Vital Signs: How Early Can Resident Evaluation Predict Acquisition of Competency in Surgical Pathology?

BARBARA S. DUCATMAN, MD
H. JAMES WILLIAMS, MD
GERALD HOBBS, PhD
KYMBERLY A. GYURE, MD

Abstract

Objectives To determine whether a longitudinal, case-based evaluation system can predict acquisition of competency in surgical pathology and how trainees at risk can be identified early.

Design Data were collected for trainee performance on surgical pathology cases (how well their diagnosis agreed with the faculty diagnosis) and compared with training outcomes. Negative training outcomes included failure to complete the residency, failure to pass the anatomic pathology component of the American Board of Pathology examination, and/or failure to obtain or hold a position immediately following training.

Findings Thirty-three trainees recorded diagnoses for 54326 surgical pathology cases, with outcome data available for 15 residents. Mean case-based performance was significantly higher for those with positive outcomes, and outcome status could be predicted as early as postgraduate year-1 ($P = .0001$). Performance on

the first postgraduate year-1 rotation was significantly associated with the outcome ($P = .02$). Although trainees with unsuccessful outcomes improved their performance more rapidly, they started below residents with successful outcomes and did not make up the difference during training. There was no significant difference in Step 1 or 2 United States Medical Licensing Examination (USMLE) scores when compared with performance or final outcomes ($P = .43$ and $P = .68$, respectively) and the resident in-service examination (RISE) had limited predictive ability.

Discussion Differences between successful- and unsuccessful-outcome residents were most evident in early residency, ideal for designing interventions or counseling residents to consider another specialty.

Conclusion Our longitudinal case-based system successfully identified trainees at risk for failure to acquire critical competencies for surgical pathology early in the program.

Introduction

In response to public concerns about physician competency, the education and evaluation of physicians has come under increasing scrutiny. A number of methods have been suggested to monitor acquisition of the Accreditation Council for Graduate Medical Education (ACGME) core competencies.¹ We previously described a longitudinal case-based evaluation system of residents and documented its methods.² However, a thorough evaluation of competence for medical practice requires integration of all 6 ACGME competencies, and is challenging. Beyond documenting competency, a primary purpose of an evaluation method is early identification of residents having difficulty mastering

the competencies, and many educators are struggling to determine the most predictive tools for their specialty.

Pathology residents are heavily involved in clinical care, including diagnosis of surgical pathology cases. Although monthly evaluations of residents are useful, faculty members are often generous in their expectations of junior residents and the opportunity for early intervention may be lost. We hypothesized that an objective patient-based system of evaluation would improve identification of trainees with deficiencies, and decided to review our cumulative data with the following measures of outcome. A negative outcome was defined as either failure to complete the training program, failure to pass the anatomic pathology (AP) component of the American Board of Pathology examination, or failure to obtain or hold a fellowship position or a job after residency. A positive outcome included success in all 3 of these measures. An implicit assumption is that the anatomic pathology component of the ABP examination tests for many skills needed for the daily practice of pathology.

Methods

Starting in August 2001, we instituted our case-based evaluation using our anatomic pathology information

All authors are at West Virginia University. **Barbara S. Ducatman, MD**, is Professor and Chair of Pathology; **H. James Williams, MD**, is Professor and Vice Chair for Educational Programs; **Kymberly A. Gyure, MD**, is Associate Professor and Vice Chair of Pathology and Residency Program Director; and **Gerald Hobbs, PhD**, is in the Department of Statistics and Community Medicine.

The authors thank Mr Ed Gray for designing the data collection program and Ms Linda Tomago for maintaining the data files.

Corresponding author: Barbara S. Ducatman, MD, Department of Pathology, West Virginia University School of Medicine, PO Box 9203, Morgantown, WV 26506-9203, 304.293.3593, bducatman@hsc.wvu.edu

DOI: 10.4300/01.01.0007

system (CoPath Client Server, Cerner, Kansas City, MO). The grading scheme for each case was simple: (1) “agree” indicated a correct diagnosis with all relevant data needed for treatment and prognosis, (2) “partial agree” reflected a somewhat inaccurate (such as correct for benign or malignant but not the exact diagnosis) or incomplete (such as missing or incorrect staging information) diagnosis, and (3) “disagree” indicated a major discrepancy between resident and faculty diagnoses. In our previous review of our data, we found that percent agreement was the most useful evaluative measure²; therefore, data on percent agreement using the agree status were the only variables analyzed. Faculty diagnosis was considered the gold standard. Consultation among our faculty members is common, and faculty performance is tracked by a parallel quality assurance program mandated by laboratory accreditation standards.

All of our residents are combined anatomic/clinical pathology residents and all start with a rotation in surgical pathology as either their first or second month. On average, our residents complete 4 months of surgical pathology per year. Residents preview their surgical pathology cases before sign-out with faculty; junior residents generally do fewer, simpler cases, and senior residents are expected to preview all cases. The faculty member grades each trainee’s diagnosis and the agreement status is entered into the quality assurance module of our Anatomic Pathology Information System. The ongoing data are automatically generated in an Excel spreadsheet (Microsoft Corporation, Redmond, WA) that is then transferred to a statistical program (JMP version 7 SAS, Cary, NC). Final resident outcome (success in all 3 measures) was then added as defined previously. After obtaining institutional review board permission to review and publish summary data, the summary data were analyzed by standard analysis of variance unless indicated.³

Results

Thirty-three trainees previewed and recorded their diagnoses for 54 326 surgical pathology cases during 92 months of data collection (August 2001 through February 2009). To date, 15 have documented outcomes, 8 successful and 7 unsuccessful. A change in program requirements by the American Board of Pathology decreased the training cycle from 5 to 4 years in 2002; thus, the majority of residents trained for 4 years. Residents finishing with a 5-year requirement used the fifth year as a surgical pathology fellowship. For the purpose of the study, postsophomore fellows, medical students acting as junior residents following their second year of medical school, were designated as postgraduate year (PGY)-0, and the surgical pathology fellows were designated as PGY-5 for the data analysis. Several individuals (predominantly postsophomore fellows and PGY-4 and PGY-5 students) left or finished the

program during this period, and 2 entered from other programs.

The number of surgical pathology rotations per resident surveyed ranged from 1 for our new PGY-1 students and postsophomore fellows to 28 for those residents completing surgical pathology fellowships. The total number of cases reviewed by each resident ranged from 34 to 4 968 cases (median, 1 274) and the overall percent agreement ranged from 33% to 100% (median, 81.95%). Although our residents rotate at 2 other institutions, there is only 1 month of surgical pathology training elsewhere during a typical total of 13 months of surgical pathology.

We previously demonstrated that faculty grading is variable regarding resident training,² especially because 27 faculty members participated (with faculty entering and leaving the department). Thus, there is variation in the grading between different academic years (FIGURE 1). In order to adjust for this variability, we analyzed not only percent agreement per resident, but also the difference between each trainee’s mean percent agreement for each rotation in an academic year and the overall mean percent agreement for all residents that academic year (FIGURE 2).

Outcome data were available for 15 residents, and the results for first attempt of the AP board examination were available for 13 individuals. Of these, 8 residents passed and 5 residents failed in their first AP board examination attempt. In addition, we counted 2 residents who never attempted the board examination as negative outcomes. The first resident was dismissed from the program for failure to progress academically after the third year. The second

FIGURE 2

OVERALL SUMMARY PERFORMANCE OF TRAINEES EXPRESSED AS THE MEAN DIFFERENCE BETWEEN INDIVIDUAL PERFORMANCE AND MEAN PERFORMANCE FOR ALL RESIDENTS DURING THE CORRESPONDING ACADEMIC PERIOD ($P < .0001$).

resident was dismissed for failure to handle the gross pathology workload. Although considered to be weak academically, diagnostic acumen was not the major problem. Another resident did not take the board examination during the period of eligibility and also experienced difficulty in the first postresidency position (as reported to us by the employer). Of the 5 residents who failed the board examination, 2 trained entirely (PGY-1–4) in our program, 1 trained for 1 year (PGY-1) before transferring to another institution for personal reasons (while in good standing), and 2 trained for 2 years (PGY-3–4), both leaving while in good standing, 1 because of a program closure and the other for personal reasons. All 8 residents with positive outcomes trained entirely in our program (PGY-1–4). In the negative outcomes group were 4 international medical graduates and 3 graduates of Liaison Committee on Medical Education-accredited programs. The 8 residents who passed the board examination were 3 international medical graduates and 5 graduates of Liaison

Committee on Medical Education-accredited medical schools. All had completed medical school within the past 10 years, and none held a joint degree. Finally, those trainees with negative experiences were more likely to have matriculated in the program in the early years of this study (the last resident to fail the first attempt of the board examination started in the second academic year of the evaluation program).

Additionally, 3 residents have completed training within the past year and have not yet taken the board examination, 4 postsophomore fellows did not enter our training program, 1 surgical pathology fellow had already passed the AP board examination, and 11 residents are still in training.

Relationship Between Program Performance and Outcomes

The review of results for residents with known outcome status showed striking differences. First, the mean overall performance, as defined by summary data on percent

agreement, was significantly higher for those with positive outcomes (FIGURE 3, $P = .0023$). Two residents with negative outcomes exited early (after PGY-1 and PGY-3) and 2 entered late (both as PGY-3 students).

Because we have previously shown that agreement status rises with level of training,² and the resident dismissed had 1 less year of data collection (having left after PGY-3), we analyzed only the data for the residents with unsuccessful outcomes. There was a difference between the dismissed resident who had an overall mean percent agreement of 65.9% versus all the other residents with unsuccessful outcomes, with an overall mean percent agreement of 73.0%. This was not statistically significant. The dismissed resident performed at the level of the (poor outcomes) peer group when the data were analyzed by PGY. When we reanalyzed the data to exclude the resident who was dismissed, the result was not statistically significantly different ($P = .0027$), as was the case for those who did not complete 4 years in our program ($P = .0056$). Additionally, we excluded the 2 residents with negative outcomes who did not take the board examination and found similar results ($P = .0049$). We performed an analysis of variance in a mixed model with repeated measures of variance and found that percent agreement status correlated with outcomes ($P = .0006$).

Thereafter, the analysis included evaluation of overall summary data for the residents with known outcomes including board passage and otherwise, except when

indicated. We also adjusted for differences in variability by comparing the differences between resident performance and the academic year means; a significant difference between residents with positive and negative outcomes persisted. There was no difference in the number of cases analyzed per rotation between the two groups.

The most surprising finding was that outcome status could be predicted as early as PGY-1, whether expressed as percent agreement (FIGURE 4, $P = .0001$) or difference between the individual resident performance and overall academic year mean ($P = .0001$). When we analyzed only residents with unsuccessful outcomes, there was no significant difference when the dismissed resident was compared with the other residents with subsequent poor outcomes in PGY-1. In fact, performance on the first rotation of PGY-1 (analyzed for 11 residents with a known outcome who did their first rotation in surgical pathology in our program) was significantly associated with the outcome ($P = .02$). At the same time, individual performance in the first rotation after adjustment for the yearly overall mean was not statistically significant ($P = .07$).

The difference between residents' successful and unsuccessful outcomes narrowed during subsequent postgraduate years, but remained statistically significant until the PGY-5 year, when the total number with outcomes dropped to 4. When we reviewed only residents with poor outcomes, there were no statically significant differences between the dismissed resident and other residents with unsuccessful outcomes in subsequent PGY-2–3 (the dismissed resident left after PGY-3).

FIGURE 5

ACQUISITION OF COMPETENCY OVER ROTATIONS FOR RESIDENTS ENROLLED FOR THE ENTIRE 4 YEARS IN OUR PROGRAM. TRAINEES WITH UNSUCCESSFUL OUTCOMES (L) ($r^2 = .52$, $P < .0001$) AND SUCCESSFUL OUTCOMES (R) ($R^2 = .24$, $P < .0001$) OVER 4 YEARS. NOTE THAT THOSE TRAINEES WITH EVENTUAL UNSUCCESSFUL OUTCOMES STARTED FAR BELOW THOSE WITH POSITIVE OUTCOMES, AND IMPROVED MORE RAPIDLY, BUT HAD NOT QUITE CAUGHT UP BY THE FOURTH YEAR

When the agreement status was plotted against the rotation months completed for residents doing their entire residency here, it is clear that the residents with unsuccessful outcomes were improving more rapidly (FIGURE 5). However, residents with unsuccessful outcomes started far below those with successful outcomes and did not make up the difference in 4 years. This finding suggests that although the unsuccessful-outcome residents had deficiencies, these were at least partially overcome during training.

Relationship to Other Assessment Data

We were also able to retrieve data on the resident in-service examination (RISE) and United States Medical Licensing Examination (USMLE) scores for our residents and compare these results with previously stated outcomes. We used the RISE surgical pathology score only. There was no significant difference between RISE scores of successful versus unsuccessful outcomes for PGY-1 residents (chi-square test, $P = .8076$), as well as for PGY-3 and PGY-4, and the difference at the PGY-2 level was the only one that was statistically significant (chi-square, $P = .0256$).

There was no association between USMLE scores and the overall performance by our evaluation process of our trainees. There was no statistically significant difference in Step 1 USMLE scores, which would cover basic pathology knowledge, for those residents with positive outcomes as compared with the negative outcomes group (mean 3-digit scores of 208 vs 201, respectively, $P = .43$), nor was there a statistically significant difference between these groups on USMLE Step 2 (mean 3-digit score of 216 for those with positive outcomes vs 211 for those with negative outcomes, $P = .68$).

Interestingly, there was a statistically significant association between the mean USMLE Step 3 score and a positive outcome for 14 trainees (mean 3-digit score of 191.5 for trainees with unsuccessful outcomes vs mean 3-digit score of 204 in those with successful outcomes, $P = .02$), excluding 1 trainee dismissed before passing USMLE Step 3. Ten of our trainees had no history of a clinical internship, whereas 4 had served a clinical internship (3 in internal medicine and 1 in a combined psychiatry/internal medicine program). When we reviewed

the results for USMLE Step 3 in terms of a clinical internship, there was a statistically significant difference ($P = .05$, student's t test) between those with and without previous clinical training. The history of previous clinical training showed no significant association with final outcome for our pathology trainees, although again the numbers are small.

As a last step, we undertook a decision tree analysis (data-mining). A RISE score for the PGY-1 level of 374 was somewhat predictive of further success, with all individuals below that score having unsuccessful results, while 85% of residents with a score of 374 or more successfully passed the examination. For PGY-2 students, all with a RISE score 418 or more passed, while two-thirds of those with scores under 418 failed. For PGY-3 students, the cutoff mean RISE score was 394. All residents below this failed, and 77% of those with 394 or more passed. Finally, for PGY-4, all residents with scores of 524 or more passed, while about half failed. Although cutoff scores could be determined, these were not cleanly predictive and declined over time.

We repeated the same decision tree analysis for percent agreement. All residents with a mean PGY percent agreement for PGY-1 at or above 70.09% had successful outcomes, whereas none of those below had successful outcomes. For PGY-2 and PGY-3, the same pattern was repeated, with cutoff percent agreement of 78.45% and 83.53%, respectively, predicting success or failure with no overlap. The cutoff for PGY-4 was not as predictive: 77% of the residents with a percent agreement of 81.67% or more had successful outcomes and residents below this number had unsuccessful outcomes.

Because of the small sample size, the results of the "cutoff scores" should be viewed critically and applied with very great caution to as a predictor of future success. Still, these findings are very interesting. The data for percent agreement were more predictive than those for RISE scores, with cleaner discriminatory values between successful and unsuccessful outcomes. This appears to validate the analyses, showing more statistical significance with our summary data than with the RISE score data.

Discussion

An important purpose of a training program evaluation process is the earliest possible identification of individuals with deficiencies. Although there were statistically significant differences in the overall performance of residents with successful and unsuccessful outcomes, this is not surprising and was not the focus of our study. Early identification of risk for failure can assist remediation of these deficiencies or removal of the individuals from the training program, depending on the nature and severity of the differences. Our data indicate that a longitudinal case-based evaluative process like that presented here can identify individuals at risk as early as the first postgraduate year and possibly as early as the first surgical pathology

rotation. Although the goal is to evaluate each resident using the same objective criteria for agreement, postsophomore fellows and PGY-1 students are graded more leniently and finishing trainees are graded most stringently. However, significant and predictive differences emerge among trainees, particularly because we can adjust for the norms in grading cases for each year.

The results are still preliminary in that only 15 individuals have an outcome, albeit each with a large set of data. Historically, residents with significant deficiencies were also apparent to our faculty and evident in subjective assessments by the end of their training. Because the utility of this evaluative process and those cutoffs predicting success or failure were not available at the time, no interventions were attempted unless difficulties came to light via standard assessments. It is fortunate that the number of dismissals is small. However, because of the limited sample, differences between residents who complete and do not complete a residency cannot be adequately analyzed. As the differences between those with positive and negative outcomes narrow over time, we believe that early intervention will be useful, and predict that either higher success rates, or when needed, alternative placements in other specialties can be achieved.

Like all medical specialties, pathology requires a defined skill set of knowledge and ability. A significant component of the AP board examination is visually based (microscopic and computer images), a unique skill set that does not fit all learning styles. We believe that the AP board examination is a valid evaluation of diagnostic acumen and a good outcome variable. Poorly performing pathology trainees may lack either what experienced pathologists call a "good eye" (ie, excellent pattern recognition or visual memory skills) or underlying medical knowledge (ability to synthesize clinical information to infer a differential diagnosis), which is especially important on a single examination event.

For this reason, we analyzed the medical knowledge competency as assessed by the USMLE examination. Although the number of individuals is still small, we were surprised to find no association between the performance on USMLE Step 1 examination (which covers basic knowledge in pathology), as might be expected. We also found no association for USMLE Step 2 with the outcome for our trainees. Although USMLE Step 3 scores were associated with outcome, this appears to be correlated with the presence or absence of a clinical internship. These results suggest to us that the deficiencies in anatomic pathology diagnostic acumen are more likely weaknesses in innate pattern recognition skills or visual memory rather than a deficit in medical knowledge. We think this hypothesis is also testable and intend to pursue it.

The RISE scores were interesting, and the RISE examination is configured differently from the AP board examination. With the exception of the PGY-2 year, RISE

scores could not distinguish, with statistic significance, residents with successful outcomes from those with unsuccessful outcomes. At first it appears counterintuitive that the RISE score is less predictive of future success. However, the RISE examination contains only photomicrographs; in contrast, a major component of the AP board examination is slide-based, more akin to normal surgical pathology practice.

The results of this study highlight areas for further investigation. First, it illustrates the need for medical students to select specialty training that is the best fit with their inherent abilities and interests. A testing and counseling process for medical students in their third year to determine areas of strengths and weaknesses would be one such step.⁴ Even if a trainee is interested in a specialty choice in which strengths are low, the trainee or program could start with this knowledge and design the educational experience with deficits in mind.

Second, the need for remediation should be identified and acted on early. However, the methodologies for remediation are more difficult and may need to be designed with respect for individual differences. As an example, one resident may benefit from more reading, but another with inherent lack of pattern recognition may be required to spend more time at the microscope and adjusting teaching to maximize visual learning. Because the interface of pattern recognition and diagnosis is considered the “art” of pathology, optimal methods for improving this will require further study.⁵

Although our program evaluated only pathology residents’ AP performance, the concepts are applicable to medical students and all training programs. Pathologists have some advantages in recording presence and degree of diagnostic agreement, but this basic concept applies to most of medicine. The increasing use of electronic medical records will facilitate this process. As an example, cognitive-based specialties such as medicine, pediatrics, and psychiatry can grade each patient encounter for history, physical examination, analysis, and treatment plans.^{6,7} Surgical specialties may wish to add operative skills.^{8,9} The large number of patient encounters for most trainees allows for robust data collection in a relatively short time frame, even in small training programs such as ours.

There are several substantial limitations to our study. First, we could not and did not control many relevant aspects of the educational experience. For example, turnover is natural and cannot be controlled; however, the influences of changing faculty members and residents are impossible to gauge. A residency program is not a prospective, randomized double-blinded trial; the data are descriptive and retrospective. The relevance of our approach to the practice experience of our trainees is not clear and is a major limitation. This study evaluated only one aspect of surgical pathology, which used predominantly the medical

knowledge and patient care competencies with a lesser utilization of communication and practice-based learning (because every case should teach the residents). Other critical competencies needed by a successful physician, such as interpersonal skills and professionalism, were not evaluated.

Furthermore, our knowledge of outcomes beyond residency is very limited. With rare exceptions, we are unaware of how our former trainees are performing in practice. Other factors critical to successful practice cannot be assessed but may play a role, such as including personal or health issues. As mandated by the ACGME, we are now working to solicit input from the practices and organizations our trainees joined. This will likely be difficult to acquire and interpret; the small number of trainees joining any organization will limit statistical inference even when data can be obtained. Finally, we do not imply that the trainees who failed the AP board examination on the first attempt are not competent. All of the trainees with an initial failure either passed the AP board on a subsequent attempt or are currently engaged in that process. It should be noted that although the individuals with poor outcomes started significantly below their peers, they improved at a more rapid rate, indicating that interventions do succeed, given time and effort. These individuals may continue to improve in the first few years of practice.

Increasing public calls for competency testing requires new, innovative methods to train and to evaluate our trainees. It is in the interest of the trainees themselves, training programs, the medical profession, and the public to evaluate acquisition of competencies on a case-based system rather than through a small number of large testing events. The former quickly identifies potential problems, allowing remediation or for the trainee to switch to another specialty. As a result of this study, we will scrutinize our starting residents more carefully and, if needed, intervene within the first few months of training. We also plan to review the success of our intervention strategies.

References

- 1 Accreditation Council for Graduate Medical Education. Toolbox of assessment methods. Available at: <http://www.acgme.org/outcome/assess/toolbox.asp>. Accessed March 27, 2009.
- 2 Ducatman BS, Ducatman AM. Longitudinal case-based evaluation of diagnostic competency among pathology residents: a statistical approach. *Arch Pathol Lab Med*. 2006;130(2):188–193.
- 3 Larson D. Analysis of variance with just summary statistics as input. *Am Stat*. 1992;46:151–152.
- 4 Stefanidis D, Korndorffer JR Jr, Black FW, et al. Psychomotor testing predicts rate of skill acquisition for proficiency-based laparoscopic skills training. *Surgery*. 2006;140(2):252–262.
- 5 Reed RJ. On the imposition of virtual images and the art of pathology: a philosophy recognizing the arbitrariness of both segmented neoplastic continua and the diagnosis of malignancy. *Semin Diagn Pathol*. 2008;25(3):136–139.
- 6 Davis LE, King MK, Skipper BJ. Education research: assessment of neurology resident clinical competencies in the neurology clinic. *Neurology*. 2009;72(1):e1–3.

- Available at: <http://www.neurology.org/cgi/content/abstract/72/1/e1>; accessed March 26, 2009.
- 7 Zimmer KP, Solomon BS, Siberry GK, Serwint JR. Continuity-structured clinical observations: assessing the multiple-observer evaluation in a pediatric resident continuity clinic. *Pediatrics*. 2008;121(6):e1633–1645. Available at: <http://pediatrics.aappublications.org/cgi/content/full/121/6/e1633>; accessed March 29, 2009.
 - 8 Chou B, Bowen CW, Handa VL. Evaluating the competency of gynecology residents in the operating room: validation of a new assessment tool. *Am J Obstet Gynecol*. 2008;199(5):e571–575. Available at: [http://www.ajog.org/articles/S0002-9378\(08\)0074-X/abstract](http://www.ajog.org/articles/S0002-9378(08)0074-X/abstract); accessed March 29, 2009.
 - 9 Kenton K. How to teach and evaluate learners in the operating room. *Obstet Gynecol Clin North Am*. 2006;33(2):325–332.