Appeals in Gynecologic Cytology Proficiency Testing

Review and Analysis of Data From the 2006 College of American Pathologists
Gynecologic Cytology Proficiency Testing Program

Barbara A. Crothers, DO; Ann T. Moriarty, MD; Lisa A. Farberee, SCT(ASCP); Christine N. Booth, MD; William D. Tench, MD; David C. Wilbur, MD

Context.—In 2006, 9643 participants took the initial College of American Pathologists (CAP) Proficiency Test (PT). Failing participants may appeal results on specific test slides. Appeals are granted if 3 referee pathologists do not unanimously agree on the initial reference diagnosis in a masked review process.

Objectives.—To investigate causes of PT failures, subsequent appeals, and appeal successes in 2006.

Design.—Appeals were examined, including patient demographic information, Centers for Medicare and Medicaid Services category (A, B, C, or D), exact reference diagnosis, examinee's Centers for Medicare and Medicaid Services category, referee's Centers for Medicare and Medicaid Services category, slide preparation type, and slide field validation rate.

Results.—There was a 94% passing rate for 2006. One hundred fifty-five examinees (1.6%) appealed 86 slides of all preparation types. Forty-five appeals (29%) were granted on 21 slides; 110 appeals (72%) were denied on 65 slides. Reference category D and B slides were most often appealed. The highest percentage of granted appeals occurred in category D (35% slides; 42% of participants) and the lowest occurred in category B (9% slides; 8% of participants). The field validation rate of all appealed slides was greater than 90%.

Conclusions.—Despite rigorous field validation of slides, 6% of participants failed. Thirty percent of failing participants appealed; most appeals involved misinterpretation of category D as category B. Referees were never unanimous in their agreement with the participant. The participants and referees struggled with the reliability and reproducibility of finding rare cells, “overdiagnosis” of benign changes, and assigning the morphologically dynamic biologic changes of squamous intraepithelial lesions to static categories.

(Arch Pathol Lab Med. 2009;133:44–48)
Table 1. Clinical Laboratory Improvement Amendments of 1988 Grading System by Participant Type*

<table>
<thead>
<tr>
<th>Correct Response</th>
<th>Examinee Response</th>
<th>Pathologist (Technical Supervisor) 10-Slide Test</th>
<th>Cytotechnologist 10-Slide Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A: UNSAT</td>
<td>B: NEG</td>
<td>C: LSIL</td>
</tr>
<tr>
<td>A: UNSAT</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B: NEG</td>
<td>5</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>C: LSIL</td>
<td>5</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>D: HSIL</td>
<td>0</td>
<td>-5</td>
<td>5</td>
</tr>
</tbody>
</table>

* UNSAT indicates unsatisfactory; NEG, negative; LSIL, low-grade squamous intraepithelial lesion; and HSIL, high-grade squamous intraepithelial lesion.

RESULTS

A total of 9643 individuals participated in the initial PAP PT. In the present analysis, 574 participants were from Veterans Administration and Department of Defense sites, which are not reported to CMS. The passing rate (including the Department of Defense and Veterans Administration populations) of the initial testing cycle can be seen in Table 2. Of the 589 failures, 155 examinees (26% of failures and 1.6% of total participants) appealed the CMS reference category on 86 slides. Forty-five appeals (29%) were granted, and 43 of those participants received passing scores based upon review of 21 appealed slides. The 2 failures after successful slide appeals were the result of other denied slide appeals. A total of 110 appeals were denied based upon review of 65 slides. Thirty-eight slides had multiple appeals; individuals from the same laboratory appealed their interpretation, and in all of these appeals, the examinees’ answers were the same.

MATERIALS AND METHODS

All appeals to the CAP PAP PT program in 2006 were examined. Data collected on each appealed slide consisted of CMS category (A, B, C, D), exact CAP reference diagnosis (eg, unsatisfactory, negative for intraepithelial lesion or malignancy [NILM]; NILM-Candida, LSIL, HSIL, squamous carcinoma, adenocarcinoma), number of examinees per appeal, examinee’s CMS category, referee’s CMS category, whether appeal was denied or granted, slide preparation type (conventional, ThinPrep, SurePath), and field validation rate (FVR). All slide characteristics are maintained in the CAP SCORES database. Each slide has a history of performance maintained during its circulating lifetime. All slides that validate in the field will have an FVR, which is the response rate of participants in the field compared with the reference category.

This study details the appeals process and examines the results of all appealed cases from the 2006 PAP PT program. Analysis of the appeals may provide insight into any trends occurring in appeals cases and may indicate flaws in program design or in certain reference (target) diagnostic categories.
nosis and the status of slides appealed from category B can be seen in Table 4. All appealed category B slides that were denied had an FVR of greater than 95%. Of cases appealed in category B, 36 (61%) were interpreted as category C, and 20 (34%) were interpreted as category D. Two participants considered the slides “unsatisfactory.” Changes consistent with Candida species were most often confused with LSIL, whereas in the other categories, both LSIL and HSIL were confused with the target diagnosis. Not unexpectedly, repair was confused with HSIL or above (HSIL+).

Category C (LSIL) slide appeals were granted 37% of the time; 8 participants (5 slides) answered category D (HSIL), and 3 participants (3 slides) answered category B (NILM). The 3 appeals granted (1 slide called NILM and 2 called HSIL) were granted because a single referee in each case interpreted the slide as HSIL (category D). The average FVR for category C was 96%, indicating high reproducibility for these slides.

Category D appeals included a variety of reference diagnoses (Table 5). The reference interpretation on 24 slides was HSIL; 6 slides were squamous cell carcinoma, 9 slides were adenocarcinoma, and 1 slide was HSIL/carcinoma. Seventy-five (94%) of the participants appealed category D slides based upon their interpretation of the slide as category B (NILM), which resulted in an automatic failure; 4 participants (5%) appealed an answer of category A (unsatisfactory). Only 1 case, whose target diagnosis was squamous cell carcinoma, was interpreted as category C (LSIL). Only 1 referee disagreed with the target diagnosis in 10 of 40 slides appealed in category D. Two referees disagreed with the target diagnosis in 3 of the 40 slides appealed, and in all 3 cases, the referees concurred with the participant’s interpretation of category B (NILM).

All slide preparation types were included in the appeals process. There was no difference in the field validation rate between the preparation types (Table 6). Bentz et al. have previously demonstrated that there was no statistically significant difference in failure rates for slide set modules used in the 2006 PAP PT. There are more ThinPrep slides in the PAP PT program (64%) than either SurePath (17%) or conventional preparations (19%). More appeals were granted for SurePath slide appeals (42%) than for ThinPrep slides (21%) or conventional slides (8%). The appeals granted on SurePath slides were divided between categories B (2 slides), C (3 slides), and D (3 slides). The appeals granted on ThinPrep slides were primarily in category D (10 slides), with 1 appeal each granted from category A and category B. There were no appeals granted on ThinPrep slides from category C. The only appeal granted on a conventional slide was from category D. There was no statistically significant difference ($P = .11$) in the granted appeals rate by slide preparation type using Fisher exact test at a .05 significance level. Despite the appearance that appealed ThinPrep slides in category D are granted more frequently than other preparations appealed in the same category, there is no significant difference between the granted appeals ($P = .69$) using the same statistical method.

**COMMENT**

Appeals to the 2006 CAP PT program illustrate the difficulties in designing and administering proficiency testing based upon a 10-slide test that lacks the usual team approach of the pathologist and the cytotechnologist. Despite rigorous field validation (beyond the requirement of the CLIA PT testing mandates), 30% of failing participants requested a review of their test slide because of arguable differences between their interpretation and the reference diagnosis. Thirty-eight slides were appealed by more than 1 participant, which most likely represents a laboratory bias (individuals in the same laboratory interpret the slide in a similar fashion). All slide preparation types were represented in the appeals process in proportions representative of their relative frequencies in the testing program.

All category A (unsatisfactory) appeals were liquid-based specimens. Although there are well-established criteria for the definition of unsatisfactory specimens in liquid-based preparations, some individuals apparently are not applying currently accepted adequacy criteria for liquid-based preparations.

*FVR indicates field validation rate; LSIL, low-grade squamous intraepithelial lesion; and HSIL+, high-grade squamous intraepithelial lesion or carcinoma.*
requiring triage to colposcopy. In actual practice, some of these cases would likely be interpreted as “atypical squamous cells—undetermined significance” (ASC-US) and be referred for high-risk human papillomavirus DNA testing. In order to prevent a false-negative interpretation, expectation bias in the testing environment may have caused participants to err on the safe side of overinterpretation. Most of the appeals in this category were denied, reflecting excellent reproducibility of the benign category.

The difficulty in category C (LSIL) interpretation demonstrates the morphologic and biologic continuum seen in intraepithelial lesions. There is not a definite morphologic dividing line between LSIL and HSIL, and the initial management guidelines for patients having either of these lesions is identical, except in women of special circumstances. All appeals granted in this group were based on 1 of 3 referees identifying morphologic changes consistent with category D (HSIL+) on the appealed slides. This reflects the difficulty of applying a binary, static diagnosis to a dynamic, evolving process. A previous study from the Cytopathology Resource Committee has shown that a subset of slides in the CAP PAP educational program with a biopsy diagnosis of LSIL do not perform well due to the presence of some HSIL cells. In the PT environment, individuals are unable to assign cases to an intermediate category of “SIL of undetermined significance.” When some cells suggest a higher-grade lesion, participants may choose to err on the side of the higher degree of abnormality present to prevent the penalty of missing a high-grade lesion. In recognition of the potential morphologic overlap between LSIL and HSIL, many practices have adopted the term low-grade squamous intraepithelial lesion, cannot exclude high-grade squamous intraepithelial lesion (LSIL-H) when LSIL slides show features suggestive of HSIL. Studies have shown that these women have a higher incidence of HSIL on subsequent cervical biopsies than do women with an interpretation of LSIL alone.

The most commonly appealed slides were from reference category D. Most category D appeals (94%) were associated with a participant indicating an answer of category B (NILM), and most of these granted appeal slides were ThinPrep slides. The apparent higher rate of granted category D appeals for ThinPrep slides as opposed to other preparations is likely due to the higher percentage of ThinPrep slides in the program. A Fisher test for low-frequency values showed no statistical significance between preparation types in category D appeals ($P = .69$). The single conventional slide appeal granted was also from category D. This finding may reflect the increased difficulty of locating single HSIL cells in these preparations. Further investigation is required to determine whether these slides have innate qualities predisposing observers to error, such as very bland or rare HSIL cells. When paucicellular, or composed of hyperchromatic crowded groups, it has been documented that slides with a reference interpretation of HSIL are not reliably and reproducibly interpreted. Appeals were usually granted in category D because the referees agreed with the category B (NILM) interpretation or deemed the slides technically unsatisfactory. Only 1 referee categorized a slide as category C (LSIL) in a granted appeal. In categories other than D, the appeals were usually granted as the result of a single referee disagreement with the reference category. In category D, there was more divergence of opinion by the referees, although in no category did the referees agree unanimously with the opinion of the participant appealing the slide. For this category, only 1 slide appeal was granted because of technical issues.

Although the stain quality of slides in the program deteriorates over time, most slides of poor quality fail to make it to the PT program or are retired upon discovery. Overall, the quality of the slides did not appear to adversely influence participant interpretation.

The appeals in the PAP PT program reflect the difficulties of constructing a fair and reproducible PT program. Both participants and referees struggled to reliably and reproducibly apply static categories to dynamic biologic processes. Problem areas include application of criteria for satisfactory liquid-based specimens, finding rare cells in HSIL+ cases, overinterpreting benign changes as intraepithelial lesions, and identifying the low-high dividing line between the dynamic morphologic spectrum of SIL. Proficiency testing also eliminates categories of uncertainty (ASC-US, ASC-H [ASC, cannot exclude HSIL], LSIL-H) that are applied clinically, thus forcing artificial choices. In addition, the usual consensus approach to medical decision making is not practiced in this test setting.

Analysis of future PAP PT appeals data will be necessary to determine whether these findings remain consistent. Are the findings associated with inherent test-taking variability of both participants and slides, or will examiners’ test-taking abilities improve over time (pseudoproficiency)? Alternatively, with exclusion of successfully appealed slides from the PT sets, the challenges may become more straightforward, a process that may improve examiners’ outcomes without improving individual proficiency or patient care. Interpretation of Pap tests is subjective, and the success of cervical screening depends in part upon the locator and interpretive skills of the individual.

### Table 5. Exact Match and Reference Diagnosis of Category D

<table>
<thead>
<tr>
<th>Reference Diagnosis</th>
<th>Slides Appealed, No.</th>
<th>Participants, No.</th>
<th>Slide Appeals Granted, No.</th>
<th>Participants Granted Appeals, No.</th>
<th>Average FVR, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSIL</td>
<td>24</td>
<td>47</td>
<td>10</td>
<td>25</td>
<td>95</td>
</tr>
<tr>
<td>Squamous cell carcinoma</td>
<td>6</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>95</td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>9</td>
<td>21</td>
<td>2</td>
<td>7</td>
<td>95</td>
</tr>
<tr>
<td>HSIL/carcinoma</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>95</td>
</tr>
</tbody>
</table>

* FVR indicates field validation rate; HSIL, high-grade squamous intraepithelial lesion.

### Table 6. Slide Type of Appealed Challenges*

<table>
<thead>
<tr>
<th>Slide Type</th>
<th>Slides Appealed, No.</th>
<th>Appeals Granted, No.</th>
<th>Average FVR, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional</td>
<td>12</td>
<td>1</td>
<td>95</td>
</tr>
<tr>
<td>ThinPrep</td>
<td>55</td>
<td>12</td>
<td>95</td>
</tr>
<tr>
<td>SurePath</td>
<td>19</td>
<td>8</td>
<td>95</td>
</tr>
</tbody>
</table>

* FVR indicates field validation rate.
Success in identification of an abnormal slide is also dependent upon the type of abnormality, the number of cells demonstrating abnormality, and the presence of accompanying cellular changes. In this setting, designing a reproducible test using biologic material that is able to reliably measure individual proficiency is nearly impossible. Pap PT uses a field validation process that is more rigorous than CMS validation requirements. Field validation requires consensus of opinion on a slide from a majority of practitioners reviewing that slide, not just consensus from 3 expert cytologists. Although Pap PT field validation rates of appealed slides average 95%, reproducibility in a test setting remains problematic. More often, the slides themselves are being tested, not the locator or interpretive skills of the individual. The current government-mandated PT is based upon knowledge of cervical disease clinical management, and laboratory practices that are more than 17 years old. Our understanding of the biologic continuum of cervical carcinogenesis, current patient management, and the operating characteristics of the Pap test (sensitivity, specificity, negative and positive predictive values, and the inherent variability of interpretation) have all increased substantially during the long hiatus between the initial conceptualization and the final implementation of PT.

Medical care is a team effort. No greater collaboration is necessary than for the interpretation of Pap tests, which are, by their nature, subjective. During evaluation of Pap tests, cytotechnologists and pathologists routinely consult with one another, contact providers to gain additional information, review previous patient material, and may review medical records. Proficiency testing does not allow for this collaboration. Historically the CAP has encouraged laboratory participation in an educational program that measures the performance of a practice group, which more closely simulates the team approach to the final interpretation of the specimen. Participation in laboratory-based educational interlaboratory comparison programs has been one requirement for successful accreditation in the CAP Laboratory Accreditation Program.

In our opinion, the CLIA ‘88–prescribed PT is a flawed model that does not accurately reflect the actual practice of cytopathology. The regulations, as approved, are punitive to individuals, impose upon the states’ right to confer licenses to practice medicine, and are not scientifically proven to assess an individual’s proficiency in gynecologic cytology. Furthermore, there is no evidence that they improve patient outcomes. Despite these failings, the CAP has designed a proficiency program that includes only superior-quality slides that are highly field validated with a low standard of error. As a result, very few of the slides are challenged, and even fewer challenges are granted. A recent CAP summary of the 2006 Cytology Proficiency Testing Program indicated that 99.6% of participants enrolled achieved passing scores using the current system. While the quality of the CAP program slides may help to ensure a successful passing score for most participants, it does nothing to educate them on the subtleties of actual practice or to encourage them to seek assistance from within their practice on difficult cases. The current CMS protocol for Pap proficiency testing requires confidentiality of results and disallows discussion of the case complexities either before or after result submission. An individual can pass a test but is unable to review missed slides. One of the failings of this model is that it is not educational; it lacks the potential to produce improvement in performance. In the situation where multiple appellants contested the same slide, the opportunity to correct systemic laboratory bias was lost. Individuals may even pass the test but continue to repeat the same screening or interpretive error. In clinical laboratory models of proficiency testing, the laboratory director can review the results and detect systematic trends for further investigation. A similar model implemented in cytopathology would allow for the detection of diagnostic discrepancies. The additional opportunity for group review of slides imparts the ability to correct the deficiency. It is this practice that enriches and promotes medical efforts to improve diagnostic concordance and quality among individuals, even those who pass the test. Eliminating a few individuals who fail a 10- or 20-slide test only serves to further decrease the declining numbers of individuals willing to evaluate Pap tests. Future efforts to improve the current CMS regulations should address the existing shortcomings and provide an educational environment conducive to performance improvement if the nation is to truly prepare pathologists and cytotechnologists to recognize challenging Pap test cases.

References

1. Medicare, Medicaid and CLIA programs; extension of certain effective dates for clinical laboratory requirements and personnel requirements for cytopathology, HCFA. Final rule with comment period. Fed Regist. 1994;59:62606–62609.