Performance Evaluation of Medical Expert Systems

D. Christine Georgakis  
Northeastern Illinois University  
Chicago, Illinois 60625  

Martha Evens  
Illinois Institute of Technology  
Chicago, Illinois 60616  

Frank Naeymi-Rad  David A. Trace  
University of Health Sciences/The Chicago Medical School  
North Chicago, Illinois 60064

Abstract

The major problem in the evaluation of expert systems is the selection of the appropriate statistical measures of performance consistent with the parameters of the system domain. The objective of this paper is to develop the statistical evaluation methodology needed to assess the performance of medical expert systems including MEDAS - the Medical Emergency Decision Assistance System. The measures of performance are selected so as to have an operational interpretation and also reflect the predictive diagnostic capacity of a medical expert system. Certain summary measures are used that represent the sensitivity, specificity, and system response of a medical expert system. Measures of agreement such as the kappa statistic and the measure of conditional agreement are used to measure the agreement between the medical expert system and the physician. Goodman and Kruskal’s lambda and tau measures of predictive association are introduced to evaluate the predictive capacity of a medical expert system. This methodology has been partially implemented in the performance evaluation of MEDAS.

1 Introduction

Medical expert systems require comprehensive evaluation of their diagnostic accuracy at every stage of development. Without established evaluation methods, the
usefulness of medical expert systems is limited. Acceptance in the clinical arena is contingent on the verification of diagnostic accuracy first and foremost.

This paper addresses the need for formal evaluation in the development of medical expert systems. We develop a statistical evaluation methodology that assesses the performance of medical expert systems. This methodology includes standard summary measures, measures of agreement such as the kappa statistic and the measure of conditional agreement, and Goodman and Kruskal's λ and τ measures of predictive association. The various statistical measures are developed using the medical expert system MEDAS as an example.

2 The MEDAS Diagnostic System

MEDAS (the Medical Emergency Decision Assistance System) is a multimembership Bayesian diagnostic expert system designed to assist the clinician in determining diagnoses, ordering tests, and prescribing treatment. It can handle patients with one or more disorders.

The input to the diagnostic system consists of the patient's features. A feature is any piece of clinical data, age, sex, race, complaints, signs, symptoms, findings from the physical examination, results of tests and diagnostic medical procedures. The system's output consists of an ordered list of the possible disorder patterns ranked by their probability. Detailed description of the system is given by Ben-Bassat et al. [1].

MEDAS is the outgrowth of fourteen years of research on pattern recognition and expert systems carried on originally at the University of Southern California and continued for the past seven years at the Chicago Medical School. The knowledge base of the system consists of disorder patterns, each of which is composed of a list of features which may be relevant for the diagnosis of the disorder, coupled with two probability estimates which represent the sensitivity and specificity of these features with respect to this disorder.

3 Summary Measures

In our methodology we broaden the use of the sensitivity, specificity, and system response measures used by Ben-Bassat [2]. We utilize these measures through stratified random sampling and devise confidence limits according to established statistical practice [3].

Sensitivity Measure. This measure is the probability \( p_{ij} \) that the \( i \)th test disorder diagnosed by the physician is included in the list of the top \( j \) disorders given by the MEDAS output, where \( i = 1, 2, \ldots, 10 \) and \( j = 1, 2, 3 \). Note that \( p_{11} \) is the number of times that MEDAS correctly predicts that the \( i \)th test disorder is the primary disorder present in the patient over the total number of times the \( i \)th disorder occurs in the \( i \)th test population group. In the traditional error rate approach for