Available online at www.sciencedirect.com
ScienceDirect
Neural Networks 21 (2008) 387–397
Neural Networks
www.elsevier.com/locate/neunet

2008 Special Issue

Reader studies for validation of CAD systems☆

Brandon D. Gallas*, David G. Brown

NIBIB/CDRH Laboratory for the Assessment of Medical Imaging Systems, FDA, Silver Spring, MD, 20993-0002, United States

Received 22 August 2007; received in revised form 7 December 2007; accepted 11 December 2007

Abstract

Evaluation of computational intelligence (CI) systems designed to improve the performance of a human operator is complicated by the need to include the effect of human variability. In this paper we consider human (reader) variability in the context of medical imaging computer-assisted diagnosis (CAD) systems, and we outline how to compare the detection performance of readers with and without the CAD. An effective and statistically powerful comparison can be accomplished with a receiver operating characteristic (ROC) experiment, summarized by the reader-averaged area under the ROC curve (AUC). The comparison requires sophisticated yet well-developed methods for multi-reader multi-case (MRMC) variance analysis. MRMC variance analysis accounts for random readers, random cases, and correlations in the experiment. In this paper, we extend the methods available for estimating this variability. Specifically, we present a method that can treat arbitrary study designs. Most methods treat only the fully-crossed study design, where every reader reads every case in two experimental conditions. We demonstrate our method with a computer simulation, and we assess the statistical power of a variety of study designs.
Published by Elsevier Ltd

Keywords: ROC; Reader studies; Study design; Multi-reader multi-case (MRMC) variance analysis

1. Introduction

Many computational intelligence systems are designed to improve the performance of a human operator or *reader*. For these systems, measuring stand-alone performance is inadequate. It is not sufficient to determine, for example, the mean and variance of the sensitivity and specificity of the computational intelligence on a test set: it is necessary to demonstrate that the human reader benefits from the use of the computational intelligence. *Has access to the computational intelligence agent permitted the reader to improve his or her sensitivity, specificity, or receiver operating characteristic (ROC) curve?* The determination of this improvement entails the use of reader studies with the complications and uncertainties inherent to the fickle and fallible human.

In the field of medicine, computational intelligence agents already have a rich tradition. They are commonly referred to as computer-assisted diagnosis (CAD) software devices. Examples of such systems in current clinical use are the breast cancer detection and Pap slide reader aids introduced into clinical use in the mid-1990s. At the present time, there is great interest in colon polyp and lung cancer screening CAD systems used to help in the analysis of the vast quantities of imaging data collected by modern computed tomography x-ray systems. These CAD systems employ a wide variety of classifier constructs: linear and quadratic discriminants, classification trees, clustering algorithms, and neural networks. Rather than elaborating on specific algorithms, this paper is about evaluating the CAD systems.

For CAD systems used in medical imaging, a medical professional, e.g., radiologist or pathologist, views an image of some portion of the anatomy and must decide whether or not a pathological condition is present. Generically, we shall refer to the medical professional who interprets the image as the reader. For these systems, three kinds of studies could be performed: (1) measurement of the performance of the reader without CAD, (2) measurement of the stand-alone performance of the CAD device, and (3) measurement of the performance of the reader as aided by the CAD device. The first kind of study provides a benchmark against which any improvement provided by the CAD is measured. The second is useful during the CAD development stage but is only determinative if the

☆ An abbreviated version of some portions of this article appeared in Brown and Brandon (2007) as part of the IJCNN 2007 Conference Proceedings, published under IEE copyright.

* Corresponding author. Tel.: +1 301 796 2531; fax: +1 301 796 9925.
E-mail address: brandon.gallas@fda.hhs.gov (B.D. Gallas).

0893-6080/$ - see front matter Published by Elsevier Ltd
doi:10.1016/j.neunet.2007.12.013

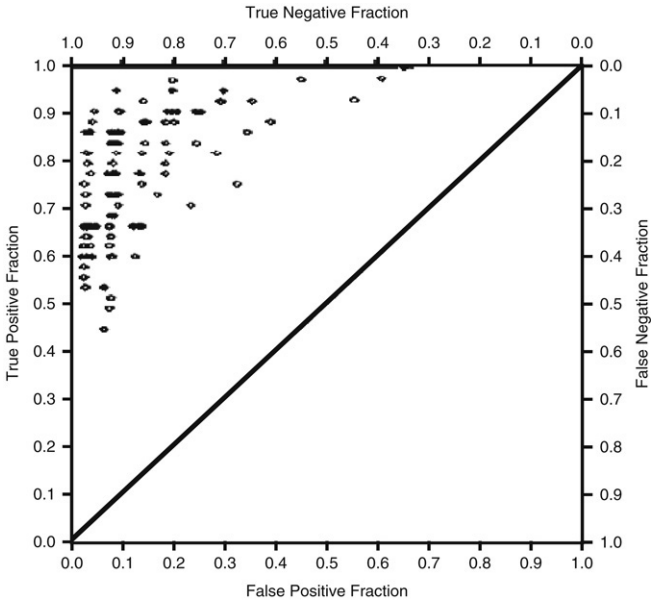

Fig. 1. (Beam et al., 1996): In a study of 108 US radiologists all reading the same set of 79 mammograms (34 normal/benign, 45 breast cancer), Beam et al. (1996) found that there was a wide range of radiologist skill and that radiologists operate at a wide range of thresholds.

CAD performance is so good that the human reader may be dispensed with, and CAD moves beyond being an *assist*. The third provides the necessary evidence that the CAD system provides a benefit.

Reader variability presents a serious problem for both studies with and without the CAD. Fig. 1 from the work of Beam, Layde, and Sullivan (1996) is the iconic illustration of the problem presented by human reader variability. The figure presents performance in terms of sensitivity and specificity (jointly establishing the *operating points* on the ROC curve) of 108 experienced mammographers reading 79 identical mammograms for cancer. Not only are there tremendous differences in the radiologists' operating points or how aggressive each radiologist is in making a positive diagnosis, but clearly the points do not trace out a single ROC curve. Some radiologists perform remarkably close to the chance level (45 degree line) in comparison to others who are nearly perfect (upper left corner) on the exact same set of cases.

Reader variability plays an important role in the sizing of clinical studies. Large variability implies a need for large studies, though the required size of a study is unknown without some initial data, generally collected in a pilot study. Given the great cost and difficulty of clinical studies, we recommend a pilot study to flesh out the sources of variability and their size. The information from a pilot study can be used to investigate the size and sampling strategy for a pivotal trial.

In what follows, we shall outline a typical ROC reader study. The ROC curve characterizes the diagnostic capacity of a reader in a way that is objective and meaningful (Green & Swets, 1966; Metz, 1986; Swets & Pickett, 1982; Van Trees, 1968). This ROC curve is often summarized with the area under the ROC curve (AUC). We shall review some methods for multi-reader multi-case (MRMC) variance analysis that treat the fully-crossed study design (every reader reads every case in every condition being studied) (Dorfman, Berbaum, & Metz, 1992; Gallas, 2006; Obuchowski et al., 2004; Roe & Metz, 1997a, 1997b; Swets & Pickett, 1982). We then present an extension of the moment-based approach (Gallas, 2006) that can treat a general study design, because a fully-crossed study design may not be possible or practical to implement. We demonstrate our extension with a computer simulation, and we assess the statistical power of a variety of study designs.

2. Methods

2.1. Background

An MRMC ROC experiment consists of readers and cases. We assume the existence of a *gold* standard to divide the cases into two classes: normal (signal-absent) and diseased (signal-present). Establishing ground truth, that is, deciding which cases truly are normal or diseased, is far from trivial but outside the scope of this paper. Then, based on some study design, cases are assigned to readers, who then score each according to their level of suspicion of the presence of the specified disease.

The ROC curve for each reader characterizes the separation between the distribution of scores for normal cases and the distribution of scores for diseased cases. It does so with the concept of a threshold. The threshold can take on every possible score and is also allowed between scores when the scores are discrete. For every possible threshold, scores are dichotomized: scores below the threshold are *called* normal, or "negative" for disease, and those above are *called* diseased, or "positive" for disease. The ROC data may now be given in terms of the number of correct and incorrect calls on the diseased cases (TP = the number of true positives and FN = the false negatives) and on the normal cases (TN = the number of true negatives and FP = the false positives). In terms of fractions of normal or diseased cases, the four numbers reduce to two independent ones, most notably,

- sensitivity (Se), the fraction of the correctly diagnosed diseased cases, $\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, and
- specificity (Sp), the fraction of the correctly diagnosed normal cases, $\mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$.

The empirical ROC curve is the collection of all the (SE, SP) pairs generated with the ROC data as the threshold takes all possible values from $-\infty$ to ∞. There also exists a theoretical population-based ROC curve defined by the areas of the distributions of normal and diseased scores above and below the possible thresholds. The empirical ROC curve is a non-parametric estimate of the population curve; there are also parametric, or semi-parametric, estimates (Alonzo & Pepe, 2002; Dorfman & Alf, 1969; Tosteson & Begg, 1988).

While the entire ROC curve gives a complete picture of the diagnostic capacity of a reader, we shall summarize this capacity with the area under the ROC curve (AUC). AUC is a common summary measure of the ROC curve that is equivalent to the reader's sensitivity averaged over all specificities, as