9

Diagnostic Tests

Gregory E. Gray, M.D., Ph.D.

Questions about diagnostic tests are most commonly related to the accuracy of the test, which is the focus of this chapter. For information on topics such as the clinical and economic impact of screening, clinical decision rules, and the differential diagnosis process, readers are referred to recent monographs (Hunink et al. 2001; Knottnerus 2002) and standard textbooks of clinical epidemiology (Fletcher et al. 1996; Sackett et al. 1991).

Evaluating Diagnostic Tests

The accuracy of a diagnostic test is generally assessed in a cross-sectional study, in which patients are evaluated with both the "gold standard" diagnostic procedure and the diagnostic test under evaluation (Knottnerus and Muris 2002; Knottnerus et al. 2002; Newman et al. 2001). Several important issues can affect the validity of such evaluations (Fletcher et al. 1996; Knottnerus and Muris 2002; Knottnerus et al. 2002; Newman et al. 2001; Sackett et al. 1991), but the two major issues are the choice of the gold standard and the choice of subjects.

Choice of Gold Standard

The first issue is related to the choice of the gold standard. In some branches of medicine, the gold standard might be a pathological diagnosis based on a biopsy or an autopsy; in psychiatry, we usually rely on the DSM-IV-TR criteria (American Psychiatric Association [APA] 2000).

Although the DSM-IV-TR diagnostic criteria are often used as the gold standard for evaluating diagnostic tests, they have limitations. First, there are unresolved issues concerning the validity of the criteria themselves (Goldstein and Simpson 2002; Kendell 1989; Kendler 1990).

The second issue relates to the method of eliciting symptoms from patients and of using this information to arrive at a diagnosis. In unstructured clinical interviews, clinicians may ignore or not inquire about certain symptoms and may choose not to follow the DSM criteria in arriving at a diagnosis (Robins 2002). As a result, unstructured interviews may not be reliable. For this reason, a variety of diagnostic instruments have been developed to standardize the process of psychiatric diagnosis, including structured clinical interviews (e.g., the Structured Clinical Interview for DSM-IV [SCID]) and structured interviews that may be administered by either a lay interviewer or a computer (e.g., the Composite International Diagnostic Interview [CIDI] and the Diagnostic Interview Schedule [DIS]) (Kobak et al. 2008). Although these standardized instruments increase the reliability of the diagnoses made, issues of validity remain (Murphy 2002; Narrow et al. 2002; Regier et al. 1998; Robins 2002).

For the assessment of particular symptoms or specific measures of cognitive function (e.g., memory or intelligence), well-established instruments are generally used as the gold standard. Descriptions of many of these are included in the *Handbook of Psychiatric Measures* (Rush et al. 2008)

Choice of Subjects

The other major issue relates to the choice of subjects. We administer diagnostic tests because we have uncertainty about a diagnosis. A cross-sectional

study that includes a spectrum of patients who are similar to those to whom the test is expected to be administered in clinical practice is the most appropriate design. All too often, a test is evaluated in a population composed of a mix of very ill patients and a healthy control group. In such a population, the test performs much better in distinguishing the ill from the healthy than in actual practice. This is referred to as *spectrum bias* (Knottnerus et al. 2002; Newman et al. 2001).

Measures of Test Performance: Dichotomous Results

A variety of terms are used to describe the performance of a diagnostic test (Habbema et al. 2002). As an aid in discussing these terms, Table 9–1 displays the results of a hypothetical comparison of a new diagnostic test with an appropriate gold standard. In this section, we consider test results to be dichotomous (i.e., either positive or negative).

True Positive, True Negative, False Positive, and False Negative

If both the diagnostic test being evaluated and the gold standard yield positive results, the result of the diagnostic test is considered true positive (Table 9–1, cell A). Likewise, if both yield negative results, the result of the diagnostic test is considered true negative (Table 9–1, cell D). If the diagnostic test under evaluation gives a positive result, but the gold standard gives a negative result, the result of the diagnostic test is considered false positive (Table 9–1, cell B). Similarly, if the diagnostic test result is negative, but the gold standard result is positive, the result of the diagnostic test is considered false negative (Table 9–1, cell C).

Sensitivity and Specificity

Sensitivity refers to the proportion of patients with the disease (as assessed by the gold standard) who are detected by the diagnostic test. In Table 9–1, this is calculated as A/(A+C). A highly sensitive test is one that detects most cases of disease.

Specificity refers to the proportion of patients without the disease (as assessed by the gold standard) who are identified by the diagnostic test as *not* having the disease. In Table 9–1, this is calculated as D/(B+D). A highly specific test is one that does not misidentify healthy individuals as having disease.

Although it may seem counterintuitive, highly sensitive diagnostic tests are most useful for ruling *out* diseases. This is because such tests seldom miss cases of disease. Sackett et al. (2000) have proposed the mnemonic SnNout (**Sn**sitive test, **N**egative result, rules **out** disease) as a teaching aid. Similarly, highly specific tests are most useful for ruling *in* diseases, because they seldom misidentify healthy individuals. The mnemonic for such tests is SpPin (**Sp**ecific test, **P**ositive test, rules **in** disease).

Confidence intervals (CIs) for specificity and sensitivity can be calculated using the tables given by Habbema et al. (2002).

Positive and Negative Predictive Values

The positive predictive value (PPV) is the proportion of positive test results that is true positives. Using Table 9–1, PPV is calculated as A/(A+B). The negative predictive value (NPV) is the proportion of negative test results that is true negatives. Using Table 9–1, NPV is calculated as D/(C+D).

PPV and NPV can also be looked at in terms of posttest probabilities of disease (Fletcher et al. 1996; Habbema et al. 2002). PPV represents the probability that an individual with a positive test result has the disease, whereas NPV represents the probability that an individual with a negative test result does not have the disease.

PPV and NPV are dependent on the prevalence of disease in the population being tested, which is also referred to as the *pretest probability of disease* (Fletcher et al. 1996). Mathematically, the relationship is as follows:

$$\text{PPV} = \frac{\text{sensitivity} \times \text{prevalence}}{(\text{sensitivity} \times \text{prevalence}) + [(1 - \text{specificity})(1 - \text{prevalence})]}$$

Because of this dependence on disease prevalence, screening for diseases in low-prevalence populations yields few true-positive test results, regardless of the sensitivity and specificity of the test (Fletcher et al. 1996). Baldessarini et al. (1983) have provided an excellent discussion of the effects of disease prevalence on PPV in psychiatric practice, using as their example the dexamethasone suppression test as a diagnostic test for depression.

TABLE 9–1. Possible results of a diagnostic test

	Disease present	Disease absent	Totals
Test results positive	A True positive	B False positive	A+B
Test results negative	C False negative	D True negative	C+D
Totals	A+C	B+D	

Sensitivity=A/(A+C)
Specificity=D/(B+D)
Positive predictive value=A/(A+B)
Negative predictive value=D/(C+D)
Likelihood ratio of positive test=[A/(A+C)]/[B/(B+D)]
Likelihood ratio of negative test=[C/(A+C)]/[D/(B+D)]
Error rate=(B+C)/(A+B+C+D)
Accuracy=(A+D)/(A+B+C+D)

Likelihood Ratios

Although PPV and NPV allow the clinician to estimate the probability of disease in a patient with a particular test result, they are both dependent on the prevalence of disease in the study population. Likelihood ratios (LRs), in contrast, are independent of disease prevalence.

The likelihood ratio of a positive test (LR+) is the ratio of the likelihood (probability) of a positive test result in the population with disease divided by the likelihood of a positive test result in the population without disease. This can be calculated as follows, using the data in Table 9–1:

$$\text{LR}+ = \frac{A/(A+C)}{B/B+D} = \frac{\text{sensitivity}}{1-\text{specificity}}$$

The likelihood ratio of a negative test result (LR–) is similarly the ratio of the likelihood of a negative test result in the population with disease divided by the likelihood of a negative test result in the population without disease. Once again, using the data in Table 9–1:

$$\text{LR}- = \frac{C/(A+C)}{D/(B+D)} = \frac{1-\text{sensitivity}}{\text{specificity}}$$

Formulas for calculating CIs for LRs are given by Habbema et al. (2002).

One of the most useful features of LRs is that they can be used to estimate the probability that a given patient has an illness, given a particular test result. The following formula can be used for this estimate:

$$\text{posttest odds} = \text{pretest odds} \times \text{LR}$$

The odds of an event occurring is the ratio of the probability of an event occurring (P) divided by the probability of the event not occurring ($1-P$), or the odds$=P/(1-P)$.

The pretest probability of disease in a patient can be estimated from the prevalence of a disease in your own or similar patient populations, as derived from papers evaluating diagnostic tests, epidemiologic studies, or hospital statistics (Mant 1999; Sackett et al. 2000). For some medical conditions, clinical prediction rules may be used to better refine the estimated pretest probability (McGinn 2002). Sometimes published data are not available and a subjective estimate of pretest probability must be made, although this is subject to numerous biases (Elstein and Schwartz 2002; Hunink et al. 2001). When there is uncertainty about the pretest probability, it is best to estimate an upper and lower limit and to do the calculations with both values to see if it will change your management of the patient.

Once you have a pretest probability of disease, it is converted to the pretest odds, using the formula above. The pretest odds are then multiplied by the appropriate likelihood ratio (LR+ if the test is positive or LR– if the test is negative). The resulting posttest odds can then be converted to a posttest probability, using the following formula:

$$P = \frac{\text{odds}}{1+\text{odds}}$$

For those wishing to go directly from pretest probability to posttest probability, without having to go back and forth between probabilities and odds, nomograms and computer programs are available (Fagan 1975; Glasziou 2001).

Using diagnostic tests in this way to quantify the risk of disease in a patient and to then use these probabilities in clinical decision making has become an important field of study, but it is outside the scope this chapter. For further details, readers are referred to Hunink et al. (2001) and Sackett et al. (1991, 2000).

Error Rate and Accuracy

The error rate of a diagnostic test is the percentage of test results that is either false positives or false negatives (Habbema et al. 2002). Using the data in Table 9–1, error rate can be calculated as follows:

$$\frac{B+C}{A+B+C+D}$$

The term *accuracy* is sometimes used to describe the percentage of test results that is either true positives or true negatives (Fletcher et al. 1996; Greenhalgh 2001). Using the data in Table 9–1, accuracy is calculated as follows:

$$\frac{A+D}{A+B+C+D}$$

Accuracy can also be calculated as 1−error rate.

Diagnostic Odds Ratio

The diagnostic odds ratio (DOR) is sometimes used as a measure of overall test performance (Habbema et al. 2002). Using the data in Table 9–1, DOR is calculated as follows:

$$\frac{A \times D}{B \times C}$$

The DOR is sometimes used in systematic reviews and in meta-analyses because of its mathematical properties; however, it is not a statistic that is easy to interpret clinically (Deeks 2001).

Example of Calculations

Table 9–2 provides data from a study by Watkins et al. (2001) that evaluates the accuracy of a single question ("Do you often feel sad or depressed?") in screening for depression in stroke patients. In the study, answers to the question ("yes" or "no") were compared with the results of the Montgomery-Åsberg Rating Depression Rating Scale (MADRS), with a MADRS score >6 indicating depression. As can be seen from Table 9–2, the single question had a sensitivity of 86% and a specificity of 78%.

We can use LRs to see the impact of a positive or negative response to the question on the probability of a patient being depressed. Depression typically occurs in 25%–40% of patients with stroke and other neurologic conditions (APA 2000). Thus, the pretest odds for such a patient are between 25/75=0.33 and 40/60=0.67. If the patient answers "yes" to the question, the posttest odds are between 0.33×3.87=1.28 and 0.67×3.87=2.59, corresponding to posttest probabilities of 56%–72%. If the patient answers "no," the posttest odds are between 0.33×0.18=0.06 and 0.67×0.18=0.12, corresponding to posttest probabilities of 6%–11%.

Measures of Test Performance: Ordinal or Continuous Results

Although we often think of diagnostic test results as positive or negative, rating scales and many laboratory tests have more than just two values. Rating scales, for example, can have outcomes that are ordered categories (e.g., none, minimal, mild, moderate, or severe) or that yield a numerical score. Although a cutoff value is often chosen to convert such an outcome to a dichotomous measure, information is lost in the process.

Choosing a Cutoff Score

When a test result can take more than two values as described above, a decision must be made about a cutoff score if the results are to be viewed as either positive or negative. The choice of the cutoff value involves a trade-off between sensitivity and specificity (Fletcher et al. 1996).

Table 9–3 shows the results of a study of the CAGE questionnaire as a screening tool for diagnosing alcohol abuse in general medical patients (Buchsbaum et al. 1991). (CAGE is an acronym for certain key symptoms of alcohol abuse.) As can be seen, as the cutoff score is increased, the specificity increases, but the sensitivity decreases. At the usual cutoff score of 2 (Martino et al. 2008), the sensitivity is 74% and the specificity is 91%.

Receiver Operator Characteristic Curves

The data in Table 9–3 can also be displayed graphically, in the form of a receiver operator characteristic

TABLE 9–2. Accuracy of a single question ("Do you often feel sad or depressed?") in screening for depression in stroke patients

Answer to question	Depressed (MADRS>6)	Not depressed (MADRS ≤6)	Totals
Yes	37	8	45
No	6	28	34
Totals	43	36	79

Sensitivity=37/43=86%
Specificity=28/36=78%
Positive predictive value=37/45=82%
Negative predictive value=28/34=82%
Likelihood ratio of positive test=0.86/0.22=3.9
Likelihood ratio of negative test=0.14/0.78=0.18
Error rate=14/79=18%
Accuracy=65/79=82%

Note. MADRS=Montgomery-Åsberg Depression Rating Scale.
Source. Data from Watkins et al. 2001.

(ROC) curve (Figure 9–1). An ROC curve plots the sensitivity against (1–specificity) for each cutoff value (Fletcher et al. 1996; Sackett et al. 1991, 2000). The ROC curve can be used to identify a cutoff value, where a point near the upper left corner is the appropriate cutoff. ROC curves can also be used to compare two diagnostic tests; the better test is the one with its ROC curve closer to the upper left corner (Fletcher et al. 1996).

Likelihood Ratios

Intuitively, it would seem that converting numerical or ordered test results into just two categories would result in a loss of information. It would seem that a value greatly over the cutoff would be more indicative of disease than a value that is barely more than the cutoff. LRs allow these differences to be taken into account.

If a test result takes multiple values, LRs can be calculated for multiple scores, not just LR+ and LR–. This is illustrated in Table 9–4, using the same CAGE data found in Table 9–3 (Buchsbaum et al. 1991). It can be seen that as the CAGE score increases from 0 to 4, LRs increase from 0.14 to 100. Assuming a prevalence (pretest probability) of alcohol dependence of 5% (APA 2000), a CAGE score of 0 decreases the posttest probability to 1%. Using the same figures, a CAGE score of 2 increases the posttest probability to 19% and a CAGE score of 5 increases the posttest probability to 84%. Clearly, this provides more information than simply treating the results as *positive* (2+) or *negative* (0–1).

Systematic Reviews and Meta-Analyses of Diagnostic Tests

As with the evaluation of therapies, systematic reviews that are related to diagnostic tests may be conducted. However, there are several approaches to meta-analysis (Deeks 2001; Glasziou et al. 2001). One approach is to use one forest plot for sensitivities and their CIs and another for specificities and their CIs (see Chapter 6). Another approach is to plot the sensitivity against (1–specificity) on an

TABLE 9–3. Sensitivity and specificity of the CAGE questionnaire in screening for alcohol abuse in general medical patients

CAGE score cutoff	Sensitivity (%)	Specificity (%)
0 vs. 1+	89	81
<2 vs. 2+	74	91
<3 vs. 3+	44	98
<4 vs. 4+	25	100

Source. Data from Buchsbaum et al. 1991.

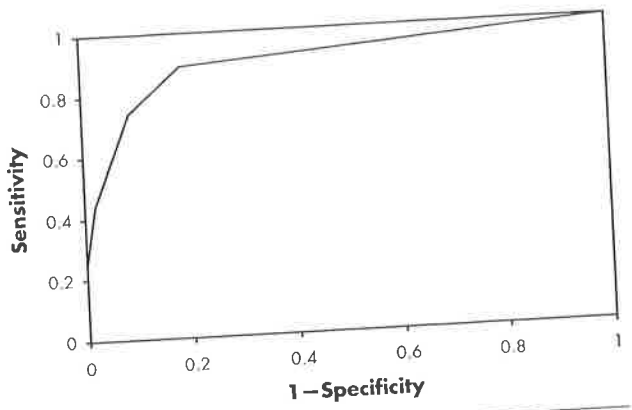

FIGURE 9–1. Receiver operator characteristic curve for CAGE questionnaire.

Source. Data from Buchsbaum et al. (1991).

ROC plot, with each study representing one point on the plot. Methods deriving pooled estimates of sensitivity, specificity, and LRs are also available. Finally, DORs can be pooled and this can be used to derive a summary ROC curve. Although such methods can demonstrate that a test is useful, if there is heterogeneity in results the summary statistics may not give a valid estimate of the probability associated with a specific test outcome (Deeks 2001).

Critical Appraisal Guide for Studies of Diagnostic Test Accuracy

Having discussed how diagnostic tests are evaluated and measures of their accuracy, we now turn to the critical appraisal of reports of such evaluations. A guide for critically appraising studies of diagnostic test accuracy is given in Table 9–5.

Is the Study Valid?

As discussed earlier in this chapter, the appropriate study design to assess the accuracy of a diagnostic test is a cross-sectional study in which patients are evaluated with both the gold standard diagnostic procedure and the diagnostic test under evaluation (Knottnerus and Muris 2002; Knottnerus et al. 2002; Newman et al. 2001). Both tests should be administered blindly to avoid observer bias (Fletcher et al. 1996), and the results of one test should not influence the decision to administer the other test. The gold standard itself should be appropriate, and an appropriate spectrum of patients should be chosen to avoid spectrum bias (Knottnerus et al. 2002; Newman et al. 2001).

TABLE 9–4. Likelihood ratios for specific CAGE scores in screening for alcohol abuse in general medical patients

CAGE score	Likelihood ratio
0	0.14
1	1.5
2	4.5
3	13
4	100

Source. Data from Buchsbaum et al. 1991.

Are the Results Important?

Sensitivity and specificity are the usual measures of diagnostic test accuracy. Which of the two is more important will depend on the purpose of the test. Recall that highly sensitive diagnostic tests are most useful for ruling *out* diseases (SnNout: **Sen**sitive test, **N**egative result, rules **out** disease), whereas highly specific tests are most useful for ruling *in* diseases (SpPin: **Sp**ecific test, **P**ositive test, rules **in** disease) (Sackett et al. 2000).

PPV and NPV will provide estimates of the probabilities of disease in patients with positive or negative test results, respectively, in the study population. However, LR+ and LR– are more useful measures, because they can be generalized to other populations with different disease prevalences (Fletcher et al. 1996; Mant 1999).

Can I Apply the Results to My Patient?

The first question to ask is whether the test is feasible in your setting. Does it require specialized equipment or is it very expensive? Does it require expertise to administer or to interpret the test, and is such expertise available?

The next question is specific to your patient and relates to whether the results of the test will change your patient management. As described in more detail elsewhere (Hunink et al. 2001; Sackett et al. 1991, 2000), physicians make decisions about clinical management, based on their (often unstated) as-

TABLE 9–5. Critical appraisal guide for studies of diagnostic accuracy

Is the study valid?
Is it a cross-sectional study?
Was the test evaluated in an appropriate spectrum of patients?
Was there an independent blind comparison with a "gold standard"?
Were both tests administered, regardless of outcome?
Are the results important?
What is the sensitivity?
What is the specificity?
What is the positive predictive value?
What is the negative predictive value?
What is the likelihood ratio for a positive test result?
What is the likelihood ratio for a negative test result?
Can I apply the results to my patient?
Is the test available or feasible in my setting?
Do I have enough information to apply the test and to interpret the results?
Given reasonable pretest estimates of disease probability in my patient, what are the posttest probabilities if the test is positive? If it is negative?
Will these probabilities change my management of the patient?

Source. Based on Gilbert et al. 2001; Glasziou et al. 2001; Sackett et al. 2000.

sessments of the probability of disease in a given patient. If that probability is high, physicians treat the patient; if that probability is intermediate, physicians order more diagnostic tests; and if it is low, physicians do neither. The probability at or above which tests are ordered is termed the *test threshold*, and the probability at or above which treatment is begun is termed the *treatment threshold* (Hunink et al. 2001; Sackett et al. 2000).

If your patient's pretest probability is already above the treatment threshold, there is no point in ordering the diagnostic test. For example, if your diagnostic interview shows that your patient met the DSM-IV-TR criteria for major depression, there is no need to also use a rating scale developed to screen for depression. You have already established the diagnosis with enough certainty to begin treatment.

Conversely, if the posttest probability remains low even if the test result is positive, the test should probably not be performed because it is likely that any positive test result will be a false positive. Ordering the test under these circumstances would be both a waste of money and an unnecessary cause of anxiety for patients with false-positive results (Fletcher et al. 1996).

To do these sorts of calculations, you will need the LR+ and LR– from the article reporting the performance of the test in question and an estimate of your patient's pretest probability of disease. If your patient is similar to patients in the study population, the pretest probability can be derived from the study evaluating the diagnostic test. If not, the pretest probability can be estimated from epidemiologic data or from patient data from your hospital or health maintenance organization (Mant 1999; Sackett et al. 2000). If published data are not available, you may have to make a subjective estimate of the pretest. In this case, it is probably best to estimate an upper and lower limit and to do the calculations with both values.

If the test is valid and feasible and if the results will make a difference in patient management, you should use it.

References

American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision. Washington, DC, American Psychiatric Association, 2000

Baldessarini RJ, Finkelstein S, Arana GW: The predictive power of diagnostic tests and the effect of prevalence of illness. Arch Gen Psychiatry 40:569–573, 1983

Buchsbaum DG, Buchanan RG, Centor RM, et al: Screening for alcohol abuse using CAGE scores and likelihood ratios. Ann Intern Med 115:774–777, 1991

Deeks JJ: Systematic reviews of evaluation of diagnostic and screening tests. BMJ 323:157–162, 2001

Elstein AS, Schwartz A: Clinical problem solving and diagnostic decision making: a selective review of the cognitive research literature, in The Evidence Base of Clinical Diagnosis. Edited by Knottnerus JA. London, BMJ Books, 2002, pp 179–195

Fagan TJ: Nomogram for Bayes' theorem. N Engl J Med 293:257, 1975

Fletcher RH, Fletcher SW, Wagner EH: Clinical Epidemiology: The Essentials, 3rd Edition. Baltimore, MD, Williams & Wilkins, 1996

Gilbert R, Logan S, Moyer VA, et al: Assessing diagnostic and screening tests, II: how to use the research literature on diagnosis. West J Med 175:37–41, 2001

Glasziou P: Which methods for bedside Bayes? ACP J Club 135:A11–A12, 2001

Glasziou P, Irwig L, Bain C, et al: Systematic Reviews in Health Care: A Practical Guide. Cambridge, UK, Cambridge University Press, 2001

Goldstein JM, Simpson JC: Validity: definitions and applications to psychiatric research, in Textbook in Psychiatric Epidemiology, 2nd Edition. Edited by Tsuang MT, Tohen M. New York, Wiley-Liss, 2002, pp 149–163

Greenhalgh T: How to Read a Paper: The Basics of Evidence Based Medicine, 2nd Edition. London, BMJ Books, 2001

Habbema JDF, Eijkemans R, Krijnen P, et al: Analysis of data on the accuracy of diagnostic tests, in The Evidence Base of Clinical Diagnosis. Edited by Knottnerus JA. London, BMJ Books, 2002, pp 117–143

Hunink M, Glasziou P, Siegel J, et al: Decision Making in Health and Medicine: Integrating Evidence and Values. New York, Cambridge University Press, 2001

Kendell RE: Clinical validity. Psychol Med 19:45–55, 1989

Kendler KS: Toward a scientific psychiatric nosology: strengths and limitations. Arch Gen Psychiatry 47:969–973, 1990

Knottnerus JA (ed): The Evidence Base of Clinical Diagnosis. London, BMJ Books, 2002

Knottnerus JA, Muris JWM: Assessment of the accuracy of diagnostic tests: the cross-sectional study, in The Evidence Base of Clinical Diagnosis. Edited by Knottnerus JA. London, BMJ Books, 2002, pp 39–60

Knottnerus JA, van Weel C, Muris JWM: Evaluation of diagnostic procedures. BMJ 324:477–480, 2002

Kobak KA, Skodol AE, Bender DS: Diagnostic measures for adults, in Handbook of Psychiatric Measures, 2nd Edition. Edited by Rush AJ, First MB, Blacker D. Washington, DC, American Psychiatric Publishing, 2008, pp 35–60

Mant J: Is this test effective? In Evidence-Based Practice: A Primer for Health Professionals. Edited by Dawes M, Davies P, Gray A, et al. New York, Churchill Livingstone, 1999, pp 133–157

Martino S, Poling J, Rounsaville BJ: Substance use disorders measures, in Handbook of Psychiatric Measures, 2nd Edition. Edited by Rush AJ, First MB, Blacker D. Washington, DC, American Psychiatric Publishing, 2008, pp 437–470

McGinn T: Practice corner: using clinical prediction rules. ACP J Club 137:A10–A11, 2002

Murphy JM: Symptom scales and diagnostic schedules in adult psychiatry, in Textbook in Psychiatric Epidemiology, 2nd Edition. Edited by Tsuang MT, Tohen M. New York, Wiley-Liss, 2002, pp 273–332

Narrow WE, Rae DS, Robins LN, et al: Revised prevalence estimates of mental disorders in the United States: using a clinical significance criterion to reconcile 2 surveys' estimates. Arch Gen Psychiatry 59:115–123, 2002

Newman TB, Browner WS, Cummings SR: Designing studies of medical tests, in Designing Clinical Research: An Epidemiologic Approach, 2nd Edition. Edited by Hulley SB, Cummings SR, Browner WS, et al. Philadelphia, PA, Lippincott Williams & Wilkins, 2001, pp 175–191

Regier DA, Kaelber CT, Rae DS, et al: Limitations of diagnostic criteria and assessment instruments for mental disorders: implications for research and policy. Arch Gen Psychiatry 55:109–115, 1998

Robins LN: Birth and development of psychiatric interviews, in Textbook in Psychiatric Epidemiology, 2nd Edition Edited by Tsuang MT, Tohen M. New York, Wiley-Liss 2002, pp 257–271

Rush AJ, First MB, Blacker D (eds): Handbook of Psychiatri Measures, 2nd Edition. Washington, DC, American Psy chiatric Publishing, 2008

Sackett DL, Haynes RB, Guyatt GH, et al: Clinical Epide miology: A Basic Science for Clinical Medicine, 2nd Ed tion. Boston, MA, Little, Brown, 1991

Sackett DL, Straus SE, Richardson WS, et al: Evidenc Based Medicine: How to Practice and Teach EBM, 2r Edition. New York, Churchill Livingstone, 2000

Watkins C, Daniels L, Jack C, et al: Accuracy of a single que tion in screening for depression in a cohort of patients ter stroke: comparative study. BMJ 323:1159, 2001