Assessing the Predictive Value of the American Board of Family Practice In-training Examination

William H. Replogle, PhD; William D. Johnson, PhD

Background and Objectives: The American Board of Family Practice In-training Examination (ABFP ITE) is a cognitive examination similar in content to the ABFP Certification Examination (CE). The ABFP ITE is widely used in family medicine residency programs. It was originally developed and intended to be used for assessment of groups of residents. Despite lack of empirical support, however, some residency programs are using ABFP ITE scores as individual resident performance indicators. This study’s objective was to estimate the positive predictive value of the ABFP ITE for identifying residents at risk for poor performance on the ABFP CE or a subsequent ABFP ITE.

Methods: We used a normal distribution model for correlated test scores and Monte Carlo simulation to investigate the effect of test reliability (measurement errors) on the positive predictive value of the ABFP ITE. Results: The positive predictive value of the eight specialty subscales ranged from .26 to .57. Conclusions: Only the composite score of the ABFP ITE has acceptable positive predictive value to be used as part of a comprehension resident evaluation system. The ABFP ITE specialty subscales do not have sufficient positive predictive value or reliability to warrant use as performance indicators.

(Fam Med 2004;36(3):185-8.)

The American Board of Family Practice In-training Examination (ABFP ITE) is a cognitive examination similar in content to the ABFP Certification Examination (CE). The ABFP ITE measures knowledge and patient management skills in eight major areas: internal medicine, surgery, obstetrics, community medicine, pediatrics, psychiatry and behavioral sciences, geriatrics, and gynecology. Lancaster et al. reported that 98% of the accredited family medicine residency programs use the ABFP ITE. The ABFP ITE was originally developed and intended to be used for overall assessment of groups of residents. Despite lack of empirical support, however, some residency programs are using ABFP ITE scores as individual resident performance indicators.

As with diagnostic tests used in clinical practice, use of ABFP ITE scores to identify residents who are at risk for failure of the ABFP CE or specialty area knowledge deficits will result in one of four outcomes. For example, a resident who fails the ABFP CE will have scored below an arbitrary cut point on the ABFP ITE (true positive) or obtained a score above the arbitrary cut point on the ABFP ITE (false negative). A resident who passes the ABFP CE will have obtained a score below the arbitrary cut point on the ABFP ITE (false positive) or obtained a score above the arbitrary cut point on the ABFP ITE (true negative). These four outcomes can be calculated from data on the sensitivity and specificity of the ABFP ITE.

Knowing the sensitivity and specificity of the ABFP ITE, however, is not sufficient to interpret a given test result. A test’s sensitivity only indicates the probability of obtaining a low ABFP ITE score when a resident is destined to fail the ABFP CE. It does not, however, tell us the positive predictive value (PPV) of the ABFP ITE—ie, the likelihood that a resident with a low ABFP ITE score will fail the ABFP CE. Although such information has been reported for the internal medicine ITE, no investigators have reported on the sensitivity, specificity, and predictive value of the ABFP ITE for identifying residents at risk to have a poor score on the ABFP CE or a subsequent ABFP ITE. Accordingly, the present study’s purpose is to estimate the positive predictive value of ABFP ITE scores for
identifying residents at risk for failure or poor performance on the ABFP CE or a subsequent ABFP ITE, respectively.

Methods

To estimate the PPV of the ABFP ITE, we assumed that a composite score or specialty subscale score that placed a resident in the lower 7% of the distribution when compared to other residents in the same year of training indicated an at-risk resident on that scale (an assumption we do not support but one that is likely made in practice). Selection of the seventh percentile is arbitrary, but this percentile rank approximates the percent of examinees that fail the ABFP CE. An ABFP ITE composite score of 300, 380, and 420 for a first-, second-, and third-year resident, respectively, places a resident at about the seventh percentile compared to other residents in the same year of training who also took the ABFP ITE.8

Depending on the reliability of the scale, a certain percentage of residents who score at or below the seventh percentile on an ABFP ITE scale would be expected to score above the seventh percentile on a test retake even without remediation. This notion is often referred to as regression toward the mean.9–11 Thus, some residents who score at or below the seventh percentile would be true positives, and some would be false positives. Similarly, some of those who score above the seventh percentile would be true negatives, and some would be false negatives. As the reliability coefficient for the test decreases, the number of false positives increases, and the positive predictive value decreases.

We used a normal distribution model for correlated test scores and Monte Carlo simulation2,6,13 to investigate the effect of test reliability (measurement errors) on the PPV of the ABFP ITE. Monte Carlo simulation is a statistical procedure for investigating difficult mathematical problems in all areas of research that otherwise might require years of experimental observation. The name comes from the city in Morocco where games of chance are played on a daily basis. The outcome of a single game cannot be predicted accurately, but the frequency of different types of outcomes in a large random sample of games can be predicted with complete accuracy.

In this study, we simulated very large random samples from a population of test score pairs that are known to be normally distributed with prespecified mean, standard deviation (SD), and test reliability. Using this approach, we cannot accurately predict the success or failure of an individual resident, but we can accurately predict the proportions of successes and failures in the population.

Our primary aim was to estimate the probability that a resident who scores at or below the seventh percentile on an ABFP scale will subsequently score at or below the seventh percentile on that scale on subsequent examination using either the ABFP ITE or ABFP CE. We assumed the scores on each test are normally distributed with a mean of 500 and an SD of 100. Additionally, we assumed that the joint distribution of scores on the two tests is bivariate normal, with tests scores having a correlation coefficient equal to the square root of a common reliability coefficient. We used the most recently published ABFP ITE reliability coefficients, which were derived by the ABFP.

For each reliability coefficient shown in Table 1, we simulated a sample of 100,000 pairs of test scores, and for each sample, we enumerated the true and false positives. That is, a resident who scored at or below the seventh percentile on the first test and subsequently at or below the seventh percentile on the second test would be a true positive. On the other hand, a resident who scored at or below the seventh percentile on the first test and subsequently above the seventh percentile on the second test would be a false positive. Next, we used the simulated test scores to compute empirical estimates of the sensitivity, specificity, and positive predictive value of each ABFP ITE scale. We repeated the simulation for various values of the test reliability coefficient to study the effect of test reliability on the PPV of the test.

The only potential limitation of this method is the possibility that the actual ITE test scores for the population of residents are not normally distributed. However, the Monte Carlo technique is robust to minor violations of this normality assumption, such that minor discrepancies in normality should have little effect on the accuracy of our results.

Results

The results of the simulation study are shown in Table 1 and Figure 1. The reliability coefficient for the composite score on the ABFP ITE is 0.87 while subscale

| Table 1 |
| Reliability Coefficients and Positive Predictive Value of ABFP ITE Scores |
| Composite score | 0.87 | 0.72 |
| Internal medicine | 0.70 | 0.57 |
| Surgery | 0.34 | 0.33 |
| Obstetrics | 0.53 | 0.44 |
| Community medicine | 0.41 | 0.37 |
| Pediatrics | 0.54 | 0.44 |
| Psychiatry | 0.23 | 0.26 |
| Geriatrics | 0.55 | 0.46 |
| Gynecology | 0.37 | 0.35 |

ABFP ITE—American Board of Family Practice In-training Examination
reliability coefficients range from 0.23 to 0.70. The composite score has the highest PPV (0.72). This implies that 72% of residents who are identified as at risk based on the composite score will actually be at risk for poor performance on the ABFP ITE or a subsequent ABFP ITE (ie., have a score at or below the seventh percentile). That is, 72% of those identified as at risk based on the composite score will be true positives; 28% will be false positives.

At the other extreme, the psychiatry subscale has a reliability coefficient of 0.23 and PPV = 0.26. This indicates that only 26% of those identified as at risk based on the psychiatry subscale will be true positives; 74% will be false positives. The PPVs reported in Table 1 and Figure 1 apply to first-, second-, and third-year residents.

Discussion

The reliability coefficient for a test is a measure of the consistency or repeatability of scores obtained by the same person on retesting with an identical test or equivalent form of a test. The closer the reliability coefficient is to a value of 1.0, the more the test is free of measurement error. If the ABFP ITE were completely free of measurement error, a resident’s score obtained on the first examination will be exactly the same as the score obtained on retesting. As measurement error increases, the reliability coefficient decreases, and the value of the examination for predicting future performance diminishes.

As shown in Figure 1, the PPV of a test increases as the reliability coefficient increases. However, the PPV of any scale does not exceed 75% until the scale’s reliability coefficient reaches 0.9. Nunnally and Bernstein cautioned that if a test score is to be used to make decisions regarding an individual, the minimally acceptable reliability coefficient should be 0.90. The composite score is the only ABFP ITE scale that approximates this requirement. Additionally, Repogle cautioned that the ABFP ITE subscales lack sufficient reliability to warrant interpretation for individual residents. In fact, in the ABFP ITE Handbook, Avant cautions against making judgments regarding individual residents based on scales with a small number of items, because of low reliability. The results presented in Figure 1 are consistent with these recommendations not to make decisions about individual residents based on ABFP ITE specialty subscales.

Using diagnostic tests with low PPVs, such as those used in community screening programs, results in a large number of false positive test results. The cost of additional testing among this group of false-positive patients needlessly drains health care resources. Similarly, when we identify 10 residents at risk because of poor performance in the pediatrics section of the exam, we correctly identify, on average, 4.4 at-risk residents (true positives) and incorrectly identify 5.6 residents (false positives). Thus, we may require that 56% of those identified as at risk needlessly devote educational time and resources to remediation of a problem that does not exist.

Slawson and Shaughnessy have cautioned clinicians against the use of inaccurate diagnostic tests, even those that may appear to be extremely useful. Physicians must protect their patients from “well-intended but misinformed policies.” As educators, we too must clearly understand the consequences of the well-intended but often misguided use of educational tests. Based on this Monte Carlo simulation, which used known reliability coefficients of the ABFP ITE, the ABFP ITE composite score has acceptable PPV to be used as part of a comprehension resident evaluation system. The ABFP ITE specialty subscales do not, however, have sufficient PPV to warrant use as performance indicators.

Figure 1
Results of Simulation Study

Corresponding Author: Address correspondence to Dr Repogle, University of Mississippi Medical Center, Department of Family Medicine, 2500 N. State Street, Jackson, MS 39216. 601-984-5415. Fax: 601-984-6835. wreplogle@familymed.umsm.edu.

REFERENCES