Journal of Clinical Epidemiology 68 (2015) 752−759

Journal of Clinical Epidemiology

An enhancement of ROC curves made them clinically relevant for diagnostic-test comparison and optimal-threshold determination

Fabien Subtil[a,b,c,d,*], Muriel Rabilloud[a,b,c,d]

[a]*Service de Biostatistique, Hospices Civils de Lyon, 162 avenue Lacassagne, Lyon F-69003, France*
[b]*Université de Lyon, 92 rue Pasteur, Lyon F-69000, France*
[c]*Université Lyon 1, 43 boulevard du 11 novembre 1918, Villeurbanne F-69100, France*
[d]*CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne F-69100, France*

Accepted 6 January 2015; Published online 12 January 2015

Abstract

Objectives: The receiver operating characteristic curves (ROC curves) are often used to compare continuous diagnostic tests or determine the optimal threshold of a test; however, they do not consider the costs of misclassifications or the disease prevalence. The ROC graph was extended to allow for these aspects.

Study Design and Setting: Two new lines are added to the ROC graph: a sensitivity line and a specificity line. Their slopes depend on the disease prevalence and on the ratio of the net benefit of treating a diseased subject to the net cost of treating a nondiseased one. First, these lines help researchers determine the range of specificities within which test comparisons of partial areas under the curves is clinically relevant. Second, the ROC curve point the farthest from the specificity line is shown to be the optimal threshold in terms of expected utility.

Results: This method was applied: (1) to determine the optimal threshold of ratio specific immunoglobulin G (IgG)/total IgG for the diagnosis of congenital toxoplasmosis and (2) to select, among two markers, the most accurate for the diagnosis of left ventricular hypertrophy in hypertensive subjects.

Conclusion: The two additional lines transform the statistically valid ROC graph into a clinically relevant tool for test selection and threshold determination.

Keywords: Diagnostic test; ROC curve; Decision analysis; Test selection; Optimal threshold; Utility

1. Introduction

Nowadays, screening, diagnosing, prognosing, and monitoring diseases rely increasingly on continuous diagnostic tests: biomarker assays or imagery signals. One very popular tool used to evaluate these tests is the receiver operating characteristic curve (ROC curve) [1,2]. This curve provides a graphic representation of the trade-off between the true-positive rate (sensitivity) and the false-positive rate (1 − specificity) over all possible thresholds for a test.

The ROC curve has, at least, two different uses. First, it allows comparing the diagnostic accuracies of several continuous tests through the areas under the curves (AUCs). The AUC can be interpreted as the probability that a case and a noncase pair of test values are correctly ranked. Second, it allows determining the optimal threshold for dichotomization of a continuous diagnostic test. This uses various methods, such as the point of the ROC curve the closest to point (0,1) or the farthest point from the first diagonal of the graph [3].

However, these two uses have been criticized because the ROC curve reflects only the test diagnostic accuracy (sensitivity and specificity) but does not consider the consequences of the clinical decisions it leads to. In fact, the choice of a test or of its threshold does not depend only on its accuracy but also on the disease prevalence and on the clinical benefits and costs, respectively, associated with correct and incorrect subject classification [4].

Regarding test comparison, let us consider the case of a minor disease for which a highly specific test is required (to avoid treating a nondiseased subject) and the ROC curves of two hypothetical tests, A and B, with B more specific than A at high sensitivity values (ie, sensitivity > 90%) but A more sensitive than B at high specificity values (ie, specificity > 90%; Fig. 1). The AUCs (0.78 for A and 0.85 for B) favor test B. However, the AUC is an overall

Conflict of interest: None.

* Corresponding author. Tel: +33-4-72-11-52-38; fax: +33-4-72-11-51-41.
E-mail address: fabien.subtil@chu-lyon.fr (F. Subtil).

http://dx.doi.org/10.1016/j.jclinepi.2015.01.003
0895-4356/

What is new?

Key findings

- With the additional "specificity line," the receiver operating characteristic curve (ROC curve) can be used to determine a threshold that takes into account the disease prevalence and the consequences of subject misclassification.
- With the "specificity line" and the "sensitivity line," the ROC graph helps defining the region within which the partial areas under the curves of two or more diagnostic tests should be calculated and compared.
- More than a mere statistical tool, the ROC curves with the two additional lines may become component of medical decision making.

measure of accuracy over all possible test thresholds and, consequently, over all sensitivity and specificity values, whereas the test comparison should focus in this example on the thresholds with high specificities. Actually, a local comparison of ROC curves in the region with specificity >0.9 (dark gray region in Fig. 1) leads to conclude that test A should be rather preferred because it has higher sensitivities than test B at comparable specificities. More generally, when the ROC curves do not intersect, total and local ROC curve comparisons through the AUCs lead to the same conclusion, but this is not necessarily true when the ROC curves do intersect. In the latter case, a local comparison is necessary, but defining the region within which the comparison is relevant requires the knowledge of the disease prevalence and the consequences of the subject misclassifications, which is not displayed on the ROC graph.

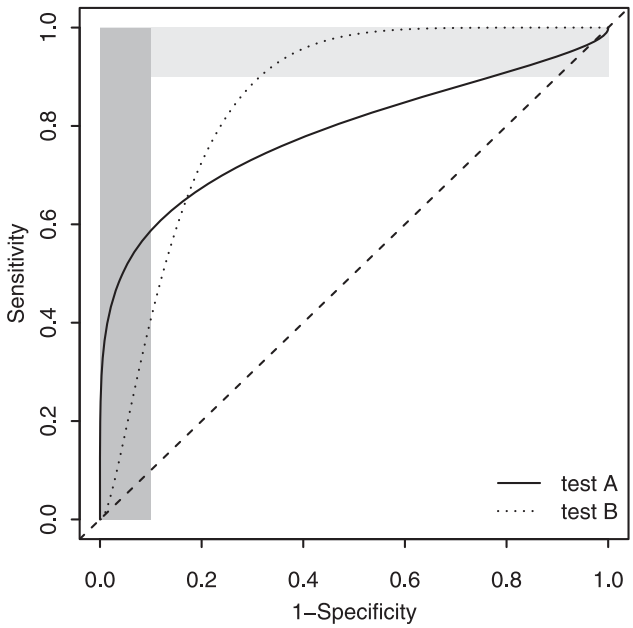

Fig. 1. ROC curves of two hypothetical diagnostic tests. The light gray strip is the region with sensitivity >0.9. The dark gray strip is the region with specificity >0.9.

Several other tools or measures have been proposed to compare continuous diagnostic tests: the predictiveness curve [5], the net reclassification improvement (and its weighted version) [6], the net benefit, and the relative utility [7]. All take into account the disease prevalence and, sometimes, the consequences of subject misclassifications. The properties of these tools have been analyzed and compared [8,9]. However, their basic concepts are far different from that of the popular ROC curve.

Regarding the determination of the optimal threshold, the point of the ROC curve the farthest from the first diagonal of the graph is the best compromise between sensitivity and specificity. However, with rare exceptions (see Methods), this method is not appropriate because the threshold depends also on the disease prevalence and on the consequences of subject misclassifications [10]. For instance, in some screening tests, such as the immunologic tests for colorectal cancer, the vast majority of screened subjects are disease-free; thus, a small defect in specificity can lead to a huge number of unnecessary colonoscopies. In such a case, the optimal threshold is not the threshold that ensures the best compromise between sensitivity and specificity but the one that favors specificity.

From a medical-decision-making perspective, the optimal threshold is the one that maximizes the expected utility or minimizes the expected cost of the test in a given population [11]. Several estimation methods have been proposed for this purpose [11−13]. However, a lot of biomarker thresholds are still determined using the ROC curve probably because the other methods are less intuitive for the physician or the applied statistician. Moreover, these methods also provide a confidence interval (CI) of the optimal threshold and hence are more complex. This CI gives important information about the optimal threshold estimate. But in moderate sample size studies (common for diagnostic tests), the CI is large and hence not really informative. Thus, methods that provide optimal thresholds without CI remain useful.

The objective of this article was to show that a slight improvement in the ROC curve graph, precisely, adding one or two lines, allows taking into account the disease prevalence and the consequences of subject misclassifications. This provides a statistically and clinically relevant tool for diagnostic-test selection and test-threshold determination in terms of expected utility. The two additional lines we propose here are graphical implementations of previously proposed mathematical methods. The roles of these lines are illustrated by two examples: one relative to the determination of the threshold of a serological test for the diagnosis of congenital toxoplasmosis and another relative to the comparison of two markers in the diagnosis of left