Computational Statistics and Data Analysis 117 (2018) 76–89

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Informativeness of diagnostic marker values and the impact of data grouping☆

Hua Ma [a,1], Andriy I. Bandos [a,*], David Gur [b]

[a] Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15216, USA
[b] Department of Radiology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15216, USA

ARTICLE INFO

Article history:
Received 17 September 2016
Received in revised form 25 July 2017
Accepted 26 July 2017
Available online 8 August 2017

Keywords:
Biomarker evaluation
Diagnostically non-informative values
Grouped marker values
Receiver Operating Characteristic (ROC) analysis
Statistical efficiency

ABSTRACT

Assessing performance of diagnostic markers is a necessary step for their use in decision making regarding various conditions of interest in diagnostic medicine and other fields. Globally useful markers could, however, have ranges of values that are "*diagnostically non-informative*". This paper demonstrates that the presence of marker values from diagnostically non-informative ranges could lead to a loss in statistical efficiency during nonparametric evaluation and shows that grouping non-informative values provides a natural resolution to this problem. These points are theoretically proven and an extensive simulation study is conducted to illustrate the possible benefits of using grouped marker values in a number of practically reasonable scenarios. The results contradict the common conjecture regarding the detrimental effect of grouped marker values during performance assessments. Specifically, contrary to the common assumption that grouped marker values lead to bias, grouping non-informative values does not introduce bias and could substantially reduce sampling variability. The proven concept that grouped marker values could be statistically beneficial without detrimental consequences implies that in practice, tied values do not always require resolution whereas the use of continuous diagnostic results without addressing diagnostically non-informative ranges could be statistically detrimental. Based on these findings, more efficient methods for evaluating diagnostic markers could be developed.

1. Introduction

Advances in medicine and other diagnostic fields heavily rely on the availability of adequate tools to assess the presence of a specific underlying condition of interest (e.g., disease) which may be related to the current health status of a patient and/or to a response to an intervention. Biomarkers, diagnostic tests and new or improved technologies require diagnostic accuracy evaluation as a part of development, optimization, and regulatory approval. Data for analyses of diagnostic accuracy consist of marker values (e.g., ratings, diagnostic results) collected for a fixed number of "diseased" (i.e., with the condition of interest) and "non-diseased" subjects. The diagnostic usefulness of a marker is determined by differences in the distributions of the marker values for diseased and non-diseased subjects, which can be captured by the Receiver Operating Characteristics (ROC) curve (Zhou et al., 2011; Pepe, 2003). The ROC curve and associated indices allow one to focus specifically on diagnostic

☆ Additional numerical results, and code for examples, are provided in Supplementary Material in the electronic version of the manuscript.

* Correspondence to: 7137 Parran Hall ,130 DeSoto, Str., Pittsburgh, PA 15146, USA
E-mail address: anb61@pitt.edu (A.I. Bandos).

[1] Present address: Translational Biomarkers, Early Clinical Development Statistics, Merck Research Laboratories, North Wales, PA 19454, USA.

http://dx.doi.org/10.1016/j.csda.2017.07.008
0167-9473/

accuracy aside from other characteristics of marker values (e.g., means, variances). Standard statistical procedures for logistic regression (e.g., proc logistic, SAS v.9.4, Cary NC) also include ROC tools that enable the assessments of diagnostic accuracy of continuous markers without the restrictions imposed by odds ratios.

The typical set-up for evaluating diagnostic accuracy includes marker values for n_0 "non-diseased" subjects x_i, $i = 1, \ldots, n_0$ and for n_1"diseased" subjects y_j, $j = 1, \ldots, n_1$. In the target population, diagnostic results can be characterized with survival distribution functions of $S_X(x) = Pr(X > x)$, and $S_Y(y) = Pr(Y > y)$, respectively. When evaluated at a given threshold ξ these functions are called correspondingly 1-*specificity*, or *False Positive Fraction* ($FPF(\xi)$) and *sensitivity*, or *True Positive Fraction* ($TPF(\xi)$). The corresponding Receiver Operating Characteristic (ROC) curve is formed by points $(FPF(\xi), TPF(\xi))$, and it has an explicit formulation of $ROC(e) = S_Y\left(S_X^{-1}(e)\right)$, where e swaps values from 0 to 1. We note that throughout this manuscript we use capital letters to emphasize either the random nature of a quantity or its functional role (as opposed to observed/fixed values).

The diagonal line ($ROC(e) = e$) represents the ROC curve of a useless marker (or a pure guessing process). The discrepancies between a marker's ROC curve and the diagonal line reflect the difference between the distributions of the marker values for diseased and non-diseased subjects. Conventional ROC approaches for evaluating the diagnostic usefulness of a marker often have analogs among standard statistical approaches for comparing two distributions. In particular, the maximum difference between the ROC curve and a diagonal line is quantified by the Youden's index which is directly related to the Kolmogorov–Smirnov statistic (Pepe, 2003). The area under the ROC curve (AUC) is the most frequently used summary index of diagnostic performance $A = \int_0^1 ROC(e)\, de$ and has a natural relationship to the Wilcoxon statistic (Hanley and McNeil 1982). The partial AUC (pAUC) often offers a more relevant, and sometimes more efficient, assessment of diagnostic performance by focusing on the range of operating points which are of interest in the specific applications being considered (McClish, 1989; Wieand et al., 1989; Ma et al., 2013).

Statistical analysis of ROC data can be implemented using a number of approaches including parametric, non-parametric, and semi-parametric types of the frequentist inferences (Pepe, 2003; Zhou et al., 2011) as well as Bayesian approaches (Peng and Hall, 1996; Zou and Hall, 2000; Erkanli et al., 2006; Gu et al., 2008). These approaches have known advantages and drawbacks. In this paper we focus on non-parametric approaches that employ minimum assumptions about the underlying data and rely only on the empirical ROC points.

Overall diagnostic accuracy of a marker is driven by the overall differences in the distributions of marker values for diseased and non-diseased subjects, however not all ranges of values of a generally useful marker are equally informative. In this work we formalize the *diagnostic informativeness* of a range of marker values and prove that the nonparametric statistical assessment of diagnostic accuracy performed using marker values from diagnostically non-informative ranges leads to a loss of statistical efficiency. Moreover, we demonstrate that a natural way to handle diagnostic non-informativeness is provided by simple grouping of marker values. In the next two sections we provide the theoretical justifications for this concept and present a simulation study illustrating the quantitative losses/gains in precision in a range of scenarios typical for diagnostic accuracy studies. We provide examples in Section 4 and we conclude with a summary of our results and several practical ramifications thereof, in Section 5.

2. Methods

In this section we first relate non-informative diagnostic results to the shape of a part of the ROC curve. Next, we derive a sufficient statistic for the ROC curve of a partially non-informative marker. In other words, we demonstrate that only a fraction of the data completely determines the distribution of the empirical ROC points. Thereafter, we demonstrate that given the derived sufficient statistic, the conditional expectation of the ROC curve is the same as the ROC curve for data with ties. Losses in efficiency of the estimates computed using continuous marker data with non-informative ranges of values immediately follow the Rao–Blackwell theorem. To simplify the theoretical derivations we focus on scenarios in which the diagnostically non-informative results occur in the range of values below a certain threshold.

The shape of the ROC curve corresponding to diagnostically non-informative marker values

It is well known that the diagonal line in the ROC space ($TPF = FPF$) represents the performance of a guessing process which randomly assigns diagnostic results to subjects, regardless of their actual disease status (Egan, 1975; Pepe, 2003; Zhou et al., 2011). This is a scenario where all diagnostic results are non-informative for discriminating between diseased and non-diseased subjects, namely, the distributions of the diagnostic results are the same $S_Y(\xi) = S_X(\xi)$, or $FPF(\xi) = TPF(\xi)$. Similarly, although not as widely recognized, the straight-line extending from any specific operating point $(fpf(\xi_0), tpf(\xi_0))$ to (1, 1) represents the performance of a guessing process applied to subjects with test results lower than ξ_0 (e.g., Wagner et al., 2001; Fawcett, 2006). Thus, the straight-line shape of a part of the ROC curve can be interpreted as an indication of the lack of diagnostic information in marker values within the corresponding range.

The straight-line shape of the part of the ROC curve from $(fpf(\xi_0), tpf(\xi_0))$ to (1, 1) is equivalent to a constant derivative for $fpf > fpf(\xi_0)$, i.e.,

$$\forall e > fpf(\xi_0), \quad \frac{d}{de} ROC(e) = constant.$$

This, in its turn, is equivalent to the constancy of the diagnostic likelihood ratio for all marker values in the corresponding range, i.e., $\forall \xi < \xi_0$, $\frac{d}{d\xi} TPF(\xi) / \frac{d}{d\xi} FPF(\xi) = constant$. In other words, all marker values below ξ_0 have the same relative