American Dietetic Association

RESEARCH

Review

Meets Learning Need Codes 9000, 9010, 9060, and 9070. To take the Continuing Professional Education quiz for this article, log in to ADA's Online Business Center at www.eatright.org/obc, click the "Journal Article Quiz" button, click "Additional Journal CPE Articles," and select this article's title from a list of available quizzes.

Publishing Nutrition Research: Validity, Reliability, and Diagnostic Test Assessment in Nutrition-Related Research

PHILIP M. GLEASON, PhD; JEFFREY HARRIS, DrPH, RD; PATRICIA M. SHEEAN, PhD, RD; CAROL J. BOUSHEY, PhD, MPH, RD; BARBARA BRUEMMER, PhD, RD

ABSTRACT

This is the sixth in a series of monographs on research design and analysis. The purpose of this article is to describe and discuss several concepts related to the measurement of nutrition-related characteristics and outcomes, including validity, reliability, and diagnostic tests. The article reviews the methodologic issues related to capturing the various aspects of a given nutrition measure's reliability, including test–retest, inter-item, and interobserver or inter-rater reliability. Similarly, it covers content validity, indicators of absolute vs relative validity, and internal vs external validity. With respect to diagnostic assessment, the article summarizes the concepts of sensitivity and specificity. The hope is that dietetics practitioners will be able to both use high-quality measures of nutrition concepts in their research and recognize these measures in research completed by others.
J Am Diet Assoc. 2010;110:409-419.

P. M. Gleason is a senior fellow, Mathematica Policy Research, Geneva, NY. J. Harris is an associate professor and didactic program director, Department of Health, West Chester University of Pennsylvania, West Chester, PA. P. M. Sheean is an instructor, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL. C. Boushey is an associate professor and director, Coordinated Program in Dietetics, Purdue University, West Lafayette, IN. B. Bruemmer is a senior lecturer, Graduate Program in Dietetics, and director, Graduate Coordinated Program in Dietetics, University of Washington, Seattle.

Address correspondence to: Philip M. Gleason, PhD, Mathematica Policy Research, 331 Washington St, Geneva, NY 14456. E-mail: pgleason@mathematica-mpr.com

Manuscript accepted: September 30, 2009.

0002-8223/10/11003-0007$36.00/0
doi: 10.1016/j.jada.2009.11.022

This is the sixth in a series of monographs on research design and analysis. The purpose of this monograph is to describe and discuss several concepts related to the measurement of nutrition-related characteristics and outcomes, including validity, reliability, and diagnostic tests. A critical aspect of implementing quality nutrition-related research involves the techniques for actually conducting the analysis, and several of the previous monographs in this series have covered various analytical techniques (1-5). No matter how sophisticated the analytical techniques, the research will not be useful unless the variables and concepts being analyzed are measured accurately and reliably and the results interpreted appropriately. For example, suppose a study of the relationship between individuals' usual intake of saturated fat and their cardiovascular health used an unreliable measure of their usual intake of saturated fat by measuring intake at a single meal on a single day. The amount of saturated fat consumed in a single meal is likely a poor indicator of day-to-day saturated fat intake. Thus, the research may fail to find a significant relationship even if a relationship exists in the population being studied.

This report describes key issues surrounding measurement and interpretation of nutrition-related outcomes and tools used in diagnostic assessment. The first section focuses on the measurement of nutrition outcomes and concepts, with particular attention on the measurement concepts of reliability and validity. The second section examines statistical tools used in nutrition-related diagnostic assessment. Figure 1 provides a glossary of relevant terminology.

THE MEASUREMENT OF NUTRITION EXPOSURE, OUTCOMES, AND CONCEPTS

In any research study, an important step involves measuring or operationalizing some set of concepts or variables of interest. These may include the main outcomes being examined in the study as well as variables measuring exposure to an intervention or thought to influence the outcomes of interest. Self-report methods are common to nutrition studies and are frequently employed to examine an individual's usual nutrient in-

Absolute validity Extent to which a measure exactly captures the concept it is intended to reflect.

Biomarker External variable, typically a component of body fluids or tissues, which has a strong direct relationship with dietary intakes of one or more dietary components.[a]

Cohen's kappa coefficient Statistical measure of the amount of agreement between two measures of the same concept, used to capture test–retest or inter-rater reliability.

Content validity Extent to which a measure covers all dimensions present in the concept it is intended to reflect.

Convergent validity Extent to which several different measures of a concept agree with each other and with a test measure of that concept; considered a type of relative validity.

Criterion validity Any type of validity based on a comparison of a test measure to a criterion intended to reflect the exact value of the concept the measure is intended to reflect.

Cronbach's α coefficient Statistical measure of the amount of agreement among a set of different items making up a scale or index; used to capture inter-item reliability or internal consistency; similar to Kuder-Richardson formula 20.

Discriminant validity Extent to which a measure of a concept disagrees with each another measure intended to reflect the opposite of that concept; considered a type of relative validity.

External validity Extent to which a measure captures the concept is it is intended to reflect not only among the sample of individuals being studied, but also in the broader population represented by that sample.

Face validity Extent to which a measure appears to most observers to capture the concept it is intended to reflect.

False negative Result of a test for the presence of a disease or condition indicating that the disease or condition is not present for a given subject when it fact the disease or condition is present.

False positive Result of a test for the presence of a disease or condition indicating that the disease or condition is present for a given subject when it fact the disease or condition is not present.

Inter-item reliability Extent to which multiple indicators of a single construct are correlated.

Internal validity Extent to which a measure captures the concept is it is intended to reflect among the sample of individuals being studied.

Inter-rater reliability Extent to which different raters or observers of a given measure come up with the same value of the measure for a given case; also known as inter-observer reliability.

Kuder-Richardson formula 20 Statistical measure of the amount of agreement among a set of different items making up a scale or index; used to capture inter-item reliability or internal consistency and similar to Cronbach's α, but used with binary items (ie, items that can take on only two possible values).

Negative likelihood ratio Measure of the usefulness of a diagnostic test for the presence of a particular condition or disease; indicates the odds of the test yielding a false negative among those with the condition or disease relative to yielding a true negative among those without the condition or disease; lower values of the negative likelihood ratio indicate that the test is effective at ruling out the condition or disease.

Negative predictive value Measure of the probability of a given case not having a condition or disease if the result of a diagnostic test for the presence of a particular condition or disease is negative (ie, indicates the condition or disease is not present); higher values of the negative predictor value indicate that the test is effective at ruling out the condition or disease.[a]

Phi coefficient Statistical measure of the extent to which two binary variables are correlated; used to capture test– retest or inter-rater reliability.

Positive likelihood ratio Measure of the usefulness of a diagnostic test for the presence of a particular condition or disease; indicates the odds of the test yielding a true positive among those with the condition or disease relative to yielding a false positive among those without the condition or disease; higher values of the positive likelihood ratio indicate that the test is effective at establishing the condition or disease.

Positive predictive value Measure of the probability of a given case having a condition or disease if the result of a diagnostic test for the presence of a particular condition or disease is positive (ie, indicates the condition or disease is present); higher values of the positive predictor value indicate that the test is effective at establishing the condition or disease.[a]

Pretest probability Measure of how common a condition or disease is in the population of interest; the percentage of all individuals who have the condition or disease; also known as prevalence.

Prevalence Measure of how common a condition or disease is in the population of interest; the percentage of all individuals who have the condition or disease; also known as pretest probability.[a]

Relative validity Extent to which a test measure of a concept agrees with a reference measure of that concept that has a greater degree of demonstrated validity, even if it is not an exact measure of the concept.[a]

Reliability Extent to which a measurement process gives the same results when repeated under similar circumstances.[a]

Sensitivity Extent to which a diagnostic test correctly identifies those who have a particular condition or disease; if a person has a condition or disease, a sensitive test will identify them as having the condition or disease nearly always.[a]

Specificity Extent to which a diagnostic test correctly identifies those who do not have a particular condition or disease; for those without the condition or disease, a specific test will nearly always be negative.[a]

Test–retest reliability Extent to which repeated measurements of the same concept for a given individual will be similar to one another.

True negative Result of a test for the presence of a disease or condition indicating that the disease or condition is present for a given subject in cases when it really is not present.[a]

True positive Result of a test for the presence of a disease or condition indicating that the disease or condition is present for a given subject in cases when it really is present.[a]

Validity Extent to which a variable or measure captures the underlying concept it is intended to reflect.[a]

Figure 1. The following terms are useful in understanding measurement and related issues in nutrition research. [a]Entries are based on reference 8. NOTE: Information from this figure is available online at www.adajournal.org as part of a PowerPoint presentation.

take. At the same time, studies often include measures not based on self-reported data; for example, anthropometric, laboratory, or clinical measures. Using the physical measurements of weight and height, body mass index is often used in research studies as a global marker of body fat. Nutrition research also relies frequently on laboratory measures, such as blood glucose levels; physical measures based on imaging methods, such as bone mass from dual-energy x-ray absorptiometry; or clinical characteristics, such as spoon-shaped nails or a swollen smooth tongue.

A critical first step in the measurement process is to define clearly and specifically what will be measured. A measurement process can only be assessed as appropriate or inappropriate if one knows exactly what is being measured because a process that is appropriate for one type of measure may do a poor job of measuring a related but slightly different concept. This point may be most obvious with anthropometric measures—a scale is an appropriate instrument for measuring weight but not height. The same principle holds with other types of measures. A single 24-hour dietary recall, for example, may provide a valid and reliable measure of a single individual's target day intake of vitamin C, but will likely be a poor measure of that individual's usual vitamin C intake. Alternatively, data collected from a sample using a 24-hour dietary recall may be used to construct a valid and reliable measure of the usual mean intake of vitamin C among that group, as described by the Institute of Medicine (6). As is clear from this example, researchers should be specific about the concept being measured—the appropriate measure depends not just on the general concept being examined (vitamin C intake), but also on the reference period over which intake was to be measured (a single day vs the longer time period necessary to measure usual intake) and the reference group to which the measure was being applied (an individual's intake level vs the mean intake of a group).

In assessing the quality of a measure, two questions are important to address. First, does the process used to generate the measure produce consistent results if repeated under similar circumstances? Second, does the measurement or variable resulting from this process actually reflect what it is intended to reflect? The answer to the first question tells us about the reliability of the measure. The answer to the second question tells us about the measure's validity.

Reliability

Reliable measurements give the same results when repeated under similar circumstances. In other words, reliable measures are subject to little random measurement error. Unreliable measures are subject to measurement error, implying that a researcher cannot be confident that the value of the measure in a given case is accurate. For example, different people measuring the triceps skinfolds of the same person will likely obtain slightly different measurements due to different measurement sites, caliper placement, or caliper readings. Nutrition researchers use different forms of reliability depending on the circumstances surrounding the concept being measured and the measurement method. Below, test–retest reliability, inter-item reliability, and interobserver or inter-rater reliability are described.

Test–Retest Reliability. Researchers sometimes use the same measurement process (ie, the same test) two or more times. Thus, for a given individual, there are repeated values of the same measure of a given concept. The extent to which these repeated values of the measure for this subject are similar to one another (or the correlation between these values) reflects the test–retest reliability of the measure. For example, a researcher might take repeated blood samples from an individual to assess the test–retest reliability of a measure of the individual's serum cholesterol level. Or a study focusing on children's obesity levels might weigh sample members more than one time to assess the reliability of the process used to measure children's weight.

The usefulness of test–retest reliability depends on measurement costs, the feasibility of replicating the measure's reference time period, and the possibility of a second measurement being influenced by the first measurement. Test–retest reliability is usually not appropriate when the measurement process is time-consuming or costly. For example, the doubly-labeled water method for measuring individuals' energy expenditure is both time-consuming and costly (7), so it would be difficult to repeat this method for a large sample. Obtaining repeated measures of a participant's weight to assess test–retest reliability, on the other hand, would be quite feasible. Test–retest reliability is also less useful when the initial test measure covers a specific time period that cannot be repeated. A researcher conducting a 24-hour recall to measure individuals' dietary intake on a specific day can conduct a second 24-hour recall the next day, but it will refer to foods and beverages consumed on a different day and so cannot be used to measure test–retest reliability. The researcher could administer the second 24-hour recall later on the same day as the first 24-hour recall, but the participant's responses on the second 24-hour recall would likely be influenced by their responses on the first 24-hour recall, assuming they remember what they had previously reported. For a true indicator of test–retest reliability, the test and retest should be independent; the results of the retest should not be influenced by the results of the initial test.

The specific statistical tools appropriate for estimating test–retest reliability depend on whether the measure is continuous (such as serum cholesterol concentration or grams of saturated fat consumed over a 24-hour period) or ordinal (such as an indicator for whether an individual exercises never, rarely, sometimes, or often). One approach involves calculating the correlation between the measure's value in the original test and the retest using a correlation coefficient. For continuous measures, Pearson's correlation coefficient is appropriate; for ordinal measures, a researcher might use the Spearman nonparametric rank correlation coefficient; and for two binary variables, the phi coefficient could be used (8). For example, the Eating Disorder Inventory-3 is a continuous measure used in the diagnosis and treatment of individuals suspected of having eating disorders (9). Suppose the Eating Disorder Inventory-3 is administered to a group of female gymnasts one day and re-administered the next day. Pearson's correlation could be calculated to examine

Table 1. Test–retest reliability of weight status measure

Test value	Retest Value: Underweight	Normal weight	Overweight or obese	Total
Underweight	5	1	0	6
Normal weight	2	134	8	144
Overweight or obese	1	1	48	50
Total	8	134	56	200
Summary	← *n* →		← % →	
No. of subjects	200		—	
Possible responses to question	3		—	
No. of subjects with exact test–retest agreement	187		93.5	
No. of subjects with test–retest values separated by 1 category	12		6	
No. of subjects with test–retest values separated by >1 category	1		0.5	
% with exact matches	—		93.5	
% within 1 category	—		99.5	
Cohen's weighted kappa	0.852		—	

the correlation between scores on the 2 days. If the value of the correlation is large and positive, test–retest reliability would be considered to be strong. Conversely, if the values were close to zero, test–retest reliability would be deemed weak, indicating substantial measurement error.

A graphic display of a test–retest comparison is useful to determine if there is a pattern to any disagreement. As an example, one might compare energy requirements between a newer energy formula and one that has been traditionally used. Even if the Pearson correlation coefficient is high, the newer measure could systematically over- or underestimate energy needs compared to the traditional method. The Bland-Altman method visually displays the results by plotting the mean of the two estimates as a line of unity (=0) and then adds a line for the estimates from the new formula. If the new estimates are consistently greater than zero, the new formula overestimates energy requirements. The data could also be displayed with the x-axis scaled by a characteristic such as body mass index to indicate how the agreement of the two measures differs by body size.

Another statistical test for assessing the test–retest reliability of a continuous measure involves comparing the mean value of the measure from the initial test with that from the retest, using a paired *t* test (or one-sample *t* test). If the paired *t* test fails to reject the null hypothesis that the two values are equal (ie, does not find a significant difference between the two values), this can be considered evidence in support of the reliability of the measure. A recent study used this approach to assess the reliability of an iron food frequency questionnaire—among seven nutrients examined, the analysis found that the intake measures in the two administrations of the questionnaire were not significantly different in six cases (10).

For a categorical variable, examining mean values is not appropriate. A researcher may instead examine the extent to which the value of the measure in the initial test agrees with that in the retest. Consider the example shown in Table 1, a hypothetical comparison of two measures of weight for a sample of college students, each of which takes on one of three possible values—underweight, normal weight, or overweight/obese. The top panel shows the number of students with each of the nine possible combinations of the test and retest values. For example, five students were underweight according to both the initial test and retest, while one was underweight according to the initial test and normal weight according to the retest. Reliability can be assessed by examining how often the two values agree with one another, with more reliable measures having a higher agreement percentage. In this example, the test and retest values agree in 93.5% of the cases ([5+134+48]/200). For some measures, a researcher may decide that the two values being within one category of each other is sufficient to be considered in agreement. If that standard were used in this example, there would be near perfect (99.5%) agreement.

When the categorical measure has a limited number of possible values, there would be agreement in a certain percentage of cases by chance. A reliability measure that accounts for the likelihood of chance agreement is Cohen's kappa coefficient (11), which first determines the expected percentage of cases of chance agreement and then determines whether there is test–retest agreement over and above this chance agreement. If P_o represents the total proportion of cases with test–retest agreement and P_c represents the proportion of cases in which chance test–retest agreement is expected, then Cohen's kappa is calculated as $(P_o–P_c)/(1–P_c)$. In the above example, there is agreement in 93.5% of cases and chance agreement was expected in 56% of cases, leading to a value of Cohen's kappa coefficient of 0.852. A Cohen's kappa of 0.852 reflects excellent agreement between tests.

Inter-Item Reliability. When a researcher measures an unobserved construct using a scale or index based on a set of indicators of the construct, inter-item reliability is appropriate. Since each indicator is measuring the same construct, their values should be highly correlated. Inter-

item reliability measures the extent to which the indicators of a construct are correlated. For example, a registered dietitian (RD) or other dietetics practitioner interested in adult males' attitudes about high-fat foods might ask a set of five questions about these attitudes of a sample of adult men. If a respondent indicates a concern about consuming high-fat foods on one question, one would expect their responses to the other questions to have values indicating a similar concern. By itself, each of the five indicators is an imperfect measure of attitudes toward high-fat foods, so the construct is measured through a scale or index that combines the values of the individual indicators. Since the five indicators are all measures of the same underlying construct, they should have a high degree of correlation, or internal consistency.

To assess a scale's inter-item reliability, both the number of indicators and the average reliability of these indicators must be factored. If the individual indicators have three or more answer options (such as "no concern," "some concern," and "much concern"), internal consistency using Cronbach's α coefficient can be measured (12). Cronbach's α represents the extent to which the items of a scale are correlated with one another—the greater the correlation of these items with one another, the greater the value of Cronbach's α. For example, Fitzgerald and colleagues (13) calculated a Cronbach's α coefficient of .813 to establish the inter-item reliability of a 24-item scale designed to measure nutrition knowledge among a Latina population. Similarly, the inter-item reliability of each of the eight subscales that make up the Eating Disorder Inventory-3 described above was calculated, with Cronbach's α values ranging from .72 to .93 (9). In general, values of the Cronbach's α coefficient $>.70$ are considered to have satisfactory inter-item reliability. For scales based on binary indicators (such as responses to true/false questions), a similar reliability coefficient known as the Kuder-Richardson Formula 20 is used. As with Cronbach's α, the Kuder-Richardson Formula 20 statistic measures the degree to which the binary indicators are correlated with one another.

Inter-Rater Reliability. To assess the reliability of a measurement process in which the specific values of the measure are determined by more than one individual, inter-rater reliability is used. Inter-rater reliability shows the extent to which different observers or raters come up with the same value of a measure for a given case. For example, multiple members of a study team might be sent out to observe the food intakes of a sample of middle school students. In this case, a researcher might wonder whether any of the variation in sample members' observed intake levels arose from variation in observers' methods for determining the intake amount, rather than from variation in the intake amount itself. To address this issue, the researcher might have multiple observers measure the intake of the same sample members—if different observers produce the same or very similar intake amounts for a the same sample members, the measure can be said to have high inter-rater reliability. The statistical tools used in determining inter-rater reliability are the same as the tools for determining test–retest reliability, including Pearson's correlation coefficient, Spearman's nonparametric rank correlation coefficient, a paired or one-sample t test of mean values, and Cohen's kappa coefficient.

Validity

A variable's validity reflects the extent to which it measures what it is intended to measure. To understand the validity of a given variable, a nutrition researcher must know what specific concept or outcome the variable is intended to measure. The researcher should also be confident that the variable is relatively free of random errors; in other words, that it meets at least some minimum threshold of reliability (8). An unreliable measure cannot be a truly valid measure. Three different ways of understanding a measure's validity are described below—content or face validity, absolute vs relative validity, and internal vs external validity.

Face or Content Validity. Face validity indicates whether a variable appears to most observers to be a sensible indicator of the concept it is trying to measure. This is a subjective measure, and there are no statistical tools to determine whether a variable has face validity—it is in the eye of the observer. A variable with high face validity might be a measure of individuals' nutrition knowledge concerning foods' fat content that was obtained by summing the number of correct answers to a series of questions about foods' fat content. These questions could ask whether one common food item has more or less energy from fat than the same amount of some other common food item. By contrast, measuring individuals' fat content knowledge by asking them to report the exact number of fat grams in a single food item would have less face validity.

Content validity is another subjective measure of validity similar to face validity, but it focuses on whether a particular measure covers all dimensions present in the concept it is intended to reflect. For example, measuring the overall nutritional quality of adult women by using only their usual energy intakes would lack content validity. Usual energy intake would reflect one dimension of the nutritional quality of the women's diets—whether there is overconsumption or underconsumption, but would not cover other important dimensions of nutritional quality such as whether their intakes of key nutrients are adequate or whether their intakes of sodium, cholesterol, or dietary fat are excessive.

Absolute vs Relative Validity. The distinction between absolute validity and relative validity depends on the extent to which the true value of the concept being measured can be determined. Absolute validity is the highest standard of validity, capturing the extent to which a measure accurately reflects the exact concept it is intended to reflect. To determine a measure's absolute validity, a perfect (or at least very good) indicator of the target concept or behavior is needed—a gold standard. This indicator is the criterion against which a measure's absolute validity is assessed; hence, researchers sometimes refer to a measure's criterion validity.

To validate a measure of the reported dietary intake of a sample, some studies have used the criterion of direct observation of what sample members consumed over a given time period. Crawford and colleagues (14) examined the validity of reported intakes of 9- and 10-year-old

girls on both a 24-hour recall and a 3-day food record by comparing these reports with intakes observed by unobtrusive observers (14). In this case, the researchers did not have perfect measures of the concepts they intended to capture, because the girls may have consumed some foods surreptitiously, or observers may have been mistaken about the amounts or types of foods consumed. They did not truly capture absolute validity. However, validity studies that really capture absolute validity are rare, and assessing criterion validity by comparing a test measurement process against a criterion believed to be highly accurate is a close substitute.

Relative validity, by contrast, is determined by comparing a test method with a reference method where the reference method has a greater degree of demonstrated validity, even if not an exact measure of the underlying concept (8). For example, modifications were made to the National Cancer Institute's food frequency questionnaire in the Black Women's Health Study to create a more ethnically appropriate questionnaire to identify important associations between dietary variables and health outcomes in this large prospective cohort (15). Three 24-hour recalls and a 3-day food diary were used as reference standards to evaluate the questionnaire's relative validity. Even though 24-hour recalls and food diaries do not measure true intake with absolute certainty, the authors established the relative validity of their instrument, addressing several design considerations important in validity studies. In particular, they used similar test and reference populations and measured intakes over the same 3-day intake period with both methods.

The notion of relative validity can also be applied to various clinical conditions. Suppose a critical care dietitian or other dietetics practitioner would like to evaluate the resting energy expenditure for a mechanically ventilated individual who requires nutrition support. Whereas most clinicians would consider a metabolic cart to procure gold-standard results using indirect calorimetry, this technology is based on theoretical principles, varies considerably by operator, and is usually performed once for a few minutes. It provides a snapshot of an individual's energy needs, which is extrapolated to reflect total 24-hour energy requirements. Because indirect calorimetry is expensive and requires specialized equipment and trained personnel, it is not readily available. Reliance on energy equations with determinants and modifiers of resting metabolic rate have been established and implemented for this population (eg, Harris-Benedict, Mifflin-St Jeor, Penn State, Ireton-Jones). Conceptually, these equations are the test method and indirect calorimetry is the reference method for comparison. Application and evaluation of these equations demonstrates their relative validity in clinical practice (16).

In designing a strategy for assessing relative validity, two key design considerations are critical. First, both the test method and reference method must measure the same underlying concept over the same time period. For example, testing the validity of a usual nutrient intake measure by comparing a test method of usual intake during the summer with a reference method of usual intake during the winter would be inappropriate. Second, measurement error from the test method and from the reference method should be independent. Independent errors will help ensure that if the test and reference measures are correlated, it is because they are both accurately measuring the underlying concept rather than because they are measured with the same type of error. For example, different methods for measuring usual intake, such as food records, food frequency questionnaires, and 24-hour dietary recalls, may all suffer from similar types of underreporting (17). By contrast, errors in the measurement of food intake from a 24-hour recall and that recorded by an observer are likely to be independent, since the former involves self-reports of intake (subject to recall error) and the latter involves reporting by an independent observer (subject to errors in observers' judgments of food types and amounts).

A variant of relative validity is convergent validity. Convergent validity exists if different measures of an underlying concept agree with each other and with a test measure. In the usual intake example, the convergent validity of a new method for measuring usual intake for a group of individuals might be assessed by comparing it with a 24-hour dietary recall, food frequency questionnaire, and diet history. None is a perfect measure, but the convergent validity of the test measure will be enhanced if there is agreement among the different measurement methods.

The problem of nonindependent measurement or reporting errors of the various measurement methods in this example might lead to spurious correlations. In this case, a researcher might turn to discriminant validity, which indicates the extent to which a method for measuring a particular construct disagrees with a reference measure intended to reflect the opposite of that construct. A researcher proposing a new scale of teenaged girls' physical activity levels, for example, might compare it with an indicator of the girls' daily screen (television, video, or computer) time. One would expect that girls who spend a lot of time each day watching television or behind a computer would be less physically active, so a negative correlation between these measures would increase the validity of the physical activity measure.

Nutrition researchers sometimes assess the relative validity of intake measures using biomarkers or intake indicators based on biological measures (8). These typically measure the presence of a nutrient in a sample of body fluids or tissues, such as blood, urine, or fat tissue. Biomarkers may not precisely estimate the underlying concept, but they are considered to accurately measure physical conditions that are correlated with the intake of the nutrient of interest. Thus, one would expect a valid measure of the intake of a given nutrient to be correlated with the biomarker for that nutrient. Further, because it is a biological measure, any measurement errors are unlikely to be correlated with errors in reported intakes of the nutrient.

In using biomarkers to examine an intake measure's validity, one must consider the particular nutrient being examined and the time period over which intake is reported. Some nutrients remain in the body for a short time and so the presence of the nutrient in serum, plasma, or urine can be used only to validate intake over a similarly short period. For example, a 24-hour urine sample tested for the presence of vitamin C might be used as a biomarker for vitamin C intake over 24 hours, but not

over a month. Biomarkers for usual nutrient intake might also be assessed by examining samples of hair, toenails, or adipose tissue for presence of the nutrient (18).

Internal vs External Validity. A particular measure has internal validity if it is valid for the individuals in the sample being studied. An internally valid measure may or may not be valid for individuals outside the study sample. If the measure is valid for the broader population, it is said to have external validity. While a given measure may be internally valid but lack external validity, measures that are not internally valid cannot be externally valid.

Internal and external validity are useful concepts for interpreting the results of studies aimed at measuring the effectiveness of an intervention in achieving its goals for a target population. Suppose a weight-loss program emphasizing physical activity is evaluated in a clinical trial whose participants volunteered for the study. The volunteers may be especially motivated to lose weight, and the program may be especially effective for this group. A well-designed study estimating the effect of participation in the program may find that the program promotes weight loss among participants. Such a measure of program effectiveness would be said to have internal validity—we would be confident that it is accurate for the population being studied. However, because the study relied on volunteers rather than a randomly selected sample, we could not be sure that the program would be similarly effective in the broader population. We would not know, for example, whether the program's effectiveness depended on the motivation level of the volunteers. Thus, the estimate of program effectiveness would lack external validity.

As in this example, the distinction between internal and external validity often rests on how study participants were selected. If a study used probability sampling methods, such as random sampling, then measures that are internally valid will likely also be externally valid. If the study used nonprobability sampling methods, such as convenience sampling, then it will not be externally valid.

Similar statistical tools may be used to assess various types of validity as were used to assess reliability. The basic approach involves comparing the values of the test measure with the values of the reference measure for the same set of individuals. This may be done by calculating a correlation coefficient measuring the relationship between the two measures, with higher values of the coefficient indicating greater validity. Alternatively, a researcher may calculate the mean value of the test measure among these individuals and compare it with the mean value of the reference measure. The closer the mean values, the greater the validity. Because of random measurement error (especially if errors in the test and reference measures are independent), one would not expect the mean values to be identical. However, a paired t test of the difference between the mean values will reveal whether they are statistically equivalent.

An alternative approach for examining validity may be to examine the specific values of the test measure and reference measure for each sample member, and count the frequency of discrepancies. Baxter and colleagues (19) studied the validity of the 24-hour dietary recall method for measuring intakes at school meals among children by comparing reported intakes with the reference method of intakes as observed by trained RDs. The authors wished to examine the frequency of and reasons for intrusions—food items reported being consumed during a specific time period that were not actually consumed. They examined foodservice production records to determine whether intrusions might have been the result of foods served at the school on a previous day. To document the validity of the test method, the authors counted the number of intrusions per interview—the fewer intrusions the more valid the method—and analyzed the relationship of the number of intrusions with various characteristics of the interview.

A related point made in a second article is that studies of the validity of nutrient intake measures that compare reported vs observed intake of a nutrient at a given meal risk overstating the validity of the reported measure if they do not account for intrusions; ie, individual food items reported that were not actually consumed. This is because the over-reported intake of these nutrients in the intrusions may cancel out under-reported intake of nutrients from foods that were consumed but not reported (20).

DIAGNOSTIC ASSESSMENT

Dietetics practitioners diagnose nutrition-related disorders using data from various sources and take advantage of statistical tools to help with this process. Two well-known tools are sensitivity and specificity, which are related to validity and reliability in that they capture how well particular indicators measure what they are intended to measure. Sensitivity and specificity, along with a set of additional parameters described below, are helpful in diagnosing nutrition-related disorders using anthropometry, nutrition-focused physical examination, food/nutrition-related or medical histories, cognitive assessment, and laboratory examination. These methods boost the probability of a correct nutrition diagnosis to effectively treat symptoms and prevent further morbidity.

To understand these concepts, it is helpful to use the terms true positives (Tp), false positives (Fp), true negatives (Tn), and false negatives (Fn). Note that to use these terms in an empirical setting, one needs gold-standard knowledge of the true condition of individuals in a given sample. A Tp occurs when the test measure indicates the presence of the disorder/condition, and the individual truly has this condition. An Fp occurs when a positive test measure indicates the presence of the disorder/condition even though the individual does not have the condition. Conversely, a Tn is a negative test result that accurately indicates that the individual does not have the condition, whereas an Fn is a negative test result that is inaccurate because the individual does have the condition. Fps and Fns indicate misclassification, and the greater their number the less accurate the test in ruling in or out the disorder/condition. Each has serious implications for patients. Fps lead to the use of treatments that may be expensive, have unnecessary side effects, and lead to undue emotional distress due to the belief that one has the nutrition-related problem. Fns may lead to a delay in treatment and further progression and seriousness of the condition.

Test Result	True Condition: Has disease/condition	True Condition: Does not have disease/condition
Positive (indicates presence of disease/condition)	A True positive (Tp)	B False positive (Fp)
Negative (indicates absence of disease/condition)	C False negative (Fn)	D True negative (Tn)

- Sensitivity (Sn) = (A/A+C)×100
- Specificity (Sp) = (D/B+D)×100
- Positive predictive value (+PV) = (A/A+B)×100
- Negative predictive value (−PV) = (D/C+D)×100
- Positive likelihood ratio (LR+) = (A/A+C)/(B/B+D) OR Sn/(1−Sp)
- Negative likelihood ratio (LR−) = (C/A+C)/(D/B+D) OR (1−Sn)/Sp
- Prevalence (pretest probability) = ((A+C)/(A+B+C+D))×100

Figure 2. Classification of true/false positives and negatives.

Parameters for Evaluating a Diagnostic Assessment

Seven related parameters can be used to determine the usefulness of a diagnostic assessment (sensitivity [Sn], specificity [Sp], positive predictive value [+PV], negative predictive value [−PV], positive likelihood ratio [LR+], negative likelihood ratio [LR−], and prevalence [pre-test probability]) (21-24). To calculate these parameters, investigators conduct studies using individuals with and without a disease such as breast cancer or a condition such as iron deficiency anemia. A diagnostic assessment related to the disease or condition (eg, needle aspiration of the breast) is tested by being applied to these individuals, with the results of the assessment compared with their known disease status. Clinical professionals are interested in discovering those diagnostic assessments that can be applied to large groups of people inexpensively and comfortably, and yet are accurate. Studies of the usefulness of diagnostic assessments are conducted so that less-expensive and -intrusive diagnostic assessments might be used instead of highly accurate but costly and intrusive procedures (eg, excision of the breast and cytological evaluation).

In such studies, each participant is classified as a Tp, Fp, Fn, or Tn, and the number in each category is summed. If the number of Tps is represented by the value *A*, the number of Fps is *B*, the number of Fns is *C*, and the number of Tns is *D*, then the parameters can be calculated as follows (Figure 2).

Prevalence measures how common the disease/condition is—it is the percentage of all individuals who have the condition. Three of the remaining parameters represent the ability of the diagnostic assessment to correctly identify those who have the condition. Sn is the percentage of those with the condition who tested positive on the diagnostic assessment; +PV is the percentage of those who tested positive who have the condition; LR+ is the proportion of those who have a positive test and have the condition compared to the proportion of those who have a positive test and do not have the condition. The remaining parameters tell us how well the diagnostic assessment correctly identifies those who are free from the disorder/condition. Sp is the percentage of those without the condition who tested negative on the assessment; −PV is the percentage of those who tested negative who do not have the condition; LR− is the proportion of those who have a negative test and have the condition compared to the proportion of those who have a negative test and do not have the condition.

LR	Interpretation
>10	Large and often conclusive increase in the likelihood of disease
5-10	Moderate increase in the likelihood of disease
2-5	Small increase in the likelihood of disease
1-2	Minimal increase in the likelihood of disease
1	No change in the likelihood of disease
0.5-1.0	Minimal decrease in the likelihood of disease
0.2-0.5	Small decrease in the likelihood of disease
0.1-0.2	Moderate decrease in the likelihood of disease
<0.1	Large and often conclusive decrease in the likelihood of disease

Figure 3. General interpretation of likelihood ratios (LR).

Sn and Sp are reported as percentages and can range from 0% to 100%. The higher the Sn, the fewer the Fns, and the better the assessment rules out the condition (for those with a negative test). The higher the Sp, the fewer the Fps, and the better the test rules in the condition (for those with a positive test). A very useful diagnostic test will have both a high Sn and Sp, ruling in and out the condition effectively, with few Fps and Fns.

It is often the case that an RD or other nutrition practitioner will have diagnostic assessment data for an individual, but no definitive knowledge about the presence of the particular nutrition condition or problem. In this case, it is more useful to know the percentage who have (or do not have) the condition given their test results (as measured by +PV and −PV), rather than the percentage who will test positive (or negative) for the condition given their true status (as measured by Sn and Sp). The +PV and –PV parameters range between 0% and 100%, and the higher these percentages the more confident the RD or other nutrition practitioner will be that the test results

Table 2. Meta analysis data for iron deficiency anemia (IDA) and serum ferritin

	Presence of IDA (Per Stainable Bone Marrow)		
Diagnostic test result	**Yes**	**No**	**Total**
Positive (<65 ng/mL)[a]	731	270	1,001
Negative (≥65 ng/mL)[a]	78	1,500	1,578
Total	809	1,770	2,579
Diagnostic Testing Evaluation Parameter	**Formula**		
Sensitivity (Sn)	(731/809)×100=90%		
Specificity (Sp)	(1,500/1,770)×100=85%		
Positive predictive value (+PV)	(731/1,001)×100=73%		
Negative predictive value (−PV)	(1,500/1,578)×100=90%		
Positive likelihood ratio (LR+)	(90 /(100–85))=6		
Negative likelihood ratio (LR−)	(100–90)/85=0.12		
Prevalence (pretest probability)	(809/2,579)×100=31%		

[a]To convert ng/mL serum ferritin to pmol/L, multiply ng/mL by 2.247. Serum ferritin of 65 ng/mL=146 pmol/L.

Table 3. Data for iron deficiency anemia (IDA) and serum ferritin with prevalence of 10%

	Presence of IDA (Per Stainable Bone Marrow)		
Diagnostic test result	**Yes**	**No**	**Total**
Positive (<65 ng/mL)[a]	232	348	580
Negative (≥65 ng/mL)[a]	26	1,973	1,999
Total	258	2,321	2,579
Diagnostic Testing Evaluation Parameter	**Formula**		
Sensitivity (Sn)	(232/258)×100=90%		
Specificity (Sp)	(1,973/2,321)×100=85%		
Positive predictive value (+PV)	(232/580)×100=40%		
Negative predictive value (−PV)	(1,973/1,999)×100=99%		
Positive likelihood ratio (LR+)	(90 /(100–85))=6		
Negative likelihood ratio (LR−)	(100–90)/85=0.12		
Prevalence (pretest probability)	(258/2,579)×100=10%		

[a]To convert ng/mL serum ferritin to pmol/L, multiply ng/mL by 2.247. Serum ferritin of 65 ng/mL=146 pmol/L.

rule in (or out) the condition. So a +PV value of 86% means that 86% of those with a positive test will have the condition; a –PV value of 86% means that 86% of those with a negative test will not have the condition. A limitation of the predictive values is that they change with the prevalence of the condition in the reference population. As the prevalence in the population rises, the +PV increases and the –PV decreases—the disorder or condition becomes easier to detect (since it is more prevalent), but harder to rule it out.

The most useful parameters for a diagnostic assessment are the likelihood ratios, which can indicate the predictive value of the test factoring in changing prevalence. Likelihood ratios can take on any positive value. Values of 1 indicate that the diagnostic test is not useful. As LR+ increases above 1, the test has greater usefulness in ruling in the disorder or condition. As LR− decreases below 1, the test has greater usefulness in ruling out the condition. Figure 3 presents a rubric for the interpretation of likelihood ratios (25).

Example: Serum Ferritin and Iron Deficiency Anemia. In the United States, approximately 4% to 8% of premenopausal women are iron deficient (26). The gold standard for diagnosing iron deficiency anemia is bone marrow aspiration from the sternum or iliac crest and whether it is not stainable for iron. This procedure is invasive, expensive, and not feasible for large groups of people. A less-costly and more feasible diagnostic assessment is serum ferritin level. Ferritin is a protein that stores iron atoms and is found in highest quantities in the liver. Serum ferritin level reflects the amount of iron storage in the liver. A 1992 meta-analysis examined the ability of serum ferritin level to accurately diagnose iron deficiency anemia, using bone marrow aspiration to determine participants' true status (27). Results from this analysis are summarized in Table 2.

With an iron deficiency anemia prevalence of 31%, a positive test rules in iron deficiency anemia 73% of the time (+PV=73%), and a negative test rules out iron deficiency anemia 90% of the time (−PV=90%). Using the cutoff value of <65 ng/mL (<146 pmol/L) for serum ferritin (ie, values <65 ng/mL [<146 pmol/L] are considered a positive test for iron deficiency anemia), the test is good at ruling out iron deficiency anemia, but has a proportion of false positives that is greater than ideal. Mast and colleagues (28) found that decreasing the cutoff value to ≤30 ng/mL (≤67 pmol/L) for serum ferritin could increase the +PV to 92% and the –PV to 98% (with a prevalence of iron deficiency anemia in the reference population of 23%). This lower cutoff value for diagnosis makes the test better at both ruling in and ruling out iron deficiency anemia. As in this example, an important factor in the usefulness of the diagnostic test is the cutoff value for deciding whether or not the disease is present.

The Problem of Shifting Prevalence

As noted previously, predictive values change with shifting prevalence, with +PV typically decreasing and –PV increasing as prevalence in the reference population decreases. Table 3 presents data for serum ferritin and iron deficiency anemia with a total population prevalence of 10% rather than 31%. The changing prevalence leads to a decrease in +PV from 73% to 40% and an increase in –PV from 90% to 99%. This highlights the value of likelihood ratios, which—unlike PVs—do not change with changing prevalence values. They can be used to determine +PV and –PV given the prevalence of the disease in the target population. If one knows or can estimate the disease's prevalence in the reference population, he or she can use the likelihood ratios to determine +PV and –PV for a suspected disease in a patient using a nomogram (Figure 4) (29). To determine +PV, a point is placed on the vertical pre-test probability line at the level of the known prevalence. Another point is placed on the vertical likelihood ratio line at the LR+ value. A straight line through these two points will intersect with the vertical line labeled post-test probability at the +PV. A similar procedure is used with a line between the pre-test probability point and LR− used to yield –PV. This nomogram and the LR+ and LR− for the serum ferritin–iron deficiency anemia examples can be used to verify the +PV and –PV for a prevalence of 31% and 10%. When the prevalence is 31%, for example, +PV should be 73% and –PV should be 90%; when the prevalence is 10%, +PV should be 40% and –PV should be 99%. When estimating –PV this way, the intersection with the post-test probability line represents (1–[−PV]). To get –PV, simply subtract 1 from the intersection with the post-test probability line.

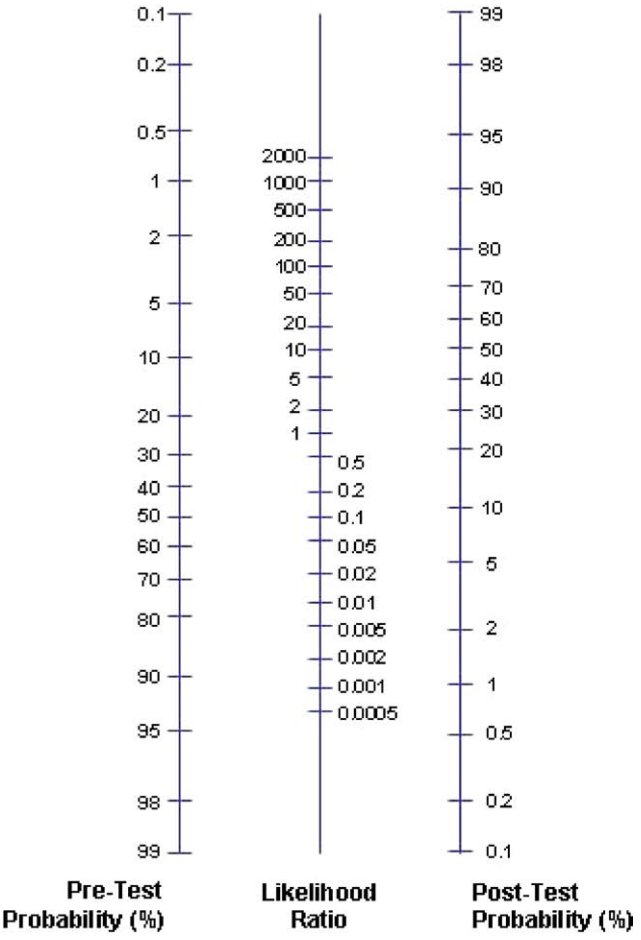

Figure 4. Nomogram for using likelihood ratios and pretest probability to determine post-test probability.

CONCLUSIONS

In conducting research or diagnostic assessment to detect the presence or absence of a nutrition- or health-related condition, it is critically important for dietetics practitioners to accurately measure the characteristics, outcomes, or conditions they are attempting to measure. In this report, reliability and validity are described—two important attributes that all measures used in nutrition research should have. In addition, the concepts of sensitivity and specificity relevant to the diagnostic assessment process are outlined.

Reliability and validity allow us to answer the question of whether the measurement process used in a study produces consistent and accurate information. If a measure is reliable, that means its value would have been approximately the same even if some aspect of the measurement had been different, such as being completed at a different time or by a different observer or rater. With a reliable measure, one can feel confident that its value is not unduly influenced by some idiosyncratic aspect of the measurement process.

It is not sufficient for a measure to just be reliable. A measurement process that yields a consistent value when repeated is not useful if that value does not accurately reflect the underlying characteristic or outcome it is intended to measure. Thus, variables or measures used in research must be valid as well as reliable. A valid measure reflects what it is intended to reflect, and one can feel confident that there is no systematic error in this measure.

Diagnostic assessment focuses on a different measurement process—measuring the presence or absence of some underlying nutrition- or health-related condition

using the value of a test for that condition. Although a dietetics practitioner conducting diagnostic assessment would not claim that the test is an exact measure of whether an individual has the condition, it is somewhat analogous to validity in that it tells us how well the test accurately reveals what it is trying to reveal—the presence or absence of the condition. The parameters upon which diagnostic assessment is based are all different ways of indicating the accuracy of the test.

One can perhaps best understand the importance of valid and reliable measures and of diagnostic assessment based on tests with high levels of sensitivity and specificity by understanding what happens in their absence. With poor measures, a research study cannot accurately describe nutrition outcomes among a population or analyze the relationship between nutrition exposure and key nutrition-related outcomes. With unreliable and/or invalid measures, the resulting analysis—no matter how sophisticated—will also be unreliable and/or invalid. Similarly, a diagnostic assessment based on a test that lacks sensitivity and specificity is likely to result in poor recommendations for nutrition intervention.

STATEMENT OF POTENTIAL CONFLICT OF INTEREST: No potential conflict of interest was reported by the authors.

References

1. Boushey C, Harris J, Bruemmer B, Archer S, Van Horn L. Publishing nutrition research: A review of study design, statistical analysis, and other key elements of manuscript preparation, Part 1. *J Am Diet Assoc.* 2006;106:89-96.
2. Boushey C, Harris J, Bruemmer B, Archer S. Publishing nutrition research: A review of sampling, sample size, statistical analysis, and other key elements of manuscript preparation, Part 2. *J Am Diet Assoc.* 2008;108:679-688.
3. Harris J, Boushey C, Bruemmer B, Archer S. Publishing nutrition research: A review of nonparametric methods, Part 3. *J Am Diet Assoc.* 2008;108:1488-1496.
4. Harris J, Gleason P, Sheean P, Boushey C, Beto J, Bruemmer B. An introduction to qualitative research for food and nutrition professionals. *J Am Diet Assoc.* 2009;109:80-90.
5. Bruemmer B, Harris J, Gleason P, Boushey C, Sheean P, Van Horn L. Publishing nutrition research: A review of epidemiological methods. *J Am Diet Assoc.* 2009;109:1728-1737.
6. Institute of Medicine. *Dietary Reference Intakes: Applications in Dietary Assessment*. Washington, DC: National Academies Press; 2000.
7. Prentice AM, ed. *The Doubly-Labelled Water Method for Measuring Energy Expenditure: A Consensus Report by the IDECG Working Group, Nahres-4.* Vienna, Austria: International Atomic Energy Agency; 1990.
8. Gibson RS. *Principles of Nutritional Assessment,* 2nd ed. New York, NY: Oxford University Press; 2005.
9. Garner DM. *Eating Disorder Inventory-3*. Lutz, FL: Psychological Assessment Resources; 2004.
10. Heath A, Skeaff CM, Gibson RS. The relative validity of a computerized food frequency questionnaire for estimating intake of dietary iron and its absorption modifiers. *Euro J Clin Nutr.* 2000; 54:592-599.
11. Cohen J. A coefficient of agreement for nominal scales. *Edu Psychol Meas*. 1960; 20:37-46.
12. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16:297-334.
13. Fitzgerald N, Damio G, Segura-Perez S, Perez-Escamilla R. Nutrition knowledge, food label use, and food intake patterns among Latinas with and without type 2 diabetes. *J Am Diet Assoc*. 2008;108:960-967.
14. Crawford PB, Obarzanek E, Morrison J, Sabry ZI. Comparative advantage of 3-day food records over 24-hour recall and 5-day food frequency validated by observation of 9- and 10-year-old girls. *J Am Diet Assoc.* 1994;94:626-630.
15. Kumanyika SK, Mauger D, Mitchell DC, Phillips B, Smicklas-Wright H, Palmer JR. Relative validity of food frequency questionnaire nutrient estimates in the Black Women's Health Study. *Ann Epidemiol.* 2003;13:111-118.
16. Frankenfield D, Hise M, Malone A, Russell M, Gradwell E, Compher C. Evidence Analysis Working Group. Prediction of resting metabolic rate in critically ill adult patients: Results of a systematic review of the evidence. *J Am Diet Assoc.* 2007;107:1552-1561.
17. Sawaya AL, Tucker K, Tsay R, Willett W, Saltzman E, Dallal GE, Roberts SB. Evaluation of four methods for determining energy intake in young and older women: Comparison with doubly labeled water measurements of total energy expenditure. *Am J Clin Nutr*. 1996;63:491-499.
18. Thompson FE, Subar AF. Dietary assessment methodology. In: Coulston AM, Boushey CJ, eds. *Nutrition in the Prevention and Treatment of Disease*. 2nd ed. New York, NY: Academic Press; 2008:3-39.
19. Baxter SD, Hardin JW, Royer JA, Guinn CH, Smith AF. Insight into the origins of intrusions (reports of uneaten food items) in children's dietary recalls, based on data from a validation study of reporting accuracy over multiple recalls and school foodservice production records. *J Am Diet Assoc.* 2008;108:1305-1314.
20. Baxter SD, Smith AF, Hardin JW, Nichols MD. Conventional energy and macronutrient variables distort the accuracy of children's dietary reports: Illustrative data from a validation study of effect of order prompts. *Prev Med.* 2007;44:34-41.
21. Altman DG, Bland JM. Statistics notes: Diagnostic tests 1: Sensitivity and specificity. *BMJ.* 1994;308:1552.
22. Altman DG, Bland JM. Statistics notes: Diagnostic tests 2: Predictive values. *BMJ.* 1994;309:102.
23. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. *BMJ.* 2004;329:168-169.
24. Halkin A, Reichman J, Schwaber M, Paltiel O, Brezis M. Likelihood ratios: Getting diagnostic testing into perspective. *Q J Med.* 1998;91: 247-258.
25. Ebell M. An introduction to information mastery: Diagnosis; likelihood ratios. Michigan State University Web site. http://www.poems.msu.edu/infomastery/Diagnosis/Diagnosis.htm. Accessed September 10, 2008.
26. Conrad ME. Iron deficiency anemia. *Emedicine.* 2006:1-13. http://www.emedicine.com/med/topic1188.htm. Accessed August 13, 2009.
27. Guyatt GH, Oxman AD, Ali M, Willan A, McIlroy W, Patterson C. Laboratory diagnosis of iron-deficiency anemia: An overview. *J Gen Intern Med*. 1992;7:145-153.
28. Mast AE, Blinder MA, Gronowski AM, Chumley C, Scott MG. Clinical utility of the soluble transferring receptor and comparison with serum ferritin in several populations. Lutz, FL *Clin Chem*. 1998;44:45-51.
29. Fagan TJ. Nomogram for Bayes's theorem. *N Engl J Med.* 1975;293: 257.