Computers in Biology and Medicine 47 (2014) 120–129

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Exploring medical diagnostic performance using interactive, multi-parameter sourced receiver operating characteristic scatter plots

Hyatt E. Moore IV[a,*], Olivier Andlauer[b], Noah Simon[c], Emmanuel Mignot[d]

[a] Department of Electrical Engineering, Stanford University, Stanford, CA, USA
[b] EA 481, Laboratoires de Neurosciences, University of Franche-Comte, 1 place du marechal Leclerc, 25030 Besancon Cedex, France
[c] Department of Statistics, University of Washington, Seattle, WA, USA
[d] Center for Sleep Sciences and Medicine, Stanford University, Palo Alto, CA, USA

ARTICLE INFO

Article history:
Received 22 October 2013
Accepted 25 January 2014

Keywords:
Diagnostic testing
Receiver operating characteristic
Bootstrap
Data visualization
Sleep

ABSTRACT

Determining diagnostic criteria for specific disorders is often a tedious task that involves determining optimal diagnostic thresholds for symptoms and biomarkers using receiver-operating characteristic (ROC) statistics. To help this endeavor, we developed softROC, a user-friendly graphic-based tool that lets users visually explore possible ROC tradeoffs. The software requires MATLAB installation and an Excel file containing threshold symptoms/biological measures, with corresponding gold standard diagnoses for a set of patients. The software scans the input file for diagnostic and symptom/biomarkers columns, and populates the graphical-user-interface (GUI). Users select symptoms/biomarkers of interest using Boolean algebra as potential inputs to create diagnostic criteria outputs. The software evaluates subtests across the user-established range of cut-points and compares them to a gold standard in order to generate ROC and quality ROC scatter plots. These plots can be examined interactively to find optimal cut-points of interest for a given application (e.g. sensitivity versus specificity needs). Split-set validation can also be used to set up criteria and validate these in independent samples. Bootstrapping is used to produce confidence intervals. Additional statistics and measures are provided, such as the area under the ROC curve (AUC). As a testing set, softROC is used to investigate nocturnal polysomnogram measures as diagnostic features for narcolepsy. All measures can be outputted to a text file for offline analysis. The softROC toolbox, with clinical training data and tutorial instruction manual, is provided as supplementary material and can be obtained online at http://www.stanford.edu/~hyatt4/software/softroc or from the open source repository at http://www.github.com/informaton/softroc.

1. Introduction

Diagnoses are made based on the presence of symptoms or the results of biological or physiological tests. No such test is perfect, and for continuous variables, it is essential to select an optimal cut-off in comparison to a gold standard evaluation. Depending on the application, one may select cut-offs with equal specificity and sensitivity, or by favoring one characteristic at the cost of the other.

Receiver-operating characteristic (ROC) curves are typically used to visualize sensitivity and specificity tradeoffs for various diagnostic cut-offs. Points located closer to the ideal test point (100% sensitivity, 100% specificity) are often seen as good candidates for cut-off values. However, sensitivity and specificity by themselves provide little meaning, as it is possible to make a test that achieves 100% sensitivity or 100% specificity by simply always giving a positive test result or always giving a negative test result. In an important variation, the quality receiver-operating characteristic curves (qROC), sensitivity and specificity values are remapped to Kappa values, or quality indices, which provide a normalized measure of the standard test ROC values. These weighted Kappa coefficients — the original ROC values adjusted to be 0% for a random test and 100% for a perfect test — make it easy to identify and lay claim to the test with optimal sensitivity or specificity [8].

While several ROC software packages exist, few are dedicated to exploring medical diagnostic criteria or incorporate the ability to group and combine multiple variables, a critical feature for medical diagnostic when multiple parameters are involved. One review of eight ROC programs, commercially available and freeware, showed mixed results [17]. While statistically sound, the programs covered (e.g. MedCalc and Chicago University's Metz ROC Software) were described as unfriendly or overly complicated to use because of the interface or statistical background required. Little validation of ROC results was provided from these programs in terms of generalization and bias, and none plotted qROCs.

* Corresponding author.
E-mail address: hyatt4@stanford.edu (H.E. Moore IV).

http://dx.doi.org/10.1016/j.compbiomed.2014.01.012
0010-4825

Another ROC program, not included in this review, which does incorporate quality indices, is ROC5 [13]. This program produces decision trees aimed at providing clinical practitioners with a plan of sequential tests to follow based on the outcome at each stage of the tree. This program was the closest to helping us achieve our goals. Decision trees can be helpful in determining the order of tests to provide to a patient and are particularly well suited when tests must be performed in a sequential manner due to increasing cost. However, the nature of decision trees to follow one branch while rejecting all others prevents the evaluation and exploration of simultaneous possibilities when available. The limitations we saw in ROC5 were its inability to draw multiple ROC plots and its user friendliness. Also, it was not possible to explore the entire space of possibilities; it was only possible to do a hierarchy of cut-offs. The program selects the best variable first, then the second, and so on. This *greedy* approach does not necessarily give the best combination of all criteria which can make a huge difference.

Considering this need, we developed softROC, a MATLAB based (and run) software package for exploratory ROC analysis. Specifically, softROC provides a GUI that allows users to quickly configure candidate diagnostic criteria combinations and evaluate them for optimal performance using test and quality ROC metrics. The software uses either bootstrapping or training-with-validation techniques to provide generalizability of selected diagnostic criteria to other populations. It is intended for medical researchers and practitioners who want to analyze the sensitivity and specificity of a combination of symptom measures for diagnosing a patient versus a gold standard. We took a user-oriented approach in designing softROC. Software requirements were taken from two intended users, and iteratively refined based on their feedback from each internal release. Initial feedback focused on the user interface and interactions, which led to softROC's dynamic and flexible configuration interface. Later feedback focused on incorporating other statistical methods like bootstrapping and convex optimization, which led to its generalization.

We chose MATLAB because of its relatively straightforward, stable, development environment suitable for both statistical and graphically interactive based software programs. Octave was also considered as a free alternative to MATLAB. MATLAB and Octave both provide excellent frameworks for rapidly prototyping statistical applications. However, as stated on its website, "Octave is principally a batch or command-line language," and does not yet directly support the interactive methods required by softROC [11]. softROC is released under a creative commons license and is available online at the open source repository http://www.github.com/informaton/softroc, http://www.stanford.edu/~hyatt4/software/softroc, and as a .zip attachment to this manuscript's supplementary material section. The supplementary material section also includes the instruction manual and tutorial dataset. We provide preliminary background on the statistical methods used and an overview of narcolepsy diagnosis in Section 2. Section 3 covers the design and implementation of softROC for investigating diagnostic test tradeoffs (e.g. sensitivity vs specificity) through interactive ROC scatterplots. Discussion of softROC's application, limitations, and extension are covered in Section 4, followed by concluding remarks in Section 5.

Table 1
The contingency table captures the four possible outcomes when comparing a diagnosis based on medical testing to its gold standard "truth": true positive (TP), false negative (FN), false positive (FP), and true negative (TN). The values, when given as fractions of the whole, sum to produce quality (Q) and prevalence (P) which are necessary to calibrate ROC values. Medical testing is frequently defined as positive or negative depending on a biological, symptomatic or physiological threshold value. For example, fasting blood sugar levels greater than 126 mg/dl (7.0 mmol/L) are abnormally high and typically used to diagnose diabetes.

		Medical test		
		+	−	
Ground truth (gold standard)	+	TP	FN	P
	−	FP	TN	$P' = 1-P$
		Q	$Q' = 1-Q$	1

2. Background

Statistical background of receiver operating characteristics and Boolean algebra are presented in this section. Information on narcolepsy and its diagnosis is presented in Section 3.1.

2.1. Receiver operating characteristics

The contingency table, or confusion matrix, shown in Table 1 contains the collection of possible outcomes, expressed as percentage, for a predicted medical diagnosis of a patient and the true diagnosis, or pathology, as revealed through an accepted gold standard. The *confusion* of Table 1 lies on the diagonal where the evaluation is different from the ground truth — a patient with true disease is missed (i.e. false negative) or one without disease is wrongly diagnosed with it (i.e. false positive).

Frequently, medical tests give a continuous value that needs to be dichotomized as positive and negative for practical reasons based on a threshold or cut-off value. Altering this threshold value modifies specificity and sensitivity that can be optimized for a given application. ROCs are frequently used to evaluate these trade-offs. These and other measures, which are derived from Table 1 and implemented in softROC, are given in Table 2. The derivations require contents of Table 1 to be given as fractions of the total count ranging in value from 0.0 to 1.0.

ROC curves plot sensitivity (true positive rate) versus one-minus specificity (false positive rate) for different thresholds used to classify a patient positively or negatively for disease. When evaluating multiple curves, points along the outermost curve — the ROC convex hull — are superior to any along the other curves. The area under the ROC curve (AUC) offers insight into a diagnostic test's overall ability to discriminate between positive and negative cases and is equivalent to the Wilcoxon test metric of ranks [7]. A random ROC curve, which places a straight line from (0,0) to (1,1), has an AUC of 0.5. A legitimate diagnostic test should have an AUC between 0.5 and 1.0.

Table 2
List and definition of softROC statistics obtained from Table 1.

Term	Notation	Definition
Sample size	N_0	
True positive	TP	$1/N_0 \cdot \sum_{i=1}^{N_0}(True\ diagnosis_i^+ \cap Test_i^+)$
False negative	FN	$1/N_0 \cdot \sum_{i=1}^{N_0}(True\ diagnosis_i^+ \cap Test_i^-)$
False positive	FP	$1/N_0 \cdot \sum_{i=1}^{N_0}(True\ diagnosis_i^- \cap Test_i^+)$
True negative	TN	$1/N_0 \cdot \sum_{i=1}^{N_0}(True\ diagnosis_i^- \cap Test_i^-)$
Prevalence	P	TP+FN
Quality	Q	TP+FP
Sensitivity	SE	TP/P
Specificity	SP	TN/P'
Positive predictive value	PPV	TP/Q
Negative predictive value	NPV	TN/Q'
Efficiency	EFF	TP+FN
Quality index 1,0	$\kappa(1,0)$	$(SE-Q)/Q'$
Quality index 0,0	$\kappa(0,0)$	$(SP-Q')/Q$
Cohen's Kappa	$\kappa(0.5,0)$	$(PQ' \cdot \kappa(1,0)+P'Q \cdot \kappa(0,0))/(PQ'+P'Q)$
Chi-square	χ^2	$N_0 \cdot \kappa(1,0) \cdot \kappa(0,0)$