A Goodness-of-Fit Test for a Receiver Operating Characteristic Curve from Continuous Diagnostic Test Data

BY KELLY H. ZOU*, JOSEPH L. GASTWIRTH†, AND BARBARA J. MCNEIL‡

Harvard Medical School and Brigham and Women's Hospital, George Washington University, and Harvard Medical School and Brigham and Women's Hospital

The receiver operating characteristic (ROC) curve is a useful way to display the performance of a medical diagnostic test for detecting whether or not a patient is diseased or healthy. The diagnostic data consist of independent random samples on continuous measurement scales from diseased and healthy populations. We propose assessing the goodness-of-fit of a model by comparing a model-based estimate with a nonparametric estimate of the area under the curve (AUC). We focus on two parametric models, so-called Bi-Normal and Bi-Weibull models, and briefly on associated semiparametric transformation models. We also consider the null hypothesis that a parametric model is valid after an unspecified monotone transformation of the measurement scales. High power of the test implies sensitivity of the AUC to model assumptions; low power implies robustness of the estimate. The test is exemplified with a data set on the diagnosis of pancreatic cancer. A simulation study of the statistical power of the test is included.

1. Introduction Diagnostic testing provides important data for medical decision making and treatment planning. The receiver operating characteristic (ROC) curve is a useful graphical and statistical tool for evaluating and comparing diagnostic tests. It is a plot of (1−specificity, sensitivity)-values at all possible two-state decision thresholds (for definitions, see [3]). Much of the ROC literature deals with ordinal rating data methods, where the values indicate the degree of certainty about the disease. For example, for cancer detection, a five-point rating scale is often employed, with 1 = definitely benign, 2 = possibly benign, 3 = probably benign, 4 = possibly malignant, and 5 = definitely malignant. Recently, diagnostic tests that yield continuous results are increasingly used. Examples of such tests are those based on tumor volume or laboratory assay such as the ELISA test for HIV infection. Note that for the ordinal rating data, it is usually assumed that there is a latent continuous variable. In this article, we confine attention to ROC curves derived from continuous tests with a moderately large number of samples of both healthy (H) and diseased (D) individuals.

There are several ways of estimating an ROC curve, along with its summary measures: First, nonparametrically, a plot of pairs of observed (1−specificity, sensitivity)-values at each possible decision threshold forms an empirical ROC curve. This is equivalent to plotting two empirical survival curves against each

*Supported in part by grants NIH RO3HS13234-01 and NCI PO1CA67165
†Supported in part by grant NSF SBR9507926
‡Supported in part by grant NIH UO1CA62462