191
(75,75,75,75)	0.946	0.952	0.015	0.030	0.040	0.019	0.154	0.156
(100,100,100,100)	0.952	0.950	0.012	0.027	0.037	0.024	0.133	0.135
(20,30,20,20)	0.954	0.930	0.005	0.048	0.041	0.022	0.285	0.277
$Y_1 \sim N(0,1); Y_2 \sim N(1,1); Y_3 \sim N(3,1.2); Y_4 \sim N(4,1.4);\ ETAUC = 0.8551$								
(15,15,15,15)	0.947	0.919	0.004	0.075	0.050	0.007	0.302	0.257
(30,30,30,30)	0.952	0.938	0.006	0.049	0.043	0.014	0.196	0.184
(50,50,50,50)	0.955	0.943	0.010	0.043	0.036	0.015	0.146	0.143
(75,75,75,75)	0.946	0.952	0.014	0.033	0.041	0.016	0.118	0.118
(100,100,100,100)	0.948	0.949	0.011	0.031	0.041	0.020	0.101	0.102
(20,30,20,20)	0.951	0.922	0.004	0.065	0.046	0.014	0.225	0.204

CP: coverage probability; LT: one side coverage probability (lower tail); UT: one side coverage probability (upper tail); MIL:mean of the interval lengths.

Table 5. Estimated coverage probabilities for 95% confidence intervals for $ETAUC$ under $(Y_1, Y_2) \preceq (Y_3, Y_4, Y_5)$ with normal distribution settings based on 2000 simulations.

$(Y_1, Y_2) \preceq (Y_3, Y_4, Y_5)$								
	GI	NB	GI	NB	GI	NB	GI	NB
Sample size	CP		LT		UT		MIL	
$Y_1 \sim N(0,1); Y_2 \sim N(0.5,1); Y_3 \sim N(1,1); Y_4 \sim N(1.5,1); Y_5 \sim N(1.6,1);\ ETAUC = 0.3781$								
(15,15,15,15,15)	0.956	0.950	0.006	0.024	0.039	0.026	0.340	0.357
(30,30,30,30,30)	0.953	0.956	0.012	0.021	0.036	0.024	0.240	0.252
(50,50,50,50,50)	0.954	0.951	0.011	0.021	0.035	0.028	0.186	0.195
(75,75,75,75,75)	0.954	0.958	0.013	0.018	0.034	0.024	0.152	0.159
(100,100,100,100,100)	0.952	0.947	0.014	0.022	0.035	0.031	0.132	0.138
(20,30,20,25,30)	0.949	0.945	0.011	0.025	0.041	0.030	0.266	0.278
$Y \sim 1\ N(0,1); Y_2 \sim N(0.5,1); Y_3 \sim N(2,1.1); Y_4 \sim N(2.5,1.2); Y_5 \sim N(3,1.3);\ ETAUC = 0.6822$								
(15,15,15,15,15)	0.943	0.943	0.004	0.041	0.054	0.017	0.376	0.360
(30,30,30,30,30)	0.951	0.946	0.006	0.036	0.044	0.019	0.257	0.255
(50,50,50,50,50)	0.952	0.956	0.007	0.029	0.042	0.016	0.196	0.198
(75,75,75,75,75)	0.947	0.950	0.011	0.028	0.043	0.023	0.159	0.162
(100,100,100,100,100)	0.950	0.950	0.010	0.029	0.041	0.021	0.137	0.140
(20,30,20,25,30)	0.946	0.935	0.003	0.043	0.052	0.023	0.289	0.284
$Y_1 \sim N(0,1); Y_2 \sim N(1,1.1); Y_3 \sim N(3,1.2); Y_4 \sim N(4,1.3); Y_5 \sim N(5,1.4);\ ETAUC = 0.8437$								
(15,15,15,15,15)	0.942	0.920	0.004	0.071	0.054	0.010	0.313	0.268
(30,30,30,30,30)	0.95	0.9405	0.004	0.046	0.047	0.014	0.203	0.191
(50,50,50,50,50)	0.955	0.947	0.008	0.040	0.038	0.014	0.152	0.149
(75,75,75,75,75)	0.945	0.946	0.013	0.034	0.043	0.020	0.122	0.122
(100,100,100,100,100)	0.951	0.946	0.008	0.034	0.042	0.021	0.105	0.106
(20,30,20,25,30)	0.943	0.918	0.003	0.067	0.054	0.016	0.233	0.212

CP: coverage probability; LT: one side coverage probability (lower tail); UT: one side coverage probability (upper tail); MIL:mean of the interval lengths.

Table 6. Estimated coverage probabilities for 95% confidence intervals for $ETAUC$ under $(Y_1, Y_2) \preceq (Y_3, Y_4)$ with gamma distribution settings based on 2000 simulations.

$(Y_1, Y_2) \preceq (Y_3, Y_4)$								
	GI	NB	GI	NB	GI	NB	GI	NB
Sample size	CP		LT		UT		MIL	
$Y1 \sim Gamma(2,6); Y2 \sim Gamma(2,5); Y3 \sim Gamma(4,6); Y4 \sim Gamma(4,5);\ ETAUC = 0.5651$								
(15,15,15,15)	0.964	0.946	0.007	0.033	0.029	0.022	0.385	0.388
(30,30,30,30)	0.956	0.950	0.010	0.032	0.035	0.019	0.268	0.273
(50,50,50,50)	0.937	0.938	0.018	0.033	0.045	0.030	0.207	0.211
(75,75,75,75)	0.942	0.934	0.025	0.039	0.033	0.028	0.168	0.173
(100,100,100,100)	0.947	0.946	0.024	0.031	0.030	0.024	0.146	0.150
(20,30,20,20)	0.962	0.944	0.009	0.034	0.030	0.023	0.303	0.306
$Y1 \sim Gamma(2,6); Y2 \sim Gamma(3,6); Y3 \sim Gamma(5,5); Y4 \sim Gamma(6,5);\ ETAUC = 0.7393$								
(15,15,15,15)	0.961	0.943	0.003	0.045	0.037	0.013	0.361	0.339
(30,30,30,30)	0.959	0.945	0.012	0.040	0.030	0.015	0.244	0.239
(50,50,50,50)	0.942	0.940	0.014	0.037	0.045	0.024	0.186	0.186
(75,75,75,75)	0.951	0.939	0.021	0.043	0.029	0.018	0.151	0.152
(100,100,100,100)	0.945	0.939	0.018	0.031	0.037	0.030	0.130	0.132
(20,30,20,20)	0.960	0.943	0.006	0.046	0.035	0.012	0.275	0.265
$Y1 \sim Gamma(2,6); Y2 \sim Gamma(4,5); Y3 \sim Gamma(7,4); Y4 \sim Gamma(8,4);\ ETAUC = 0.8631$								
(15,15,15,15)	0.955	0.915	0.001	0.076	0.045	0.010	0.293	0.249
(30,30,30,30)	0.950	0.932	0.012	0.054	0.039	0.014	0.190	0.178
(50,50,50,50)	0.945	0.935	0.014	0.046	0.042	0.019	0.143	0.139
(75,75,75,75)	0.945	0.936	0.023	0.047	0.033	0.018	0.114	0.113
(100,100,100,100)	0.955	0.946	0.016	0.039	0.030	0.016	0.098	0.098
(20,30,20,20)	0.957	0.929	0.009	0.062	0.035	0.010	0.211	0.190

CP: coverage probability; LT: one side coverage probability (lower tail); UT: one side coverage probability (upper tail); MIL:mean of the interval lengths.

Table 7. Estimated coverage probabilities for 95% confidence intervals for $ETAUC$ under $(Y_1, Y_2) \preceq (Y_3, Y_4, Y_5)$ with gamma distribution settings based on 2000 simulations.

$(Y_1, Y_2) \preceq (Y_3, Y_4, Y_5)$								
	GI	NB	GI	NB	GI	NB	GI	NB
Sample size	CP		LT		UT		MIL	
$Y_1 \sim Gamma(2,6); Y_2 \sim Gamma(2,5); Y_3 \sim Gamma(4,6); Y_4 \sim Gamma(5,6); Y_5 \sim Gamma(6,5);\ ETAUC = 0.5497$								
(15,15,15,15,15)	0.965	0.945	0.004	0.032	0.032	0.024	0.383	0.388
(30,30,30,30,30)	0.957	0.949	0.009	0.033	0.035	0.019	0.267	0.274
(50,50,50,50,50)	0.943	0.940	0.014	0.031	0.044	0.029	0.206	0.212
(75,75,75,75,75)	0.945	0.935	0.021	0.040	0.035	0.026	0.168	0.173
(100,100,100,100,100)	0.949	0.949	0.019	0.029	0.033	0.023	0.145	0.150
(20,30,20,25,30)	0.955	0.945	0.007	0.033	0.038	0.023	0.303	0.308
$Y_1 \sim Gamma(2,6); Y_2 \sim Gamma(3,6); Y_3 \sim Gamma(5,5); Y_4 \sim Gamma(5.5,5); Y_5 \sim Gamma(6,5);\ ETAUC = 0.6716$								
(15,15,15,15,15)	0.950	0.943	0.002	0.044	0.049	0.014	0.378	0.365
(30,30,30,30,30)	0.948	0.950	0.007	0.033	0.046	0.018	0.259	0.258
(50,50,50,50,50)	0.940	0.941	0.009	0.033	0.052	0.026	0.198	0.201
(75,75,75,75,75)	0.948	0.938	0.013	0.039	0.040	0.024	0.161	0.164
(100,100,100,100,100)	0.931	0.939	0.014	0.032	0.056	0.030	0.139	0.142
(20,30,20,25,30)	0.955	0.946	0.002	0.039	0.044	0.016	0.287	0.282
$Y_1 \sim Gamma(2,6); Y_2 \sim Gamma(3,5); Y_3 \sim Gamma(6,5); Y_4 \sim Gamma(7,4); Y_5 \sim Gamma(8,4);\ ETAUC = 0.8007$								
(15,15,15,15,15)	0.953	0.934	0.002	0.057	0.045	0.010	0.337	0.304
(30,30,30,30,30)	0.954	0.942	0.005	0.050	0.042	0.009	0.223	0.216
(50,50,50,50,50)	0.940	0.947	0.013	0.036	0.048	0.017	0.170	0.168
(75,75,75,75,75)	0.945	0.942	0.019	0.044	0.037	0.015	0.136	0.137
(100,100,100,100,100)	0.943	0.943	0.015	0.034	0.043	0.024	0.117	0.119
(20,30,20,25,30)	0.952	0.938	0.004	0.050	0.045	0.012	0.253	0.237

CP: coverage probability; LT: one side coverage probability (lower tail); UT: one side coverage probability (upper tail); MIL:mean of the interval lengths.