Journal of Clinical Epidemiology 68 (2015) 855–859

Journal of Clinical Epidemiology

The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases

Brice Ozenne[a,b,c,d], Fabien Subtil[a,b,c,d], Delphine Maucort-Boulch[a,b,c,d,*]

[a] *Université de Lyon, Département Biomaths-Santé, Lyon, 92 Rue Pasteur, 69007, France*
[b] *Université Lyon 1, Département Biomaths-Santé, Villeurbanne, 43 boulevard du 11 Novembre 1918, 69622, France*
[c] *Hospices Civils de Lyon, Service de Biostatistique Lyon, 165 Chemin du Grand Revoyet, Pierre-Bénite F-69310, France*
[d] *CNRS UMR5558, Laboratoire de Biométrie et Biologie Evolutive, Equipe Biostatistique-Santé, Villeurbanne F-69100, France*

Accepted 23 February 2015; Published online 28 February 2015

Abstract

Objectives: Compare the area under the receiver operating characteristic curve (AUC) vs. the area under the precision–recall curve (AUPRC) in summarizing the performance of a diagnostic biomarker according to the disease prevalence.

Study Design and Setting: A simulation study was performed considering different sizes of diseased and nondiseased groups. Values of a biomarker were sampled with various variances and differences in mean values between the two groups. The AUCs and the AUPRCs were examined regarding their agreement and vs. the positive predictive value (PPV) and the negative predictive value (NPV) of the biomarker.

Results: With a disease prevalence of 50%, the AUC and the AUPRC showed high correlations with the PPV and the NPV ($\rho > 0.95$). With a prevalence of 1%, small PPV and AUPRC values (<0.2) but high AUC values (>0.9) were found. The AUPRC reflected better than the AUC the discriminant ability of the biomarker; it had a higher correlation with the PPV ($\rho = 0.995$ vs. 0.724; $P < 0.001$).

Conclusion: In uncommon and rare diseases, the AUPRC should be preferred to the AUC because it summarizes better the performance of a biomarker.

Keywords: Area under the curve; Binary biomarker; Performance assessment; Precision-Recall curve; Rare events; Receiver operating curve

1. Introduction

Assessing the performance of a diagnostic biomarker and comparing the performances of several biomarkers are important issues in medicine. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) are recommended indicators that evaluate diagnostic performance [1]. The PPV and the NPV are particularly interesting in case of low disease prevalence [2]; however, they rely on the choice of a threshold biomarker value that classifies the subjects into diseased and nondiseased. As the optimal threshold may depend on various characteristics of the population, global measures have been proposed.

The receiver operating characteristic (ROC) curve is often considered as the standard mean for depicting biomarker performance over all biomarker thresholds. Its properties and associated indices have been extensively studied [3], but its relevance in case of rare events is still debated. Indeed, only a few works have dealt with rare diseases or medical imaging data whose prevalence is typically very low (eg, Wilson disease whose worldwide prevalence is estimated at 30 per 1 million [4] or the proportion of lesioned to nonlesioned pixels on stroke images that is frequently lower than 1% [5]).

In the field of information retrieval where there are important differences between the small number of relevant documents and the huge number of irrelevant ones, the ROC curve seems to overestimate the performance of several retrieval methods [6,7]. This is why an alternative method is used: the precision–recall (PR) curve [7,8].

At a fixed prevalence, Davis and Goadrich [7] have shown that there was a one-to-one correspondence between the ROC curve and the PR curve. However, other authors have shown that the area under the ROC curve (AUC) and the area under the PR curve (AUPRC) are not equivalent but show very large differences [9]. In the case of rare events, some authors have recommended the use of the AUPRC instead of the AUC [7]; however, up to now, no comparisons are readily available.

Funding: None.

* Corresponding author. Tel.: 33-478-865-764; fax: 33-472-115-141.
E-mail address: delphine.maucort-boulch@chu-lyon.fr (D. Maucort-Boulch).

http://dx.doi.org/10.1016/j.jclinepi.2015.02.010
0895-4356/© 2015 Elsevier Inc. All rights reserved.

What is new?

- In case of low-prevalence diseases, the area under the ROC curve (AUC) may overestimate the performance of a biomarker.
- At low prevalence, the area under the precision–recall curve (AUPRC) provides a better agreement with the positive predictive value of a biomarker.
- The AUPRC should be preferred over the AUC for the evaluating uncommon or rare disease biomarkers.
- The AUPRC should not be compared between populations with different disease prevalences because their values are prevalence dependent.

The present article compares the abilities of the AUC and the AUPRC in summarizing the classification performance of a biomarker according to the disease prevalence.

2. Methods

2.1. The ROC curve

The ROC curve is a graphical technique used to assess the diagnostic accuracy of a continuous biomarker. It displays a trade-off between the sensitivity and the specificity of the biomarker over all possible biomarker threshold values. Using the conventional abbreviations (TP, FP, FN, and TN for the number of true positive, false positive, false negative, and true negative subjects), sensitivity equals TP/(TP + FN) and specificity equals FP/(TN + FP).

The AUC is a summary index of the biomarker performance. It ranges from 0.5 to 1 and corresponds to the probability that biomarker values from a randomly selected pair of diseased and nondiseased subjects are correctly ordered [3]. In this technique, sensitivity and specificity are unaffected by the disease prevalence and so is the AUC [6].

2.2. The precision–recall (PR) curve

The PR curve is an alternative approach for assessing the performance of a biomarker. It displays the trade-off between precision (instead of specificity) and sensitivity (also called recall) over all possible biomarker threshold values. Precision is the ratio TP/(TP + FP), which corresponds to the PPV in the ROC approach. The PR curve focuses on the ability of the biomarker to identify diseased subjects; it ignores the correctly classified healthy subjects (TN), which is the dominant group in case of low-prevalence disease. Unlike a ROC curve, a PR curve is not necessarily monotonic across all biomarker thresholds because an increase in the threshold can decrease TP or FP.

The AUPRC is a summary statistic that reflects the ability of a biomarker to identify the diseased group. Denoting by x a biomarker value taken from the biomarker distribution in the diseased group, the AUPRC can be interpreted as the expectation, over all the possible x values, of the proportion of diseased subjects among all those whose biomarker value exceed x [8]. The values of the AUPRC range from 0 to 1, but whereas the expected value for random guessing of the AUC is 0.5, that of the AUPRC is prevalence dependent and tends to 0 when the prevalence decreases. Details on the estimation of the AUPRC and its confidence interval (with R codes) can be found in a recent article by Boyd et al. [8].

2.3. The simulation study

Simulations were performed on biomarker values that follow normal distributions in diseased and nondiseased subjects. The variance of the nondiseased group was fixed to $\sigma_1 = 1$, and the variance of the diseased group could range from $\sigma_2 = 0.01$ to 10. The difference between the means ($\mu_2 - \mu_1$) was allowed to range from 0 to 5 by 0.25 increments (21 steps). The size of the nondiseased group (n_1) was fixed at 10,000, but the size of the diseased group (n_2) could range from 100 (prevalence: 0.0099) to 10,000 (prevalence: 0.5).

For each n_2, σ_2, and ($\mu_2 - \mu_1$) combination, 1,000 sets of biomarker values in diseased and nondiseased subjects were generated. Then, at each prevalence value and at each variance of the diseased group, Pearson correlation coefficients (ρ) were computed between the 21 × 1000 AUC and AUPRC paired values obtained over all ($\mu_2 - \mu_1$) differences. ρ were also computed between AUC and PPV paired values as well as between AUPRC and PPV paired values.

2.4. Biomarker performance assessment

The biomarker performance was evaluated using the PPV and the NPV values calculated at the threshold at which the proportion of subjects classified as diseased is equal to the disease prevalence; that is, once the subjects are ranked in ascending biomarker values, the first n_1 subjects are considered as nondiseased and the remaining n_2 subjects as diseased. The PPV corresponds to the probability for a biomarker-positive subject to have the disease, whereas the NPV corresponds to the probability for a biomarker-negative subject to be actually nondiseased. Then, the AUC or the AUPRC should not be high for low values of the PPV or the NPV whatever the biomarker threshold.

3. Results

Table 1 lists that, at prevalence 0.5, the correlation between the AUC and the AUPRC is excellent ($\rho = 0.990$); besides, the AUC and the AUPRC reflected correctly the classification performance of the biomarker (correlations