Urologic Oncology: Seminars and Original Investigations 26 (2008) 308–315

UROLOGIC ONCOLOGY

Seminar article

Statistical issues in the evaluation of screening and early detection modalities

Ruth Etzioni, Ph.D.*

Fred Hutchinson Cancer Research Center, Seattle, WA 98109-1024, USA

Abstract

Thorough evaluation of a screening test requires conducting a series of studies to ascertain its ability to detect accurately disease, as well as its benefits and costs. In this article, I review the steps involved in evaluating a screening test, using the case of prostate-specific antigen (PSA) screening for prostate cancer as a case study. I discuss designs for quantifying the diagnostic properties of a screening test and compare several different studies that have produced quite different estimates of the diagnostic accuracy of PSA screening. I also review methods that may be used to combine other markers or tests with PSA to improve test accuracy. Determining the benefits of a screening test is complex, particularly when information from randomized trials is lacking. I review several observational studies of PSA benefit and discuss the use of computer models for inferring the impact of screening from trends in population mortality.

Keywords: Prostate-specific antigen; Mass screening; Sensitivity; Specificity; ROC curve; Case-control study; Verification bias

Introduction

The concept of early detection (i.e., finding tumors early before they spread and become incurable) has tantalized cancer control researchers for many years. Recent advances in genomics and proteomics promise to expand vastly the pool of potentially useful early detection approaches. However, the evaluation of a new screening test is challenging and can be prone to a multitude of biases. Even if the test can be shown to detect preclinical disease and advance diagnosis, it does not automatically follow that the test will reduce disease-specific mortality, which is the ultimate goal of any early detection intervention. Moreover, even if the test can be shown conclusively to impact disease-specific mortality, the costs caused by false-positive results and overdiagnosis must be assessed, and evaluated relative to its benefits.

The case of prostate-specific antigen (PSA) screening for prostate cancer provides an excellent illustration of the many steps involved and the challenges that occur in the evaluation of new screening tests. Initial studies of PSA showed promising diagnostic performance, with a sensitivity of more than 70% among cases within 4 years before diagnosis [1], and a clear shift in stage toward clinically localized disease after detection by PSA screening [2]. The test rapidly disseminated in the United States in the early 1990s, more than doubling the incidence of prostate cancer [3], but the false-positive rate was not sufficiently low, particularly in older men and men with benign disease, and the issue of overdiagnosis soon became a grave concern. With large numbers of men undergoing PSA screening, the need to determine appropriate PSA cutoffs and quantify screening benefit became urgent. However, because randomized screening trials are not expected to yield results before 2008, researchers have been forced to rely on observational and population studies for evidence of benefit. As I shall demonstrate, it is extremely difficult to make unbiased, conclusive inferences about screening benefit from observational data. Consequently, there is still no consensus about whether PSA screening is associated with a reduction in disease-specific mortality. In the meantime, the diagnostic performance of PSA has been called into question as recent studies have shown that a nontrivial fraction of men with normal PSA levels have occult cancer [4].

In this article, I summarize the issues that occur when evaluating new screening tests and consider how they may be addressed through appropriate analytic techniques. I divide the steps involved in test evaluation into 3 broad areas:

* Corresponding author. Tel.: +1-206-667-4145; fax: +1-206-667-7004.

E-mail address: retztioni@fhcrc.org.

1078-1439/08/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.urolonc.2006.11.008

(1) measuring test accuracy and performance, (2) estimating the benefits of the test, and (3) quantifying the costs of the test. Investigators in the Early Detection Research Network have developed a more formal description of the steps involved in evaluating new screening tests and have partitioned the different types of studies involved into 5 phases of biomarker development [5]. In my concluding section, I show how the different types of studies discussed fit into this scheme. In my treatment of the steps involved, I shall most often refer to studies of PSA screening, although, where appropriate, I will also cite examples from other diseases and screening modalities. It is hoped that my review will not only assist researchers in developing new screening modalities but also consumers of screening studies, whose task is to interpret and apply the results in the clinical setting.

Measuring test accuracy and performance

When introducing a new screening test, the first question asked is whether the test is reliably able to detect latent disease. In this section, I review the most commonly used measures of test accuracy and discuss different study designs used in evaluating test performance. Under each study design, I briefly summarize techniques for estimating test accuracy and for comparing the performance of competing tests.

Measures of test accuracy

There are a number of different ways to measure test accuracy and performance [6]. The most commonly used are sensitivity (true-positive rate [TPR]) and specificity (1 minus false-positive rate [FPR]). A third measure, the positive predictive value (PPV), estimates the likelihood that an individual with a positive screen is, in fact, a disease case. The positive predictive value depends on the true-positive rate and false-positive rate, as well as the prevalence (p) of latent and undiagnosed disease in the population:

$$\mathrm{PPV} = \mathrm{p} * \mathrm{TPR} / [\mathrm{p} * \mathrm{TPR} + (1 - \mathrm{p}) * \mathrm{FPR}]$$

The positive predictive value is a useful measure because the ratio of 1 minus the positive predictive value to the positive predictive value is interpretable as the number of cancers detected per biopsy performed. For example, a positive predictive value of one third is synonymous with 1 cancer detected for every 3 biopsies performed, or, equivalently, 2 unnecessary biopsies per cancer detected. Thus, the positive predictive value provides a sense of the clinical implications of sensitivity and specificity.

All 3 of the aforementioned accuracy measures pertain to tests that produce a positive or negative result. For continuous tests, the receiver operating characteristic (ROC) curve is a plot of the sensitivity versus the false-positive rate of the test as the cutoff for declaring a positive test varies (Fig. 1). A summary measure, the area under the curve (AUC), is frequently used to compare continuous tests. The AUC is the probability that for a randomly select case and control, the case will have the higher value of the test variable. A larger AUC is considered to reflect a test with more desirable diagnostic properties.

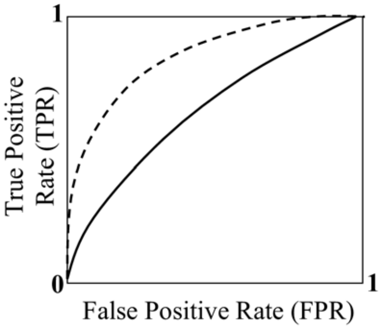

Fig. 1. ROC curves for 2 tests, showing the true-positive rate versus the false-positive rate as the threshold for positivity varies. The test with the higher curve (dashed curve) is preferred because it has a higher true-positive rate for every value of the false-positive rate. It also has a higher area under the curve.

Both sensitivity and specificity are critical in developing appropriate cutoffs for use of a continuous screening test. Low sensitivity (low true-positive rate) implies that a large proportion of latent cases will remain undiagnosed with potentially adverse consequences. However, low specificity (high false-positive rate) is problematic because of the costs and morbidity associated with false-positive diagnoses. Studies that propose cutoffs for continuous tests on the basis of (apparently) cancer-free individuals are effectively considering only test specificity. The ROC curve is particularly useful in determining an appropriate cutoff because it displays the trade-offs between the true-positive and false-positive rates associated with the test as the cutoff varies. In principle, the curve can be used to identify the cutoff that yields an optimal combination of sensitivity and specificity. However, this requires weighing the costs of false-positive tests against those of false-negative tests. Because these costs are naturally measured on very different metrics, this can be challenging. In practice, a target false-positive rate (or false-positive rate range) may be defined, possibly based on a value for disease prevalence and an acceptable positive predictive value; the goal then becomes to identify whether the value of the ROC curve at that false-positive rate is sufficiently high, reflecting a sufficiently sensitive test.

Study designs for measurement of test accuracy

Studies of test accuracy generally are in 1 of 2 categories: retrospective or prospective. In retrospective studies, patients are selected based on their disease status, with both cases (positive for disease) and controls (without disease) being included. Thus, these studies are also referred to