1

Introduction

1.1 Diagnostic Test Accuracy Studies

Diagnostic medicine is the process of identifying the disease or condition that a patient has and ruling out conditions that the patient does not have, through assessment of the patient's signs, symptoms, and results of various diagnostic tests. *Diagnostic accuracy studies* are research studies that examine the ability of diagnostic tests to discriminate between patients with and without the condition; these studies are the focus of this book.

A diagnostic test has several purposes: (1) to provide reliable information about the patient's condition, (2) to influence the health care provider's plan for managing the patient (Sox et al. 1989), and possibly, (3) to understand disease mechanism and natural history through research (e.g. the repeated testing of patients with chronic conditions) (McNeil and Adelstein 1976). A test can serve these purposes only if the health care provider knows how to interpret it. Diagnostic test studies are conducted to tell us how diagnostic tests perform and, thus, how they should be interpreted. There are several measures of diagnostic test performance. Fryback and Thornbury (1991) described a hierarchical model for studying diagnostic performance for imaging tests. The model starts with image quality and progresses to diagnostic accuracy, effects on treatment decisions, impact on patient outcome, and finally, costs to society. A key feature of the model is that for a diagnostic test to be efficacious at a higher level, it must be efficacious at all lower levels. The reverse is not true; for example, a new test may have better accuracy than a standard test but may be too costly (in terms of monetary expense and/or patient morbidity due to complications) to be efficacious. In this book, we deal exclusively with the assessment of diagnostic *accuracy* (level 2 of the hierarchical model), recognizing that it is only one step in the complete assessment of a diagnostic test.

Diagnostic test accuracy is simply the ability of the test to discriminate among alternative states of health (Zweig and Campbell 1993). If a test's results do not differ between alternative states of health, then the test has negligible accuracy; if the results do not overlap for the different health states, then the test has perfect accuracy. Most test accuracies fall between these two extremes. It's important to recognize that a test result is not a true representation of the patient's condition (Sox et al. 1989). Most diagnostic information is imperfect; it may influence the health care provider's thinking, but uncertainty remains about the patient's true condition. If the test is negative for the condition, should the health care provider assume that the patient is disease-free and thus send him or her home? If the test is positive, should the health care provider assume the patient has the condition and thus begin treatment? Finally, if the test result requires interpretation by a trained reader (e.g. a radiologist), should the health care provider seek a second interpretation?

To answer these critical questions, the health care provider needs to have information on the test's absolute and relative capabilities and an understanding of the complex interactions between the test and the trained readers who interpret the imaging data (Beam 1992). The health care provider must ask: How does the test perform among patients with the condition (i.e. the test's sensitivity)? How does the test perform among patients without the condition (i.e. the test's specificity)? Does the test serve as a replacement for an older test, or should multiple tests be performed? If multiple tests are performed, how should they be executed (i.e. sequentially or in parallel)? How reproducible are interpretations by different readers? These sorts of questions are addressed in diagnostic test accuracy studies.

Diagnostic test accuracy studies have three common features: a sample of subjects who have, or will, undergo one or more of the diagnostic medical tests under evaluation; some form of interpretation or scoring of the test's findings; and a reference, or *gold standard*, to which the test findings are compared. This may sound simple enough, but diagnostic accuracy studies are difficult to design. Here are three common misperceptions about diagnostic test accuracy.

The first misperception involves the interpretation of diagnostic tests. Investigators of new diagnostic tests sometimes develop criteria for interpreting their tests based only on the findings from healthy volunteers. For example, in a new test to detect pancreatitis, investigators measure the amount of a certain enzyme in healthy volunteers. A typical decision criterion, or *cut point*, is three standard deviations (SDs) below the mean of the normals. New patients with an enzyme level of three SDs below the mean of the healthy volunteers are labeled "positive" for pancreatitis; patients with enzyme levels above this cut point are labeled "negative." In proposing such a criterion, investigators fail to recognize (1) the relevance of the natural distributions of the test results (i.e. are they really Gaussian [normal]?), (2) the magnitude of any overlap between the test results of patients with and without pancreatitis (i.e. are the test results from most pancreatitis patients three SDs below the mean?), (3) the clinical significance of diagnostic errors (i.e. falsely labeling a patient without pancreatitis as "positive" for the condition and falsely labeling a patient with pancreatitis as "negative"), and (4) the poor generalization of results from studies based on healthy volunteers (i.e. healthy volunteers may have very different enzyme levels than sick patients without pancreatitis who might undergo the test). In Chapter 2, we discuss factors involved in determining optimal cut points for diagnostic tests; in Chapter 4, we discuss methods of finding optimal cut points and estimating diagnostic errors associated with them.

Another common misperception in diagnostic test studies is the notion that a rigorous assessment of a patient's true condition – with the exclusion of patients for whom a less rigorous assessment was made – allows for a scientifically sound study. An example comes from literature on the use of ventilation-perfusion lung scans for diagnosing pulmonary emboli. The ventilation-perfusion lung scan is a noninvasive test used to screen high-risk patients for pulmonary emboli; its accuracy in various populations is unknown. Pulmonary angiography, on the other hand, is a highly accurate but invasive test. It is often used as a reference for assessing the accuracy of other tests. (See Chapter 2 for the definition and examples of *gold standards*.) To assess the accuracy of ventilation-perfusion lung scans, patients who have undergone both a ventilation-perfusion lung scan and a pulmonary angiogram are recruited, while patients who did not undergo the angiogram are excluded. Such a design usually leads to biased estimates of test accuracy. The reason is that the study sample is not representative of the patient population undergoing ventilation-perfusion lung scans – rather, patients with a positive scan are often recommended for angiograms, while patients with a negative scan are often not sent for an angiogram because of the risk of complications with it. In Chapter 3, we define *work-up bias*, its most common form – *verification bias* – and discuss strategies to avoid them. In Chapter 10, we present statistical methods developed specifically to correct for verification bias.

A third error common in diagnostic test accuracy studies involves confusion between accuracy and agreement. Investigators sometimes draw incorrect conclusions about a new test's diagnostic accuracy because it agrees well with a conventional test; however, what if the new and conventional tests do not agree? We cannot simply conclude that the new test has inferior accuracy. In fact, a new test with superior accuracy will definitely sometimes disagree with the conventional test. Similarly, the two tests may have the same accuracy but make mistakes on different patients, resulting in poor agreement. A more valid approach to assessing a new test's diagnostic accuracy is to compare both tests against a gold standard reference. Assessment of diagnostic accuracy is usually more difficult than assessment of agreement, but it is a more relevant and valid approach (Zweig and Campbell 1993). In Chapter 5, we present methods for comparing the accuracy of two tests when the true diagnoses of the patients are known; in Chapter 11, we present methods for comparing two tests' accuracies when the true diagnoses are unknown.

There is no question that studies of diagnostic test accuracy are challenging to design and require specialized statistical methods for their analysis. We will present and illustrate concepts and methods for designing, analyzing, interpreting, and reporting studies of diagnostic test accuracy. In Part I (Chapters 1–7), we define various measures of diagnostic accuracy, describe strategies for designing diagnostic accuracy studies, and present the basic statistical methods for estimating and comparing test accuracy, calculating sample size, and synthesizing the literature for meta-analysis. In Part II (Chapters 8–16), we present more advanced statistical methods for describing a test's accuracy when patient characteristics affect it, for analyzing multi-reader studies, studies with verification bias or imperfect gold standards, and for performing meta-analyses.

1.2 Case Studies

We introduce three diagnostic test accuracy studies to illustrate the kinds of designs, questions, and statistical issues that arise in diagnostic medicine. These case studies, along with many other examples, are used throughout the book to illustrate various statistical methods. The datasets for these case studies are given in appendix at the end of the book.

1.2.1 Case Study 1: Parathyroid Disease

Parathyroid glands are small endocrine glands usually located in the neck or upper chest that produce a hormone that controls the body's calcium levels. Most people have four parathyroid glands. In the most common form of parathyroid disease, one of these glands grows into a benign tumor, called a parathyroid adenoma, which produces excess amounts of parathyroid hormone. In a less common condition, called parathyroid hyperplasia, all four parathyroid glands become enlarged and secrete excess parathyroid hormone. In both conditions, a patient's serum calcium levels become elevated, and the patient experiences loss of energy, depression, kidney stones, and headaches. Surgical removal of the offending parathyroid lesion is considered curative in most cases.

Single-photon emission computed tomography (SPECT) using the radiopharmaceutical Tc-99m sestamibi is a nuclear medicine imaging test used to detect and localize parathyroid lesions prior to surgical intervention. In this prospective study (D. Neumann, MD, PhD, Cleveland Clinic, Ohio, personal communication, 2007), 61 consecutive patients with hyperparathyroidism were imaged using a hybrid SPECT/CT instrument in an attempt to localize the diseased parathyroid glands preoperatively. Each patient underwent SPECT imaging, both with and without attenuation correction, as well as SPECT combined with CT imaging. Following imaging, the patients went to surgery to remove the diseased glands. The goal of the study was to compare the accuracy of these three tests.

One expert nuclear radiologist, blinded to the surgical findings, interpreted the images. On the SPECT imaging, each gland was scored on a scale from 1 to 7, with 1 = definitely no disease, 2 = probably no disease, 3 = indeterminate, 4 = maybe diseased, and 5 = definitely diseased. Scores of 5, 6, and 7 were all considered definitely diseased but were distinguished by the intensity of the attenuation: 5 = low, 6 = medium, and 7 = high, respectively. The SPECT/CT images were scored using just the 1–5 part of the scale. For this study, SPECT images scored as 1–3 were considered negative, and scores of 4–7 were considered positive. For SPECT/CT, scores of 1–3 were considered negative, and scores of 4–5 were considered positive. A total of 97 glands in 61 patients were localized by imaging prior to undergoing parathyroid surgery, the results of which were considered the gold standard.

The investigators wanted to compare the sensitivity and specificity of these three tests to determine which single test should be used for future patients. In Chapter 2, we show that one of these tests appears more sensitive than the others, while another test appears more specific. A comparison of the tests' receiver operating characteristic (ROC) curves gives us a complete understanding of the strengths and weaknesses of the three tests and thus allows us to identify the most suitable test for preoperative patients.

The data from this study are complicated by the fact that many of the 61 patients had multiple glands visualized at screening, so-called "clustered data." Observations from the same patient, even if from different glands, are usually correlated, at least to some small degree. If we ignore this correlation, then the resulting confidence intervals and p-values can be misleading. In Chapters 4 and 5, we describe a simple analysis method that can be used for clustered data so that confidence intervals and p-values are correct.

1.2.2 Case Study 2: Colon Cancer Detection

Polyps that form in the colon or rectum can progress to cancer without any signs or symptoms. Computed tomography colonography (CTC) is an imaging test that can detect polyps before they develop into cancer. Radiologists sometimes overlook polyps on the CTC images; however, and these missed polyps ("false negatives") can develop into cancer, which can lead to symptoms, even death. Investigators have developed a computer algorithm, called computer-aided detection (CAD), to help radiologists detect polyps on the CTC. The CAD utilizes tissue intensity, volumetric and surface shape, and texture characteristics to identify suspicious areas. The CAD marks the suspicious areas for the reader to examine more closely. Often, the CAD identifies multiple suspicious areas on the same image. The radiologist must distinguish marked areas that contain a polyp ("true positive") from marked areas that do not contain a polyp, for example a folded bowel lining ("false positive").

In this study (Baker et al. 2007), the investigators wanted to compare radiologists' accuracy without CAD to their accuracy with CAD to determine if CAD improves radiologists' accuracy. Seven radiologists from two institutions participated in the study. The readers had varying levels of overall experience with abdominal imaging, as well as varying levels of training with CTC imaging technology. Overall, the seven were considered inexperienced CTC readers.

Two hundred seventy patients from six institutions were compiled in this retrospective design. These 270 patients had undergone CTC for the following reasons: screening, follow-up exams for polyps detected in a prior exam, and failed prior colonoscopy, including patients at risk for colon polyps/carcinoma, but who were deemed not suitable candidates for a colonoscopy. An expert abdominal imager with extensive CTC experience and with knowledge about each patient's follow-up (clinical, imaging, pathologic, and surgical), stratified the 270-patient sample into presence versus absence of a polyp; cases with a polyp were further stratified by polyp size (less than 6 mm

"small," 6–9 mm "medium," or 10 mm or larger "large"). One hundred forty-one *training cases* were randomly sampled from the different strata to improve the CAD algorithm and train the readers. From the remaining 119 *test cases*, 30 were randomly selected to be used in this reader performance study; the study sample was composed of 25 positive cases with at least one polyp of middle to large size (a total of 39 polyps) and 5 cases with no polyps.

The seven readers were each given a unique order for reading the 30 images. First, without CAD, the reader marked all findings. The reader used a pull-down window to identify the location of each finding according to one of eight colon segments. The reader then scored each finding according to their confidence that a polyp was present: 1 = definitely not a polyp; 2 = probably not a polyp; 3 = indeterminate; 4 = maybe a polyp; and 5 = definitely a polyp. When the reader's interpretation without CAD was completed, the reader was given a list of potential polyps detected by the CAD. Any CAD marks that coincided with a lesion found by the reader without CAD were not presented to the reader and were discarded. New CAD marks were scored by the reader using the 1–5 rating scale. The investigators in this study want to know if the CAD improves inexperienced radiologists' accuracy over their accuracy without CAD ("unaided setting"). The seven-reader design helps us to get a better estimate of reader accuracy but also complicates the analyses because the readers' findings are correlated by the fact that they all interpreted the same sample of 30 patients. Sensitivity was defined for this study as correct detection of a polyp in a patient with polyps and, in addition, required that the reader identify the correct location of the polyp. If the wrong location was chosen, then the missed polyp was considered a false negative. In this study, patients could have multiple true positives (i.e. multiple correctly located polyps in the same patient), as well as a mixture of true positives, false positives, false negatives, and true negatives. Clustered data complicates the statistical analyses, but statistical methods are presented in Chapters 4 and 5 to handle these data appropriately.

1.2.3 Case Study 3: Carotid Artery Stenosis

Excessive plaque formation, or stenosis, in the carotid (neck) artery can lead to transient ischemic attacks (TIAs) or even stroke. Conventional catheter angiography is an invasive diagnostic test used by physicians to examine the carotid arteries in patients who have suffered a TIA or stroke. Because the test is invasive, there are risks associated with the test, including stroke and death. Magnetic resonance angiography (MRA) is a noninvasive test that may help physicians examine the carotid arteries without risk. Patients with other cardiovascular problems who are at high risk for plaque formation in the carotid arteries can also benefit from such a noninvasive screening test.

In this study, investigators (T. Masaryk, MD, Cleveland Clinic, Ohio, personal communication, 2007) wanted to assess the accuracy of MRA for detecting carotid artery plaque. Patients scheduled for a conventional catheter angiogram because they had suffered a recent stroke (symptomatic) or because they were at high risk for suffering a stroke in the future (asymptomatic) were asked to participate in this study. One hundred sixty-three patients were prospectively recruited for the study. These patients first underwent an MRA, then a conventional catheter angiogram.

Four radiologists from three institutions independently interpreted the conventional catheter angiograms, and the same four radiologists independently interpreted the MRA images. At least two weeks passed between the catheter angiogram and MRA interpretations; the study ID numbers were changed so that there was no obvious connection between the catheter angiogram and MRA images.

A significant stenosis requiring surgical intervention was defined as stenosis that blocked 60–99% of the carotid vessel. Note that arteries that are completely blocked (100% stenosis, or occlusions) are not considered good surgical lesions. The radiologists were asked to grade their confidence

that a significant stenosis was present using a 5-point scale: 1 = definitely no significant stenosis, 2 = probably no significant stenosis, 3 = equivocal, 4 = probably significant stenosis, and 5 = definitely significant stenosis. They were also asked to indicate the percentage of stenosis present (a number between 0 and 100). The radiologists responded to these questions for both the left and right sides for both MRA and conventional catheter angiography.

In this study, the investigators want to know the accuracy of MRA and whether or not it can replace the conventional invasive test, catheter angiography. The data are complicated by the multiple-reader design, as well as by the fact that the data are clustered (i.e. findings from both the left and right carotid arteries in the same patient). There are several patient characteristics, such as gender, age, and symptoms, which the investigators suspect may affect the accuracy of MRA. In Chapter 3, we discuss the kinds of effects that covariates can have on diagnostic test accuracy; in Chapter 8, we discuss various regression methods to handle covariate data. Finally, we note that the gold standard for this study, catheter angiography, is not a perfect test, and radiologists often disagree in their interpretations of its findings. Fortunately, there are statistical methods, which we describe in Chapter 11, that deal with studies with imperfect reference standards.

1.3 Software

A variety of software has been written to implement many of the statistical methods discussed in this book. These programs can be found in FORTRAN, SAS macros (SAS Institute, Cary, North Carolina, USA), Stata (Stata Data Analysis and Statistical Software, Stata Corp LP, College Station, Texas), and R (free software at http://www.r-project.org/). The authors have prepared a Website (https://xiaohuazhou.pku.edu.cn/info/1010/1008.htm) that contains the R code of various methods introduced in this book. The website will be maintained and updated periodically after this book's publication date.

1.4 Topics Not Covered in This Book

Although this book covers the main themes in statistical methods for diagnostic medicine, it does not cover several related topics, as follows.

Decision analysis, cost-effectiveness analysis, and cost-benefit analysis are methods commonly used to quantify the long-term, or downstream, effects of a test on the patient and society. In Chapters 2 and 4, we discuss how these methods can be applied to find the optimal cut point on the ROC curve. Description of how to perform these methods, however, is beyond the scope of this book. There are many excellent references on these topics, including Gold et al. (1996), Pauker and Kassirer (1975), Russell et al. (1996), Weinstein et al. (1996), Drummond et al. (2005), Glick et al. (2007), and Willan and Briggs (2006).

Most of the methods we present for estimation and hypothesis testing are from a frequentist perspective. Bayesian methods can also be used, whereby one incorporates into the assessment of the diagnostic test some previously acquired information or expert opinion about a test's characteristics or information about the patient or population. Examples of Bayesian methods used in diagnostic testing include Gatsonis (1995), Joseph et al. (1995), Peng and Hall (1996), Hellmich et al. (1988), O-Malley and Zou (2001), Broemeling (2007).

Finally, when there are multiple diagnostic tests performed on a patient, we may want to combine the information from the tests in order to make the best possible diagnosis. See, for example, Pepe and Thompson (2000), Zhou et al. (2011), and Lin et al. (2011) for various methods for combining tests' results to optimize diagnostic accuracy.