

Principles of Analytic Validation of Immunohistochemical Assays

Guideline From the College of American Pathologists Pathology and Laboratory Quality Center

Patrick L. Fitzgibbons, MD; Linda A. Bradley, PhD; Lisa A. Fatheree, BS, SCT(ASCP); Randa Alsabeh, MD;
Regan S. Fulton, MD, PhD; Jeffrey D. Goldsmith, MD; Thomas S. Haas, DO; Rouzan G. Karabakhtsian, MD, PhD;
Patti A. Loykasek, HT(ASCP); Monna J. Marolt, MD; Steven S. Shen, MD, PhD; Anthony T. Smith, MLS; Paul E. Swanson, MD

• **Context.**—Laboratories must validate all assays before they can be used to test patient specimens, but currently there are no evidence-based guidelines regarding validation of immunohistochemical assays.

Objective.—To develop recommendations for initial analytic validation and revalidation of immunohistochemical assays.

Design.—The College of American Pathologists Pathology and Laboratory Quality Center convened a panel of pathologists and histotechnologists with expertise in immunohistochemistry to develop validation recommendations. A systematic evidence review was conducted to address key questions. Electronic searches identified 1463 publications, of which 126 met inclusion criteria and were extracted. Individual publications were graded for quality,

and the key question findings for strength of evidence. Recommendations were derived from strength of evidence, open comment feedback, and expert panel consensus.

Results.—Fourteen guideline statements were established to help pathology laboratories comply with validation and revalidation requirements for immunohistochemical assays.

Conclusions.—Laboratories must document successful analytic validation of all immunohistochemical tests before applying to patient specimens. The parameters for cases included in validation sets, including number, expression levels, fixative and processing methods, should take into account intended use and should be sufficient to ensure that the test accurately measures the analyte of interest in specimens tested in that laboratory. Recommendations are also provided for confirming assay performance when there are changes in test methods, reagents, or equipment.

(*Arch Pathol Lab Med.* 2014;138:1432–1443; doi: 10.5858/arpa.2013-0610-CP)

Accepted for publication February 3, 2014.

Published as an Early Online Release March 19, 2014.

Supplemental digital content is available for this article at www.archivesofpathology.org in the November 2014 table of contents.

From the Department of Pathology, St. Jude Medical Center, Fullerton, California (Dr Fitzgibbons); the Department of Pathology and Laboratory Medicine, Women & Infants Hospital/Brown University, Providence, Rhode Island (Dr Bradley); the College of American Pathologists, Northfield, Illinois (Ms Fatheree and Mr Smith); the Department of Pathology, Kaiser Permanente - Los Angeles Medical Center, Los Angeles, California (Dr Alsabeh); PhenoPath Laboratories, Seattle, Washington (Dr Fulton); the Department of Pathology, Beth Israel Deaconess Medical Center, Boston, Massachusetts (Dr Goldsmith); the Department of Pathology, Mercy Hospital, Janesville, Wisconsin (Dr Haas); the Department of Pathology, Montefiore Medical Center, New York, New York (Dr Karabakhtsian); Regional Medical Laboratory, St John's Medical Center, Tulsa, Oklahoma (Ms Loykasek); the Department of Pathology, University of Minnesota Medical Center, Fairview, Minneapolis (Dr Marolt); the Department of Pathology, The Methodist Hospital, Houston, Texas (Dr Shen); and the Department of Pathology, University of Washington Medical Center, Seattle (Dr Swanson).

Authors' disclosures of potential conflicts of interest and author contributions are found in the appendix at the end of this article.

Reprints: Patrick L. Fitzgibbons, MD, Department of Pathology, St. Jude Medical Center, 101 E. Valencia Mesa Dr, Fullerton, CA 92835 (e-mail: Patrick.Fitzgibbons@stjoe.org).

For additional questions and comments, contact the Pathology and Laboratory Quality Center at center@cap.org.

Immunohistochemical (IHC) testing is an essential component of the pathologic evaluation of many specimens and increasingly provides key information that helps determine how patients are treated. As with any test, laboratories must ensure that IHC test results are accurate and reproducible and that the test performs as intended. Laboratories subject to US regulations are required by the Clinical Laboratory Improvement Amendments of 1988 (CLIA) to verify the performance characteristics of any assay used in patient testing before it is placed into clinical service.^{1,2}

Before reporting patient results for unmodified US Food and Drug Administration (FDA)–cleared or FDA–approved tests, laboratories must demonstrate performance characteristics for accuracy, precision, and reportable range of test results that are comparable to those established by the manufacturer. The laboratory medical director must determine the extent to which these performance specifications are verified, based on the method, testing conditions, and personnel performing the test. Manufacturers of FDA–approved or FDA–cleared test kits may provide the user with recommendations and directions for verifying that the

kit is performing according to the manufacturer's specification. Typically, this is performed by testing known positive and negative samples that either are supplied by the manufacturer or have been tested by a validated reference-laboratory method.

Laboratories that introduce non-FDA-approved or non-FDA-cleared tests (laboratory-developed tests) or modify FDA-cleared or FDA-approved test systems (laboratory-modified tests) must, before reporting patient test results, establish performance specifications for accuracy, precision, analytic sensitivity, analytic specificity, reportable range, and reference intervals.¹ For tests that are reported qualitatively or semi-quantitatively (most IHC tests), reportable range and reference intervals are generally not applicable.

Good laboratory practice requires establishing optimal antibody concentration and antigen retrieval and detection methods. Analytic validation follows assay optimization and is done by testing an appropriate tissue set to determine analytic sensitivity and specificity. For tests without a gold standard referent test, this usually involves determining overall concordance with an appropriate comparator. Validation procedures are intended to reasonably assure that the test performs as expected. Once validation has been completed, assays must be regularly monitored to detect changes in analytic performance, usually by daily quality control, periodic proficiency testing, and comparing positivity rates for selected markers (eg, hormone receptors, *HER2/neu*) with expected positivity rates. Ongoing monitoring of assay performance is as important as initial assay validation.

Although IHC test methods have steadily improved with the introduction of automated staining platforms and improved antigen retrieval and detection systems, results are still affected by various preanalytic and analytic factors, and the need for assay validation and ongoing monitoring has not diminished. Assay validation is particularly important when a polymer-based detection system is used and a negative reagent control is omitted. The College of American Pathologists (CAP) Laboratory Accreditation Program (LAP) accepts omission of this control, but only if the assay has been properly validated (LAP checklist ANP.22570).³

Unfortunately, recent studies^{4,5} have found significant interlaboratory variation in validation practices and revealed that many laboratories do not follow consistent procedures when validating IHC assays. Comments received during the open comment period for this guideline also revealed a surprising lack of understanding among some respondents of requirements for analytic validation. To address this important shortfall in laboratory practice, the CAP convened representatives to systematically review the published data and develop evidence-based recommendations for analytic validation of IHC assays.

METHODS

A detailed description of the methods and systematic review (including the 7 key questions, quality assessment, and complete analysis of the evidence) used to create this guideline can be found in the supplemental digital content available at www.archivesofpathology.org in the November 2014 table of contents.

Panel Composition

The CAP Pathology and Laboratory Quality Center (the Center) convened expert and advisory panels consisting of members with expertise in immunohistochemistry. Panel members included

pathologists, histotechnologists, methodologists, and CAP staff. CAP approved the appointment of the project chair (P.L.F.) and panel members.

Conflict of Interest Policy

Before acceptance on the expert or advisory panel, potential members completed the CAP conflict of interest disclosure process, whose policy and form (in effect April 2010) require disclosure of material financial interest in or potential for benefit of significant value from the guideline's development or its recommendations 12 months prior through the time of publication. Potential members completed the conflict of interest disclosure form, listing any relationship that could be interpreted as constituting an actual, potential, or apparent conflict. Everyone was required to disclose conflicts before beginning and continuously throughout the project's timeline. One expert panel member (R.S.F.) was recused from discussion and voting on the recommendation pertaining to tissue microarrays, and one (T.S.H.) was recused from voting on recommendations pertaining to potential increased antibody usage. Expert panel members' disclosed conflicts are listed in the Appendix. The CAP provided funding for the administration of the project; no industry funds were used in the development of the guideline. All panel members volunteered their time and were not compensated for their involvement. Please see the supplemental digital content for full details on the conflict of interest policy.

Objective

The panel addressed the overarching question, "What is needed for initial analytic assay validation before placing any IHC test into clinical service and what are the revalidation requirements?" The scope questions are as follows:

1. When and how should validation assess analytic sensitivity, analytic specificity, accuracy (assay concordance), and precision (interrun and interoperator variability)?
2. What is the minimum number of positive and negative cases that need to be tested to analytically validate an IHC assay for its intended use(s)?
3. What parameters should be specified for the tissues used in the validation set?
4. How do certain preanalytic variables influence analytic validation?
5. What conditions require assay revalidation?

Literature Search and Selection

Electronic searches of the English language-published literature in Ovid MEDLINE, US National Library of Medicine PubMed, and Elsevier Scopus databases were initially conducted for the time period spanning January 2004 to May 2012; an update was conducted through May 2013. In addition to peer-reviewed journal articles, the search identified books, book chapters, and published abstracts from English-language sources. Bibliographies of included articles were hand searched, and additional information was sought through targeted grey literature electronic searches (eg, Google) and review of laboratory compliance and guidance Web sites (eg, Clinical and Laboratory Standards Institute, FDA, National Guideline Clearinghouse, Wiley Cochrane Library).

Inclusion Criteria

Published studies were selected for full-text review if they met each of the following criteria:

1. English-language articles/documents that addressed IHC and provided data or information relevant to 1 or more key questions;
2. Study designs that included validation, method comparison, cohort or case-control studies, clinical trials, and systematic reviews, as well as qualitative information from consensus guidelines, regulatory documents, and US or international proficiency testing reports; and

- Articles/documents focused on the clinical use of IHC for identification of predictive and nonpredictive markers and analytic variables.

Exclusion Criteria

Editorials, letters, commentaries, and invited opinions were not included in the study. Articles were also excluded if the full article was not available in English, did not address any key question, and/or focused primarily on assay optimization, quality control or quality assurance, basic or nonhuman research, nontissue immunoassays, preanalytic and postanalytic variables, or clinical validation only.

Quality Assessment

Grading the quality of individual studies was performed from study design-specific criteria by the methodology consultant (L.A.B.), with input as needed from the expert panel. The aim of analytic validation is to determine a test's ability to accurately and reliably detect the antigen or marker of interest in specimens consistent with those to be tested in clinical practice.⁶ Analytic validity studies have a different design, compared to studies of diagnostic accuracy or therapeutic interventions. For this reason, the criteria needed to assess the quality of analytic validity studies are different. Quality in this context is considered to be essentially equivalent to internal validity and is assessed on the basis of study design and execution, analyses, and reporting.⁶ The strength of evidence for individual key questions or outcomes was assessed by using published criteria.⁶ The criteria included the quality and execution of studies, the quantity of data (number and size of studies), and the consistency and generalizability of the evidence across studies.⁶ Strength of evidence was graded *convincing*, *adequate*, or *inadequate* (Table 1).

Assessing the Strength of Recommendations

Development of recommendations requires that the panel review the identified evidence and make a series of key judgments. Grades for strength of recommendations were developed by the CAP Pathology and Laboratory Quality Center and are described in Table 2.

Guideline Revision

This guideline will be reviewed every 4 years, or earlier in the event of publication of substantive and high-quality evidence that could potentially alter the original guideline recommendations. If necessary, the entire panel will reconvene to discuss potential changes. When appropriate, the panel will recommend revision of the guideline to CAP for review and approval.

Disclaimer

The CAP developed the Pathology and Laboratory Quality Center as a forum to create and maintain evidence-based practice guidelines and consensus statements. Practice guidelines and consensus statements reflect the best available evidence and expert consensus supported in practice. They are intended to assist physicians and patients in clinical decision making and to identify questions and settings for further research. With the rapid flow of scientific information, new evidence may emerge between the time a practice guideline or consensus statement is developed and when it is published or read. Guidelines and statements are not continually updated and may not reflect the most recent evidence. Guidelines and statements address only the topics specifically identified therein and are not applicable to other interventions, diseases, or stages of diseases. Furthermore, guidelines and statements cannot account for individual variation among patients and cannot be considered inclusive of all proper methods of care or exclusive of other treatments. It is the responsibility of the treating physician or other health care provider, relying on independent experience and knowledge, to determine the best course of treatment for the patient. Accordingly, adherence to any practice guideline or consensus statement is voluntary, with the ultimate

Table 1. Grades for Strength of Evidence

Grade	Description
Convincing	Two or more level 1 ^a or level 2 ^b studies (study design and execution) that had an appropriate number and distribution of challenges ^c and reported consistent ^d and generalizable ^e results. One level 1 or level 2 study that had an appropriate number and distribution of challenges and reported generalizable results.
Adequate	Two or more level 1 or level 2 studies that lacked the appropriate number and distribution of challenges OR were consistent but not generalizable.
Inadequate	Combinations of level 1 or level 2 studies that show unexplained inconsistencies OR 1 or more level 3 ^f or level 4 ^g studies OR expert opinion.

From Teutsch et al.⁶ Reprinted with permission from Macmillan Publishers Ltd.

^a Level 1 study: Collaborative study using a large panel of well-characterized samples; summary data from external proficiency-testing schemes or interlaboratory comparisons.

^b Level 2 study: High-quality peer-reviewed studies (eg, method comparisons, validation studies).

^c Based on number of possible response categories and required confidence in results.

^d Consistency assessed by using central estimates/ranges or testing for result homogeneity.

^e Generalizability is the extension of findings and conclusions from 1 study to other settings.

^f Level 3 study: Lower-quality peer-reviewed studies OR expert panel-reviewed US Food and Drug Administration summaries.

^g Level 4 study: Unpublished or non-peer-reviewed data.

determination regarding its application to be made by the physician in light of each patient's individual circumstances and preferences. CAP makes no warranty, express or implied, regarding guidelines and statements and specifically excludes any warranties of merchantability and fitness for a particular use or purpose. CAP assumes no responsibility for any injury or damage to persons or property arising out of or related to any use of this statement or for any errors or omissions.

RESULTS

Of the 1463 studies identified by electronic searches, 126 met inclusion criteria and underwent data extraction. These included 122 published peer-reviewed articles, 2 book chapters, and 2 grey literature documents. Among the extracted documents, 43 did not meet minimum quality standards, presented incomplete data or data that were not in useable formats, or included only information based on expert opinion. These articles were not included in analyses or narrative summaries. The expert panel met 28 times by teleconference Webinar from June 2010 through September 2013 and met in person on May 11 and May 12, 2013, to review evidence to date and draft recommendations. Additional work was completed via electronic mail. An open comment period was held from July 8 through July 29, 2013. Eighteen draft recommendations and 5 methodology questions were posted online on the CAP Web site.

A total of 1071 comments were received from 263 respondents ("agree" and "disagree" responses were also captured). Twelve of 18 draft recommendations achieved more than 80% agreement; only 2 had less than 70% agreement. Each expert panel member was assigned 1 to 2 draft recommendations for which to review all comments

Table 2. Grades for Strength of Recommendations

Designation	Recommendation	Rationale
Strong recommendation	<i>Recommend for or against</i> a particular analytic validation practice (can include must or should).	Strength of evidence is <i>convincing</i> , based on consistent, generalizable, good-quality evidence; further studies are unlikely to change the conclusions.
Recommendation	<i>Recommend for or against</i> a particular analytic validation practice (can include should or may).	Strength of evidence is <i>adequate</i> , based on limitations in the quality of evidence; further studies may change the conclusions.
Expert consensus opinion	<i>Recommend for or against</i> a particular analytic validation practice (can include should or may).	Important validation element to address but strength of evidence is <i>inadequate</i> ; gaps in knowledge may require further studies.

received and provide an overall summary to the rest of the panel. Three draft recommendations were maintained with the original language; 5 were modified with minor changes for clarification and/or further explanation within the manuscript, and 6 were considered extremely discordant with major revisions made accordingly for a total of 14 final recommendations. Resolution of all changes was obtained by majority consensus of the panel. The final recommendations were approved by the expert panel with a formal vote (with specific abstentions from R.S.F. and T.S.H.). The panel considered laboratory redundancy, efficiency, and feasibility throughout the whole process. Formal cost analysis or cost effectiveness was not performed.

An independent review panel, masked to the expert panel and vetted through the conflict of interest process, provided final review of the guideline and recommended it for approval by the CAP. The final recommendations are summarized in Table 3.

Guideline Statements

1: Recommendation.—Laboratories must validate all immunohistochemical tests before placing into clinical service.

Note: Such means include (but are not necessarily limited to):

1. Correlating the new test's results with the morphology and expected results;
2. Comparing the new test's results with the results of prior testing of the same tissues with a validated assay in the same laboratory;
3. Comparing the new test's results with the results of testing the same tissue validation set in another laboratory using a validated assay;
4. Comparing the new test's results with previously validated non-immunohistochemical tests; or
5. Testing previously graded tissue challenges from a formal proficiency testing program (if available) and comparing the results with the graded responses.

The strength of evidence was *adequate* to support when analytic validation should be done and that it should include determination of analytic sensitivity and specificity (or concordance in the absence of a gold standard referent test) and precision (eg, interrater and interoperator) as part of validation. The evidence was *inadequate* (ie, evidence was not available or did not permit a conclusion to be reached) to assess the precision of IHC assays in practice or how validation should be done with regard to the listed approaches, but did show that these approaches have been used. The panel found that analytic validation provides a net benefit for the overall performance and safety of IHC tests

by contributing to the avoidance of potential harms related to analytic false-positive and false-negative test results.

Laboratories are required by CLIA (section 493.1253) to validate the performance characteristics of all assays used in patient testing, in order to ensure that the results are accurate and reproducible.⁷ This includes establishment of the analytic validity of all non-FDA-cleared/approved (or "laboratory-developed") tests.⁷ For qualitative assays such as IHC, validation usually requires comparing a new assay's results with a reference standard and calculating estimates of analytic sensitivity and specificity; however, because there are no gold standard referent tests for most IHC assays, laboratories must use another means of demonstrating that the assay performs as expected.^{8–10} Publications addressing IHC validation include independent comparisons of a new test's results to clinical outcomes, other validated IHC tests (intralaboratory or interlaboratory), or previously characterized tissue validation sets.^{9,11–19} Non-immunohistochemical tests may include in situ hybridization, flow cytometry, and molecular, cytogenetic, or microbiologic studies. Laboratories may use a combination of comparison methods when appropriate.

When correlating the new test's results with expected results, positive and negative tissues pertinent to each intended clinical use must be included in the validation set. Normal tissues (with 100% positive staining expected) cannot comprise the entire validation set for markers primarily used in diagnosing neoplasms, but may be used in conjunction with neoplastic and lesional tissue as appropriate. In some cases a section of tissue may contain both antigen-positive cells and negative internal control cells, and therefore serve as both a positive and negative validation challenge. The laboratory medical director must determine the most appropriate selection of tissues in the validation set, but the validation set must not consist solely of the same tissues used for antibody optimization.

Although not currently available for many markers, excess tissue previously used in a proficiency testing or interlaboratory comparison program could also be used for assay validation. Tissue from previously graded proficiency-testing challenges could be tested and the results compared with the graded responses from the program.

This recommendation applies to all assays in clinical use (including those for pathogen-specific antigens such as cytomegalovirus and *Helicobacter pylori*) irrespective of the regulatory status of the primary antibody (eg, in vitro diagnostic, analyte-specific reagent).

2: Recommendation.—For initial validation of every assay used clinically, with the exception of HER2/*neu*, estrogen receptor (ER), and progesterone receptor (PgR)

Table 3. Guideline Statements and Strength of Recommendations

Guideline Statement	Strength of Recommendation
<p>1. Laboratories must validate all IHC tests before placing into clinical service. <i>Note:</i> Such means include (but are not necessarily limited to): Correlating the new test's results with the morphology and expected results; Comparing the new test's results with the results of prior testing of the same tissues with a validated assay in the same laboratory; Comparing the new test's results with the results of testing the same tissue validation set in another laboratory using a validated assay; Comparing the new test's results with previously validated non-immunohistochemical tests; or Testing previously graded tissue challenges from a formal proficiency testing program (if available) and comparing the results with the graded responses.</p>	Recommendation
<p>2. For initial validation of every assay used clinically, with the exception of HER2/<i>neu</i>, ER, and PgR (for which established validation guidelines already exist), laboratories should achieve at least 90% overall concordance between the new test and the comparator test or expected results. If concordance is less than 90%, laboratories need to investigate the cause of low concordance.</p>	Recommendation
<p>3. For initial analytic validation of nonpredictive factor assays, laboratories should test a minimum of 10 positive and 10 negative tissues. When the laboratory medical director determines that fewer than 20 validation cases are sufficient for a specific marker (eg, rare antigen), the rationale for that decision needs to be documented. <i>Note:</i> The validation set should include high and low expressors for positive cases when appropriate and should span the expected range of clinical results (expression levels) for markers that are reported quantitatively.</p>	Expert consensus opinion
<p>4. For initial analytic validation of all laboratory-developed predictive marker assays (with the exception of HER2/<i>neu</i>, ER, and PgR), laboratories should test a minimum of 20 positive and 20 negative tissues. When the laboratory medical director determines that fewer than 40 validation tissues are sufficient for a specific marker, the rationale for that decision needs to be documented. <i>Note:</i> Positive cases in the validation set should span the expected range of clinical results (expression levels). This recommendation does not apply to any marker for which a separate validation guideline already exists.</p>	Expert consensus opinion
<p>5. For a marker with both predictive and nonpredictive applications, laboratories should validate it as a predictive marker if it is used as such.</p>	Recommendation
<p>6. When possible, laboratories should use validation tissues that have been processed by using the same fixative and processing methods as cases that will be tested clinically.</p>	Recommendation
<p>7. If IHC is regularly done on cytologic specimens that are not processed in the same manner as the tissues used for assay validation (eg, alcohol-fixed cell blocks, air-dried smears, formalin-postfixed specimens), laboratories should test a sufficient number of such cases to ensure that assays consistently achieve expected results. The laboratory medical director is responsible for determining the number of positive and negative cases and the number of predictive and nonpredictive markers to test.</p>	Expert consensus opinion
<p>8. If IHC is regularly done on decalcified tissues, laboratories should test a sufficient number of such tissues to ensure that assays consistently achieve expected results. The laboratory medical director is responsible for determining the number of positive and negative tissues and the number of predictive and nonpredictive markers to test.</p>	Expert consensus opinion
<p>9. Laboratories may use whole sections, TMAs, and/or MTBs in their validation sets as appropriate. Whole sections should be used if TMAs/MTBs are not appropriate for the targeted antigen or if the laboratory medical director cannot confirm that the fixation and processing of TMAs/ MTBs is similar to clinical specimens.</p>	Recommendation
<p>10. When a new reagent lot is placed into clinical service for an existing validated assay, laboratories should confirm the assay's performance with at least 1 known positive case and 1 known negative case.</p>	Expert consensus opinion
<p>11. Laboratories should confirm assay performance with at least 2 known positive and 2 known negative cases when an existing validated assay has changed in any one of the following ways: Antibody dilution; Antibody vendor (same clone); Incubation or retrieval times (same method).</p>	Expert consensus opinion
<p>12. Laboratories should confirm assay performance by testing a sufficient number of cases to ensure that assays consistently achieve expected results when any of the following have changed: Fixative type; Antigen retrieval method (eg, change in pH, different buffer, different heat platform); Antigen detection system; Tissue processing or testing equipment; Environmental conditions of testing (eg, laboratory relocation); Laboratory water supply. The laboratory medical director is responsible for determining how many predictive and nonpredictive markers and how many positive and negative tissues to test.</p>	Expert consensus opinion
<p>13. Laboratories should run a full revalidation (equivalent to initial analytic validation) when the antibody clone is changed for an existing validated assay.</p>	Expert consensus opinion
<p>14. The laboratory must document all validations and verifications in compliance with regulatory and accreditation requirements.</p>	Expert consensus opinion

Abbreviations: ER, estrogen receptor; IHC, immunohistochemistry; MTBs, multitissue blocks; PgR, progesterone receptor; TMAs, tissue microarrays.

(for which established validation guidelines already exist), laboratories should achieve at least 90% overall concordance between the new test and the comparator test or expected results. If concordance is less than 90%, laboratories need to investigate the cause of low concordance.

Strength of evidence was *adequate* to support a 90% (versus 95%) overall concordance benchmark for analytic validation of IHC tests (excepting HER2/*neu*, ER, PgR).

Supporting evidence for this recommendation is obtained from published IHC validation studies, method comparisons, and proficiency testing or interlaboratory comparisons. Examples include the following:

1. Median overall concordance in a 2-year interlaboratory comparison of CD117 IHC and target results was 87.6%.²⁰
2. Median overall concordance in 5 comparisons of different HER2/*neu* IHC tests was 89.0% (range, 74%–92%), with 2 of 5 studies greater than 90% concordant.^{13–16,19}
3. Median overall concordance in 5 comparisons of HER2/*neu* IHC tests to HER2/*neu* in situ hybridization tests was 88.2% (range, 66%–94%), with 2 of 5 comparisons greater than 90% concordant.^{17,20–22}
4. Median overall concordance in 6 comparisons of IHC tests (PTEN [phosphatase and tensin homologue deleted on chromosome 10], ER, PR, HER2/*neu*, MPT64, p16) to alternative referent tests (eg, RNA expression, clinical diagnosis) was 91.4% (range, 74%–99%), with 3 of 6 studies greater than 90% concordant.^{12,17,21–23}

Summary concordance estimates (using a random effects model) provided similar concordance estimates, but heterogeneity was high ($I^2 > 75\%$ in all cases; $P < .001$) and could not be explained by analysis of selected covariates (eg, tissue type, antibody, study quality grade). The number of studies was too small to allow analysis of the many possible covariates.

These data illustrate the challenge of achieving an overall concordance of 95%, even in large studies of IHC tests with guidance recommending stringent protocol standards (ie, HER2/*neu*, ER, PgR).^{10,24–26} Overall concordance of 90% was achieved in nearly half of the above analyzed comparisons, all of which were subject to many sources of variation (eg, sample type; ischemic time; fixation, antigen retrieval, and staining protocols; scoring). Therefore, laboratory validation studies designed to minimize differences in such variables would have a higher probability of meeting a 90% concordance benchmark.

If the overall concordance estimate in an assay validation study is less than 90%, laboratories should calculate positive and negative concordance rates as well as the discordance (using the McNemar test when sample size is appropriate) to help investigate the cause of low concordance. The McNemar test assesses the significance of the difference between the discordant results (false positives and negatives) in a 2×2 contingency table. Refer to the supplemental digital content for more information and link to available resources.

3: Expert Consensus Opinion.—For initial analytic validation of nonpredictive factor assays, laboratories should test a minimum of 10 positive and 10 negative tissues. When the laboratory medical director determines that fewer than 20 validation cases are sufficient for a specific marker (eg,

rare antigen), the rationale for that decision needs to be documented.

Note: The validation set should include high and low expressors for positive cases when appropriate and should span the expected range of clinical results (expression levels) for markers that are reported quantitatively.

Strength of evidence was *inadequate* to support the recommended number of validation samples, but was *adequate* to support distinguishing nonpredictive from predictive IHC tests and using different numbers of validation samples for each.

A key criterion for determining the number of samples needed to validate an IHC assay is the test's intended use: whether it is used alone or as part of a test panel and interpreted only in the context of other morphologic and clinical data (most nonpredictive markers) or as a stand-alone test reported to physicians as independent diagnostic information that may directly determine treatment (most predictive markers and selected pathogen-specific assays, such as viral antigens in transplant patients), for which the risk of an incorrect result must be minimized.^{5,8,27} Some tests can fall into both categories. Other criteria for determining the number of validation samples include the complexity of interpretation (ie, multiple test outcomes and result categories require more samples) and the number and range of control materials available.⁸ For example, an IHC test with 3 or more result categories would require a larger number of samples to ensure validation than one interpreted only as positive or negative.⁸

Validity in laboratory practice must be based on objective observations. The most practical objective guidance for determining the size of a validation set is statistical analysis. Not surprisingly, the more samples that are run in a validation set, the higher the likelihood that the concordance estimate reflects the test's "true" concordance; increasing the number of samples in a validation set increases the confidence that the assay performs as expected. Table 4 illustrates overall concordance estimates with 95% confidence interval (CI) for 10 and 20 sample validation sets with 0 to 2 observed discordant results.

Using a 10-sample validation set, the overall concordance estimate (ie, the level of agreement between 2 tests) reaches the 90% concordance benchmark with only 1 discordant result. This concordance estimate has a 95% CI (the range of values that has a 95% chance of including the "true" concordance) of 57% to 100%. Using a 20-sample validation set, overall concordance meets the 90% benchmark with 2 or fewer discordant results and a 95% CI of 69% to 98%.

Both the "true" concordance and the number of validation samples have an impact on the probability that a test will reach or exceed the overall concordance benchmark of 90%. For example, if the 95% concordance estimate (1 discordant result) in the 20-sample validation set is a "true" representation of the relationship between the 2 tests, the probability of achieving the 90% benchmark would be very high (92%). The probability of achieving the benchmark if the 90% concordance estimate in the 20-sample set is a "true" representation would be 68% (Stat Trek Binomial Calculator, <http://stattrek.com/online-calculator/binomial.aspx>; accessed November 7, 2013).²⁸

With this in mind, the panel determined that use of 10 samples (5 negative and 5 positive) in a validation set for a nonpredictive marker assay provides unacceptably broad CIs with either 100% (CI, 68%–100%) or 90% (CI, 57%–100%) concordance estimates. For predictive markers,

Table 4. Validation Using 10- and 20-Tissue Validation Sets Against a 90% Concordance Benchmark

No.	Concordance Estimate, % (95% CI)		
	0 Discordant	1 Discordant	2 Discordant
10	100 (68–100)	90 (57–100)	80 (48–95)
20	100 (81–100)	95 (75–100)	90 (69–98)

Abbreviations: CI, confidence interval; No., number of validation tissues.

however, the critical relationship between the antibody/testing method and the *actual* presence of the target analyte for purposes of guiding specific therapeutic intervention or predicting treatment response requires an even higher level of confidence (see recommendation No. 4).

Although analytic assay validation principles are independent of the frequency of testing or the availability of appropriate validation samples, the panel recognized that it may be difficult for some laboratories to obtain the recommended minimum number of positive validation specimens for rare antigens. Working with other laboratories to pool positive cases or using validation sets prepared by other laboratories may allow laboratories to meet this recommendation.

The laboratory medical director is ultimately responsible for demonstrating the validity of each assay and in selected instances may determine that a validation set smaller than 20 samples is sufficient. In such cases, the medical director must also provide and document an objective rationale for this determination.

For validation results that do not meet the 90% standard, the medical director will be responsible for determining both the basis for this result and the appropriate mitigation (testing of additional tissues, change in test conditions, or use of a different antibody). In general, assays that cannot be validated against this standard should not be used in clinical practice.

Some nonpredictive markers are reported quantitatively. Examples include, but are not limited to, immunoglobulin G4 (IgG4) in sclerosing inflammatory disorders, activated caspase 3 or Microtubule-associated protein 1 light chain 3 in ischemia or sepsis, and Phosphohistone H3 as a surrogate of mitotic figure count. For such markers, we recommend that the validation set include high and low expressors to ensure test accuracy over the analytic range.

4: Expert Consensus Opinion.—For initial analytic validation of all laboratory-developed predictive marker assays (with the exception of HER2/*neu*, ER, and PgR), laboratories should test a minimum of 20 positive and 20 negative cases. When the laboratory medical director determines that fewer than 40 validation cases are sufficient for a specific marker, the rationale for that decision needs to be documented.

Note: Positive cases in the validation set should span the expected range of clinical results (expression levels). This recommendation does not apply to any marker for which a separate validation guideline already exists.

Strength of evidence was *inadequate* to support the recommended number of validation samples, but was *adequate* to support distinguishing nonpredictive from predictive IHC tests and using different numbers of validation samples for each.

The statistical argument is updated here for predictive factor assays. Table 5 provides overall concordance estimates with 95% CIs for a 40-tissue validation set and for a

20-tissue set for those who will compute positive and negative concordance estimates.

Using a 40-sample validation set, the overall concordance estimates meet the 90% benchmark with 4 or fewer discordances. The “true” concordance between the 2 assays has only a 5% chance of falling outside the 95% CIs of the concordance estimates, and can be lower or higher than the estimate. If the 95% to 100% concordance estimates for the 40-sample validation set are a “true” representation of the relationship between the 2 tests, the validation results would meet the benchmark more than 95% of the time with 0 to 2 observed discordant results. The probabilities of meeting the benchmark if the 92.5% or 90% concordance estimates are a “true” representation would be 82% (approximation) and 63%, respectively (Binomial Calculator, Stat Trek; <http://stattrek.com/>).

In a 40-sample validation that does not meet the benchmark, analyses such as the McNemar test may help determine whether an observed difference in the off-diagonal represents a significant bias between the new and referent tests. Table 6 provides an example. In this case, the κ statistic showed “substantial” agreement, but the overall concordance estimate (87.5%) missed the benchmark by a small margin. The positive concordance of 75% suggests false negatives could be occurring in the new test, but the McNemar test is not significant, indicating that the 5 discordant results all in a single cell could have happened by chance.

Some laboratories may choose to validate predictive tests with tissue sets larger than the recommended minimum. For validation sets of 80 samples or more, the McNemar test is more useful in documenting whether observed differences/biases between the tests are significant. For example, for an 80-tissue validation set in which the numbers in each of the 4 cells in Table 6 are doubled, the McNemar result for 10 to 0 asymmetry on the off-diagonal would be significant ($P = .004$).

For validation results that do not meet the 90% standard, the laboratory medical director will be responsible for determining both the basis for this result and the appropriate mitigation (testing of additional tissues, change in test conditions).

5: Recommendation.—For a marker with both predictive and nonpredictive applications, laboratories should validate it as a predictive marker if it is used as such.

Strength of evidence was *adequate* to support the use of the higher validation standard (eg, number of samples) in the case of a marker with both nonpredictive and predictive intended uses.

Immunohistochemical assays have a variety of clinical applications including cell, tissue, or microbiologic identification, tumor diagnosis and prognosis, genetic and cancer risk assessment, and prediction of response to targeted therapies (predictive markers).

Although most IHC assays are interpretable only within the context of the clinical and histologic evaluation of the

Table 5. Validation Using a 40-Tissue Validation Set (20 Positive and 20 Negative) Against a 90% Concordance Benchmark

No.	Concordance Estimate, % (95% CI)				
	0 Discordant	1 Discordant	2 Discordant	3 Discordant	4 Discordant
20	100 (81–100)	95 (75–100)	90 (69–98)	85 (63–96)	80 (58–92)
40	100 (90–100)	97.5 (86–100)	95 (83–99)	92.5 (79–98)	90 (76–97)

Abbreviations: CI, confidence interval; No., number of validation tissues.

specific case, the results of predictive factor testing often directly influence how patients are managed. Some IHC assays are used for more than 1 purpose—the same antigen may be assessed to determine a patient’s eligibility for a targeted therapy as well as part of a panel in determining tumor type.

Assay validation procedures must take into account the test’s intended uses. When a marker will be used in both predictive and nonpredictive applications, assay validation should follow the recommendation for predictive markers because of its greater stringency.

When assessing the analytic validity of a predictive marker, cases should be selected to ensure that the new assay is concordant with its comparator over the expected range of clinical results. When validating the same marker for nonpredictive uses, cases should be selected to ensure that the test has acceptable concordance. Assays, such as ER or CD117 (c-KIT), that have been optimized to detect low levels of antigen for predictive uses could have high false-positive results (low negative concordance) when used as a lineage marker. Laboratories may choose to perform separate validations for the marker’s predictive and non-predictive applications.

6: Recommendation.—When possible, laboratories should use validation tissues that have been processed with the same fixative and processing methods as cases that will be tested clinically.

Strength of evidence was *inadequate* to address the influence of fixation, the type of decalcification solution, the time in decalcification solution, or validation tissues processed in another laboratory on analytic validation; however, the strength of evidence was *adequate* to support that laboratories should, whenever possible, use the same fixative and processing methods as cases tested clinically, in order to validate using representative specimens.

Fixative type, fixation time, tissue processing, and other preanalytic variables significantly affect the performance characteristics of IHC assays. To reduce the risk of false-negative and false-positive comparisons, validation materials should be handled in a manner similar to clinical specimens. Reference laboratories that test tissues from outside facilities usually cannot control differences in specimen handling and processing but should consider such differences when interpreting results.

Key criteria in grading the quality and strength of evidence for analytic validation include the internal validity of the studies and the consistency and generalizability of the results.^{6,29} To generalize the laboratory’s analytic validation results, the tissues included in a validation set must be representative of the specimens received in routine practice and must provide a representative range of expression intensities and patterns.

Although it is ideal if validation materials are identical to patient test specimens (eg, formalin-fixed tissue sections;

cell blocks from cytologic specimens initially fixed in alcohol; decalcified tissues), it is generally not practical to maintain complete validation sets specific for all possible specimen types, fixatives, and times in decalcification solution. It is reasonable for laboratories to test a selected panel of common markers to show that specimens of different type or processed differently exhibit equivalent immunoreactivity (LAP checklist ANP.22550).³

Note that there have been reports of false-positive and false-negative reactions for some markers after alcohol fixation. Although there are currently few data on this subject and more evidence is needed, the laboratory medical director should consider this possibility when selecting markers for the panel.

7: Expert Consensus Opinion.—If IHC is regularly done on cytologic specimens that are not processed in the same manner as the tissues used for assay validation (eg, alcohol-fixed cell blocks, air-dried smears, formalin-postfixed specimens), laboratories should test a sufficient number of such cases to ensure that assays consistently achieve expected results. The laboratory medical director is responsible for determining the number of positive and negative cases and the number of predictive and nonpredictive markers to test.

The strength of evidence was *inadequate* to address the criteria and number of samples needed for validation with cytology specimens.

Laboratories typically optimize and validate their IHC assays by using formalin-fixed, paraffin-embedded tissues but may use cytologic specimens in some circumstances; however, cytologic specimens usually have different fixation and processing methods and these factors may have unknown effects on IHC test results. Although separate validation of all markers on all potential cytologic specimens is generally not feasible, laboratories should determine

Table 6. 2 × 2 Contingency Table of a 40-Tissue Validation Set That Did Not Meet the Benchmark With Associated Statistical Tests^{a-c}

New Test	Comparator Test		Total
	Positive	Negative	
Positive	15	0	15
Negative	5	20	25
Total	20	20	40

^a Overall concordance: 35 of 40 = 87.5% (does not meet 90% benchmark); positive concordance: 15 of 20 = 75%; negative concordance: 20 of 20 = 100%.

^b κ: 0.75; McNemar test: *P* = .13.

^c The κ statistic shows “substantial” agreement, but the overall concordance estimate misses the 90% benchmark. Positive concordance of 75% could suggest that false negatives are occurring in the new test, but the McNemar test is not significant, indicating that the 5 discordant results all in a single cell could have happened by chance.

whether cytologic specimens have equivalent immunoreactivity to routinely processed, formalin-fixed tissue.

To assess the extent to which differences in cytologic specimen types and processing steps influence IHC test results, laboratories should test a selected set of commonly ordered markers (eg, keratin, CD45, S100, ER) in a set of cytologic specimen types used for IHC staining. The results should be correlated with expected results in routinely processed (control) tissues and with other applicable test results (eg, surgical specimen of primary neoplasm). The laboratory medical director must determine the number of cases and markers to test, bearing in mind the possibility of spurious results in alcohol-fixed materials. This assessment should be repeated when there is a change in cytologic fixative, collection media, sample preparation, or processing.

If an assay has not been fully validated on cytologic specimens, laboratories may include a disclaimer in their report that results should be interpreted with caution.

No primary studies, systematic evidence reviews, or qualitative documents were identified that addressed the specific question regarding the number and type of cytology specimens that are needed in a validation set for a new IHC assay. Studies^{30–36} were identified that compared cytology specimens to formalin-fixed tissue sections for ER, PgR, and/or HER2/*neu* IHC testing. Most concordance estimates were high ($\geq 90\%$), but the studies were small and used different fixatives, fixation times, and cytology specimen types (eg, smears, thin-layer, cell blocks). No two studies could be directly compared.

8: Expert Consensus Opinion.—If IHC is regularly performed on decalcified tissues, laboratories should test a sufficient number of such tissues to ensure that assays consistently achieve expected results. The laboratory medical director is responsible for determining the number of positive and negative tissues and the number of predictive and nonpredictive markers to test.

The strength of evidence was *inadequate* to address the criteria and number of samples needed for validation with decalcified specimens.

Decalcifying solutions vary in their effects on retention and integrity of nucleic acids and proteins. Results of IHC testing on decalcified specimens are unpredictable because of wide variations in specimen types and sizes, the length of time specimens are held in decalcification solution, and the particular solution(s) used. Although separate validation of all markers on all potential decalcified specimen types is not feasible, laboratories should determine the extent to which their decalcification procedures affect test results, particularly among specimen types that commonly have IHC testing, such as bone marrow biopsy samples.

No primary studies, systematic evidence reviews, or qualitative documents (eg guidelines, consensus meeting reports) were identified that address the specific question regarding the number of decalcified bone marrow specimens from positive and negative cases needed in a validation set for a new IHC assay. Nine articles and documents^{25,26,37–43} addressed the potential influence of decalcification as a modifier in the analytic validation process. Some authors^{26,38–40} report variability in decalcification protocols and in preservation of antigenicity in IHC tests. Two IHC guidelines recommend interpreting IHC results on decalcified samples with caution because of the possibility of antigen (and tissue) loss, but others report good morphology and successful staining with protocols using different fixatives, acid or EDTA decalcification, and paraffin or resin

embedding.^{37,40,42,43} Although the evidence was *inadequate*, these observations emphasize the need for a defined protocol and a validation plan that will ensure robust and reproducible IHC results in decalcified specimens.

Compared with other specimens, bone marrow biopsy samples are more consistent in size and in the time needed for decalcification, and are usually subject to standardized processing and decalcification protocols. To assess the influence of their decalcification procedure on IHC test results in bone marrows, laboratories should test a selected set of commonly ordered markers (eg, CD3, CD20, CD138) in a series of cases. The results may be correlated with expected results in routinely processed (control) tissues and with other applicable test results (eg, flow cytometry, IHC testing of lymph node in same patient). The laboratory medical director must determine the number of cases and markers to test. This assessment should be repeated when there is a change in decalcifying solution or fixative type.

For specimen types other than bone marrow samples, laboratories may include a disclaimer in their reports that the assay has not been fully validated on decalcified tissues and that results should be interpreted with caution given the possibility of false negativity on decalcified specimens (LAP checklist ANP.22985).³

9: Recommendation.—Laboratories may use whole sections, tissue microarrays (TMAs), and/or multitissue blocks (MTBs) in their validation sets as appropriate. Whole sections should be used if TMAs/MTBs are not appropriate for the targeted antigen or if the laboratory medical director cannot confirm that the fixation and processing of TMAs/MTBs is similar to clinical specimens.

Strength of evidence was adequate to support TMA usage; however, there are many variables to be considered and thorough validation is needed for each marker. Strength of evidence was *inadequate* to recommend the *routine* use of TMA samples.

Whole sections usually provide more antigen-positive cells and negative internal control cells within each section than TMAs/MTBs, but the latter can be designed to contain multiple previously tested positive and negative tissues. This allows for comparison of results in multiple tissues tested with an identical assay protocol and, when properly selected, a cost-effective validation strategy. Because of the small size of each tissue sample, however, TMAs and MTBs may be inappropriate for antigens with limited tissue expression, heterogeneous distribution, or restricted compartmentalization within tissues. The laboratory director must use information from the literature and clinical judgment to determine if TMAs or MTBs are useful for validating a given assay.

Comparisons of overall concordance between IHC assays performed on whole sections and TMAs have been done with at least 9 markers, but primarily with ER, PgR, and HER2/*neu*.^{44–55} Summary estimates of concordance (random effects model) were computed, but heterogeneity was high across the studies ($I^2 > 75$; $P < .001$), and specific sources of heterogeneity could not be identified. Consequently, concordance is reported as ranges and median values for specific markers, all in breast cancer tissues.

Median overall concordance estimates for ER, PgR, and HER2/*neu* were 95% (range, 84%–99%), 91% (range, 81%–93%), and 93% (range, 73%–100%), respectively, but concordance estimates in our review only met or exceeded the 90% standard in about two-thirds of cases. Comparisons

of overall concordance for ER and PgR from an earlier systematic review were 97% and 93%, respectively.⁵²

10: Expert Consensus Opinion.—When a new reagent lot is placed into clinical service for an existing validated assay, laboratories should confirm the assay's performance with at least 1 known positive case and 1 known negative case.

The strength of evidence was *inadequate* to address conditions requiring assay revalidation and whether revalidation should be the same as initial validation.

Confirmation that assay performance has not changed is necessary when a new lot of primary antibody or antigen retrieval or detection reagent is used. For predictive markers, testing both high and low expressors may be useful. Including a weakly positive sample is recommended when there is a specified cut point for positivity (eg, ER) (LAP checklist COM.30450).³ Including 2 positive cases (1 weak and 1 strong) should be considered for new reagent lots of predictive marker antibodies.

11: Expert Consensus Opinion.—Laboratories should confirm assay performance with at least 2 known positive and 2 known negative cases when an existing validated assay has changed in any one of the following ways:

1. Antibody dilution;
2. Antibody vendor (same clone);
3. Incubation or retrieval times (same method).

The strength of evidence was *inadequate* to address conditions requiring assay revalidation and whether revalidation should be the same as initial validation.

Confirmation that assay performance has not changed is necessary when there are minor changes to the assay method. Public comments received on this recommendation were more contentious than for most other recommendations. Some argued that these changes fundamentally change the nature of the assay and therefore should require full assay revalidation, while others noted that the number of cases needed to ensure the assay is performing as expected will vary by antibody. The importance of not replacing the pathologist's judgment with arbitrary minimum numbers was also stressed. From the comments received, the panel concluded that re-assessing assays with at least 2 positive and 2 negative cases was a reasonable compromise in ensuring assay performance and provides the laboratory medical director flexibility to increase the number as needed.

For predictive markers, laboratories testing both high and low expressors may be useful. Including weakly positive samples is recommended when there is a specified cut point for positivity (eg, ER). Major changes in antibody dilution or incubation times (as defined by the laboratory) may warrant testing more than 2 negative and 2 positive cases.

12: Expert Consensus Opinion.—Laboratories should confirm assay performance by testing a sufficient number of cases to ensure that assays consistently achieve expected results when any of the following have changed:

1. Fixative type;
2. Antigen retrieval method (eg, change in pH, different buffer, different heat platform);
3. Antigen detection system;
4. Tissue processing or testing equipment;
5. Environmental conditions of testing (eg, laboratory relocation);

6. Laboratory water supply.

The laboratory medical director is responsible for determining the number of positive and negative cases and the number of predictive and nonpredictive markers to test.

The strength of evidence was *inadequate* to address conditions requiring assay revalidation and whether revalidation should be the same as initial validation.

Recommendations 10 and 11 apply to changes in 1 antibody or assay, but this recommendation applies to changes that affect most or all of a laboratory's assays. Full revalidation of every assay in this situation is not practical, but an assessment is needed to ensure that results of testing under new conditions are comparable to the results of prior testing. The laboratory medical director must determine the extent of this testing based on the nature of the change. A representative panel of predictive and nonpredictive markers could be selected to assess the impact of the change. Based on those results, more thorough testing may be needed, particularly for predictive markers, but if results on this panel are acceptable, remaining assays could be verified less rigorously. Markers selected for testing should include those with different immunolocalizations (ie, nuclear, membranous, cytoplasmic) as appropriate for the laboratory.

When feasible, comparing the results of staining after the change with the slides from initial assay validation may help to determine if the intensity of staining has changed. Laboratories are required to verify method performance specifications after an instrument is moved to ensure that the test system was not affected by the relocation process or environmental changes (LAP checklist COM.40000).³

13: Expert Consensus Opinion.—Laboratories should run a full revalidation (equivalent to initial analytic validation) when the antibody clone is changed for an existing validated assay.

The strength of evidence was *inadequate* to address conditions requiring assay revalidation and whether revalidation should be the same as initial validation.

Although a limited re-assessment of assay performance is sufficient when there are minor changes in assay conditions (eg, antibody dilution or incubation time), introduction of a different antibody clone represents a fundamental change to the assay and requires complete revalidation. This is because different antibody clones are raised against different epitopes on the target protein and their performance characteristics may significantly vary. This phenomenon is exemplified by the expression of TTF-1 (thyroid transcription factor 1) in carcinomas other than those of thyroid or pulmonary origin. Multiple studies^{56–58} have shown low levels of expression in metastatic and primary colorectal carcinomas, carcinomas of gynecologic origin, and glial neoplasms, using the SPT24 clone. By contrast, the 8G7G3/1 clone is uniformly negative in these tumor types. Similar data exist for CDX2.⁵⁹

14: Expert Consensus Opinion.—The laboratory must document all validations and verifications in compliance with regulatory and accreditation requirements.

For laboratories subject to US regulations, CLIA specifies that "records of the laboratory's establishment and verification of method performance specifications must be retained for the period of time the test system is in use by the laboratory, but not less than 2 years."¹ Laboratories accredited by CAP must retain records of method performance specifications while the method is in use and for at

least 2 years after discontinuation of the method (LAP checklist COM.40000).³

In addition to written procedures that describe their validation and revalidation processes, laboratories should have documentation, signed by the laboratory medical director, of the validation, verification, or revalidation studies and approval of each test for its intended clinical use(s).

Note on Evidence Analysis for Revalidation Recommendations (No.10–No.13).—No objective evidence was identified that addressed requirements for revalidating IHC assays when there are changes to an existing validated assay (eg, new reagent lot, change in antibody dilution, changes in equipment). Refer to the full analysis of key question 6 and key question 7 regarding revalidation in the supplemental digital content for further discussion of the evidence.

CONCLUSION

Physicians and patients rely on accurate diagnostic and prognostic testing in the clinical laboratory. Established guidelines for validating and revalidating immunohistochemistry tests used on clinical specimens are important in ensuring accuracy, reproducibility, and consistency of test results. The potential harms of false-positive and false-negative results due to inadequate validation need to be recognized and addressed. This guideline is intended to help laboratories improve the accuracy of testing and reassure clinicians and patients that accepted procedures from evidence-based and expert consensus-based recommendations are being followed. Direction for re-assessing assays when changes have occurred or when results are not as expected is also provided.

We thank the Center advisors Raouf Nakhleh, MD, Sandi Larsen, MBA, MT(ASCP), and John Olsen, MD, as well as advisory panel members Richard W. Brown, MD, Richard N. Eisen, MD, and Hadi Yaziji, MD.

References

1. US Department of Health and Human Services. Clinical laboratory improvement amendments of 1988: final rule. *Fed Regist*. 1992;57(40):7001–7186. Codified at 42 CFR §1405–494.
2. Immunology Branch, Division of Clinical Laboratory Devices, Office of Device Evaluation. 3.9: Manufacturers' recommendations for verification of IHC performance by the user. In: *Guidance for Submission of Immunohistochemistry Applications to the FDA*. Center for Devices and Radiological Health, US Food and Drug Administration; 1998. <http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm094015.pdf>. Accessed September 1, 2013.
3. College of American Pathologists. CAP Laboratory accreditation checklists. <http://www.cap.org/apps/cap.portal>. Accessed September 1, 2013.
4. Hardy LB, Fitzgibbons PL, Goldsmith JD, et al. Immunohistochemistry validation procedures and practices: a College of American Pathologists survey of 727 laboratories. *Arch Pathol Lab Med*. 2013;137(1):19–25.
5. Nakhleh RE, Grimm EE, Idowu MO, Souers RJ, Fitzgibbons PL. Laboratory compliance with the American Society of Clinical Oncology/College of American Pathologists guidelines for human epidermal growth factor receptor 2 testing: a College of American Pathologists survey of 757 laboratories. *Arch Pathol Lab Med*. 2010;134(5):728–734.
6. Teutsch SM, Bradley LA, Palomaki GE, et al. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group. *Genet Med*. 2009;11(1):3–14.
7. US Department of Health and Human Services. Medicare, Medicaid and CLIA programs: regulations implementing the Clinical Laboratory Improvement Amendments of 1988: final rule. *Fed Regist*. 1992;57:7002–7186.
8. Clinical Laboratory Standards Institute. Quality assurance for design control and implementation of immunohistochemistry assays: approved guideline, second edition. In: *CLSI Document I/LA28-A2*. Wayne, PA: Clinical and Laboratory Standards Institute; 2011.
9. Dowsett M, Hanna WM, Kockx M, et al. Standardization of HER2 testing: results of an international proficiency-testing ring study. *Mod Pathol*. 2007;20(5):584–591.

10. Wolff AC, Hammond ME, Schwartz JN, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *Arch Pathol Lab Med*. 2007;131(1):18–43.

11. Allred DC, Carlson RW, Berry DA, et al. NCCN Task Force report: estrogen receptor and progesterone receptor testing in breast cancer by immunohistochemistry. *J Natl Compr Canc Netw* [quiz in *J Natl Compr Canc Netw*. 2009;7(suppl 6):S22–S23]. 2009;7(suppl 6):S1–S21.

12. Baba K, Dyrhol-Riise AM, Sviland L, et al. Rapid and specific diagnosis of tuberculous pleuritis with immunohistochemistry by detecting Mycobacterium tuberculosis complex specific antigen MPT64 in patients from a HIV endemic area. *Appl Immunohistochem Mol Morphol*. 2008;16(6):554–561.

13. Boers JE, Meeuwissen H, Methorst N. HER2 status in gastro-oesophageal adenocarcinomas assessed by two rabbit monoclonal antibodies (SP3 and 4B5) and two in situ hybridization methods (FISH and SISH). *Histopathology*. 2011;58(3):383–394.

14. Mayr D, Heim S, Werhan C, Zeindl-Eberhart E, Kirchner T. Comprehensive immunohistochemical analysis of Her-2/neu oncoprotein overexpression in breast cancer: HercepTest (Dako) for manual testing and Her-2/neuTest 4B5 (Ventana) for Ventana BenchMark automatic staining system with correlation to results of fluorescence in situ hybridization (FISH). *Virchows Arch*. 2009;454(3):241–248.

15. Moelans CB, Kibbelaar RE, van den Heuvel MC, Castigliogio D, de Weger RA, van Diest PJ. Validation of a fully automated HER2 staining kit in breast cancer. *Cell Oncol*. 2010;32(1–2):149–155.

16. O'Grady A, Allen D, Happerfield L, et al. An immunohistochemical and fluorescence in situ hybridization-based comparison between the Oracle HER2 Bond Immunohistochemical System, Dako HercepTest, and Vysis PathVysion HER2 FISH using both commercially validated and modified ASCO/CAP and United Kingdom HER2 IHC scoring guidelines. *Appl Immunohistochem Mol Morphol*. 2010;18(6):489–493.

17. Phillips T, Murray G, Wakamiya K, et al. Development of standard estrogen and progesterone receptor immunohistochemical assays for selection of patients for antihormonal therapy. *Appl Immunohistochem Mol Morphol*. 2007;15(3):325–331.

18. Rhodes A, Jasani B, Anderson E, Dodson AR, Balaton AJ. Evaluation of HER-2/neu immunohistochemical assay sensitivity and scoring on formalin-fixed and paraffin-processed cell lines and breast tumors: a comparative study involving results from laboratories in 21 countries. *Am J Clin Pathol*. 2002;118(3):408–417.

19. van der Vegt B, de Bock GH, Bart J, Zwartjes NG, Wesseling J. Validation of the 4B5 rabbit monoclonal antibody in determining Her2/neu status in breast cancer. *Mod Pathol*. 2009;22(7):879–886.

20. Dorfman DM, Bui MM, Tubbs RR, et al. The CD117 immunohistochemistry tissue microarray survey for quality assurance and interlaboratory comparison: a College of American Pathologists Cell Markers Committee study. *Arch Pathol Lab Med*. 2006;130(6):779–782.

21. Jordan RC, Lingen MW, Perez-Ordóñez B, et al. Validation of methods for oropharyngeal cancer HPV status determination in US cooperative group trials. *Am J Surg Pathol*. 2012;36(7):945–954.

22. Lotan TL, Gurel B, Sutcliffe S, et al. PTEN protein loss by immunostaining: analytic validation and prognostic indicator for a high risk surgical cohort of prostate cancer patients. *Clin Cancer Res*. 2011;17(20):6563–6573.

23. Lehmann-Che J, Amira-Bouhidel F, Turpin E, et al. Immunohistochemical and molecular analyses of HER2 status in breast cancers are highly concordant and complementary approaches. *Br J Cancer*. 2011;104(11):1739–1746.

24. Fitzgibbons PL, Murphy DA, Hammond ME, Allred DC, Valenstein PN. Recommendations for validating estrogen and progesterone receptor immunohistochemistry assays. *Arch Pathol Lab Med*. 2010;134(6):930–935.

25. Hammond ME, Hayes DF, Dowsett M, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *Arch Pathol Lab Med*. 2010;134(6):907–922.

26. Torlakovic EE, Naresh K, Kremer M, van der Walt J, Hyjek E, Porwit A. Call for a European programme in external quality assurance for bone marrow immunohistochemistry; report of a European Bone Marrow Working Group pilot study. *J Clin Pathol*. 2009;62(6):547–551.

27. US Department of Health and Human Services. Medical devices: classification/reclassification of immunochemistry reagents and kits. *Fed Regist*. 1998;63(106):30132–30142. Codified at 21 CFR §864. Doc. No. 94P–0341.

28. Wolff AC, Hammond EH, Hicks DG, et al. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. *Arch Pathol Lab Med*. 2014;138:241–256. doi:10.5858/arpa.2013-0953-SA.

29. Sun F, Bruening W, Erinoff E, Schoelles KM. *Addressing Challenges in Genetic Test Evaluation: Evaluation Frameworks and Assessment of Analytic Validity*. Methods research report (prepared by the ECRI Institute Evidence-Based Practice Center under contract No. HHS 290-20007-10063-l). Rockville, MD: Agency for Healthcare Research and Quality; June 2011. AHRQ Publication No. 11-EHC048-EF.

30. Ferguson J, Chamberlain P, Cramer HM, Wu HH. ER, PR, and Her2 immunocytochemistry on cell-transferred cytologic smears of primary and

metastatic breast carcinomas: a comparison study with formalin-fixed cell blocks and surgical biopsies. *Diagn Cytopathol*. 2013;41(7):575–581.

31. Gong Y, Symmans WF, Krishnamurthy S, Patel S, Sneige N. Optimal fixation conditions for immunocytochemical analysis of estrogen receptor in cytologic specimens of breast carcinoma. *Cancer*. 2004;102(1):34–40.

32. Hanley KZ, Birdsong GG, Cohen C, Siddiqui MT. Immunohistochemical detection of estrogen receptor, progesterone receptor, and human epidermal growth factor receptor 2 expression in breast carcinomas: comparison on cell block, needle-core, and tissue block preparations. *Cancer*. 2009;117(4):279–288.

33. Kumar SK, Gupta N, Rajwansi A, Joshi K, Singh G. Immunocytochemistry for oestrogen receptor, progesterone receptor and HER2 on cell blocks in primary breast carcinoma. *Cytopathology*. 2012;23(3):181–186.

34. Nishimura R, Aogi K, Yamamoto T, et al. Usefulness of liquid-based cytology in hormone receptor analysis of breast cancer specimens. *Virchows Arch*. 2011;458(2):153–158.

35. Pegolo E, Machin P, Riosa F, Bassini A, Deroma L, Di Loreto C. Hormone receptor and human epidermal growth factor receptor 2 status evaluation on ThinPrep specimens from breast carcinoma: correlation with histologic sections determination. *Cancer Cytopathol*. 2012;120(3):196–205.

36. Shabaik A, Lin G, Peterson M, et al. Reliability of Her2/neu, estrogen receptor, and progesterone receptor testing by immunohistochemistry on cell block of FNA and serous effusions from patients with primary and metastatic breast carcinoma. *Diagn Cytopathol*. 2011;39(5):328–332.

37. Adegboyega PA, Gokhale S. Effect of decalcification on the immunohistochemical expression of ABH blood group isoantigens. *Appl Immunohistochem Mol Morphol*. 2003;11(2):194–197.

38. Arber JM, Arber DA, Jenkins KA, Battifora H. Effect of decalcification and fixation in paraffin-section immunohistochemistry. *Appl Immunohistochem*. 1996;4(4):241–248.

39. Bussolati G, Leonardo E. Technical pitfalls potentially affecting diagnoses in immunohistochemistry. *J Clin Pathol*. 2008;61(11):1184–1192.

40. Fend F, Tzankov A, Bink K, et al. Modern techniques for the diagnostic evaluation of the trephine bone marrow biopsy: methodological aspects and applications. *Prog Histochem Cytochem*. 2008;42(4):203–252.

41. Hsi ED. A practical approach for evaluating new antibodies in the clinical immunohistochemistry laboratory. *Arch Pathol Lab Med*. 2001;125(2):289–294.

42. Wittenburg G, Volkel C, Mai R, Lauer G. Immunohistochemical comparison of differentiation markers on paraffin and plastic embedded human bone samples. *J Physiol Pharmacol*. 2009;60(suppl 8):43–49.

43. Zustin J, Boddin K, Tsourlakis MC, et al. HER-2/neu analysis in breast cancer bone metastases. *J Clin Pathol*. 2009;62(6):542–546.

44. Batistatou A, Televantou D, Bobos M, et al. Evaluation of current prognostic and predictive markers in breast cancer: a validation study of tissue microarrays. *Anticancer Res*. 2013;33(5):2139–2145.

45. Drev P, Grazio SF, Bracko M. Tissue microarrays for routine diagnostic assessment of HER2 status in breast carcinoma. *Appl Immunohistochem Mol Morphol*. 2008;16(2):179–184.

46. Fons G, Hasibuan SM, van der Velden J, ten Kate FJ. Validation of tissue microarray technology in endometrioid cancer of the endometrium. *J Clin Pathol*. 2007;60(5):500–503.

47. Graham AD, Faratian D, Rae F, Thomas JSJ. Tissue microarray technology in the routine assessment of HER-2 status in invasive breast cancer: a prospective study of the use of immunohistochemistry and fluorescence *in situ* hybridization. *Histopathology*. 2008;52:847–855.

48. Gulbahce HE, Gamez R, Dvorak L, Forster C, Varghese L. Concordance between tissue microarray and whole-section estrogen receptor expression and intratumoral heterogeneity. *Appl Immunohistochem Mol Morphol*. 2012;20:340–343.

49. Henriksen KL, Rasmussen BB, Lykkesfeldt AE, Moller S, Ejlersen B, Mouridsen HT. Semi-quantitative scoring of potentially predictive markers for endocrine treatment of breast cancer: a comparison between whole sections and tissue microarrays. *J Clin Pathol*. 2007;60(4):397–404.

50. Jones S, Prasad ML. Comparative evaluation of high-throughput small-core (0.6-mm) and large-core (2-mm) thyroid tissue microarray: is larger better? *Arch Pathol Lab Med*. 2012;136(2):199–203.

51. Kwon MJ, Nam ES, Cho SJ, et al. Comparison of tissue microarray and full section in immunohistochemistry of gastrointestinal stromal tumors. *Pathol Int*. 2009;59(12):851–856.

52. Nofech-Mozes S, Vella ET, Dhesy-Thind S, et al. Systematic review on hormone receptor testing in breast cancer. *Appl Immunohistochem Mol Morphol*. 2012;20(3):214–263.

53. Soiland H, Skaland I, van Diermen B, et al. Androgen receptor determination in breast cancer: a comparison of the dextran-coated charcoal method and quantitative immunohistochemical analysis. *Appl Immunohistochem Mol Morphol*. 2008;16(4):362–370.

54. Thomson TA, Zhou C, Chu C, Knight B. Tissue microarray for routine analysis of breast biomarkers in the clinical laboratory. *Am J Clin Pathol*. 2009;132(6):899–905.

55. Warnberg F, Amini RM, Goldman M, Jirstrom K. Quality aspects of the tissue microarray technique in a population-based cohort with ductal carcinoma *in situ* of the breast. *Histopathology*. 2008;53(6):642–649.

56. Comperat E, Zhang F, Perrotin C, et al. Variable sensitivity and specificity of TTF-1 antibodies in lung metastatic adenocarcinoma of colorectal origin. *Mod Pathol*. 2005;18(10):1371–1376.

57. Kristensen MH, Nielsen S, Vyberg M. Thyroid transcription factor-1 in primary CNS tumors. *Appl Immunohistochem Mol Morphol*. 2011;19(5):437–443.

58. Zhang PJ, Gao HG, Pasha TL, Litzky L, Livolsi VA. TTF-1 expression in ovarian and uterine epithelial neoplasia and its potential significance, an immunohistochemical assessment with multiple monoclonal antibodies and different secondary detection systems. *Int J Gynecol Pathol*. 2009;28(1):10–18.

59. Borrisholt M, Nielsen S, Vyberg M. Demonstration of CDX2 is highly antibody dependant. *Appl Immunohistochem Mol Morphol*. 2013;21(1):64–72.

APPENDIX Disclosed Interests and Activities June 2010 to September 2013		
Name	Interest/Activity Type	Entity
Linda A. Bradley, PhD	Consultancy	Blue Cross Blue Shield Association American College of Medical Genetics Foundation
Regan S. Fulton, MD, PhD	Board or advisory board	Center for Medical Technology Policy
	Consultancy	Gerson Lehrman Group
	Grants	National Institutes of Health-Small Business Innovation Research Grant (application pending)
	Patent received or pending	United States Patent and Trademark Office Application (application pending)
	Ownership or beneficial ownership of stock	Array Science, LLC
Jeffrey D. Goldsmith, MD	Lecture fee paid by entity	United States and Canadian Academy of Pathology
	Expert witness	Various
Thomas S. Haas, DO	Consultancy	Biocare Medical, Concord, California Newcomer Histology Supply, Middleton, Wisconsin
	Board or advisory board	Biocare Medical, Concord, California Leica Microsystems, Buffalo Grove, Illinois
	Lecture fee paid by entity	Leica Microsystems, Buffalo Grove, Illinois Biocare Medical, Concord, California National Society for Histotechnology
Patti A. Loykasek, HTL(ASCP), QIHC	Board or advisory board	Clover Park College
Paul E. Swanson, MD	Consultancy	PhenoPath Laboratory
	Consultancy	American Society of Clinical Pathology
	Lecture fee paid by entity	College of American Pathologists