1-1-2011

Statistical analysis and interpretation of prenatal diagnostic imaging studies, part 1: Evaluating the efficiency of screening and diagnostic tests

Katherine R. Goetzinger
Washington University School of Medicine in St. Louis

Anthony O. Odibo
Washington University School of Medicine in St. Louis

Statistical Analysis and Interpretation of Prenatal Diagnostic Imaging Studies, Part 1

Evaluating the Efficiency of Screening and Diagnostic Tests

Katherine R. Goetzinger, MD, Anthony O. Odibo, MD, MSCE

Invited paper

Screening and diagnostic testing play fundamental roles in all fields of clinical medicine, with obstetric imaging and prenatal diagnosis being no exceptions. With advances in maternal serum screening and ultrasound technology, much research effort in the field of prenatal diagnosis has actually been dedicated to the development and refinement of screening tests for fetal aneuploidy and other congenital disorders. This article aims to review the differences between screening and diagnostic tests, describe the accepted criteria for an efficient screening test, and provide an overview of the calculation and interpretation of test performance characteristics in relation to prenatal imaging studies.

Key Words—diagnostic tests; receiver operating characteristic curves; screening tests; test efficiency

Received February 3, 2011, from the Department of Obstetrics and Gynecology, Washington University, St Louis, Missouri USA. Revision requested March 3, 2011. Revised manuscript accepted for publication March 16, 2011.

Address correspondence to Anthony O. Odibo, MD, MSCE, Department of Obstetrics and Gynecology, Washington University School of Medicine, 4911 Barnes-Jewish Hospital Plaza, Campus Box 8064, St Louis, MO 63110 USA.

E-mail: odiboa@wudosis.wustl.edu

Abbreviations

AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value; ROC, receiver operating characteristic

Screening and diagnostic testing play fundamental roles in all fields of clinical medicine, with obstetric imaging and prenatal diagnosis being no exceptions. The term *screening test* refers to a test typically performed in an asymptomatic population to identify or assign a probability of the risk of disease to individuals in that population. In contrast, a *diagnostic test* typically refers to a test performed in a symptomatic or high-risk population meant to identify or confirm an affected individual.[1] With this distinction in mind, the term *prenatal diagnosis* should really be expanded to *prenatal screening and diagnosis*. With advances in maternal serum screening and ultrasound technology, much research effort in the field of prenatal diagnosis has actually been dedicated to the development and refinement of screening tests for fetal aneuploidy and other congenital disorders. When these screening tests yield positive results or confer a high probability of disease, the patient is then often referred for invasive diagnostic testing. Understanding the features of both screening and diagnostic testing is essential to clinicians, not only for critically interpreting study results in the literature, but also for the practical application of these results to their patient populations. This article aims to review the differences between screening and diagnostic tests, describe the accepted criteria for an efficient screening test, provide an overview of the calculation and interpretation of test performance characteristics in relation to prenatal imaging studies, and discuss relevant aspects of study design to consider when evaluating these tests.

What Makes a Screening Test Efficient?

As is inherent in the name, a screening test is a "screen" of an apparently healthy population to identify individuals at high risk for disease. If a positive result is obtained, results of that screening test are typically confirmed by a subsequent diagnostic test. For a screening test to be efficient, there are certain criteria, initially established by the World Health Organization, that should be met.[2] First, the disease of interest should be important, be relatively common, and place a substantial burden on health. Next, the screening test should be relatively easy to perform, cost-effective, and acceptable to patients. In contrast, diagnostic tests are often more complex, potentially invasive, and costly. Another criterion for a successful screening test is that an effective intervention for the disease of interest should also exist with a generally agreed-on plan of action if a positive test result should occur. Particularly related to prenatal genetic screening, the test should be able to be performed early enough in gestation so that the option for pregnancy termination remains available. Finally, test results should be reliable and valid.[1–6]

Noninvasive first-trimester screening for aneuploidy is an example of a screening test that meets these criteria. Fetal chromosomal abnormalities are relatively common and important obstetric problems, placing a substantial burden on the health and well-being of families and affected fetuses. Screening with a maternal serum analyte and nuchal translucency measurement is a relatively easy technique that is acceptable to most patients, and the screening efficiency of this process has been well validated in the literature.[7,8] Although there is no curative intervention available, parents do have the option to proceed with invasive diagnostic testing via chorionic villus sampling or amniocentesis in the event of a positive screening test result.

Measures of Screening Efficiency

To establish the accuracy of a screening or diagnostic test, it is essential to understand how the test performs in relation to an accepted reference standard of diagnosis. To do so, the diagnostic indices of sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) should be determined. These indices can easily be calculated by creating a traditional 2 × 2 table, with rows representing positive and negative test results and columns representing the presence or absence of disease as diagnosed by a reference standard method (Table 1). By forming a table in this manner, 4 mutually exclusive cells are created, representing true-positive results (a), false-positive results (b), false-negative results (c), and true-negative results (d).

Sensitivity and specificity are inherent test characteristics that describe the performance of the screening test in the population being studied. Sensitivity is defined as "the proportion of people with the disease who have a positive test-result for the disease."[3] This parameter can be calculated by dividing the number of patients with the disease who test positive by the total number of patients with the disease: sensitivity = $a/(a+c)$, or true-positive results/total population with the disease.

Specificity is defined as "the proportion of people without the disease who have a negative test result."[3] Specificity can be calculated by dividing the number of patients without the disease who actually test negative by the total number of patients without the disease: specificity = $d/(b+d)$, or true-negative results/total population without the disease.

Tests with high sensitivity tend to perform well for screening because they rarely miss patients who have the disease, although sometimes at the expense of a higher false-positive rate. Alternatively, tests with high specificity tend to perform well for diagnosis. Highly specific tests rarely misclassify patients as having the disease of interest when, in reality, they do not. Highly specific tests are often useful for confirming the diagnosis in an individual who initially has a positive screening test result.[1,3,9]

The use of highly specific tests is especially necessary when a false-positive test result can place the patient at risk for harm. The false-positive rate for any test can be calculated by the formula 1 – specificity. In prenatal diagnosis, the accepted false-positive rate for most sonographic markers of aneuploidy is approximately 5%. Above this threshold, it is suggested that an unacceptable number of patients would be subjected to both the emotional burden of anxiety over a positive test result and the associated risk, albeit low, of potentially invasive diagnostic testing.

The above-mentioned attributes of the sensitivity and specificity of ideal screening and diagnostic tests are mentioned as broad general principles. As with most general principles, there are exceptions. Although it is important

Table 1. Calculating Measures of Screening Efficiency Using a 2 × 2 Table

	Disease-Positive	Disease-Negative
Test-Positive	True-positive (a)	False-positive (b)
Test-Negative	False-negative (c)	True-negative (d)

Sensitivity = $a/(a+c)$; specificity = $d/(b+d)$; positive predictive value = a/(a + b); and negative predictive value = $d/(c+d)$.

for screening tests to pick up most cases of the disease of interest, high specificity is essential, particularly if the disease is of relatively low prevalence. In such cases, a small drop in specificity has a dramatic impact on the predictive value of a positive test result. False-positive results are not without impact for patients and also influence health care costs. Similarly, ideal diagnostic tests must have high sensitivity as well as high specificity.

Although sensitivity and specificity are inherent performance characteristics of a test that may aid clinicians with the dilemma of whether to order that particular test, they do not provide information on how to proceed if the result is positive. What many clinicians want to know is the predictive value of that test or the probability that a positive result represents an affected patient. These predictive values can be calculated by returning to our 2 × 2 table. The PPV is defined as "the probability of disease in a patient with a positive test result" and can be calculated by dividing the number of patients who test positive and actually have the disease (true-positive results) by the total number of patients with a positive result: PPV = $a/(a + b)$, or true-positive results/total number with a positive result.

The NPV is defined as "the probability of not having the disease when the test result is negative" and can be calculated by dividing the number of patients who test negative and do not have the disease by the total number of patients with a negative result[3]: NPV = $d/(c + d)$, or true-negative results/total number with a negative result.

A fundamental principle in understanding predictive values is that, unlike sensitivity and specificity, the PPV and NPV are dependent on the prevalence of the disease of interest. Holding sensitivity and specificity constant, as the prevalence of a disease increases, the PPV will increase and the NPV will decrease. Conversely, as the prevalence of a disease decreases, the PPV will also decrease whereas the NPV will increase.[1,3,5] In interpreting studies and attempting to generalize them to your population, it is then essential to understand both the setting of the research as well as the characteristics of the population being studied. For example, a study of a new sonographic marker of trisomy 21 in a high-risk referral center of patients all older than 35 years may yield a PPV that is quite high, but when this same marker is tested in a low-risk community setting, one may find a substantial decrease in its predictive value, which could limit the utility of this new marker in this type of patient population.

A practical application of these test performance characteristics can be observed in a study by Odibo et al,[10] which compared the efficiency of second-trimester nasal bone hypoplasia to increased nuchal fold in screening for Down syndrome. To calculate the diagnostic indices, 2 × 2 tables were constructed with the rows representing the presence or absence of the sonographic marker of interest (eg, nasal bone present or nasal bone absent) and the columns representing the presence or absence of Down syndrome, diagnosed by the reference standard method of fetal or neonatal karyotyping. An absent nasal bone yielded sensitivity of 29% and specificity of 99%, whereas a nasal bone measuring less than 0.75 multiples of the median yielded higher sensitivity of 47% at the expense of lower specificity of 94%. Considering the 2 × 2 table, we can see why this finding is true. An absent nasal bone was present in 14 (a) of 49 ($a + c$) cases of Down syndrome, whereas the less discriminatory definition of a nasal bone measurement of less than 0.75 multiples of the median was found in 23 (a) of 49 ($a + c$) cases of Down syndrome, thereby yielding higher sensitivity. However, given that the definition of a nasal bone measurement of less than 0.75 multiples of the median is less discriminatory, this finding was also observed in a higher proportion of patients who did not have Down syndrome compared to the more discriminatory finding of an absent nasal bone, thereby yielding lower specificity.

The study by Odibo et al[10] also shows the effect of disease prevalence on predictive values. Given that the study cohort was of "mixed risk" for aneuploidy, a stratified analysis was then performed after dividing the population into high- and low-risk groups based on referral for the indication of maternal age older than 35 years. The prevalence of Down syndrome was 1.8% in the high-risk group compared to 0.49% in the low-risk group. Using a nasal bone measurement of less than 0.75 multiples of the median as an example, we can see that the PPV was higher (12%) in the high-risk group compared to the low-risk group (3%).

Likelihood Ratios to Refine Risk

Likelihood ratios are another measure of screening efficiency commonly reported in the prenatal diagnostic literature and serve a function similar to that of predictive values. Unlike predictive values, likelihood ratios can be used across varying ranges of disease prevalence and can actually be practically used to alter an individual patient's risk for a given condition. Likelihood ratios are reported separately for positive and negative results and are defined as "the ratio of the likelihood of that result in someone with the disease to the likelihood of that result in someone without the disease".[3] A likelihood ratio of 1.0 indicates that the test result is nondiscriminatory in determining those with the disease and those without. Likelihood ratios can be cal-

culated if both the sensitivity and specificity of a test are known: positive likelihood ratio = sensitivity/(1 – specificity), or the proportion of affected individuals with a positive result/proportion of unaffected individuals with a positive result; and negative likelihood ratio = (1 – sensitivity)/specificity, or the proportion of affected individuals with a negative result/proportion of unaffected individuals with a negative result.

After a likelihood ratio is calculated, it can then be multiplied by a patient's pretest odds of having the disease to determine the patient's posttest odds of having the disease (PPV). In other words, a likelihood ratio can be used to alter an individual patient's a priori risk for having the disease or outcome of interest. The further the likelihood ratio is from 1.0 (in either direction), the greater the effect of the test result on the individual's posttest probability of having or not having the disease of interest. Of note, the equation to calculate posttest probability actually requires the use of odds rather than probability; therefore, pretest probability must first be converted to odds and then posttest odds converted back to probability[3,5,6,11]: odds = probability of the event/(1 – probability of the event); and probability = odds/(1 + odds).

Again, these concepts can best be explained using an example. In 2001, Nyberg et al[12] published a study that used likelihood ratios to estimate the degree of risk of isolated second-trimester sonographic markers for the detection of trisomy 21. The marker of nuchal thickening (≥5 mm) had the highest likelihood ratio of 11.0, meaning that the presence of this marker on second-trimester sonography was 11 times more likely to be found in a fetus with trisomy 21 than in a fetus without trisomy 21. In clinical practice, this finding could mean that a patient could present with a theoretical age-related risk of 1 per 1000 for trisomy 21. If the finding of nuchal thickening is observed during second-trimester sonography, it would raise her risk of trisomy 21 to 1 per 91 ($1/1000 \times 11.0 = 1/91$). These results may substantially influence this patient's decision of whether to undergo amniocentesis. This example illustrates how a sonographic marker could increase the posttest odds of a young patient. Alternatively, an older patient with high pretest odds (eg, 1 per 168) can have lower posttest odds if no notable sonographic marker is present. In the latter scenario, it is more complex to calculate the posttest odds because the individual markers that are absent have specific negative likelihood ratios. Computerized software in sonographic reporting packages is helpful in these situations when patients want specific risk calculations. Alternatively, another report by Nyberg et al[13] suggested using a likelihood ratio of 0.4 when a normal sonographic finding is obtained. For the above example, the posttest odds after a normal sonographic finding would be $1/168 \times 0.4 = 1/420$.[13] Conveniently, obstetricians typically discuss aneuploidy risk in terms of odds; therefore, conversion back and forth between odds and probability is unnecessary for these particular calculations.

Determining a Threshold for a Screening Test: The Receiver Operating Characteristic Curve

Up to this point, we have discussed the calculation of screening test efficiency measures using dichotomous test results. In reality, many tests used in clinical medicine yield results that fall on a continuum; therefore, a threshold for what is considered positive or negative needs to be determined. In the field of prenatal diagnosis, this issue may be relevant in determining the threshold at which an umbilical artery pulsatility index is considered elevated or at which a femur or humerus diaphysis length is considered shortened. To establish these thresholds or cutoffs, a receiver operating characteristic (ROC) curve can be used. To create an ROC curve, the sensitivity and false-positive rates of several different potential threshold values are plotted (sensitivity on the y-axis and 1 – specificity [false-positive rate] on the x-axis). By creating this curve, one can visualize the trade-off between sensitivity and specificity at each cutoff point. Often, the inflection point (or shoulder) of the graph is chosen as the cutoff value because at this value, there are an equal number of false-positive and -negative results.[3,11] In reality, the chosen threshold should depend on the purpose of the test of interest. As mentioned earlier, a screening test should miss as few cases of a disease as possible; therefore, choosing a threshold with high sensitivity, possibly at the expense of a higher false-positive rate, may be warranted. Choosing a threshold with a low false-positive rate at the expense of a lower sensitivity may be more appropriate for a confirmatory diagnostic test.[1,6,9]

An example of establishing thresholds using an ROC curve can be observed in a study by Mari et al,[14] which evaluated the value of peak systolic velocity in the middle cerebral artery for the detection of fetal anemia in cases of maternal alloimmunization. Sensitivity values and false-positive rates of various peak systolic velocities in the middle cerebral artery were plotted for the detection of mild, moderate, and severe anemia. Because the goal of this particular screening test would be to not miss any cases of moderate or severe fetal anemia, the threshold at which sensitivity reached 100% was chosen, corresponding to 1.50 multiples of the median for moderate anemia and 1.55 for severe anemia. To achieve 100% sensitivity, the authors

had to accept a higher false-positive rate of approximately 12% (Figure 1).

Not only can an ROC curve be useful in establishing thresholds for positive and negative test results, but it can also evaluate the overall accuracy of the test. In general, a test that performs well will have an ROC curve that falls near the top left corner of the graph, whereas a test that performs poorly will have an ROC curve that falls closer to the 45° diagonal line (Figure 2). One can then calculate the area under the curve (AUC), which will provide a numeric value for the overall accuracy of the test. For reference, an AUC of 1.0 indicates a perfect test, whereas an AUC of 0.5 indicates a test that performs no better than chance. Because the AUC is a measure of test accuracy, it can also be used to compare different screening or diagnostic tests, with the higher AUC representing the more accurate test in the comparison.[6,9,11]

Returning to our previous example of using nasal bone hypoplasia as a second-trimester marker of aneuploidy, a study by Odibo et al[15] evaluated the test performance characteristics of various definitions of nasal bone hypoplasia and then used these characteristics to construct an ROC curve (Figure 3A). The definition of a biparietal diameter to nasal bone ratio of greater than 11 provided the optimal trade-off between sensitivity and specificity, although at the cost of a 15% false-positive rate compared to the 7% false-positive rate observed using a biparietal diameter to nasal bone ratio of greater than 12. The AUC was 0.7761, correlating with the overall accuracy of nasal bone hypoplasia alone as a marker of fetal aneuploidy. Figure 3B shows a similar ROC curve, although this curve incorporates various definitions of nasal bone hypoplasia combined with other markers of trisomy 21. When combining nasal bone hypoplasia with other markers, one can see that the overall screening efficiency increases (AUC = 0.8089).

Using the AUC to compare test accuracy is most valuable when comparing ROC curves that do not cross at any point. Comparing two curves that cross at one or more points becomes more complicated because two such curves can actually have the same AUC value. Despite having the same AUC value, one curve may have higher false-positive rates for a particular range of values and lower false-positive rates for another range.[16] Rather than merely comparing AUC values for these types of curves, the use of "partial-area" indices has been suggested. These partial-area indices restrict the comparison to a range of sensitivity values or false-positive rates that are most clinically relevant to the question at hand.[17,18] An even more restrictive approach is to compare ROC curves at a preselected value for either sensitivity or the false positive rate. This approach has not been well accepted in the literature for two main reasons. First, it is uncommon for clinicians and researchers to actually agree on a single value at which to compare screening or diagnostic tests. Second, there is a theoretical conflict over comparing ROC curves on the basis of a single value.[19] When evaluating the literature, it is essential to know not only which type of index is being used for comparison but also the advantages and disadvantages of the method.

The AUCs of several ROC curves can be compared statistically by scrutinizing the 95% confidence intervals of the individual AUCs. A simple approach is to consider the ROC curves to be significantly different if the 95% confi-

Figure 1. Receiver operating characteristic curves for the peak systolic blood flow velocity in the middle cerebral artery for the prediction of mild, moderate, and severe fetal anemia. Reproduced with permission from Mari et al.[14]

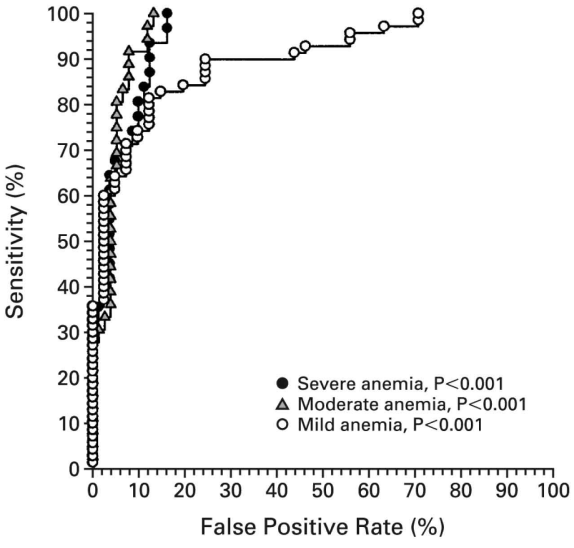
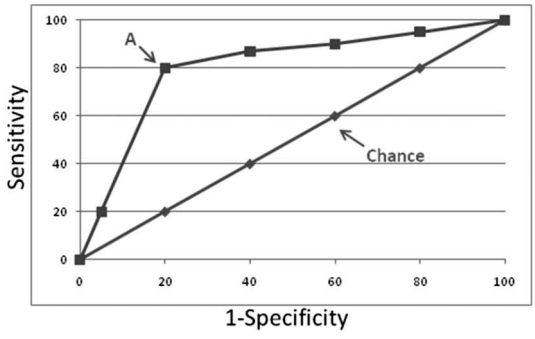


Figure 2. Receiver operating characteristic curve. "A" indicates inflection point (optimal trade-off between sensitivity and specificity).



dence intervals do not overlap. For more advanced analysis including the limitations and misuses of ROC curves, the reader is referred to works by Pepe et al[20] and Cook.[21]

Issues in Study Design and Interpretation

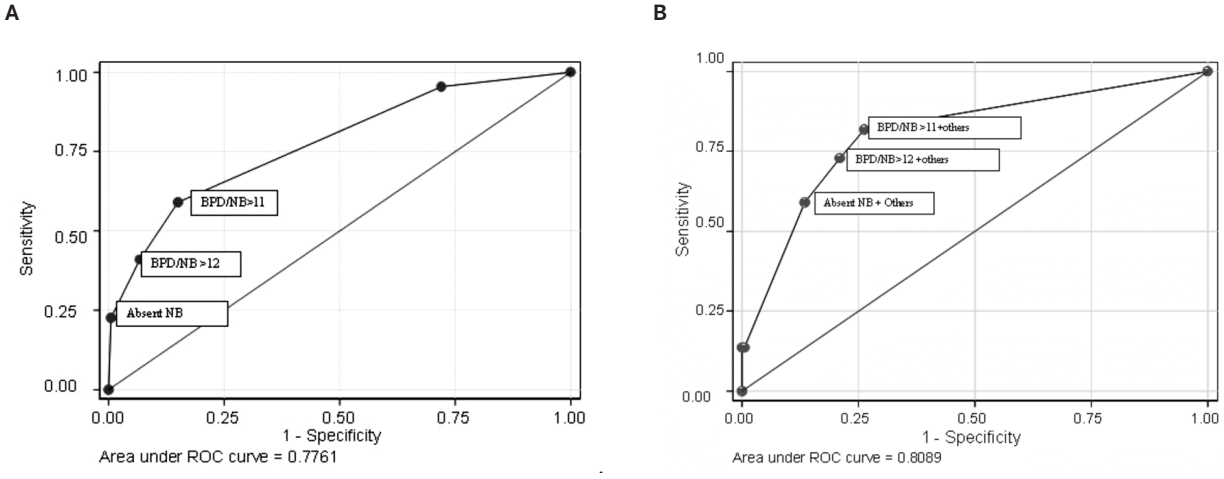
Understanding these key concepts and calculations is only one step in interpreting studies of screening and diagnostic efficiency. As with any type of study, critical evaluation of the study design and methods is necessary before implementing results into clinical practice. There are multiple ways that bias can produce falsely elevated estimates of sensitivity and specificity. First, determination of disease should ideally be obtained through a single reference standard of diagnosis. Also, positive and negative test results should not influence how aggressively or by what means a diagnostic workup is pursued. In addition, the test result itself should not be part of the diagnostic criteria for the disease of interest. Blinding is also essential, especially when subjectivity is involved in the interpretation of a result. Those who interpret the test results should not be aware of the diagnosis, and similarly, those who make the diagnosis should not be aware of the test results. For example, many consider fetal echogenic bowel to be somewhat of a subjective sonographic diagnosis. If conducting a retrospective cohort study to determine the efficiency of echogenic bowel for the detection of trisomy 21, those reviewing the sonograms and assigning a positive or negative finding of echogenic bowel should be blinded to the actual diagnosis of trisomy 21.[1,3,9,11]

It is also important to consider the spectrum of patients included in the study population. A test designed to evaluate the efficiency of a screening or diagnostic test should encompass a broad spectrum of patients both with and without the disease of interest. This population of patients should be similar to the population in whom the test would be used in clinical practice. It is important to remember that the goal of a screening test is not to distinguish very sick patients from very healthy patients; therefore, the study population should attempt to resemble a population of patients with varying disease severities, medical comorbidities, and even other diseases that may closely mimic the disease of interest.[3,11]

Finally, to determine reproducibility, the methods of the study should offer a detailed explanation as to how the test was performed and by whom it was administered. The skill level of the operator (potentially the sonographer) and the type of institution (community center versus tertiary care or referral center) will also influence how the results will generalize to another population.

Studies of screening and diagnostic efficiency are typically of an observational nature and may be performed using a range of designs, including prospective cohort, retrospective cohort, cross-sectional, case-control, and secondary analysis of data from prior studies. Each of these designs has its own benefits and pitfalls, which are beyond the scope of this article. It is important to highlight that the case-control study design cannot be used to determine predictive values because these values are influenced by disease prevalence. Because cases and controls are selected for inclusion, the

Figure 3. Receiver operating characteristic (ROC) curves evaluating the efficiency of nasal bone (NB) hypoplasia in the detection of Down syndrome. **A**, Receiver operating characteristic curve of a model using different definitions of nasal bone hypoplasia only for detection of Down syndrome. **B**, Receiver operating characteristic curve of a model using combinations of different definitions of nasal bone hypoplasia and other markers for detection of Down syndrome. BPD indicates biparietal diameter. Reproduced with permission obtained from Odibo et al.[15]



prevalence of disease is, therefore, "fixed" by the study design. Reproducing a generalizable spectrum of patients also becomes difficult with this type of study design.[10]

When evaluating screening or diagnostic tests using a combination of variables, it is important to provide some statistical approach to weighting of the variables because it could improve the performance of an overall set of variables. This process is important because different components of the set of variables may not be equally effective predictors of the outcome of interest. One method for handling this situation is by using principal components analysis, which would provide the weighting factor for each variable. Principal components analysis is a mathematical procedure using orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of uncorrelated variables called principal components. The transformation is defined to ensure that the first principal component has as high a variance as possible and each succeeding component in turn has the highest variance possible under the constraint that it be uncorrelated with the preceding components.[22] This approach has rarely been used in obstetric imaging literature but should be encouraged given the increasing use of multiple-parameter screening paradigms.

Conclusions

Studies evaluating screening and diagnostic efficiency are common in the sonographic and prenatal diagnostic literature. This article aimed to provide an overview of some of the key concepts frequently encountered in such studies. Whether designing your own study or analyzing results from a study in the literature, understanding how to calculate and interpret test performance characteristics in the context of a variety of different study designs is essential to the practical application of these results to any particular patient population.

References

1. Boardman LA, Peipert JF. Screening and diagnostic testing. *Clin Obstet Gynecol* 1998; 41:267–274.
2. Wilson JMG, Jungner G. *Principles and Practice of Screening for Disease*. Geneva, Switzerland: World Health Organization; 1968.
3. Fletcher RH, Fletcher SW. *Clinical Epidemiology*. 4th ed. Baltimore, MD: Lippincott Williams & Wilkins; 2005.
4. Wapner RJ, Jenkins TM, Khalek M. Prenatal diagnosis of congenital disorders. In: Creasy RK, Resnik R, Iams JD, Lockwood CJ, Moore TR (eds). *Maternal-Fetal Medicine: Principles and Practice*. 6th edition. Philadelphia, PA: Saunders Elsevier; 2009:221–274.
5. Macones GA. Evidence-based practice in perinatal medicine. In: Creasy RK, Resnik R, Iams JD, Lockwood CJ, Moore TR (eds). *Maternal-Fetal Medicine: Principles and Practice*. 6th ed. Philadelphia, PA: Saunders Elsevier; 2009:207–218.
6. Schulz KF, Grimes GA. *The Lancet: Handbook of Essential Concepts in Clinical Research*. Philadelphia, PA: Elsevier; 2006.
7. Malone FD, Canick JA, Ball RH, et al. First-trimester or second-trimester screening, or both, for Down's syndrome. *N Engl J Med* 2005; 353:2001–2011.
8. American College of Obstetricians and Gynecologists. ACOG practice bulletin No. 77: screening for fetal chromosomal abnormalities. *Obstet Gynecol* 2007; 109:217–227.
9. Peipert JF, Sweeney PJ. Diagnostic testing in obstetrics and gynecology: a clinician's guide. *Obstet Gynecol* 1993; 82:619–623.
10. Odibo AO, Sehdev HM, Gerkowicz S, Stamilio DM, Macones GA. Comparison of the efficiency of second-trimester nasal bone hypoplasia and increased nuchal fold in Down syndrome screening. *Am J Obstet Gynecol* 2008; 281:e1–e5.
11. Newman TB, Browner WS, Cummings SR, Hulley SB. Designing studies of medical tests. In: Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB (eds). *Designing Clinical Research*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2007:183–205.
12. Nyberg DA, Souter VL, El-Bastawissi A, Young S, Luthhardt F, Luthy DA. Isolated sonographic markers for detection of fetal Down syndrome in the second trimester of pregnancy. *J Ultrasound Med* 2001; 20:1053–1063.
13. Nyberg DA, Luthy DA, Resta RG, Nyberg BC, Williams MA. Age-adjusted ultrasound risk assessment for fetal Down's syndrome during the second trimester: description of the method and analysis of 142 cases. *Ultrasound Obstet Gynecol* 1998; 12:8–14.
14. Mari G, Deter RL, Carpenter RL, et al. Noninvasive diagnosis by Doppler ultrasonography of fetal anemia due to maternal red-cell alloimmunization. Collaborative Group for Doppler Assessment of the Blood Velocity in Anemic Fetuses. *N Engl J Med* 2000; 342:9–14.
15. Odibo AO, Sehdev HM, Sproat L, et al. Evaluating the efficiency of using second-trimester nasal bone hypoplasia as a single or a combined marker for fetal aneuploidy. *J Ultrasound Med* 2006; 25:437–441.
16. Metz CE. Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems. *J Am Coll Radiol* 2006; 3:413–422.
17. McClish DK. Analyzing a portion of the ROC curve. *Med Decis Making* 1989; 9:190–195.
18. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology* 1996; 201:745–750.
19. Halpern EJ, Alpert M, Krieger AM, Metz CE, Maidment AD. Comparisons of receiver operating characteristic curves on the basis of optimal operating points. *Acad Radiol* 1996; 3:245–253.
20. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* 2004; 159:882–890.
21. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007; 115:928–935.
22. Jolliffe IT. Springer Series in Statistics: *Principal Component Analysis*. Vol 29. 2nd ed. New York, NY: Springer; 2002.