Performance Validation in Anatomic Pathology

Successful Integration of a New Classification System Into the Practice Setting Using the Updated Lung Non–Small Cell Carcinoma Recommendations

Paari Murugan, MD; Michael E. Stevenson, MD; Lewis A. Hassell, MD

Context.—The new, international, multidisciplinary classification of lung adenocarcinoma, from the International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society, presents a paradigm shift for diagnostic pathologists.

Objective.—To validate our ability to apply the recommendations in reporting on non–small cell lung cancer cases.

Design.—A test based on the new non–small cell lung cancer classification was administered to 16 pathology faculty members, senior residents, and fellows before and after major educational interventions, which included circulation of articles, electronic presentations, and live presentations by a well-known lung pathologist. Surgical and cytologic (including cell-block material) reports of lung malignancies for representative periods before and after the educational interventions were reviewed for compliance with the new guidelines. Cases were scored on a 3-point scale, with 1 indicating incorrect terminology and/or highly inappropriate stain use, 2 indicating correct diagnostic terminology with suboptimal stain use, and 3 indicating appropriate diagnosis and stain use. The actual error type was also evaluated.

Results.—The average score on initial testing was 55%, increasing to 88% following the educational interventions (60% improvement). Of the 54 reports evaluated before intervention, participants scored 3 out of 3 points on 15 cases (28%), 2 of 3 on 31 cases (57%), and 1 of 3 on 8 cases (15%). Incorrect use of stains was noted in 23 of 54 cases (43%), incorrect terminology in 15 of 54 cases (28%), and inappropriate use of tissue, precluding possible molecular testing, in 4 out of 54 cases (7%). Of the 55 cases after intervention, participants scored 3 out of 3 points on 15 of 55 cases (28%), 2 of 3 on 31 cases (57%), and 1 of 3 on 1 case (2%). Incorrect use of stains was identified in 9 of 55 cases (16% of total reports), and inappropriate use of tissue, precluding possible molecular testing, was found in 1 of the 55 cases (2%).

Conclusions.—The study results demonstrated marked improvement in the pathologists’ understanding and application of the new non–small cell lung cancer classification recommendations, which was sufficient to validate our use of the system in routine practice. The results also affirm the value of intensive education on, and validation of, pathologists’ use of a classification or diagnostic algorithm.

classification, we felt it was important to accelerate that adoption process and to verify our ability to perform acceptably. This thinking represented a novel shift toward documented reproducibility and quality, which could effectively provide a model for future process changes.

We present here a process scenario we used successfully to document both thorough understanding and appropriate use of the new IASLC/ATS/ERS lung adenocarcinoma classification.

METHODS

Understanding Assessment

Shortly following the initial publication of the IASLC/ATS/ERS classification of lung adenocarcinoma and the presentation at national meetings in October 2011, we decided our department should adopt the classification guidelines into routine use when reporting and managing patients with lung adenocarcinoma. We designed an assessment tool, based on the classification and presentations, which was administered to all willing members of the department. Nine senior pathology faculty and 7 senior residents completed the initial survey. The questionnaire consisted of 26 multiple choice and true/false, graphic- and text-based questions that representatively sampled the recommended modifications and were subcategorized into (1) diagnostic terminology, (2) diagnostic criteria, (3) ancillary testing (special stains and molecular studies), and (4) clinical correlation.

Answers to the survey questions were not distributed to the participants on completion of the initial assessment. All surveys were scored by one of us (PM), and the results were tabulated and classified according to the area of knowledge or issue being addressed by the questions. Results of these classifications were distributed to the participants. Individuals who did not obtain an overall score of greater than 80% were asked to review any interim cases of lung adenocarcinoma they encountered with a faculty member who acceptably passed the initial survey. The former individual then participated in one or more of several learning options directed at increasing understanding and correct usage of the IASLC/ATS/ERS classification of lung adenocarcinoma. Those options included personal study of journal articles, review of didactic presentation materials from national meetings, and attendance at a live seminar featuring a national expert on lung cancer talking about use of the new IASLC/ATS/ERS classification. Within 3 weeks of the live presentation and 5 to 6 weeks after the initial assessment, the survey instrument was readministered, scored, and tabulated. Successful passage was communicated to the participants, and the review proviso was lifted.

Validation of Practice

Surgical pathology reports of non–small cell lung cancer (NSCLC) specimens from 2 periods were reviewed for compliance with the terminology and workup process of the IASLC/ATS/ERS classification of lung adenocarcinoma. Electronic search of case files returned 54 cases of NSCLC evaluated between October 1, 2011, and March 31, 2012 (preintervention), and 55 cases of NSCLC evaluated between June 1, 2012, and November 30, 2012 (postintervention). Those cases included 55 surgical biopsies or resections and 54 cytologic samples.

Review of the accompanying reports was performed to evaluate various quality markers and compliance with the new classification guidelines and terminology. Cases were scored on a 3-point scale, with 1 indicating incorrect terminology and/or highly inappropriate immunohistochemical and/or histochemical stain use, 2 indicating correct diagnostic terminology with suboptimal stain use, and 3 indicating appropriate diagnosis and stain use. Error type was also evaluated and tabulated.

Comparison of the groups was done using an unpaired Student t test.

RESULTS

Test Results

Participants were evaluated using a comprehensive questionnaire based on the 2011 IASLC/ATS/ERS classification. The test scores were evaluated for each of the following categories: (1) diagnostic terminology, (2) diagnostic criteria, (3) ancillary testing (special stains and molecular studies), and (4) clinical correlation. Of the 16 participants in the initial survey, a passing score of greater than 80% was achieved by 3 participants (19%). Following educational intervention, the test was readministered, excluding the 3 participants who had passing scores and 2 others who withdrew from the study. Average percentages were calculated, and the results were compared after excluding the scores of the 5 participants mentioned above. The overall average score was 55% in the initial survey. In general, the best performance category in the first survey (excluding the 5 participants) was diagnostic terminology (62%), and the poorest category was ancillary testing (special stains and molecular studies) (48%). The overall average score during the second survey increased to 88% (a 60% improvement). The performance improvement was most significant in the ancillary testing category (special stains, 88%; molecular studies, 79%; average, 84%). Passing scores were achieved by 10 of the 11 remaining participants (91%). We compared pretest and posttest scores using a paired Student t test. The 11 participants who retook the test after the educational intervention all (100%) scored higher on the second survey. The improvement in scores for these 11 participants was statistically significant ($P < .001$) (Figure).

Prior experience or attendance at educational events on the topic was not captured initially, although anecdotal reports following the initial survey indicated that about one-half of the respondents were familiar with the new classification and its implications.

Reporting Compliance Results

The 54 cases reviewed from the preintervention period (October 1, 2011, to March 31, 2012) scored 3 of 3 points in 15 instances (28%), 2 of 3 in 31 instances (57%) and 1 of 3 in 8 instances (15%). The 55 cases reviewed from the postintervention period (June 1, 2012, to November 30, 2012) scored 3 of 3 points in 46 instances (84%, a 200% increase in the number of cases where terminology and stain use were correct), 2 of 3 in 8 instances (15%, a 74% decrease), and 1 of 3 points in 1 instance (2%, a 87% decrease). The average score for the pathology reports was 2.13, and the average score for the postintervention period was 2.82 (a 32% improvement in compliance). An unpaired Student t test shows that to be a significant difference ($P < .001$) (Table 1). Twelve different pathologists reported the cases included in the preintervention group, whereas 8 different pathologists reported the cases in the postintervention group.

Incorrect use of stains was the predominant error category in preintervention cases (23 of 54 cases; 43%), followed by incorrect diagnostic terminology (15 of 54 cases; 28%) and molecular diagnostic issues (4 of 54 cases; 7%). Incorrect use of stains was identified in 9 of 55 postintervention cases (16%, reflecting a 63% decrease in the number of incorrect cases), and molecular diagnostic issue errors was noted in 1 of 55 postintervention cases (2%, a 71% decrease). No diagnostic terminology errors (0%) were uncovered after intervention (Table 2). The number of errors overall dropped...
from 42 errors in 54 cases (78%) before intervention to 10 errors in 55 cases (2%) after intervention (a 97% decrease in the number of errors).

Molecular testing was ordered on 7 out of 43 cases (16%) of adenocarcinoma or NSCLC, not otherwise specified, before intervention, and was ordered on 9 out of 39 cases (23%) after intervention.

Comments from participants in the survey were also pertinent to our study. One person withdrew following the initial survey possibly because of the potential implications of a perceived restriction on their practice. Another person complained to the chair of the hospital’s credentials committee. Others praised the effort as being beneficial to them in revealing gaps in their knowledge of the new classification and its application. One colleague noted that having taken the initial survey, their attention to, and understanding of, the live presentation was enhanced, resulting in an improved learning experience.

Coincident with the intervention phase of our study, the department changed sign-out procedures from a generalist to sub-specialist team model for surgical cases, including lung samples. (Cytology specialty sign-out did not change.) We, therefore, looked at the distribution of results from report scoring among 2 different groups, those achieving a pretest passing score greater than 80% and those who did not, to determine whether this more-selective sign-out cohort could alone account for the differences in results observed. As noted above, one member of the lung specialty team was among the former, and the other participants were not. After excluding that person’s reports from the groups, the percentage of cases scoring 3 of 3 points increased from 29% (14 of 49) preintervention to 86% (25 of 29) after intervention, and the number of cases scored 2 of 3 points decreased from 28 of 49 (57%) to 3 of 29 (10%). Similarly, cases scoring 1 of 3 points decreased from 7 of 49 (14%) to 1 of 29 (3%). These values were statistically significant at \( P < .001 \).

**COMMENT**

Documentation of quality in the diagnostic efforts of individual pathologists is traditionally retrospective or, at best, limited by the few cases that can be assessed in real-time, before sign-out reviews or is limited in initial focus to practice-evaluation exercises. Departmental or group performance may be measured by various retrospective means, such as amended report rates, or by monitoring extramural-review discrepancies. Rarely are overall efforts of a diagnostic service assessed in a prospective, confirmatory manner. This study demonstrates a model that can be used to assess both learning effectiveness and performance quality of the overall service in a particular discipline.

In part, this model was based on viewing the pathologist or the pathology service encountering a new classification or practice guideline as one might view a new instrument-reagent combination in the clinical laboratory. Because the new classification for lung adenocarcinoma was not trivial and represented important new data on outcomes, new or revised terminology, and significant shifts in application of testing modalities, we felt it offered an ideal setting to test our process. In doing so, we hoped to provide greater assurance to our clinical colleagues and ourselves that we could successfully apply the new classification in a repeatable and consistent manner.

A host of authors have contributed to our understanding of diagnostic and other errors in anatomic pathology, but none, to our knowledge, have addressed those errors in the context of the shifting knowledge base that informs diagnostic criteria and classification systems. In addition, most metrics used to attempt to demonstrate quality, such

<table>
<thead>
<tr>
<th>Case Score, ( n = 3 )</th>
<th>Cases Before, ( n = 54 ), No. (%)</th>
<th>Cases After, ( n = 55 ), No. (%)</th>
<th>% Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>15 (28)</td>
<td>46 (84)</td>
<td>200</td>
</tr>
<tr>
<td>2</td>
<td>31 (57)</td>
<td>8 (15)</td>
<td>-74</td>
</tr>
<tr>
<td>1</td>
<td>8 (15)</td>
<td>1 (2)</td>
<td>-87</td>
</tr>
</tbody>
</table>

**Table 1. Case Scoring Before and After Intervention**

Survey questions answered correctly by concept tested and before and after intervention.
as amended report rates, are entirely retrospective or, at best, a partial, concurrent sample. A review of classification in pathology recognized this activity as a fundamental skill for diagnostic pathologists and emphasized the dynamic nature of that process, both because of the changes in understanding and because of the progress in technology or tools available. The author also identified precision and ease of use as critical factors in classification methods and, in reviewing the performance of various systems then in use, commented on reproducibility, liability, and improvement in patient safety. Our study adds a prospective method for evaluating the fidelity of the classification process, particularly in the adoption phase.

We acknowledge some inherent and some avoidable weaknesses in our study. We did not begin with completely unbiased, naïve participants in their knowledge of the IASLC/ATS/ERS classification system publications. Instead, the knowledge base was uneven. We even considered abandoning the effort after one pathologist in the intended study group took time in a consensus conference setting to discuss the points of the classification changes when most department members were present. Our testing instrument was not independently validated, although it was developed with careful attention to critical points in the existing publications. Thus, our choice of an 80% cutoff for successful passage was arbitrary but comparable to the cutoffs used for most Maintenance of Certification Self-Assessment Module programs. However, the improvement in scoring by every person who took the subsequent test would suggest that the instrument represented the educational content well.

Our choice to use the same assessment tool might also be considered a weakness in design. Recollection of the questions might still be a possibility, despite the 4 to 8 weeks between assessments. Selective recognition and attention to the question issues might have been given when educational options were studied or meetings attended, thus augmenting the relative improvement we demonstrated. Because we also included an audit of the actual cases in the study, which demonstrated a consonant improvement in performance, we do not feel the repeated use of the assessment tool was a significant weakness. In addition, because the participants were unaware of the intent to reuse the assessment, their attention to gaps in their knowledge base exposed during the initial assessment would have only been a desired educational outcome, rather than an attempt to game the system. The selection of cases used in the preintervention and posttest groups might also have been a source of bias in our study, depending on the distribution among responsible pathologists. The number of pathologists signing out cases in the postintervention group was smaller, but included only one pathologist who successfully passed the evaluation on the first attempt and who did not account for a disproportionate share of the cases included in that group.

The shift to a specialist sign-out model of care was an unplanned variable in our study and may have introduced selection bias by reducing the number of individuals reporting lung samples. However, our subgroup analysis suggests that the change did not account for the improvements observed in reporting practices. Although specialist sign-out practices are generally thought to improve quality and to reduce educational inefficiencies, such as might be associated with matters like the adoption of new classifications or staining algorithms, our data suggest that an efficient transformation of reporting practice is not dependent on such a setting alone.

The survey instrument represents a novel design for evaluating participants’ understanding and application of the IASLC/ATS/ERS classification of lung adenocarcinomas. Given the lack of previous, similar investigations, the validity of this type of evaluation method in adequately assessing proper integration of new classifications or recommendations into routine practice can be questioned. The ability of the survey instrument design to accurately quantify participants’ conceptual understanding of the new classification before and after educational intervention, as well as the lack of standardization of educational intervention options, may also lessen the effect of the study results. Nevertheless, the substantial improvement in correct answers on the survey instrument before and after educational intervention does lend credibility to the conclusion that participants did improve their understanding of the new classification as the study progressed. Participants were also able to correctly employ the new classification in everyday practice, as shown by a substantial decrease in application errors and by improvement in case scoring during the study period.

We experienced one negative response regarding the experience from a department member who withdrew from the study. In part, that response was due to a cultural misunderstanding of the nature of the Focused and Ongoing Professional Practice Evaluation processes (of which this study was not explicitly a part) and perhaps, in part, due to the perceived risk of a change in privileges resulting from continuing participation. The age and experience differential between that individual and the pathologists who successfully passed the initial assessment was also a barrier. Ironically, a subsequent shift to a specialty sign-out model in our department has meant that the individual no longer deals regularly with lung cases, so we were not forced to impose a restriction for unsuccessful completion of the activity.

Our validation process has some potential implications for individuals and groups proposing new classification and management schemes in the future, as well as offering an approach for the day-to-day users who need to follow those changes. It would have been advantageous for Travis et al, as authors of the study, to develop an assessment and validation tool that could have been widely and easily applied by individuals in practice. Such evaluation tools are

<table>
<thead>
<tr>
<th>Error Categories</th>
<th>Cases Before Intervention, n = 54, No. (%)</th>
<th>Cases After Intervention, n = 55, No. (%)</th>
<th>% Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect use of immunostains</td>
<td>23 (43)</td>
<td>9 (16)</td>
<td>-63</td>
</tr>
<tr>
<td>Incorrect diagnostic terminology</td>
<td>15 (28)</td>
<td>0 (0)</td>
<td>-100</td>
</tr>
<tr>
<td>Inappropriate tissue use for molecular testing</td>
<td>4 (7)</td>
<td>1 (2)</td>
<td>-71</td>
</tr>
</tbody>
</table>

Table 2. Error Categories Before and After Intervention
commonly applied in Maintenance of Certification programs requiring self-assessment modules, which often also require a passing score of 80%. Consensus development efforts, such as the Bethesda conferences on cervical and thyroid cytology have resulted in highly useful image atlases to enhance education in the new classification categories. Offering an image-based assessment tool is another useful adjunct for new users to verify their ability to correctly use such classification systems, such as that available with the online Bethesda gynecologic atlas. The comparable online atlas for thyroid cytology does not include any self-assessment tools. Earlier studies demonstrating that review alone of one of these atlases did not improve the reproducibility of diagnosis could have hinged on the absence of a valid self-assessment tool or other missing elements in the education and adoption process we have employed. Systems that include specimen-management elements in the education and adoption process we have employed. Systems that include specimen-management choices, such as the lung adenocarcinoma system studied here, would also need to include such issues in the assessment tool.

As also revealed during this study, the implementation of new or updated classification schemes can be disruptive or disconcerting for practitioners who may have used previously accepted guidelines for a number of years. The pathologists may resist altering their practices because of disagreements with the new classification criteria, because of reluctance to change the way they interpret cases, or because of resistance to learning the new information. Such resistance may result in the continued application of outdated classification guidelines with important patient-care implications. Overcoming such impediments is of prime importance for the generalized proper application of new classification systems and may be an issue of patient safety as well as quality of care.

**SUMMARY**

We have demonstrated that following routine (early), diffusion-based adoption practices our department’s knowledge of the new lung classification system was deficient and that this correlated with poor performance in our reporting. Likewise, our educational interventions combined with the assessments themselves resulted in improved performance, both from a test-score standpoint and from a clinical-reporting standpoint. We feel this model provides a reasonable means of validating department-wide successful adoption and performance in using a new classification system and related guidance.

**References**

4. Cagle P. Paper presented at: 2012 Annual Scientific Meeting of the Oklahoma State Association of Pathologists; May 19, 2012; Oklahoma City, OK.