Computational Statistics and Data Analysis 82 (2015) 186–198

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Empirical likelihood ratio confidence interval estimation of best linear combinations of biomarkers

Xiwei Chen, Albert Vexler*, Marianthi Markatou

Department of Biostatistics, University at Buffalo, USA

ARTICLE INFO

Article history:
Received 3 April 2014
Received in revised form 5 September 2014
Accepted 9 September 2014
Available online 18 September 2014

Keywords:
Area under the ROC curve
Best linear combination
Empirical likelihood
Kernel
Receiver operating characteristic curve (ROC)

ABSTRACT

A novel smoothed empirical likelihood (EL) approach that incorporates kernel estimation of the area under the receiver operating characteristic curve (AUC) to construct nonparametric confidence intervals of AUC based on the best linear combination (BLC) of biomarkers is proposed. The method has several advantages including the feasibility to use gradient-based techniques for fast computation of BLC coefficients and to employ powerful likelihood methods without specification of underlying data distributions. Simulation results show that the new method performs well even when the distribution of biomarkers is skewed, a situation commonly met in practice. A data set from a clinical experiment related to atherosclerotic coronary heart disease is used to illustrate the efficiency of the proposed method.

1. Introduction

The receiver operating characteristic (ROC) curve methodology is often applied to evaluate the performance of diagnostic markers that classify subjects into two populations: diseased and nondiseased. Suppose that a biomarker, measured on a continuous scale, is performed on n diseased subjects, yielding independent, identically distributed (i.i.d.) measurements $\mathbf{X}_i,\ i = 1, \dots, n$, with distribution function F, and on m nondiseased subjects, yielding i.i.d. measurements $\mathbf{Y}_j,\ j = 1, \dots, m$, with distribution function G. The ROC curve is a plot of $R(t) = 1 - F\{G^{-1}(1-t)\}$ against t, $0 \leq t \leq 1$ (Pepe, 1997; Faraggi and Reiser, 2002). The area under the ROC curve (AUC), as a common measure of the diagnostic performance of a biomarker, is equal to $P(X > Y)$ (Bamber, 1975).

In practice, different markers are usually related to the disease in various magnitudes and different directions. For example, low levels of high density lipoprotein (HDL)-cholesterol and high levels of thiobarbuturic acid reacting substances (TBARS) are biomarkers of oxidative stress and antioxidant status, as indicators of coronary heart disease (Schisterman et al., 2002). When multiple biomarkers are available, we are interested in seeking a simple best linear combination (BLC) of biomarkers such that the combined score achieves the maximum AUC over all possible linear combinations. The maximum AUC measures the ability to discriminate between the control and the disease groups (Pepe and Thompson, 2000; McIntosh and Pepe, 2002).

A variety of publications have focused on the BLC of multiple biomarkers in both parametric and nonparametric cases. Su and Liu (1993) investigated the BLC and the corresponding maximum AUC under multivariate normal assumptions. Based on the point estimator given by Su and Liu (1993), Reiser and Faraggi (1997) derived confidence intervals for the BLC-based AUC under normal assumptions, which is useful when the sample size in the two groups is small and equal or moderate

* Correspondence to: 715 Kimball Tower, 3435 Main St, Buffalo, NY 14214, USA. Tel.: +1 716 829 2754; fax: +1 716 829 2200.
E-mail addresses: xiweiche@buffalo.edu (X. Chen), avexler@buffalo.edu (A. Vexler), markatou@buffalo.edu (M. Markatou).

http://dx.doi.org/10.1016/j.csda.2014.09.010
0167-9473/© 2014 Elsevier B.V. All rights reserved.

and unequal, respectively. In practice, however, the distributions of biomarker measurements commonly deviate from the normal assumptions (Limpert et al., 2001). For example, epidemiological studies have demonstrated that TBARS has a distribution with heavy tails (Schisterman et al., 2001). In this case, loss of efficiency of the BLC coefficients constructed under normality assumptions can be expected and alternative approaches are needed to accommodate different data distributions.

Pepe et al. (2006) considered the empirical AUC and noticed that the empirical BLC estimator is consistent under the generalized linear model assumption. They noted that sophisticated computational algorithms are required because the empirical AUC is not a continuous function. Ma and Huang (2005, 2007) also noticed the computational complexity in maximizing the empirical AUC and proposed a corresponding sigmoid approximation. Under the generalized linear model assumption, it is proved that the sigmoid AUC converges to the maximized theoretical AUC and the corresponding BLC coefficients are asymptotically normally distributed (Ma and Huang, 2007). Instead of the specific sigmoid AUC, Vexler et al. (2006) demonstrated consistency, uniqueness, and ease of computation of the kernel-smoothed AUC estimator. Furthermore, they provided an upper confidence bound for the maximum AUC. The distribution-free confidence interval estimation of the maximum AUC remains unsolved.

As a nonparametric alternative to the optimal parametric likelihood method (Markatou et al., 1998), the EL methodology (Owen, 1990, 2001) is an efficient way to construct confidence intervals. In the context of a single biomarker, Qin and Zhou (2006) proposed an EL-based confidence interval for the AUC and demonstrated that it outperforms the existing normal approximation-based intervals and bootstrap intervals, particularly when the AUC is close to one.

In this paper, we extend and modify the method of Qin and Zhou (2006) to adjust for multiple biomarkers. Incorporating a kernel distribution estimation (Azzalini, 1981; Nadaraya, 1964) into the EL-based confidence interval estimation, we propose smoothed EL-based confidence intervals for BLC-based AUCs. The advantages of the proposed method include: (1) the smooth estimate of the AUC is differentiable, making gradient-based methods feasible and the computation of BLC coefficients simple, fast and unique; (2) it is robust to underlying distributions of biomarkers' measurements; (3) it employs the EL methodology, that is known to be an approximate nonparametric most powerful tool based on likelihood ratios (Lazar and Mykland, 1998); (4) the EL-based confidence interval is range preserving (Hall and La Scala, 1990) and therefore always lies between 0.5 and 1. The proposed method is illustrated with a study of the accuracy of biomarkers related to the atherosclerotic coronary heart disease. An EL interval for AUC is derived based on a linear combination of measurements of four important biomarkers related to the atherosclerotic coronary heart disease including lutein, TBARS, HDL cholesterol and uric acid, obtained from a 12-hour fasting blood specimen for biochemical analysis at baseline.

The paper is organized as follows. Section 2 develops the smoothed EL-based confidence interval for the BLC-based AUC. Section 3.2 presents simulation studies to compare the relative performance of the proposed smoothed EL-based interval with the existing intervals for the BLC-based AUC. In Section 4, we apply our method to a real example related to atherosclerotic coronary heart disease. Section 5 presents a broader discussion on deriving linear combinations of biomarkers to improve the diagnostic accuracy. Conditions and proofs are deferred to Appendix.

2. Methods

Consider a study with d continuous-scale biomarkers yielding measurements $\mathbf{X}_i = (X_{1i}, \dots, X_{di})^T,\ i = 1, \dots, n$, on n diseased subjects and measurements $\mathbf{Y}_j = (Y_{1j}, \dots, Y_{dj})^T,\ j = 1, \dots, m$, on m nondiseased patients, respectively. We are interested in constructing effective one-dimensional combined scores of biomarkers' measurements, say, $X(\mathbf{a}) = \mathbf{a}^T\mathbf{X}$ and $Y(\mathbf{a}) = \mathbf{a}^T\mathbf{Y}$, such that the AUC based on these scores is maximized over all possible linear combinations of biomarkers. Define $A(\mathbf{a}) = P(X(\mathbf{a}) > Y(\mathbf{a}))$; the statistical problem is to estimate the maximum AUC defined as $A = A(\mathbf{a}_0)$, where $\mathbf{a}_0$ are the BLC coefficients satisfying $\mathbf{a}_0 = \arg\max_{\mathbf{a}} A(\mathbf{a})$. For simplicity, we assume that the first component of the vector $\mathbf{a}$ equals 1 throughout the paper; see Pepe et al. (2006) and Ma and Huang (2007).

2.1. Confidence interval for a single biomarker-based AUC

Qin and Zhou (2006) developed an EL-approach for constructing confidence intervals for the AUC using a single biomarker (in our notation $d = 1$ and $\mathbf{a} = \mathbf{a}_0 = 1$). The confidence interval estimation is executed via construction of the empirical likelihood ratio (ELR) test statistic for testing the hypothesis $H_0 : A = A_0$ versus $H_a : A \neq A_0$.

Qin and Zhou (2006) based the ELR test statistic on the concept of placement value of a diseased subject (Pepe and Cai, 2004). The placement value is defined as $U = 1 - G(X)$, where U can be interpreted as the proportion of non-diseased subjects with their biomarker measurements greater than X. Noticing that $E(1 - U) = A$, where A denotes the AUC based on a single marker, EL inference for the AUC for the case of a single marker is developed. Specifically, let $\mathbf{p} = (p_1, p_2, \dots, p_n)^T$ be a probability weight vector, $\sum_{i=1}^n p_i = 1$ and $p_i \geq 0$ for all $i = 1, \dots, n$. The profile EL for the AUC, evaluated at the true value A_0 of AUC, can be defined as

$$\tilde{L}(A_0) = \sup\left\{\prod_{i=1}^n p_i : \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i(1 - U_i) = A_0\right\}, \tag{1}$$

where $U_i = 1 - G(X_i),\ i = 1, \dots, n$. When the distribution function G of the nondiseased population is unknown, the empirical distribution $\hat{G}$ is used. Accordingly, replacing U_i by its estimator $\hat{U}_i = 1 - \hat{G}(X_i)$ in Eq. (1), and using Lagrange