Form Approved
OMB No. 0920-1067
Exp. Date: xx/xx/20xx

COLLEGE of AMERICAN PATHOLOGISTS

Principles of Analytic Validation of Immunohistochemical Assays

Summary of Recommendations

Guideline Statement	Strength of Recommendation
1. Laboratories must validate all IHC tests before placing into clinical service. *Note*: 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.	Recommendation
2. For initial validation of every assay used clinically, with the exception of HER2/*neu*, 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.	Recommendation
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. *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.	Expert Consensus Opinion
4. 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 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. *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.	Expert Consensus Opinion
5. For a marker with both predictive and nonpredictive applications, laboratories should validate it as a predictive marker if it is used as such.	Recommendation

23685.060215

Guideline Statement	Strength of Recommendation
6. When possible, laboratories should use validation tissues that have been processed using the same fixative and processing methods as cases that will be tested clinically.	Recommendation
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.	Expert Consensus Opinion
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 issues and the number of predictive and nonpredictive markers to test.	Expert Consensus Opinion
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.	Recommendation
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.	Expert Consensus Opinion
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).	Expert Consensus Opinion
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.	Expert Consensus Opinion
13. Laboratories should run a full revalidation (equivalent to initial analytic validation) when the antibody clone is changed for an existing validated assay.	Expert Consensus Opinion
14. The laboratory must document all validations and verifications in compliance with regulatory and accreditation requirements.	Expert Consensus Opinion

Abbreviations: IHC, immunohistochemistry; ER, estrogen receptor; PgR, progesterone receptor; TMA, tissue microarray; MTB, multitissue block

Source: Fitzgibbons PL, Bradley LA, Fatheree LA, et al. Principles of analytic validation of immunohistochemical assays: Guideline from the College of American Pathologists Pathology and Laboratory Quality Center. *Arch Pathol Lab Med*. 2014;138(11):1432–1443.

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)

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

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.

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

3. 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 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.

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.

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	*Recommend for* or *against* a particular analytic validation practice (can include must or should).	Strength of evidence is *convincing*, based on consistent, generalizable, good-quality evidence; further studies are unlikely to change the conclusions.
Recommendation	*Recommend for* or *against* a particular analytic validation practice (can include should or may).	Strength of evidence is *adequate*, based on limitations in the quality of evidence; further studies may change the conclusions.
Expert consensus opinion	*Recommend for* or *against* a particular analytic validation practice (can include should or may).	Important validation element to address but strength of evidence is *inadequate*; 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, interrun 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**
1. Laboratories must validate all IHC tests before placing into clinical service. *Note:* 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.	Recommendation
2. For initial validation of every assay used clinically, with the exception of HER2/*neu*, 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.	Recommendation
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. *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.	Expert consensus opinion
4. 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 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. *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.	Expert consensus opinion
5. For a marker with both predictive and nonpredictive applications, laboratories should validate it as a predictive marker if it is used as such.	Recommendation
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.	Recommendation
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.	Expert consensus opinion
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.	Expert consensus opinion
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.	Recommendation
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.	Expert consensus opinion
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).	Expert consensus opinion
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.	Expert consensus opinion
13. Laboratories should run a full revalidation (equivalent to initial analytic validation) when the antibody clone is changed for an existing validated assay.	Expert consensus opinion
14. The laboratory must document all validations and verifications in compliance with regulatory and accreditation requirements.	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			
	Concordance Estimate, % (95% CI)		
No.	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					
	Concordance Estimate, % (95% CI)				
No.	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 nonpredictive 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]			
	Comparator Test		
New Test	Positive	Negative	Total
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 (≥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, Castigliego 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-Ordonez 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 Evience-Based Practice Center under contract No. HHSA 290-20007-10063-I). 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, Rajwanshi A, Joshi K, Singh G. Immunochemistry 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, Ejlertsen 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
	Board or advisory board	Center for Medical Technology Policy
Regan S. Fulton, MD, PhD	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
	Consultancy	PhenoPath Laboratory
Paul E. Swanson, MD	Consultancy	American Society of Clinical Pathology
	Lecture fee paid by entity	College of American Pathologists

COLLEGE of AMERICAN PATHOLOGISTS

Supplemental Digital Content* | Methodology | February 2015

Principles of Analytic Validation for Immunohistochemical Assays

Guideline from the Pathology and Laboratory Quality Center

Corresponding Author:
Patrick L. Fitzgibbons, MD

Authors:
Linda A. Bradley, PhD
Lisa A. Fatheree, SCT(ASCP)
Anthony T. Smith, ML

Archives Early Online Release: Principles of Analytic Validation of Immunohistochemical Assays

* The Supplemental Digital Content was not copyedited by *Archives of Pathology and Laboratory Medicine.*

METHODS USED TO PRODUCE THE GUIDELINE

Panel Composition

The College of American Pathologists (CAP) Pathology and Laboratory Quality Center (the Center) convened an expert and advisory panel consisting of pathologists and histotechnologists with expertise in implementing and performing immunohistochemical (IHC) assays. CAP approved the appointment of the project chair (PLF) and panel members. These panel members served as the Technical Expert Panel (TEP) for the systematic evidence review.

Conflict of Interest (COI) Policy

Prior to acceptance on the expert or advisory panel, potential members completed the CAP conflict of interest (COI) disclosure process, whose policy and form (in effect April 2010) requires 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. The potential members completed the COI disclosure form, listing any relationship that could be interpreted as constituting an actual, potential, or apparent conflict. The CAP Center uses the following criteria:

Nominees who have the following conflicts may be excused from the panel:

a. Stock or equity interest in a commercial entity that would likely be affected by the guideline or white paper
b. Royalties or licensing fees from products that would likely be affected by the guideline or white paper
c. Employee of a commercial entity that would likely be affected by the guideline or white paper

Nominees who have the following potentially manageable direct conflicts may be appointed to the panel:

a. Patents for products covered by the guideline or white paper
b. Member of an advisory board of a commercial entity that would be affected by the guideline or white paper
c. Payments to cover costs of clinical trials, including travel expenses associated directly with the trial
d. Reimbursement from commercial entity for travel to scientific or educational meetings

Everyone was required to disclose conflicts prior to beginning and continuously throughout the project's timeline. One expert panel member (RSF) was recused from discussion and voting on the recommendation pertaining to tissue microarrays (TMAs). One expert panel member (TSH) was recused from voting on the recommendations pertaining to potential increased antibody usage. Expert panel members' disclosed conflicts are listed in the appendix of the manuscript. 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.

CAP Expert Panel Literature Review and Analysis

The expert panel met 28 times through teleconference webinars from June 2010 through September 2013. Additional work was completed via electronic mail and the panel met in person May 11-12, 2013 to review evidence to date and draft recommendations.

All expert panelists participated in the systematic evidence review (SER) level of title-abstract and full-text review. Chair PLF and panelists PES and RSF performed the audit of data extraction. Panelist RSF was recused from performing any audit on articles pertaining to TMAs. All articles were available as discussion or background references. All members of the expert panel participated in developing draft recommendations, reviewing open comment feedback, finalizing and approving

recommendations and writing/editing of the manuscript except as noted for RSF and TSH.

Peer Review

An open comment period was held from July 8 through July 29, 2013. Eighteen draft recommendations and five methodology questions were posted online on the CAP Web site www.cap.org. An announcement was sent to the following societies deemed to have interest:

American Society for Clinical Pathology (ASCP) Association for Molecular Pathology (AMP) Society for Immunohistochemistry
National Society for Histotechnology (NSH) American Society of Cytopathology (ASC)
Association of Directors of Anatomic and Surgical Pathology (ADASP) Association of Pathology Chairs (APC)
Clinical Laboratory Management Association (CLMA)
US Food and Drug Administration (FDA)
Centers for Medicare and Medicaid Services (CMS) Canadian Association of Pathologists (CAP-APC)
United States & Canadian Academy of Pathology (USCAP)
United Kingdom National External Quality Assessment Service (UK NEQAS) Nordic IHC Quality Control (NordiQC)
Canadian IHC Quality Control (CIQC)

The website received 1,071 comments in total (Agree and Disagree responses were also captured). Twelve of 18 recommendations achieved more than 80% agreement; only 2 had less than 70% agreement. Each expert panel member was assigned 1-2 draft recommendations for which to review all comments received and provide an overall summary to the rest of the panel. Following panel discussion, a secondary internal review by the CAP Surgical Pathology and Immunohistochemistry Resource Committees and the final quality of evidence assessment, the panel members determined whether to maintain the original draft recommendation as is, revise it with minor language change, or consider it as a major recommendation change. Three draft recommendations were maintained with the original language; five were modified with minor changes for clarification and/or further explanation within the manuscript and six 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 using nominal group technique (rounds of email discussion and multiple edited recommendations) amongst the panel members. The final recommendations were approved by the expert panel with a formal vote (minus RSF on the recommendation regarding TMAs and TSH on potential increased antibody usage). 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 (IRP) was assembled to review the guideline and recommend approval to the CAP. The IRP was masked to the expert panel and vetted through the COI process.

Assessing the Strength of Recommendations

The central question that the panel addressed in developing the guideline was "*What is needed for initial analytic assay validation before placing any immunohistochemical test into clinical service, and what are the revalidation requirements?*"

Development of recommendations requires that the panel review the identified evidence and make a series of key judgments:

1) What are the significant findings related to each KQ or outcome? Determine which components of analytic validation (*e.g.,* overall and positive/negative concordance from comparisons, precision, robustness) have a regulatory requirement and/or evidence that support a specific action and/or method for the validation process.

2) What is the overall strength of evidence supporting each KQ or outcome? Strength of evidence is graded as Convincing, Adequate or Inadequate, based on four published criteria (SER, Figure 2). Strength of evidence is a key element in determining the strength of a recommendation.

3) What is the strength of each recommendation? There are many methods for determining the strength of a recommendation based on the strength of evidence and the magnitude of net benefit or harm. However, such methods have rarely (if ever) been applied to analytic validity, and certainly not to recommendations on component parts of the analytic validation process. Therefore, the method for determining strength of recommendation has been modified for this application (Table 1), and is based on the strength of evidence and the likelihood that further studies will change the conclusions. Recommendations not supported by evidence (*i.e.,* evidence was missing or Insufficient to permit a conclusion to be reached) may be made based on consensus expert opinion. Another potential consideration is the likelihood that additional studies need to fill gaps in knowledge will be conducted.

4) What is the net balance of benefits and harms? The consideration of net balance of benefits and harms will focus on the core recommendation to perform analytic validation before offering a test in practice.

Table 1: Grades for Strength of Recommendations*

Designation	Recommendation	Rationale
Strong Recommendation	Recommend For or Against a particular analytic validation practice (Can include must or should)	Strength of evidence is Convincing based on consistent, generalizable, good quality evidence; further studies are unlikely to change the conclusions
Recommendation	Recommend For or Against a particular analytic validation practice (Can include should or may)	Strength of evidence is Adequate based on limitations in the quality of evidence; further studies may change the conclusions
Expert Consensus Opinion	Recommend For or Against a particular analytic validation practice (Can include should or may)	Important validation element to address but strength of evidence is Inadequate; gaps in knowledge may require further studies

*Modified by the CAP Pathology and Laboratory Quality Center

Dissemination Plans

CAP will host an IHC Validation Resource web page which will include a link to manuscript and supplemental digital content; summary of recommendations, teaching PowerPoint, frequently asked question (FAQ) document and a free archived webinar. The guideline will be promoted and presented at various professional society meetings including the College of American Pathologists, the United States and Canadian Academy of Pathology (USCAP), the National Society for Histotechnologists (NSH), the American Society of Clinical Pathology (ASCP) and the American Society of Cytopathology (ASC).

SYSTEMATIC EVIDENCE REVIEW (SER)

The objectives of the SER were to investigate the optimal performance characteristics of IHC tests and determine how they can be achieved and measured. If of sufficient quality, findings from this review could provide an evidence base to support development of the clinical guideline. The scope of the SER and the key questions (KQs) were established by the TEP in consultation with a methodologist.

Search and Selection

Electronic searches of the English language published literature in Ovid MEDLINE®, U.S. National Library of Medicine PubMed, and Elsevier Scopus databases were initially conducted for the time period January 2004 to May 2012; an update was conducted through May 2013. The search utilized the following MeSH terms and keywords:

MeSH Terms	Keywords
Immunohistochemistry, Immunoenzyme Techniques, Validation Studies as Topic, Reproducibility of Results, Sensitivity and Specificity, Validation Studies, Evaluation Studies as Topic, Observer Variation, Clinical Laboratory Techniques, Laboratories, Hospital, Pathology, "Tumor Markers, Biological", Ki-67 Antigen, Cyclin-Dependent Kinase Inhibitor p16, "Receptor, erbB-2", "Receptors, Progesterone", "Receptors, Estrogen", Vimentin	Immunohistochemistry, IHC, Immunocytochemistry, Immunoperoxidase, Antigen retrieval, Antigen detection, Validation, Standardization, Inter-run variance, Inter-operator variance, Controls, Analytic variance, Signature molecules, Molecular tests and assays, Cytokeratin, CK 5/6, CK7, CK20, CD5, CD10, CD20, CD45, CD99, CD117, p63, Cyclin D1, bcl1, bcl2, actin, desim, chromogranin, cadherin, estrogen recepto progesterone receptor, HER2, erbB2, S10 TTF-1, vimentin, MIB-1, PTEN, Ki-67.

Bibliographies of included articles were hand searched, and additional information was sought through targeted grey literature electronic searches (*e.g.,* Google) and review of laboratory compliance and guidance websites (*e.g.,* Clinical and Laboratory Standards Institute, US Food and Drug Administration (FDA), National Guidelines Clearinghouse, Wiley Cochrane Library).

Two reviewers were used at all levels of review (*e.g.,* title/abstract, full article) and for data/information extraction. Conflicts were resolved by discussion or referred to the panel Chair for a decision. When article abstracts or document summaries were not available or a conflict was not resolved, full articles were reviewed.

Selection at all levels was based on predetermined inclusion/exclusion criteria. Included were:

- English-language articles/documents that addressed IHC and provided data or information relevant to one or more KQs;
- Study designs included validation, method comparison, cohort, or case-controlled studies, clinical trials, and systematic reviews, as well as qualitative information from consensus guidelines, regulatory documents or US and international proficiency testing reports; and
- Articles/documents focused on the clinical use of IHC for identification of non-FDA approved predictive and non-predictive markers and analytic variables.

Not included were:

- Non-English-language article/document or an English-language abstract or summary without a full article/document available in English;
- Article/document involves IHC but does not address any KQ;
- Publications with high risk of bias, such as editorials, letters, commentary, invited opinion; and
- Article/documents focused on non-human research, non-tissue IHC (immunoassays, serologic studies), assay optimization or quality control/quality assurance, pre- or post- analytic variables, or clinical validation.

Outcomes of Interest
Outcomes of interest for assessing analytic validity include analytic sensitivity (detection rate), analytic specificity (1-false positive rate), reliability (*e.g.,* repeatability of test results) and assay robustness (*e.g.,* resistance to small changes in pre-analytic or analytic variables). Computing estimates of analytic sensitivity and specificity requires a "gold standard" or well-characterized referent assay (or set of referent specimens with antigen status characterized by previous testing) against which to compare the index, or new, IHC test.[1-3]

Among IHC assays, such "gold standard" referent assays are likely to be the exception rather than the rule.[1] Even HER2 IHC and FISH assays have no "gold standard" at present, as no assay currently available is perfectly accurate in identifying overexpression of this protein.[3]
Consequently, the metric for IHC validation results is most often overall concordance between the results of the new and referent assay(s) for a specific set of validation tissues, or between the results of the new test with previous results for a characterized set of validation tissues. Estimates of positive and negative concordance may also be computed.

We sought quantitative data from primary studies (*e.g.,* validation studies, method comparisons), and systematic reviews of such studies, on concordance, repeatability, reproducibility, and robustness factors (*e.g.,* sample types, fixation). In addition, we sought qualitative information relevant to IHC validation or validation standards from regulatory materials, existing evidence- informed and/or consensus guidelines, and referenced review articles from credible sources.

Data Extraction and Management
The data elements from an included article/document were extracted by one reviewer into standard data formats and tables developed using systematic review database software (DistillerSR, Evidence Partners Inc., Ottawa, Canada); a second reviewer confirmed accuracy and completeness. In all cases, the methodologist acted as either the primary or secondary reviewer. Any discrepancies in data extraction were resolved by discussion with the Methodologist. A bibliographic database was established in EndNote (Thomson Reuters, Carlsbad, CA) to track all literature identified and reviewed during the study.

Environmental Scan
In 2009, CAP recommended strengthening the oversight of laboratory developed tests (LDTs). CAP's proposed changes would incorporate oversight of claims of clinical validity, and specify scientific and regulatory standards to be applied to all LDTs. Risk would be determined based on claims made, potential risk to patients, and the extent to which a test's results could be used in the determination of diagnosis or treatment. The FDA convened a public meeting in July 2010 to discuss issues and stakeholder concerns surrounding LDT oversight. As of submission date of the manuscript (October 2013), no further information is available.[4,5]

Quality Assessment
Grading the quality of individual studies was performed based on study design-specific criteria by the methodology consultant, with input as needed from the TEP. Quality assessments were summarized for each study and recorded in the database. 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."[2,6] 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 based on study design, execution, analyses and reporting.[2] Discordant decisions were resolved through discussion or third-party adjudication.

The hierarchy of data sources and criteria for grading quantitative studies were based on published methods (Appendix, Table 1).[2,7] Studies were rated: Good (no features that suggest flaws or bias); Fair (susceptible to some bias, but flaws not sufficient to invalidate results); or Poor (significant flaws suggesting bias of various types that might invalidate results)(Appendix, Table 2).

Qualitative articles/documents were also assessed using published methods.[8-11] The quality criteria included credibility (*e.g.*, sources, level of review, potential for bias), transferability (*i.e.,* potential for broader application) dependability (*e.g.*, findings stable over time or and/or different methods) and confirmability (*i.e.*, findings consistent and/or verified). Documents were rated: Good (*e.g.,* published/peer-reviewed, from an informed consensus process or professional/advisory committee report); Fair (*e.g*., from credible source with unknown level of peer review, report/guideline from known expert(s) with no observed bias, otherwise Good documents with a flaw or bias); or Poor (*e.g.,* document lacking information on source, peer review, potential bias, referencing, or updating; or having multiple flaws or possible biases).

The strength of evidence for individual KQs or outcomes was assessed using published criteria.[2] 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.[2] Strength of evidence was graded Convincing, Adequate or Inadequate (Table 2).

Table 2. Grades for Strength of Evidence

Convincing

Two or more Level 1[a] or 2 studies (study design and execution) that had an appropriate number and distribution of challenges[b] and reported consistent[c] and generalizable[d] results.

One Level 1 or 2 study that had an appropriate number and distribution of challenges and reported generalizable results.

Adequate

Two or more Level 1 or 2 studies that lacked the appropriate number and distribution of challenges OR were consistent but not generalizable.

Inadequate

Combinations of Level 1 or 2 studies that show unexplained inconsistencies OR one or more lower quality studies (Level 3 or 4) OR expert opinion.

a Table 1 in the Appendix provides the hierarchy of data sources for analytic validation that define Level 1 through Level 4.

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

c Consistency can be assessed formally by testing for homogeneity, or, when data are limited, less formally using central estimates and range of values.

d Generalizability is the extension of findings and conclusions from one study to other settings. Reprinted by permission from Macmillan Publishers Ltd: Genetics in Medicine[2], copyright 2009

Data Analysis

Both quantitative and qualitative methods could be used. Qualitative analysis focuses on identification of themes and patterns within and among non-study related articles and documents, descriptive narrative, content and/or logical analysis.[10,12,13] Quantitative analyses were involved collection of data from validation or method comparison studies into simple data tables or contingency tables (2x2 or 3x3).

Estimates of overall and positive and negative concordance with 95% confidence intervals (CI) can be computed from the contingency tables (Figure 1, Table 3). Overall concordance, also known as percent agreement, is a measure used for comparison of the results of the new test to those obtained using a non-gold standard referent assay (or an "imperfect standard").[14] This measure is based on the major diagonal (Figure 1, upper left cell to lower right cell). The Kappa statistic can be used to test if the major diagonal counts are significantly larger than those expected by chance alone (BMDP Statistical Software, Los Angeles, CA). *Negative concordance* measures the proportion of "negative" samples in which the index test is negative.[14] *Positive concordance* measures the proportion of "positive" samples in which the index test is positive.[14] These last two measures are analogous to analytic sensitivity and specificity, but are used in situations in which the "true" status (marker negative or positive) is not known.

Discordance is a measure based on the "off" diagonal (Figure 1, upper right to lower left) of the contingency table that focuses on discrepancies between results from different assays. In data sets of sufficient size, McNemar's test may be used to determine whether a discordant result between the two tests in one direction (*e.g.,* referent negative and new test positive) is equal to a discordant result in the other direction. A significant value ($p < 0.05$) indicates a lack of symmetry and a potential bias between the two assays. McNemar's test can be performed on data from a 2x2 table (GraphPad Quick Calc, http://www.graphpad.com/quickcalcs/McNemar1.cfm) or extended to three dimensions for a 3x3 table (BMDP Statistical Software).

Assay robustness may be tested by comparison of results between a "standard" IHC component (*e.g.,* fixative 10% neutral buffer formalin) and an alternative (*e.g.,* other fixative) and is generally measured by concordance with a 95% CI. For all comparisons, summary estimates of concordance (random effects model) may be possible, with assessment of heterogeneity and potential for publication bias (Comprehensive Meta-Analysis, Biostat Inc). Precision, or repeatability, is a measure of result agreement between specimens tested on different days.[14,15]
Reproducibility is a measure of agreement between a set of test results interpreted by different pathologists (*i.e.*, inter-rater) or performed in different laboratories.[14,15] Both are generally reported as percent concordance with a 95% CI and/or Kappa statistic.

Figure 1. Comparison of a new or index IHC to a validated IHC or alternative method in a 2x2 contingency table

	Referent IHC Positive	Referent IHC Negative	
Index IHC positive	TP	FP	Total index positive
Index IHC negative	FN	TN	Total index negative
	Total positive	Total negative	Total N

Abbreviation: IHC=Immunohistochemical; TP=True Positive; TN= True Negative; FP=False Positive; FN= False Negative; N= Number

Results

Among the 1,463 citations identified by electronic and hand searches, 126 were selected for inclusion. These included 122 published peer-reviewed articles, 2 book chapters and 2 grey literature documents (Appendix – Figure 1). Among the extracted documents, 43 articles/documents did not meet minimum quality standards, presented incomplete data or data that were not in useable formats, and included only information based on expert opinion. These articles were not included in analyses or narrative summaries. Three general categories of articles/documents were identified.

The first category was published validation and/or method comparison studies on clinical IHC assays. The second category included published, web-based and proprietary guidelines addressing IHC standardization or best practices in general, or guidance on validation and standardization of specific IHC assays (*e.g.,* HER2, ER, PgR). These guidelines were largely qualitative reports based on varying combinations and levels of evidence review and expert opinion. The third category consisted of reported studies on inter-laboratory comparisons, external proficiency testing for common IHC assays or laboratory surveys reporting current laboratory validation practices.

Table 3. Measures of Analytic Validity

Measure	Computation from 2x2 Table	Computation from 3x3 Table
Overall concordance or percent agreement	TP + TN / TP + FP + FN + TN	Sum of concordant cells (major diagonal) / Total N[1]
Overall discordance	FP + FN / TP + FP + FN + TN	Sum of 5 discordant cells / Total N
Positive and negative concordance or percent agreement	Positive = TP / (TP + FN) Negative = TN / (TN + FP)	Not applicable unless 3x3 table can be collapsed[2] to 2x2 or all 2+ samples are excluded

[1] Some studies using tests that report equivocal results (*e.g.,* 3+ positive, 2+ equivocal and 0-1+ negative) include all results as relevant to understanding the relationship between the two tests. However, a major guideline notes that equivocal cases are not expected to be 95% concordant, and cells with discordant results may be omitted. [2] Collapsed by authors' classification of equivocals as positive or negative.
Abbreviation: TP=True Positive; TN= True Negative; FP=False Positive; FN= False Negative; N= Number

KQ 1: When and how should IHC validation assess analytic sensitivity, analytic specificity and precision (e.g., inter-run, inter-operator)?
Note: 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 and comparing the results with the graded responses.

Laboratories are required by the Clinical Laboratory Improvement Amendments of 1988 (Sec. 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.[16] "*Validation* means confirmation by examination and provision of objective evidence that the particular requirements for a specific intended use can be consistently fulfilled.[17] This includes establishment of the analytic validity of all non FDA-cleared/approved (or "laboratory developed") tests.[16]

Analytic validity has been defined as the ability to accurately and reliably identify or measure the marker of interest in specimens that are representative of the clinical population to be tested.[2,6] The concept of validation specimens that are "representative of the patients to be tested" is a key accepted premise or "first principle" of assay validation.[18] The 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.[2,19] To achieve generalizability of the laboratory's analytic validation results, the tissues included in a validation set must be typical of the specimens received in routine practice and must provide a representative range of expression intensities and patterns.

The strength of evidence was Adequate to support Recommendation 6: that laboratories should, whenever possible, use the same fixative and processing methods as cases tested clinically, in order to validate using representative specimens.

Components of analytic validity applicable to IHC assays are accuracy, analytic sensitivity (detection rate) and specificity (1-false positive rate), concordance (overall, positive, negative) and precision (repeatability, reproducibility).[2,6,15,16] Analytic sensitivity and specificity are estimated by comparing a new assay's results with a "gold" standard referent test or validated tissue set. However, "gold" standard" referent tests for IHC assays are rare. For example, no confirmatory or "gold standard" test currently exists for HER2, ER and PgR IHC and these results do not represent "truth".[1,3,15,20] A HER2 *in-situ* hybridization assay (*e.g.,* FISH, CISH, SISH) can only indirectly validate a HER2 IHC test, because a nucleic acid based assay does not measure the same analyte.

Therefore, laboratories must use other approaches to demonstrate assay performance. Primary validation and method comparison studies and key published professional guidelines described IHC validation approaches.[3,15,18,21-39] They included comparisons of a new test's results to: clinical outcomes; to other validated IHC tests, to or other referent tests (intra- or inter- laboratory); or to tissue validation sets previously characterized by consensus.[20,22,30-32,34,40-51] Based on these studies, the standard metrics for IHC validation results are overall concordance between the results of the new and referent assay(s), the Kappa statistic, and positive and negative concordance for assays with binary results (positive, negative) that can be entered into a 2x2 table (Table 3). Quality grades for studies referenced here were 2 Good, 22 Fair, and 6 Poor; grades for 8 other articles/documents were 2 Good and 6 Fair.

The strength of evidence was Adequate to support the KQ 1 outcome of when analytic validation should be done, and that it should include analytic sensitivity and specificity (or concordance in absence of a "gold" standard referent test).

The evidence was Inadequate (*i.e.,* evidence was not available or did not permit a conclusion to be reached) for the KQ 1 outcome of how validation should be done with regard to the listed approaches, but did show that these approaches have been used.

The precision of an IHC assay, or result repeatability, is the extent of agreement among results (*i.e*., positive/negative results, staining patterns/localization, level of expression) obtained by replicate testing of tissue specimens under specified conditions.[14,15] Reproducibility assesses the extent of agreement among results obtained by replicate testing of specimen sets between laboratories, testing platforms or readers.[14,15] Evaluation of precision is an element required by CLIA, and CLSI IHC-specific guidance states that IHC assay validation requires acceptable precision in the analytical (*e.g.,* result repeatability over days) and postanalytical/interpretive (*e.g.,* inter-operator reproducibility) phases.[15,16]

However, no studies were identified that provided data on assay repeatability over two or more days. One guidance document recommended running validation samples over multiple days, with no more than 20 samples tested in one day.[37] Based on a recent CAP survey, the proportion of laboratories that agree with"…validation cases tested on multiple days to assess between-run precision" was 53% and 57% for non-predictive and predictive assays, respectively.[52] Since over half of laboratories support this, a possible reason for lack of identified studies may be that this step is considered too routine for inclusion in publications. Another possibility is that studies containing this information were published in the early years of IHC testing and were not captured in the post-2004 search.

A small number of studies and guidance documents addressed reproducibility. Two guidance documents have called for ongoing monitoring of the competency of histotechnologists and pathologists by measuring inter-rater reproducibility.[3,37] One recommended that the laboratory director determine the timing and standards for competency testing, while another called for 95% concordance as the standard for inter-operator or inter-laboratory reproducibility.[3,38] Five studies were identified that reported inter-rater and/or inter-laboratory reproducibility.[49,53-56] However, the differences between the study protocols were so numerous that no conclusions were possible. For example, the studies tested different markers (HER2, PTEN, multiple), compared different numbers of raters (2 to 6) and laboratories (2-3), and variably expressed results as coefficients of variation, percent concordance, Kappa statistic, weighted Kappa statistic and "composite ratings." No raw data were available to allow reanalysis.

Quality grades for studies referenced here were 3 Fair and 2 Poor; 1 document was graded Good and 3 Fair.

The strength of evidence for the KQ 1 outcome of precision was Adequate to support inclusion of precision (e.g., inter-run and inter-operator) as part of validation. The evidence was Inadequate to assess the precision of IHC assays in practice.

The strength of evidence was Adequate to support Recommendation 1: "Laboratories must validate all immunohistochemical tests before placing into clinical service."

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.

KQ 2 and KQ 3: What is the minimum number of positive cases (KQ 2) and negative cases (KQ 3) that need to be tested to analytically validate an immunohistochemical assay? Does the minimum number differ depending on whether the IHC assay:

- Is primarily used to identify cell lineage (*i.e.,* non-predictive markers)?
- Is used to direct patient treatment (*i.e.,* predictive markers)?
- Is used to identify an infectious organism?
- Is used to identify rare antigens?
- Is done on cytology specimens?
- Is done on decalcified specimens?

"The perennial question is, 'How many samples do I need to run to validate a given test?' Unfortunately, the answer is always the same—**it depends**. It depends on "…how the test is to be used, which performance criteria are most critical for the intended use, and the confidence level that is required for good medical practice, implying that medical judgment is required."[57]

A first step in addressing this question is to consider what criteria are most likely to impact the number of samples needed to validate IHC assays overall, and for the specific intended uses and specimen types listed above.

Intended Use

Class I tests have been defined as interpreted by pathologists in the context of histomorphologic, cytomorphologic and clinical data and reported as one part of a panel of tests or clinical evaluation.[15,58-60] Class I tests may also be referred to as *non-predictive* or *qualitative*, though they may have a quantitatively defined threshold (*e.g.*, >10% reactive cells).[59] In contrast, Class II tests are generally stand-alone tests with no routine morphologic correlates.[58] Class II test results are reported to physicians as independent diagnostic information, and may influence treatment decisions.[15,59,60] *Predictive IHC tests* fall into Class II.

Based on intended use, tests could be classified as predictive or non-predictive for purposes of validation standards. Of course, some tests can fall into both categories, depending on intended use. For example, CD117 can be considered Class I as an acute leukemia marker of myeloid differentiation, and Class II in assessing a stromal gastroesophageal tumor to determine the patient's eligibility for imatinib treatment.[61] Other criteria for determining number of validation samples include the complexity of interpretation (*i.e.,* multiple outcomes require more samples) and feasibility (*i.e.,* the number and range of control materials may be limited, especially for some non-

predictive tests).[15] In addition, the observed concordance and possible bias between tests in the initial validation may necessitate further testing and, possibly, additional validation specimens.[59]

No studies were identified that addressed the four specific intended uses listed in KQ 2 and KQ 3, but classifying tests' intended use as predictive or non-predictive provides a rationale for determining the number of samples needed for validation. Due to the potential for direct impact on clinical management, it is not surprising that predictive tests appear to require higher certainty in the quality of validation results.[18,37]

Strength of evidence was Adequate to support an outcome of KQ 2 and KQ 3, the decision to distinguish between non-predictive (Class I) and predictive (Class II) IHC tests in determining the recommended number of validation samples.

Strength of evidence was Adequate to support the separation of Recommendation 3 and Recommendation 4 in order to distinguish between non-predictive and predictive IHC tests for determining the recommended number of validation samples.

Strength of evidence was Adequate to support Recommendation 5, regarding use of the higher validation standard (*e.g.*, number of samples) in the case of a marker with both non-predictive and predictive intended uses.

Information on Numbers of Samples for Validation

Available information on the recommended number of samples needed for validation was limited. Suggested numbers were found in four professional society clinical guidelines (quality grade Fair), two consensus meeting reports (grade Fair), and one CLSI approved guideline (grade Fair).[3,15,18,37,38,59,62] Note that four of these documents focused on specific predictive tests (HER2, ER, PgR), and three on IHC assays in general.[3,15,18,37,38,59,62] Guidance on numbers of samples:

Minimum 25 samples, 10 high, 10 intermediate, 5 negative[38]

25-100 samples (no breakdown)[3,62]

50-100 samples, 25-50 positive with an unspecified mix of weak positives, 25-50 negative[59]

≥ 80 samples, ≥ 40 positive (10 weak positive), ≥ 40 negative[15,18,37]

In the absence of clear guidance on the number of validation samples to run, the Methodologist requested help from Women & Infants Hospital statistician (Glenn E Palomaki, PhD) to develop tables to assist the panel in discussing this important question. Practical guidance on the size of a validation set can be provided by statistical analysis. Simply put, the more samples that are run in a validation set, the higher the likelihood that the concordance estimate reflects the test's "true" concordance. But to apply and test this approach, it was necessary to determine what concordance benchmark would be used. The concordance benchmarks commonly mentioned in guidance documents are 90% and 95%. We reviewed available validation and method comparison studies to identify data that might support the selection of a benchmark.

Determining a Concordance Benchmark

Supporting evidence was identified in studies and documents reporting "real world" concordance data from IHC validation studies, method comparisons and proficiency testing or interlaboratory comparisons. The following is a summary of analyses. More detailed data can be found in the Appendix, Tables 3-5.

Data were analyzed from a two-year inter-laboratory comparison of CD117 IHC testing.[61] Ten blinded tissues were run in 2004 by 63 laboratories, and again in 2005 by 90 laboratories. The set included

four gastrointestinal stromal tumors (GIST) positive for CD117 and six tumors that were negative by histopathologic diagnosis. For the combined 1,530 challenges, the concordance estimate between the laboratory responses and the target diagnosis was 88% (95% CI 86-89; k=0.75). Results for 2004 and 2005 were not statistically different. Positive concordance was 98% and negative concordance was 81%. The McNemar's statistic was p<0.001, confirming that the observed asymmetry in discordant results (12 false negatives and 177 false positives) was significant. Possible explanations included the presence of necrotic foci or CD-117 positive mast cells in normally CD117 negative tumors (*e.g.,* leiomyosarcoma) or the variability in primary antibodies and antigen retrieval methods for tests between laboratories.

Data from comparisons of HER2 IHC assays were analyzed. Median overall concordance in 5 comparisons of different HER2 IHC tests was 89% (range 74–93%), with 2 of 5 studies greater than 90% concordant (Appendix, Table 3).[22,30-32,34] Note that concordance estimates and associated Kappa and McNemars statistics were computed from 3x3 contingency tables (BMDP Statistical Software, Los Angeles, CA).

The summary concordance estimate (random effects model) was similar at 88.1% (95% CI 81.3-92.7), but heterogeneity was high (I^2 =89, p < 0.001), and could not be explained by analysis of selected covariates (*e.g.,* tissue type, study size, study quality grade). The number of studies was too small to allow analysis for the many possible covariates. One study was rated Good and 4 Fair. The McNemar's p values < 0.05 indicate a significant difference/bias between the false positive and false negative discordant results in a number of these comparisons. Such information can be helpful for next steps in validation.

Data were analyzed from comparisons between HER2 IHC assays and *in situ* hybridization tests (*e.g.,* FISH). Median overall concordance in 7 comparisons from the four identified studies in breast cancer tissue was 89% (range 66–94%), with 2 of 7 studies > 90% concordant (Appendix, Table 4).[31,34,42,49] Three studies used The HER2 4B5 primary antibody and three used CB11. Within the limitations of the small number of studies, the results for each antibody were consistent with the overall estimate. The summary concordance estimate (random effects model) was similar at 88% (95% CI 81-93), but heterogeneity was high (I^2 =89, p < 0.001), and could not be explained by analysis of selected covariates (*e.g.,* tissue type, study size, study quality grade). The number of studies was too small to allow analysis for the many possible covariates. There was a suggestion of publication bias (Egger's p=0.002) that became insignificant when the largest study was removed (a LDT with the lowest concordance of 66%, k=0.37 and McNemar's p<0.001).[42] The quality grade for all studies was Fair.

The median concordance estimate for 4 comparisons in 3 studies of HER2 IHC and *in situ* hybridization in gastric cancers was 95% (range 88-98%), with 3 of 4 studies >90% concordant.[22,43,44] The grade for the studies was 2 Good, 1 Fair and 1 Poor.

Analyses of data from comparisons between HER2 IHC tests and alternative referent tests. Median overall concordance from 4 studies of IHC tests (ER, PR, HER2, p16) compared to alternative referent tests (*e.g.,* RNA expression, clinical diagnosis, consensus results) was 87% (range 72–95%), with 1 of 4 studies >90% concordant (Appendix, Table 5).[20,40,45,46]

These data illustrate the challenge of achieving an overall concordance of 95%, even in relatively large studies almost entirely made up of IHC tests with guidance recommending stringent protocol standards (*i.e.,* HER2, ER, PgR).[3,37,39,59] An overall concordance standard that is too stringent could have the effect of delaying or preventing successful validation, particularly for non-predictive tests. Overall concordance of 90% was achieved in nearly half of the above analyzed

comparisons, all of which were subject to many sources of variation (*e.g.,* 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.

Strength of evidence was considered Adequate to support the adoption of a 90% (versus 95%) overall concordance benchmark as an outcome for KQ 2 and KQ 3.

Strength of evidence was Adequate to support Recommendation 2 for a 90% overall concordance benchmark for analytic validation of IHC tests (excepting HER2, ER, PgR).

Considering the number of tissues needed for a validation set
The basic statistical premise is that the more samples that are run in a validation set, the higher the likelihood that the concordance estimate reflects the "true" performance of the test. As an example, 3 discordant results would be expected in a 10 sample validation set for a test with a "true" concordance of 70%. However, only 1 discordant result could be observed by chance, resulting in a concordance over-estimate of 90%. In a 20 sample validation set, 6 discordant results would be expected for the test with a "true" concordance of 70%. Observation of only 2 discordant samples could occur by chance, but the likelihood would be low.

Of course, the premise of "..the more samples the better.." has to be balanced by laboratory feasibility issues such as costs and resources. It is also important to keep the goal in mind – to keep false validation failures low while identifying assays that are truly not performing well.

Table 6 in the Appendix is an example of those considered by the panel. With a 10 sample validation set, the benchmark is reached with only 1 discordant result. The concordance estimate is 90% with a lower 95% confidence limit (L95%) of 57%. The "true" concordance could be lower or higher than 90%, but there is only a small chance (about 5%) that it will be lower than 57%. The validation fails with 2 discordant results. Even with a "true" concordance of 80%, a 10 sample validation set has a greater than 1 in 3 chance of meeting the 90% benchmark, compared to a 1 in 5 chance in a 20 sample validation set. A 20 sample validation set allows 2 discordant results for a 90% concordance estimate with a L95% of 74%, a more confident result.

Consideration of a 20 sample (10 positive, 10 negative) validation set for non-predictive tests
Overall concordance estimates meet the benchmark with 0, 1 or 2 observed discordant results among the total set of 20 tissues (Table 4). The "true" concordance between the two assays has only a 5% chance of falling outside the 95% CI of each concordance estimate, and can be lower or higher than the estimate. If the 100% or 95% concordance estimates (0, 1 observed discordant results) are a "true" representation of the relationship between the two tests, the validation result would meet the benchmark more than 92% of the time (Table 5). If the 90% concordance estimate is "true", the probability of meeting the benchmark would be 68%.

For validation results that do not meet the benchmark, it may not be useful to perform the McNemar's test in a small validation set (*e.g.,* 20 tissues). The McNemar's test is based solely on discordant results, which are likely to be few in a small validation set. Therefore, a non- significant McNemar's test could be due to true symmetry between the number of discordant results, or to asymmetry on the off-diagonal but with insufficient numbers to show statistical significance (*i.e.,* underpowered to find even important differences between the tests). In many cases, a visual inspection of the results in a 2x2 or 3x3 table will identify a potential explanation for the validation failure.

The laboratory medical director will determine any corrective action and how many additional tissues should be tested.

Table 4. Validation Using a 20 Tissue Validation Set (10 Positive and 10 Negative) against a 90% Concordance Benchmark[a]

Number of validation tissues	*0 discordant* Concordance estimate (95% CI)	*1 discordant* Concordance estimate (95% CI)	*2 discordant* Concordance estimate (95% CI)
20 Total	100% (81-100)	95% (75-100)	90% (69-98)

[a] Concordance estimates with 95% CI stratified by number of observed discordant samples
Abbreviation: CI= confidence interval

Consideration of a 40 sample (20 positive, 20 negative) validation set for predictive tests The statistical argument is updated here for predictive factor assays. Table 6 provides overall concordance estimates with 95% CIs for the 40 tissue validation set, as well as the 20 tissue sets for those who will compute positive and negative concordance estimates. Overall concordance estimates (Table 6, shaded row) meet the benchmark with 0 to 4 observed discordant results among the total set of 40 tissues. The "true" concordance between the two assays can be lower or higher than the estimate, but has only a 5% chance of falling outside the 95% CI of the concordance estimate (L95% is 76% for a 90% concordance estimate).

If the 95-100% concordance estimates (0, 1, 2 observed discordant results) are a "true" representation of the relationship between the two tests, the validation results would meet the benchmark more than 95% of the time (Table 5). The probabilities of meeting the benchmark if the 92.5% and 90% concordance estimates are "true" would be 82% (approximation) and 68%, respectively. The positive (or negative) concordance estimates among 20 tissues (bottom row) meet or exceed the same benchmark with 0, 1, or 2 discordant results.

Table 5. The percent probability of meeting or exceeding a specified benchmark concordance rate based on the number of specimens in the validation set and the "true" concordance rate of the assay[a]

Tissues in the Validation Set		"True" concordance rate	Benchmark Concordance Rate
20	**40**		
21	8	80	**90%**
40	26	85	
68	63	90	
92	95	95	
99	>99	98	
7	1	80	**95%**
18	5	85	
39	22	90	
74	68	95	
94	95	98	

[a] StatTrek.com Binomial Calculator and consistent with Wolff et al., 2013[18]

Table 6. Validation Using a 40 Tissue Validation Set (20 Positive and 20 Negative) against a 90% Concordance Benchmark[a]

Number of validation tissues	*0 discordant* Concordance estimate (95% CI)	*1 discordant* Concordance estimate (95% CI)	*2 discordant* Concordance estimate (95% CI)	*3 discordant* Concordance estimate (95% CI)	*4 discordant* Concordance estimate (95% CI)
40 Total	100% (90-100)	97.5% (86-100)	95% (83-99)	92.5% (79-98)	90% (76-97)
20 Positive or Negative	100% (81-100)	95% (75-100)	90% (69-98)	85% (63-96)	80% (58-92)

[a] Concordance estimates with 95% CI stratified by number of observed discordant samples Abbreviation: CI= confidence interval

In a 40 sample validation that does not meet the benchmark, analyses such as the McNemar's test and kappa statistic may help determine whether an observed difference in the off-diagonal represents a significant bias between the new and referent tests (Figure 2). In this case, the kappa statistic showed "substantial" agreement, but the overall concordance estimate missed the benchmark by a small margin. The positive concordance of 75% suggests false negatives could be occurring in the new test. The McNemar's p was 0.13 (not significant), indicating that the 5 discordant results all in a single cell could have happened by chance. Alternatively, the test could be underpowered.

Figure 2. A 2x2 contingency table of a 40 tissue validation set that did not meet the benchmark (results entered into a 2x2 contingency table) with associated statistical tests

New IHC Result	Referent Result Positive	Referent Result Negative	
Positive	**15**	**0**	16
Negative	**5**	**20**	24
	20	20	**40**

Overall concordance = 35/ 40 = 87.5% - Does not meet the 90% benchmark k = 0.75
McNemar's p = 0.13, not significant
Positive concordance = 15/20 = 75%
Negative concordance = 20/20 = 100%
Abbreviation: IHC= immunohistochemical

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's 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 Figure 2 are doubled, the McNemar's result for 10 to 0 asymmetry on the off-diagonal would be significant (P=0.004).

The laboratory medical director will determine any corrective action and how many additional tissues should be tested.

Strength of evidence was Inadequate to support Recommendation 3 and Recommendation 4 in determining the recommended number of validation samples.

Number of specimens in a validation set for IHC tests performed on cytologic specimens.
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. One guideline did recommend that each laboratory should validate IHC assays for cytological specimens separately from those for surgical specimens.[15]

However, studies were identified that compared cytology specimens to FFPE histologic sections for ER, PgR and/or HER2 IHC testing (Appendix, Tables 7-9).[63-68] Concordance estimates and Kappa statistics were consistently high at ≥ 90% and >0.75, respectively. The lack of a significant finding by the McNemar's test may be partly related to small sample size (4 of 5 data sets had 50 or less samples), but positive and negative concordance rates were also reasonably consistent. However, the studies were few, generally small, and used different fixatives, fixation times, and cytology specimens (*e.g.,* smears, ThinPrep, cell blocks). In 3 studies only about 90% of samples were assessable. No two studies could be directly compared.

The strength of evidence was Inadequate (*i.e.,* evidence was not available or did not permit a conclusion to be reached) to address the KQ 2 and KQ 3 outcome of number of samples needed for validation with cytology specimens.

Number of specimens in a validation set for IHC tests performed on decalcified specimens
No primary studies, systematic evidence reviews or qualitative documents (*e.g.,* guidelines, consensus meeting reports) were identified that addressed 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 addressed the potential influence of decalcification as a modifier in the analytic validation process.[15,39,48,69-74] Some reported significant variability in decalcification protocols (*e.g.*, decalcification solutions, time in solution) and in preservation of antigenicity in IHC tests.[70-73] One inter-laboratory survey in Europe reported that 68% of laboratories used the same protocols for decalcified bone biopsies as for non-decalcified tissues.[73] Two IHC guidelines recommend interpreting IHC results on decalcified samples with caution regarding the possibility of antigen (and tissue) loss.[15,39] However, others reported good morphology and successful staining with protocols using different fixatives, acid or EDTA decalcification, and paraffin or resin embedding.[48,69,72,74]

These variable observations emphasize the need for a defined protocol and a validation plan that will ensure robust and reproducible IHC results in decalcified specimens.

The strength of evidence was Inadequate to address the KQ 2 and KQ 3 outcome of number of samples needed for validation with decalcified specimens.

KQ 4. What parameters should be specified for the tissues used in the validation set?
Set ratio of immunoreactive versus non-immunoreactive? Set ratio of high expressors versus low expressors?
Set ratio of neoplastic versus non-neoplastic (when appropriate)?
Should a minimum tissue size or minimum quantity of cells be specified?

No primary studies, systematic evidence reviews or qualitative documents (*e.g.,* guidelines, consensus meeting reports) were identified that addressed the specific question regarding the parameters that should be specified in validation sets with regard to neoplastic versus non- neoplastic tissues.

Several guidelines have suggested a 50:50 ratio of immunoreactive versus non-immunoreactive tissues.[3,15,18,37] Information on number of low or weak expressors versus high expressors is similarly unspecified. In a recent CAP survey, participating laboratories reported that the median proportion of positive validation cases that were "weakly or focally" positive was 20% for non- predictive (N=195 respondents) and predictive (N=141) assays.[52] The reported median number of positive samples run for non-predictive assay validation was 7 (10^{th}-90^{th} centiles=2-20), of which 1-2 would be weakly positive. For predictive assay validation, the median number of positives samples was 10 (10^{th}-90^{th} centiles=2-30), of which 2 would be weakly positive. It appears this approach would lead to low certainty regarding validation results.

There was no specific guidance on sample size, but of 34 reviewed studies that reported whole section size, the results were 18%, 47% and 21%, respectively, for 3 um, 4 um and 5 um; the remaining 5 studies reported ranges of 2-4 um (N=3)or 4-6 um (N=22).[23,24,26-28,30,31,42,44,46,49,56,66,67,69,75-87]

Reports from 8 studies on core size for TMAs ranged from 0.6 to 3 mm.[15,34,41,79,88-91] No other articles addressed minimum tissue size or quantity of cells. A related question was raised about the comparison of TMAs with different sizes and number of cores to whole sections.

The strength of evidence was Inadequate to address other KQ 4 outcomes regarding four specific parameters for tissues in a validation set.

Comparisons of concordance between IHC assays performed on whole sections and TMAs
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.[21,23- 29,33,35,36,92] Summary estimates of concordance (random effects model) were computed, but heterogeneity was high across the studies (I^2 >75; $p < 0.001$), and specific sources of heterogeneity could not be identified. Consequently, concordance is reported as ranges with median values.

The median overall concordance estimate was 93% (range 73-100%)(Appendix, Table 10). Data were stratified by study quality, marker (Appendix, Table 11) and core size (Appendix, Table 12) as possible sources of heterogeneity. All results were consistent between quality scores, markers and core sizes. Concordance estimates met or exceeded the 90% benchmark in about two thirds of cases. Table 13 provides limited data on other markers. The quality of studies was 8 Fair and 4 Poor.

Strength of evidence was Inadequate to recommend the routine use of TMA samples. Strength of

evidence was Adequate to support the conclusion that TMA samples have been successfully utilized in IHC tests, but there are many variables to be considered and thorough validation is needed for each marker.

The strength of evidence was Adequate to support Recommendation 9 regarding the need for careful validation to determine if TMAs are appropriate for the targeted antigen and the fixation and processing is similar to clinical specimens.

KQ 5. How do the following modifiers influence analytic validation?
Type of fixative
Type of decalcification solution Time in decalcification solution
Validation tissues processed in another laboratory

No primary studies, systematic evidence reviews or qualitative documents (*e.g.,* guidelines, consensus meeting reports) were identified that addressed the specific question regarding the potential influence on validation of tissues processed in another laboratory.

Nine articles and documents addressed the potential influence of the type and timing of decalcification as a modifier in the analytic validation process.[15,39,48,69-74] Some reported significant variability in decalcification protocols (*e.g*., decalcification solutions, time in solution) and in preservation of antigenicity in IHC tests.[70-73] Two IHC guidelines recommend interpreting IHC results on decalcified samples with caution regarding the possibility of antigen (and tissue) loss.[15,39] However, others reported good morphology and successful staining with protocols using different fixatives, acid or EDTA decalcification, and paraffin or resin embedding.[48,69,72,74] These observations emphasize the need for a defined protocol and a validation plan that will ensure robust and reproducible IHC results in decalcified specimens.

Strength of evidence was Inadequate to address the KQ5 outcomes regarding the influence of the type of decalcification solution, the time in decalcification solution, or validation tissues processed in another laboratory on analytic validation.

The influence of the type of fixative on analytic validation
The authors of a 2011 article reviewed 39 primary studies that investigated preanalytical variables identified by a literature survey.[93] Among 15 preanalytical variables with the potential to impact IHC assays were time to fixation (cold ischemic time), fixative type (*e.g.,* concentration, pH), and time in fixative. Studies have shown that fixation delay of more than 12 hours affects the extent and intensity of immunostaining, possibly leading to false negative results.[93] Another report found that delays of even 1-2 hours may decrease signal intensity in ER, PgR and HER2.[18,93] One IHC guideline recommends a less than 1 hour delay when possible, but certainly as short a delay as possible.[39]

The most commonly recommended fixative is 10% neutral buffered formalin (NBF), but most studies have focused on a narrow range of IHC assays (e.g., ER, PgR, HER2) in one tissue. The fixative used can affect the extent and intensity of staining as well as nonspecific background staining, and antigen specific effects have been reported.[93] Time in fixative can also affect the extent, distribution and intensity of staining, and may be antigen dependent. Fixation for limited periods beyond 72 hours has not resulted in a reduction in assay sensitivity in several studies assays, and effective antigen retrieval may maintain immunoreactivity even after fixation for several days.[76,92,94,95]

The available data are, with some exceptions, focused on IHC hormone markers that help inform

treatment options for women with breast cancer. However, this review is intended to provide information to inform recommendations on analytic validation for a wide range of non-predictive and predictive markers. The available data may, in fact, be applicable to a wide range of antigens. In the meantime, however, careful validation will help determine when antigen specific protocol changes may be needed for these preanalytic variables.

Strength of evidence was Inadequate (*i.e.,* evidence was not available or did not permit a conclusion to be reached) to address the KQ 5 outcome regarding the influence of fixation on analytic validation.

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 tissues.

KQ 6: Which of the following conditions require assay revalidation?

New lot of antibody
Change in antibody clone
Change in antibody dilution
Change in type of fixative
Change in antigen retrieval method
Change in antigen detection system Change in instrumentation
Change in water supply
Laboratory relocation
Assay no longer performing as expected

KQ 7: Does assay revalidation have the same requirements as initial assay validation?

Available information on the conditions or changes that require assay revalidation was limited. In general, revalidation was recommended for "any significant changes to an assay/test system" or "any deviation from a standardized method" This recommendation was found in four professional society clinical guidelines (quality grade Fair), two consensus meeting reports (grade Fair), and one CLSI approved guideline (grade Fair).[3,15,18,37-39,62] Note that four of these documents focused on specific predictive tests (HER2, ER, PgR), and three on IHC assays in general.[3,15,18,37-39,62] Some of these documents also recommended revalidation for specific changes (Table 7).

Two guidelines recommended a limited revalidation for a new primary antibody lot.[38,59] Among CAP Survey responders, 64% believed revalidation should be done for a new lot of primary antibody in predictive tests, but whether a full or limited validation was not questioned.[52] Two guidelines recommended scheduled revalidation, one semi-annually and one annually.[3,39] No guidelines addressed change in antibody dilution, change in water supply, laboratory relocation, and assay no longer performing as expected.

No primary studies with data supporting the consensus expert opinions were identified. Three of the expert consensus guidelines were informed by an evidence review, but no references supported the guidance about revalidation.[3,18,37] This guidance is based on qualitative information derived from expert opinion and principles of good laboratory practice. It is possible that studies documenting clinically significant result variation based on the effects of the listed changes predate 2004, or would need different search terms to be identified.

No specific information was identified that addressed whether the requirements of revalidation are the

same as initial assay validation. The term "revalidation" is not included in the CLSI Harmonized Terminology Database.[14]

Table 7. Referenced guidance on specific changes requiring revalidation and responses from laboratories who agreed revalidation of predictive tests should be done for those changes

Specific changes requiring IHC revalidation	2010 CAP Survey[52] Non-HER2 predictive assays % responding revalidation should be done (N)
Modification of a commercial kit[15]	NA
Primary antibody clone[15,37,39,59]	NA
Primary antibody provider[59]	NA
Change between in-house primary antibody dilution and pre-dilution[59]	NA
Fixative/fixation method[15]	74 (295)
Antigen retrieval method[15,37,39,59]	80 (294)
Detection system[15,37,39,59]	81 (293)
Instrumentation	78 (296)
Autostainer[51]	Tissue processor, 55 (292)
Addition/change in imaging system[51]	NA
Relaxation of quality management procedures [37]	NA

Abbreviation: N = number of respondents for that question; [2]NA =this change was not part of the survey

The strength of evidence was Inadequate to address KQ 6 on conditions requiring assay revalidation and KQ 7 on whether revalidation should be the same as initial validation.

The strength of evidence was Inadequate to support Recommendation 10, Recommendation 11, Recommendation 12 or Recommendation 13.

REFERENCES

1. Marchio C, Dowsett M, Reis-Filho JS. Revisiting the technical validation of tumour biomarker assays: how to open a Pandora's box. *BMC Med.* 2011;9:41.
2. 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.
3. 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.
4. Paxton A. Prepping for a firmer FDA hand in regulating LDTs. *CAP Today*. 2010;24(9):5-12.
5. US Food and Drug Administration, Center for Devices and Radiological Health. Oversight of Laboratory Developed Tests (LDTs). 2010. Available at: http://www.fda.gov/medicaldevices/newsevents/workshopsconferences/ucm212830.htm. Accessed October 4, 2013.
6. Haddow JE, Palomaki GE. ACCE: a model process for evaluating data on emerging genetic tests. In: *Human Genome Epidemiology: A scientific foundation for using genetic information to improve health and prevent disease*. Oxford:Oxford University Press; 2003:217-233.
7. Whiting PF, Weswood ME, Rutjes AW, et al. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Med Res Methodol*. 2006;6:9.
8. Agency for Healthcare Research and Quality, Robert Wood Johnson Foundation. Conference on Qualitative Methods in Health Services Research. Rockville, MD, December 4, 1998. Available at: http://archive.ahrq.gov/about/cods/codsqual.htm. Accessed October 16, 2012.
9. Bowen GA. Document Analysis as a Qualitative Research Method. *Qual Res*. 2009;9(2):27-40.
10. Leys M. Health care policy: qualitative evidence and health technology assessment. *Health Policy* 2003;65(3):217-226.
11. Murphy E, Dingwall R, Greatbatch D, Parker S, Watson P. Qualitative research methods in health technology assessment: a review of the literature. *Health Technol Assess*. 1998;2(16):iii-ix, 1-274.
12. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Info Libr J*. 2009;26(2):91-108.
13. Centre for Reviews and Dissemination. *Systematic Reviews: CRD's guidance for undertaking reviews in health care*. York, England:CRD, University of York, 2009.
14. Clinical Laboratory Standards Institute (CLSI) Harmonized Terminology Database. Available at: http://login.clsi.org/HTDatabase.cfm. Accessed May 29, 2013.
15. 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.
16. 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.
17. US Food and Drug Administration. Title 21 CFR § 820.3(z). Available at: http://www.gpo.gov/fdsys/pkg/CFR-2012-title21-vol8/pdf/CFR-2012-title21-vol8-chapI-subchapH.pdf. Accessed September 2, 2013.
18. 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. *In press*. 2013.
19. 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 Evience-based Practice Center under contract no. HHSA 290-20007-10063-I.) AHRQ Publication No. 11-EHC048-EF. Rockville, MD: Agency for Healthcare Research and Quality. June 2011.
20. 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.
21. 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.

22. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. Henriksen KL, Rasmussen BB, Lykkesfeldt AE, Moller S, Ejlertsen 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.
28. 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.
29. 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.
30. 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.
31. Moelans CB, Kibbelaar RE, van den Heuvel MC, Castigliego 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.
32. 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.
33. 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.
34. 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.
35. 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.
36. 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 Molecul Morphol*. 2008;16(4):362-370.
37. 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.
38. Goldstein NS, Hewitt SM, Taylor CR, Yaziji H, Hicks DG, Members of Ad-Hoc Committee On Immunohistochemistry Standardization. Recommendations for improved standardization of immunohistochemistry. *Appl Immunohistochem Mol Morphol*. 2007;15(2):124-133.
39. 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.
40. 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.
41. 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.

42. Grimm EE, Schmidt RA, Swanson PE, Dintzis SM, Allison KH. Achieving 95% cross- methodological concordance in HER2 testing: causes and implications of discordant cases. *Am J Clin Pathol*. 2010;134(2):284-292.
43. Hofmann M, Stoss O, Shi D, et al. Assessment of a HER2 scoring system for gastric cancer: results from a validation study. *Histopathology*. 2008;52(7):797-805.
44. Sornmayura P, Rerkamnuaychoke B, Jinawath A, Euanorasetr C. Human epidermal growth-factor receptor 2 overexpression in gastric carcinoma in Thai patients. *J Med Assoc Thai*. 2012;95(1):88-95.
45. Jordan RC, Lingen MW, Perez-Ordonez 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.
46. 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.
47. 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.
48. 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.
49. Powell WC, Hicks DG, Prescott N, et al. A new rabbit monoclonal antibody (4B5) for the immuno-histochemical (IHC) determination of the HER2 status in breast cancer: comparison with CB11, fluorescence in situ hybridization (FISH), and interlaboratory reproducibility. *Appl Immunohistochem Mol Morphol*. 2007;15(1):94-102.
50. 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.
51. 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*. 2009;7(Suppl 6):S1-S21; quiz S22-23.
52. 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.
53. Emerson LL, Tripp SR, Baird BC, Layfield LJ, Rohr LR. A comparison of immunohistochemical stain quality in conventional and rapid microwave processed tissues. *Am J Clin Pathol*. 2006;125(2):176- 183.
54. Gustavson MD, Bourke-Martin B, Reilly D, et al. Standardization of HER2 immunohistochemistry in breast cancer by automated quantitative analysis. *Arch Pathol Lab Med*. 2009;133(9):1413-1419.
55. Ruschoff J, Dietel M, Baretton G, et al. HER2 diagnostics in gastric cancer-guideline validation and development of standardized immunohistochemical testing. *Virchows Arch*. 2010;457(3):299-307.
56. Sangale Z, Prass C, Carlson A, et al. A robust immunohistochemical assay for detecting PTEN expression in human tumors. *Appl Immunohistochem Mol Morphol*. 2011;19(2):173-183.
57. Jennings L, Van Deerlin VM, Gulley ML. Recommended principles and practices for validating clinical molecular pathology tests. *Arch Pathol Lab Med*. 2009;133(5):743-755.
58. 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.
59. Torlakovic EE, Riddell R, Banerjee D, et al. Canadian Association of Pathologists-Association canadienne des pathologistes National Standards Committee/Immunohistochemistry: best practice recommendations for standardization of immunohistochemistry tests. *Am J Clin Pathol*. 2010;133(3):354- 365.
60. Hsi ED. A practical approach for evaluating new antibodies in the clinical immunohistochemistry laboratory. *Arch Pathol Lab Med.* 2001;125(2):289-294.
61. 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.
62. Hanna W, O'Malley F P, Barnes P, et al. Updated recommendations from the Canadian National Consensus Meeting on HER2/neu testing in breast cancer. *Curr Oncol*. 2007;14(4):149-153.

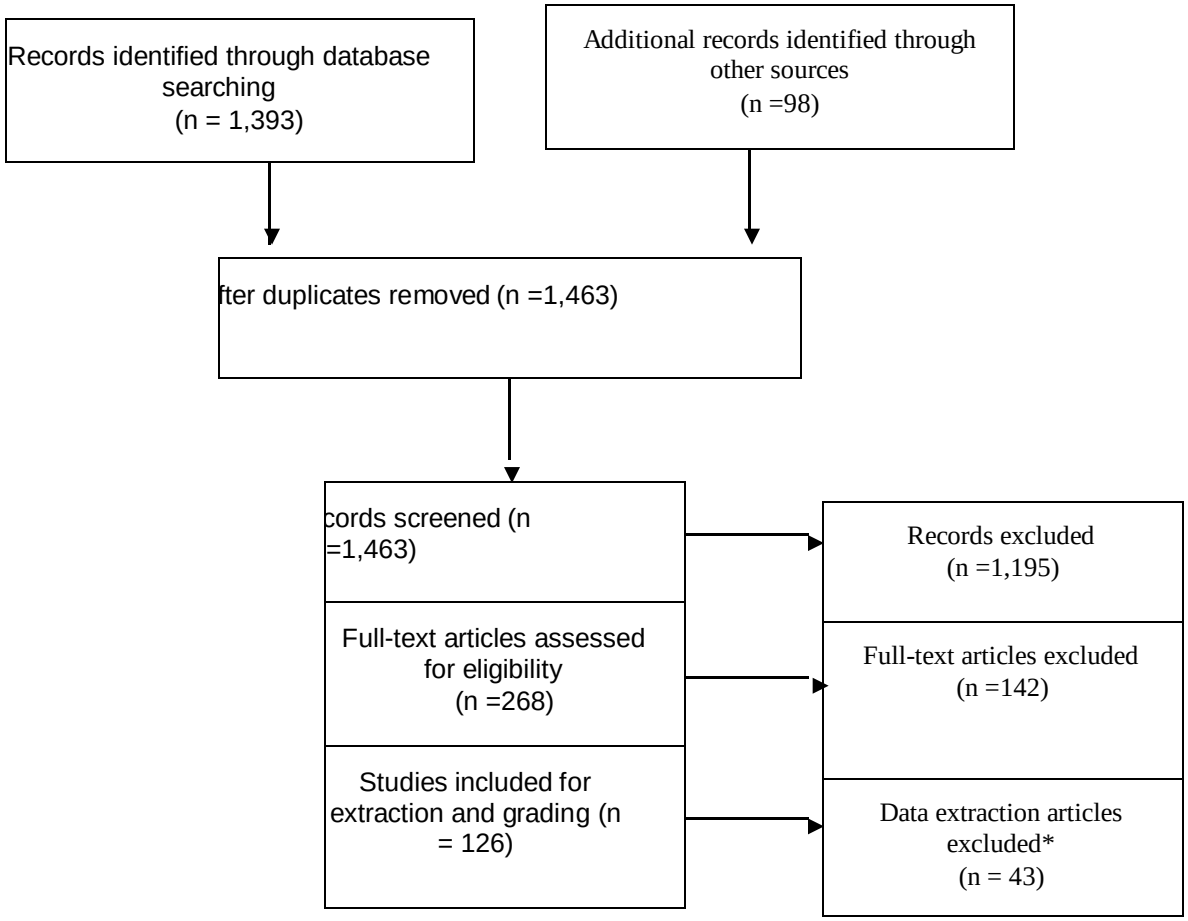
63. Ferguson J, Chamberlain P, Cramer HM, Wu HH. ER, PR, and Her2 immunocytochemistry on cell-transferred cytologic smears of primary and metastatic breast carcinomas: a comparison study with formalin-fixed cell blocks and surgical biopsies. *Diagn Cytopathol*. 2013;41(7):575-581.
64. Gong Y, Symmans 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.
65. Kumar SK, Gupta N, Rajwanshi A, Joshi K, Singh G. Immunochemistry for oestrogen receptor, progesterone receptor and HER2 on cell blocks in primary breast carcinoma. *Cytopathol*. 2012;23(3):181- 186.
66. 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.
67. 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.
68. 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.
69. 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.
70. Arber JM, Arber DA, Jenkins KA, Battifora H. Effect of decalcification and fixation in paraffin- section immunohistochemistry. *Appl Immunohistochem*. 1996;4(4):241-248.
71. Bussolati G, Leonardo E. Technical pitfalls potentially affecting diagnoses in immunohistochemistry. *J Clin Pathol*. 2008;61(11):1184-1192.
72. 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.
73. 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.
74. 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.
75. Dowsett M, Nielsen TO, A'Hern R, et al. Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *J Natl Cancer Inst*. 2011;103(22):1656-1664.
76. Arber DA. Effect of prolonged formalin fixation on the immunohistochemical reactivity of breast markers. *Appl Immunohistochem Mol Morphol*. 2002;10(2):183-186.
77. Djordjevic B, Hennessy BT, Li J, et al. Clinical assessment of PTEN loss in endometrial carcinoma: immunohistochemistry outperforms gene sequencing. *Mod Pathol*. 2012;25(5):699-708.
78. Manion E, Hornick JL, Lester SC, Brock JE. A comparison of equivocal immunohistochemical results with anti-HER2/neu antibodies A0485 and SP3 with corresponding FISH results in routine clinical practice. *Am J Clin* Pathol. 2011;135(6):845-851.
79. Vaughan MM, Toth K, Chintala S, Rustum YM. Double immunohistochemical staining method for HIF-1alpha and its regulators PHD2 and PHD3 in formalin-fixed paraffin-embedded tissues. *Appl Immunohistochem Mol Morphol*. 2010;18(4):375-381.
80. Gazziero A, Guzzardo V, Aldighieri E, Fassina A. Morphological quality and nucleic acid preservation in cytopathology. *J Clin Pathol*. 2009;62(5):429-434.
81. Hansen TP, Nielsen O, Fenger C. Optimization of antibodies for detection of the mismatch repair proteins MLH1, MSH2, MSH6, and PMS2 using a biotin-free visualization system. *Appl Immunohistochem Mol Morphol*. 2006;14(1):115-121.
82. Ikeda K, Tate G, Suzuki T, Mitsuya T. Comparison of immunocytochemical sensitivity between formalin-fixed and alcohol-fixed specimens reveals the diagnostic value of alcohol-fixed cytocentrifuged preparations in malignant effusion cytology. *Am J Clin* Pathol. 2011;136(6):934-942.
83. Kovacs A, Stenman G. HER2-testing in 538 consecutive breast cancer cases using FISH and immunohistochemistry. *Pathol Res Pract*. 2010;206(1):39-42.
84. Takai H, Kato A, Ishiguro T, et al. Optimization of tissue processing for immunohistochemistry for the detection of human glypican-3. *Acta Histochem*. 2010;112(3):240-250.

85. Roepman P, Horlings HM, Krijgsman O, et al. Microarray-based determination of estrogen receptor, progesterone receptor, and HER2 receptor status in breast cancer. *Clin Cancer Res*. 2009;15(22):7003-7011.
86. Wong SC, Chan JK, Lo ES, et al. The contribution of bifunctional SkipDewax pretreatment solution, rabbit monoclonal antibodies, and polymer detection systems in immunohistochemistry. *Arch Pathol Lab Med*. 2007;131(7):1047-1055.
87. Linderoth J, Ehinger M, Akerman M, et al. Tissue microarray is inappropriate for analysis of BCL6 expression in diffuse large B-cell lymphoma. *Eur J Haematol*. 2007;79(2):146-149.
88. Fitzgibbons PL, Murphy DA, Dorfman DM, et al. Interlaboratory comparison of immunohistochemical testing for HER2: results of the 2004 and 2005 College of American Pathologists HER2 Immunohistochemistry Tissue Microarray Survey. *Arch Pathol Lab Med*. 2006;130(10):1440-1445.
89. Lin Y, Hatem J, Wang J, et al. Tissue microarray-based immunohistochemical study can significantly underestimate the expression of HER2 and progesterone receptor in ductal carcinoma in situ of the breast. *Biotech Histochem*. 2011;86(5):345-350.
90. Lourenco HM, Pereira TP, Fonseca RR, Cardoso PM. HER2/neu detection by immunohistochemistry: optimization of in-house protocols. *Appl Immunohistochem Mol Morphol*. 2009;17(2):151-157.
91. Ricardo SAV, Milanezi F, Carvalho ST, Leitao DRA, Schmitt FCL. HER2 evaluation using the novel rabbit monoclonal antibody SP3 and CISH in tissue microarrays of invasive breast carcinomas. *J Clin Pathol*. 2007;60(9):1001-1005.
92. 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.
93. Engel KB, Moore HM. Effects of preanalytical variables on the detection of proteins by immunohistochemistry in formalin-fixed, paraffin-embedded tissue. *Arch Pathol Lab Med*. 2011;135(5):537-543.
94. Tong LC, Nelson N, Tsourigiannis J, Mulligan AM. The effect of prolonged fixation on the immunohistochemical evaluation of estrogen receptor, progesterone receptor, and HER2 expression in invasive breast cancer: a prospective study. *Am J Surg Pathol*. 2011;35(4):545-552.
95. Ibarra JA, Rogers LW. Fixation time does not affect expression of HER2/neu: a pilot study. *Am J Clin Pathol*. 2010;134(4):594-596.
96. Moher D, Liberati A, Tetzlaff J, Altman D. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. 2009;6(7):e1000097.

APPENDIX

Appendix- Figure 1: Literature Review Results

Adapted with permission from Moher et al.[96]



*Excluded based on expert opinion, did not meet minimum quality standards, presented incomplete data or data that were not in useable formats

Appendix - Table 1. Hierarchies of Data Sources for Analytic Validation[2]

Level 1

- Collaborative study using a large panel of well characterized samples
- Summary data from external proficiency testing schemes or inter-laboratory comparisons

Level 2

- High quality peer-reviewed studies (see Table 2)
  - Method comparisons
  - Validation studies

Level 3

- Lower quality peer-reviewed studies (see Table 2)
- Expert panel reviewed FDA summaries

Level 4

- Unpublished and/or non-peer reviewed research, clinical laboratory, or manufacturer data
- Studies on performance of the same basic methodology, but used to test for a different target

Reprinted by permission from Macmillan Publishers Ltd: Genetics in Medicine[2], copyright 2009

Appendix - Table 2. Criteria for Assessing Quality of Individual Analytic Validation Studies (internal validity)[2]

1. Adequate descriptions of the index test (test under evaluation)
   - Source and inclusion of positive and negative control materials
   - Reproducibility of test results
   - Quality control/assurance measures

2. Adequate descriptions of the referent test
   - Specific methods/platforms evaluated
   - Number of positive samples and negative controls tested

3. Adequate descriptions of the basis for the "right answer"
   - Comparison to a "gold standard" reference test
   - Consensus (e.g., external proficiency testing)
   - Characterized control materials (e.g., National Institute of Standards and Technology, sequenced)

4. Avoidance of biases
   - Blinded testing and interpretation
   - Specimens represent routinely analyzed clinical specimens in all aspects (e.g., collection, transport, processing)
   - Reporting of test failures and uninterpretable or indeterminate results

5. Analysis of data
   - Point estimates of analytic sensitivity and specificity with 95% confidence intervals
   - Sample size and power calculations addressed

Reprinted by permission from Macmillan Publishers Ltd: Genetics in Medicine[2], copyright 2009

Appendix –Table 3. Summary data on comparisons of concordance between IHC tests for HER2

Studies	Sample Type[a]	IHC 1	IHC 2	3x3 Table[c] Concordance	% Conc (95% CI)	Kappa, McNemars	2x2 Table (minus 2+)[d] % Conc	Grade[e]
Van der Vegt, 2009[34]	FFPE TMA	Pathway Her-2/neu, 4B5	Pathway Her-2/neu, CB11	436/467	93.4 (91-95)	0.75, <0.001	100	Fair
Boers, 2011[22]	FFPE WS	Ventana, 4B5	Ventana, SP3	134/146	92.0 (86-95)	0.66, 0.002	100	Fair
Moelans, 2010[31]	FFPE WS	Oracle Auto, CB11	Hercep Test	195/219	89.0 (84-93)	0.78 <0.001	100	Good
O'Grady, 2010[32]	FFPE WS[b]	Oracle Auto	Hercep Test	386/445	86.7 (83-90)	0.77, <0.001	100	Fair
Mayr, 2009[30]	FFPE WS	Ventana 4B5	Dako, Hercep Test	96/130	73.8 (66-81)	0.60, 0.004	97.1	Fair

[a] All breast cancer except O'Grady, 2010. [b] Gastroesophageal tumor. [c] Scoring system is 3+ positive, 2+ equivocal, 0-1+ negative; calculation of overall concordance by addition of 3 cells on the major diagonal / total N. [d] Recalculation of concordance after excluding all 2+ cells. [e] Quality grade for individual studies.
Abbreviation: Conc=concordance; FFPE=Formalin-fixed paraffin embedded; TMA= tissue microarray; WS=whole section

Appendix –Table 4. Summary data on concordance estimates from comparisons between HER IHC and *in situ* hybridization tests

Study	Grade	IHC	ISH[a]	Tissue Sample type	Data Analysis[b]	Conc cells[c] / Total N	2x2[d] [a]	[b]	[c]	[d]	Conc[e] (%)	Conc 95% CI	K[f]	McNemars p
Van der Vegt, 2009[34]	Fair	Her2/neu, CB11	FISH	BrCa TMA	2x2	444/473	62	17	12	382	94.0	91-96	0.77	0.46
Van der Vegt, 2009[34]	Fair	Her-2/neu, 4B5	FISH	BrCa TMA	2x2	436/466	54	4	19	389	93.6	91-95	0.80	0.003
Powell 2007, (Site 1+2)[49]	Fair	Pathway Her-2 CB11	FISH	BrCaWS	2x2	279/322	149	13	30	77	86.6	82-90	0.73	0.015
Powell 2007, (Site 1+2)[49]	Fair	Her-2 Benchmark auto, 4B5	FISH	BrCa WS	2x2	288/322	155	27	7	133	89.4	85-92	0.79	0.001
Moelans, 2010[31]	Fair	HercepTest Manual	CISH	BrCa WS	3x3	183/219					85.8	80-90	0.72	0.001
Moelans, 2010[31]	Fair	Oracle Auto, CB11	CISH	BrCa WS	3x3	183/219					83.6	78-88	0.66	0.004
Grimm, 2010[42]	Fair	HER2 LDT, 4B5	FISH	BrCa WS	3x3	457/697	--	--			65.6	62-69	0.37	<0.001
Boers, 2011[22]	Good	Ventana, 4B5	SISH	GI Ca WS	2x2	143/146	21	2	1	122	98.0	94-99	0.92	1.00
Boers, 2011[22]	Good	Ventana, SP3	SISH	GI Ca WS	2x2	141/146	17	0	5	124	96.6	92-99	0.85	0.07
Hofmann, 2008[43]	Poor	HercepTest Manual	FISH	Gast Ca WS	2x3	157/168	18	0	11	139	93.5	88-96	0.73	0.003
Sornmayura 2012[44]	Fair	HER2 LDT, 4B5	FISH	Gast Ca WS	2x2	171/195	15	5	19	156	87.7	82-92	0.49	0.008

[a] ISH = *In situ* hybridization. [b] Data entered into 2x2 or 3x3 contingency tables. [c] Conc=concordant cells. [d] Cells in a 2x2 table. [e] Conc=concordance.
[f] k=Kappa statistic.
Abbreviation: IHC=immunohistochemistry; BrCa=breast cancer; GICa= gastrointestinal cancer; Gast Ca= gastric cancer; TMA=tissue microarray; WS= whole section

Appendix – Table 5. Summary data on comparisons of concordance between IHC and alternative referent tests

Study	Grade	IHC	Referent	Tissue	Data Analysis[a]	Conc cells[b] / Total N	2x2[c] [a]	[b]	[c]	[d]	Conc[d] (%)	Conc 95% CI	K[e]	McNemars p
Lehmann-Che, 2011[46]	Fair	Benchmark HER2	QRT-PCR, panel consensus	BrCa	3x3	444/446					95.3	93-97	0.87	0.87
Jordan, 2012[45]	Fair	p16	QRT-PCR p16, HPV quant PCR, HPV ISH	OSCC	2x2	204/233	141	24	5	62	87.5	83-91	0.72	0.72
Baba, 2008[40]	Fair	Anti-BCG	TB diagnosis	Pleural bx	2x2	31/36	20	0	5	11	86.1	71-94	0.71	0.71
Dowsett, 2007[20]	Fair	HercepTest HER2	Consensus	BrCa WS	3x3	65/90					72.2	62-80	0.56	0.56

[a] Data entered into 2x2 or 3x3 contingency tables. [b] Conc=concordant cells. [c] Cells in a 2x2 table. [d] Conc=concordance.
[e] k=Kappa statistic
Abbreviation: IHC=immunohistochemistry; BrCa=breast cancer; OSCC=Oropharyngeal squamous cell carcinoma; bx=biopsy; WS= whole section

Appendix – Table 6. Considering the characteristics of validation sets with different numbers of samples[1]

Samples[1]	0 discordant		1 discordant		2 discordant		3 discordant	
	Conc	L 95%	Conc	L 95%	Conc	L 95%	Conc	L 95%
20	100.0	81.0	95.0	74.6	**90.0**	**68.7**	85.0	63.1
30	100.0	86.5	96.7	81.9	93.3	77.6	90.0	73.6
40	100.0	89.6	97.5	86.0	95.0	82.6	92.5	79.4
50	100.0	91.5	98.0	88.5	96.0	85.8	94.0	83.2
60	100.0	92.8	98.3	90.3	96.7	88.0	95.0	85.8
70	100.0	93.8	98.6	91.6	97.1	89.6	95.7	87.7
80	100.0	94.5	98.8	92.6	97.5	90.8	96.3	89.1
90	100.0	95.1	98.9	93.4	97.8	91.8	96.7	90.3
100	100.0	95.6	99.0	94.0	98.0	92.6	97.0	91.2

Samples[1]	4 discordant		5 discordant		6 discordant		7 discordant	
	Conc	L 95%	Conc	L 95%	Conc	L 95%	Conc	L 95%
20	80.0	57.8	75.0	52.7	70.0	47.8	65.0	43.2
30	86.7	69.7	83.3	66.0	80.0	62.3	76.7	58.8
40	**90.0**	**76.4**	87.5	73.4	85.0	70.5	82.5	67.7
50	92.0	80.7	90.0	78.2	88.0	75.8	86.0	73.5
60	93.3	83.6	91.7	81.5	**90.0**	**79.5**	88.3	77.5
70	94.3	85.8	92.9	84.0	91.4	82.2	90.0	80.5
80	95.0	87.5	93.8	85.9	92.5	84.3	91.3	82.8
90	95.6	88.8	94.4	87.3	93.3	85.9	92.2	84.6
100	96.0	89.8	95.0	88.5	94.0	87.3	93.0	86.0

Appendix – Table 7. Summary data on concordance between ER IHC performed on cytology samples and histologic sections

Study	N Pos	N Neg	Total N	Tissue	Comparator	Referent	Concordance (95% CI)	kappa	McNemar's p	Pos/Neg conc
Gong, 2004[64]	32	15	47	BrCa	Cytologic smears[a]	Histologic sections	91% (79-97)	0.79	0.13	89% 100%
Kumar, 2011[65]	20	30	50	BrCa	FNA cell block[b]	Histologic sections	90% (78-96)	0.79	0.37	80% 97%
Nishimura, 2011[66]	66	16	82	BrCa	PreserveCyt	Histologic sections	98% (91-99)	0.93	0.48	97% 100%
Ferguson, 2012[63]	22	16	38[d]	BrCa	FNA Smears[e]	Histologic sections	97% (85-99)	0.95	1.0	95% 100%
Pegolo, 2012[67]	85	16	101[f]	BrCa	Cytolyt ThinPrep	Tissue sections	98% (93-99)	0.92	0.48	100% 87%
Shabaik, 2012[68]	21	18	39[h]	BrCa	FNA cell block[g]	Tissue sections	92% (79-98)	0.85	0.25	86% 100%

[a] Abbott method (10% formalin-methanol-acetone -20C); no antigen retrieval. Addition of AR improved intensity without increasing false positives.
[b] 10% buffered formalin overnight.
[c] FNA immediately into PreserveCyt Solution, ThinPrep slides
[d] 38/47 (81%) had ≥ 50 cells
[e] FNA on alcohol fixed direct smears using cell transfer technique
[f] 101/111 (91%) assessable
[g] FNA/serous effusions FFPE cell blocks
[h] 39/42 (93%) assessable
Abbreviation: IHC=immunohistochemistry; BrCa=breast cancer

Appendix – Table 8. Summary data on concordance between PgR IHC performed on cytology samples and histologic sections

Study	N Pos	N Neg	Total N	Tissue	Comparator	Referent	Concordance (95% CI)	kappa	McNemar's p	Pos/Neg conc
Kumar, 2011[65]	17	33	50	BrCa	FNA cell block[a]	Histologic sections	94% (83-99)	0.86	1.0	88% 97%
Nishimura, 2011[66]	58	24	82	BrCa	PreservCyt/ ThinPrep[b]	Histologic sections	95% (88-98)	0.88	0.62	95% 96%
Ferguson, 2012[63]	19	23	42[c]	BrCa	FNA Smears[d]	Histologic sections	95% (83-99)	0.90	0.48	89% 100%
Pegolo, 2012[67]	75	24	99[e]	BrCa	Cytolyt ThinPrep	Tissue sections	91% (83-95)	0.76	0.50	92% 87%
Shabaik, 2012[68]	15	24	39[f]	BrCa	FNA cell block[g]	Tissue sections	92% (79-98)	0.83	0.25	80% 100%

[a] 10% buffered formalin overnight
[b] Immediately into PreserveCyt Solution, ThinPrep slides
[c] 42/47 (89%) had ≥ 50 cells
[d] FNA on alcohol fixed direct smears using cell transfer technique
[e] 99/111 (89%) assessable
[f] 39/42 (93%) assessable
[g] FNA/serous effusions FFPE cell blocks
Abbreviation: IHC=immunohistochemistry; BrCa=breast cancer

Appendix – Table 9. Summary data on concordance between HER2 IHC performed on cytology samples and histologic sections

Study	N 3+	N 2+	N Neg	Total N	Tissue	Comparator	Referent	Concordance (95% CI)	kappa	McNemar's p	Pos/Neg conc[c]
Kumar, 2011[65]	12	NR	38	50	BrCa	FNA cell block[a]	Histologic sections	90% (78-96)	0.75	0.37	92% 89%
Pegolo, 2012[67]	9	NR	91	100[b]	BrCa	Cytolyt ThinPrep	Tissue sections	100% (96-100)	1.0	NS	100% 100%

[†] 3x3 contingency table

[a] 10% buffered formalin overnight

[b] 100/111 (90%) assessable

[c] Conc=concordance

Abbreviation: IHC=immunohistochemistry; BrCa=breast cancer

Appendix-Table 10. Summary data on concordance between IHC performed on whole sections and TMA[a]

Study	Marker	Tissue	Concordance (%) between WS & TMA	kappa	McNemars p	Study Grade
Graham, 2008[25]	HER2	BrCa	73.1	0.56	<0.001	Fair
Jones, 2012[28]	CK19	Thyroid ca	83.1	0.17	0.03	Poor
Warnberg, 2008[35]	ER	BrCa	84.2	0.65	0.70	Fair
Fons, 2006[24]	ER	Endometrioid	89.5	0.78	0.13	Fair
Soiland, 2008[36]	Androgen receptor	BrCa	89.9	0.74	<0.001	Fair
Drev, 2008[23]	HER2	BrCa	91.7	0.71	<0.001	Fair
Gulbahce, 2012[26]	ER	BrCa	94.5	0.85	0.30	Poor
Kwon, 2009[29]	CD34	GIST	95.5	0.93	NR	Fair
Henriksen, 2007[27]	ER	BrCa	96.4	NR	NR	Poor
Drev, 2008 (pilot)[23]	HER2	BrCa	96.9	0.90	0.56	Fair
Thomson, 2009[33]	ER	BrCa	98.7	NR	NR	Poor
Batistatou, 2013[21]	HER2	BrCa	100.0	1.0	Not sig	Fair
			Median = 93.1%			

[a] To avoid bias in the overall concordance range and median value related to a sample set being tested for multiple markers or for multiple TMA core sizes, the, comparisons were reduced from 12 in this table. Only one comparison was included from each sample set. When multiple core sizes were reported, 0.6 mm cores were selected. When multiple markers were reported, the selection order was ER/PR, HER2 and then the most common marker.

Abbreviation: IHC=immunohistochemistry; BrCa=breast cancer; GIST=gastrointestinal stromal tumor

Appendix- Table 11. Summary data on whole section versus TMA, stratified by IHC marker

Marker	Number of studies	Tissue	Concordance Range	Median Concordance between WS & TMA	Concordance >90%
ER	5 of 6	5 BrCa, 1 endometrioid	84.2 – 98.7	5 BrCa = 95.4% 6th, k=0.97	67%
PR	4 of 5	4 BrCa, 1 endometrioid	81.5 – 92.6	4 BrCa = 90.8% 5th, k=0.90	60%
HER2 IHC	6	BrCa	73.1 - 100	92.6%	67%
HER2 FISH	2	BrCa	NA	98.6%	100%

Comparisons of overall concordance between whole sections and TMA for ER and PgR from an earlier systematic review were 97% and 93%.[92]
Abbreviation: BrCa= breast cancer; IHC=immunohistochemistry; TMA=tissue microarray; WS = whole section; NA= not applicable

Appendix- Table 12. Summary data on whole section versus TMA, stratified by TMA core size

Core sizes	Number of studies	Concordance Range	Median Concordance between WS & TMA	Concordance >90%
0.6	17	73.1 – 98.7	92.1%	59%
1.0 – 2.0	8	80.4 - 100	92.2%	50%
3.0	10	74.6 – 96.4	92.5%	60%

*These proportions are not statistically different (p >0.5; Fisher's exact test)
Abbreviation: TMA=tissue microarray; WS = whole section

Appendix- Table 13. Available data on other markers tested on whole sections versus TMA samples

Marker	Number of studies	Tissue	0.6 mm Cores	2.0 mm Cores	3.0 mm Cores
Androgen receptor	1	BrCa	--	--	--
CD 34	1	BrCa	95.5%	92.5%	89.5%
	1	GIST	74.6%	86.6%	94.0%
CK19	1	Thyroid ca	80.4%	83.1%	--
HBME1	1	Thyroid ca	92.9%	95.0%	--
Ki-67	1	BrCa	--	--	--
	1	GIST	74.6%	86.6%	94.0%
P53	1	Endometrioid	--	--	--
	1	GIST	74.6%	77.6%	92.5%

Abbreviation: TMA = tissue microarray; BrCa=breast cancer; GIST=gastrointestinal stromal tumor

COLLEGE of AMERICAN PATHOLOGISTS

Principles of Analytic Validation of Immunohistochemistry Assays

Published : *Archives of Pathology and Laboratory Medicine* March 19, 2014

Pathology and Laboratory Quality Center

Objectives

- Apply evidence-based guidelines to ensure each Immunohistochemistry (IHC) assay is validated prior to reporting on patient samples
- Recognize the requirements for revalidation
- Understand possible differences in validation requirements based on variations in fixative or specimen type
- Understand how the quality of evidence impacts the recommendations related to the validation statements

CAP

Introduction

- Laboratories are required to validate all assays before testing patient specimens.
- There is significant variation in validation practices for IHC assays.
- Current guidelines exist only for HER2 and ER/PgR.

CAP

Background

Immunohistochemistry Validation Procedures and Practices

A College of American Pathologists Survey of 727 Laboratories

Lindsay B. Hardy, MD; Patrick L. Fitzgibbons, MD; Jeffery D. Goldsmith, MD; Richard N. Fuen, MD; Mary Beth Beasley, MD; Rhona J. Soares, MS; Raouf E. Nakhleh, MD

• *Context.*—The immunohistochemistry (IHC) laboratory represents a dynamic area of surgical pathology with limited practice guidelines. Studies have shown significant interlaboratory variability in results.

Objective.—To establish baseline parameters for IHC validation procedures and practice, and to assess their feasibility of implementation.

Design.—In September 2010, a questionnaire was distributed by the College of American Pathologists. It was composed of 32 questions relating to nonpredictive assays as well as non-US Food and Drug Administration (non-FDA)–approved, predictive IHC assays other than human epidermal growth factor 2 (HER2/*neu*).

Results.—For non-FDA approved, nonpredictive IHC assays, 68% of laboratories had a written validation procedure. Eighty-six percent of laboratories validated the most recently introduced nonpredictive antibody. Seventy-five percent used 21 or fewer total cases for the validation and 40% used weakly or focally positive cases. Forty-six percent of respondents had a written procedure for validation procedures for non-FDA approved, predictive marker IHC assays other than HER2/*neu*. Seventy-five percent of laboratories validated the most recently introduced predictive antibody other than HER2/*neu*. Fewer than half used 25 or more cases for the validation, and 47% used weakly or focally positive cases.

Conclusion.—Some laboratories have written validation procedures that appear to build upon HER2/*neu* testing guidelines. Some laboratories also manage to validate new antibodies according to those standards; however, many do not. There appears to be a need for further validation guideline development for nonpredictive and non-FDA approved predictive antibody IHC assays.

(*Arch Pathol Lab Med.* 2013;137:19–25; doi: 10.5858/arpa.2011-0676-CP)

CAP

Validation Practices – Non Predictive Factor Assays

Procedures	Yes	No
Lab has written validation procedure?	68%	28%
Procedure specifies # validation cases?	54%	44%
Procedure specifies when revalidation needed?	46%	46%
Cytology specimens addressed?	37%	63%

Hardy et al. *Arch Pathol Lab Med* 2013;137:19-25

Validation Practices - Non Predictive Factor Assays

Procedures	Yes	No
Change in antigen retrieval method?	71%	25%
Change in detection method?	74%	23%
Change in instrumentation?	74%	24%
Change in fixative?	65%	30%

Hardy et al. *Arch Pathol Lab Med* 2013;137:19-25

Introduction

- CAP convened expert and advisory panels to systematically review published data and develop evidence-based recommendations
- Closely followed IOM Clinical Practice Guidelines
  - Transparency
  - Manage conflicts of interest
  - Multidisciplinary panel
  - Patient advocate (N/A for this panel)
  - Systematic Review
  - Considered judgment

Principles of Analytic Validation for IHC Assays: Expert and Advisory Panel

Chair:
Patrick Fitzgibbons, MD

Randa Alsabeh, MD
Regan Fulton, MD, PhD
Jeffrey Goldsmith, MD
Thomas Haas, DO
Rouzan Karabakhtsian, MD, PhD
Patti Loykasek, HT(ASCP)QIHC
Monna Marolt, MD
Steven Shen, MD, PhD
Paul Swanson, MD

Advisory Panel Members:
Raouf Nakhleh, MD, Center
Richard Brown, MD
Richard Eisen, MD
Hadi Yaziji, MD

Staff:
Lisa Fatheree, SCT(ASCP)
Tony Smith, MLS

Consultant Methodologist:
Linda Bradley, PhD

Systematic Evidence Review

- Identify Key Questions
- Literature search
- Data extraction
- Develop proposed recommendations
- Open comment period
- Considered judgment process

Introduction

- Overarching questions:

1. What is needed for initial analytic assay validation before placing any immunohistochemical test into clinical service?
2. What are the revalidation requirements?

Scope Questions

1. When and how should validation assess
   - analytic sensitivity
   - analytic specificity
   - accuracy (assay concordance)
   - precision (inter-run and inter-operator variability)?

Scope Questions (cont.)

2. What is the minimum number of positive and negative cases needed to analytically validate an IHC assay for its intended use(s)?
   - Non-predictive markers
   - Predictive markers
   - Identifying infectious organisms
   - Rare antigens

Should expression levels be specified for positive cases?

Scope Questions (cont.)

3. What parameters should be specified for the tissues used in the validation set?
    - Cytology specimens
    - Minimum tissue size or minimum quantity of cells
    - Neoplastic vs. non-neoplastic tissues

CAP

Scope Questions (cont.)

4. How do the following preanalytic variables influence analytic validation?
    - Type of fixative
    - Type of decalcification solution
    - Time in decalcification solution
    - Validation tissues processed in another laboratory
5. What conditions require assay revalidation?

CAP

Systematic Evidence Review

- Literature search
    - January 2004 – May 2013
    - 1,463 studies met inclusion criteria
        → Reviewed by panel
    - 126 studies identified for full data extraction

CAP

Systematic Evidence Review

- Evidence Evaluation
    - Quality (rate strength of evidence)
    - Quantity
    - Consistency

CAP

Quality Assessment

- Individual studies graded on specific criteria by the methodology consultant (LAB)
- Criteria included:
  - Quality and execution of studies
  - Quantity of data (number and size of studies)
  - Consistency and generalizability of the evidence across studies
    - Adequate descriptions of the test
    - Adequate descriptions of the basis for the "right answer"
    - Reproducibility of test results
    - Avoidance of biases
    - Analysis of data

CAP

Grades for Strength of Evidence

Grade	Description
Convincing	Level 1 or 2 studies with an appropriate number and distribution of challenges and reported consistent and generalizable results.
Adequate	Level 1 or 2 studies that lacked the appropriate number and distribution of challenges OR were consistent but not generalizable.
Inadequate	Combinations of Level 1 or 2 studies that show unexplained inconsistencies, OR one or more lower quality studies (Level 3 or 4), OR expert opinion.

Level 1: Collaborative study using a large panel of well-characterized samples; summary data from external proficiency testing schemes or inter-laboratory comparisons
Level 2: High quality peer-reviewed studies
Level 3: Lower quality peer-reviewed studies OR expert panel reviewed FDA summaries
Level 4: Unpublished or non-peer reviewed data

CAP

Grades for Strength of Recommendation

Designation	Rationale
Strong Recommendation	Strength of evidence is Convincing based on consistent, generalizable, good quality evidence; further studies are unlikely to change the conclusions
Recommendation	Strength of evidence is Adequate based on limitations in the quality of evidence; further studies may change the conclusions
Expert Consensus Opinion	Important validation element to address but strength of evidence is Inadequate; gaps in knowledge may require further studies

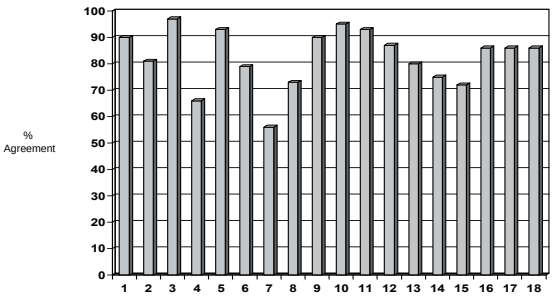
CAP

Systematic Evidence Review

- Open comment period (July 2013):
  - 18 draft recommendations and 5 methodology questions
  - 263 respondents; 1,037 comments

CAP

Open Comment Period



Original Draft Proposed Recommendation
Final Recommendations Combined/Condensed into 14 Total

CAP

Systematic Evidence Review

- Considered judgment process
  - Panel reviews and considers
    - Feedback
    - Quality/quantity/consistency of evidence
    - Benefits/harms
    - Value versus cost/burdens
    - Regulatory requirements
    - Expert opinion
  - 14 final recommendations

CAP

ASCO/CAP HER2 Guideline Recommendations Summary of Changes

Initial Test Validation

2007	2013
25–100 samples	20(+), 20(-) for FDA-approved assays 40(+), 40(-) for LDTs Not applicable if assay was previously validated and lab has successful PT performance

CAP

ASCO/CAP HER2 Guideline Recommendations Summary of Changes

Concordance

2007	2013
If <95% for any result category, cases with that test result must be automatically reflexed to alternative method	Specific concordance requirements are not required Laboratories must comply with accreditation and PT requirements

CAP

The Guidelines

Guideline 1

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

Note: 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.***

Guideline 1

- **Strength of Evidence:**
  - **Adequate to support when analytic validation should be done and that it should include determination of concordance and precision**
  - **Inadequate to assess how validation should be done with regard to the listed approaches, but did show that these approaches have been used.**
- **Rationale: 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.**

Rationale 1

- **Validation set should include:**
  - **Positive, negative, and low positive tissues**
  - **Should not be all normal tissues**
  - **Should reflect the intended use of the assay**
- **Positive and negative cell types on the same section could be used as separate challenges**

Guideline 2

***Recommendation*: For initial validation of every assay used clinically (with the exception of HER2, 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.**

CAP

Guideline 2

- **Strength of evidence**
  - **Adequate to support a 90% (versus 95%) overall concordance benchmark for analytic validation of IHC tests (except HER2, ER, PgR)**
- **Median overall concordance in a two-year inter-laboratory comparison of CD117 IHC and target results was 87.6%** (Hsi, 2001)
- **Median overall concordance in 5 comparisons of different HER2 IHC tests was 89.0% (range 74–92%), with 2 of 5 studies >90% concordant.** (Boers, 2011; Mayr, 2009; Moelans, 2010; O'Grady, 2010; van der Vegt, 2009)

CAP

Guideline 2 continued

- **Median overall concordance in 5 comparisons of HER2 IHC tests to HER2 ISH tests was 88.2% (range 66– 94%), with 2 of 5 comparisons >90% concordant** (Dorfman, 2006; Jordan, 2012; Lotan, 2011; Phillips 2007)
- **Median overall concordance in 6 comparisons of IHC tests (PTEN, ER, PR, HER2, MPT64, p16) to alternative referent tests (*e.g.,* RNA expression, clinical diagnosis) was 91.4% (range 74–99%), with 3 of 6 studies >90% concordant** (Phillips, 2007; Baba, 2008, Lehmann-Che, 2011)

CAP

Guideline 3

***Expert Consensus Opinion:* For initial analytic validation of non-predictive 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 (e.g., 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.***

CAP

Guideline 3

- **Strength of Evidence**
  - **Inadequate to support the recommended number of validation samples.**
  - **Adequate to support the distinction between non-predictive and predictive IHC tests and the use of different numbers.**

Validation Using 10 and 20 Tissue Validation Sets against a 90% Concordance Benchmark

	Concordance estimate (95% CI)		
# of validation tissues	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)

Concordance estimates with 95% confidence intervals stratified by number of observed discordant samples

Guideline 4

Expert Consensus Opinion: **For initial analytic validation of all laboratory-developed predictive marker assays, 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.**

- ***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.***

Guideline 4

- **Strength of Evidence**
  - **Inadequate to support the recommended number of validation samples.**
  - **Adequate to support the distinction between non-predictive and predictive IHC tests and the use of different numbers.**

	Concordance estimate (95% CI)				
# of validation tissues	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)

Concordance estimates with 95% confidence intervals stratified by number of observed discordant samples

2x2 contingency table of a 40 tissue validation set that did not meet the benchmark (results entered into a 2x2 contingency table) with associated statistical tests

New IHC Result	Referent Result Positive	Referent Result Negative	
Positive	15	0	15
Negative	5	20	25
	20	20	40

Overall concordance: 35/40=87.5% (does not meet 90% benchmark)
Kappa: 0.75
McNemar's p: 0.13, not significant
Positive concordance: 15/20 = 75%
Negative concordance: 20/20 = 100%

Guideline 5

Recommendation: **For a marker with both predictive and non-predictive applications, laboratories should validate it as a predictive marker if it is used as such**

- Strength of evidence:
  - Adequate to support the use of the higher validation standard (*e.g.*, number of samples) in the case of a marker with both non-predictive and predictive intended uses.

Guideline 6

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

- Strength of evidence
  - 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.

Guideline 6

- Can be difficult in reference laboratories that receive tissues with disparate fixation protocols
- Focused validation with a small number of markers may be appropriate

Guideline 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 (e.g., alcohol-fixed cell blocks, air-dried smears, formalin post-fixed 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 non-predictive markers to test.**

Guideline 7

- **Strength of evidence**
  - **Inadequate to address the criteria and number of samples needed for validation with cytology specimens.**
- **Focused validation on representative antibodies used on cytologic specimens would be appropriate**
- **A disclaimer in the report (especially in the case of negative results) may be appropriate if assays cannot be feasibly validated:**
  - **"Immunohistochemistry on cytologic specimens has not been sufficiently validated; these results should be interpreted with caution."**

Guideline 8

Expert Consensus Opinion: **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 non-predictive markers to test.**

Guideline 8

- Strength of evidence:
  - Inadequate to address the criteria and number of samples needed for validation with decalcified specimens.
- Focused validation on representative antibodies used on decalcified specimens would be appropriate
- A disclaimer in the report (especially in the case of negative results) may be appropriate if assays cannot be feasibly validated (ANP.22985)

Guideline 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.**

Guideline 9

- Strength of evidence
  - Adequate to support TMA usage; however there are many variables to be considered and thorough validation is needed for each marker.
  - Inadequate to recommend the *routine* use of TMA samples.
- TMAs / MTBs can be very useful in many circumstances. Beware of:
  - Proteins with high levels of heterogeneity (gastric Her2)
  - Limited tissue expression (e.g. bcl-6)

Revalidation Secondary to Assay Modification

Antibody Specific:

1. **Least:**
   - New antibody Lot
2. **Moderate:**
   - Antibody dilution
   - Antibody vendor (same clone)
   - Antibody incubation or antigen retrieval times (same A.R. method)
3. **Most:**
   - New antibody clone

All Assays (one tier):

- Fixative type
- Antigen retrieval method
  - pH change
  - buffer type
  - heat type
- Antigen detection system
- Tissue processing equipment
- Environmental conditions
  - location
  - water supply

Evidence for Revalidation Guidelines 10-13

- **Strength of evidence**
  - **Inadequate to address conditions requiring assay revalidation and whether revalidation should be the same as initial validation.**

CAP

Guideline 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.**

- **Laboratories may want to include low-expressors, especially with predictive markers**

CAP

Guideline 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:**

- **Antibody dilution**
- **Antibody vendor (same clone)**
- **Incubation or retrieval times (same method)**

- **Laboratories may want to include low-expressors, especially with predictive markers**

CAP

Guideline 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:**

- **Fixative type**
- **Antigen retrieval method (e.g., change in pH, different buffer, different heat platform)**
- **Antigen detection system**
- **Tissue processing or testing equipment**
- **Environmental conditions of testing (e.g. laboratory relocation)**
- **Laboratory water supply**

CAP

Guideline 12

The laboratory medical director is responsible for determining how many predictive and non-predictive markers and how many positive and negative tissues to test.

- Reasonable approach:
  - Selection of antibodies from menu with:
    - Variable clinical uses (predictive and non-predictive)
    - Variable antigen localizations
    - Variable antibody types (monoclonal / polyclonal, etc.)

CAP

Guideline 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.

CAP

Guideline 14

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

CAP

Summary

- Physicians and patients rely on accurate diagnostic and prognostic testing in the clinical laboratory.
- Analytic validation is essential to ensuring that an assay performs as expected, accurately identifies and/or quantifies the targeted analyte, and minimizes the chances of false positive or false negative results.
- Established guidelines are important to improve the reproducibility and consistency of the test results.

CAP

References

Early Online Release March 19, 2014

Archives of Pathology and Laboratory Medicine

http://www.archivesofpathology.org/doi/pdf/10.5858/arpa.2013-0610-CP

CAP

Disclaimer
IHC Validation Teaching PowerPoint Copyright

Effective March 19, 2014

CAP

Topic: **Principles of Analytic Validation of Immunohistochemical Assays**
Date: April 22, 2014

Why is this guideline needed? Is there any evidence that patients have been harmed by incorrect immunohistochemistry tests?
There is ample evidence that improper immunohistochemistry (IHC) tests have led to patient harm. In perhaps the best documented example, nearly 400 of 1,000 breast cancers tested in one laboratory in Newfoundland from 1997-2005 initially classified as ER negative were subsequently found to be ER positive. Because of the incorrect test results, these patients did not receive appropriate therapy and more than 100 died. A governmental inquiry determined that the high error rate was due to improper testing practices. The American Society of Clinical Oncology (ASCO) and the College of American Pathologists (CAP) guideline for hormone receptor and HER2 testing in breast cancer were a direct result of well documented testing inaccuracies.

How will the guideline be enforced? What happens if a laboratory doesn't follow the guideline?
As with any clinical evidence-based guideline they are not mandatory. These recommendations *may* be incorporated into future versions of the CAP Laboratory Accreditation Program (LAP) Checklist; however, they are not currently required by LAP or any regulatory or accrediting agency. It is encouraged that laboratories adopt these evidence-based recommendations.

When validating an estrogen receptor (ER) assay, must we use only breast cancers for validation tissues?
No. Since ER is most frequently used to assess eligibility for hormonal therapy in patients with breast cancer, positive and negative breast cancers should comprise at least part of the validation set, but other ER positive and negative tissue types could be included.

How do these recommendations apply to assays for pathogen-specific antigens (e.g., *Helicobacter pylori*)?
Assays for infectious organisms are similar to predictive marker assays in that the results can directly influence patient treatment, but selection of validation sets can be quite challenging when the organism is rarely encountered. The option of using normal tissues for positive cases is also not applicable. For selected organisms, including *H. pylori*, *Cryptococcus* spp, cytomegalovirus and herpes simplex I/II, histologic features may be sufficiently characteristic to provide "expected" positive cases for validation purposes, but for true analytic validation, concurrent culture evidence of specific infection or either retrospective or prospective molecular confirmation of the formalin fixed paraffin embedded sample may be required.

For rare antigens, do laboratory directors have the flexibility to use fewer validation samples as they deem appropriate?
Yes. Following public comment and independent peer review of the draft recommendations, it was determined that the guideline should not be too prescriptive and that the medical director must have the discretion to modify the recommended steps in cases where it is not possible to gather a full validation set. Several of the final recommendations include the caveat that the laboratory medical director may decide that fewer cases are sufficient for a specific marker (e.g., rare antigen); however the rationale for that decision needs to be documented. If the laboratory is unable to find sufficient cases to provide reasonable confidence that test results are valid, the director is responsible for the decision to offer that test.

Are normal tissues prohibited in validation sets?
No. Normal tissues may be used in conjunction with neoplastic and lesional tissue as appropriate, but the guideline specify that normal tissues cannot comprise the entire validation set for markers that are primarily used in diagnosing neoplasms. If the marker will be used to determine cell lineage in neoplasms, at least some of the tissues in the validation set should be neoplasms with positive and negative expression for that marker.

What is the difference between a tissue microarray (TMA) and a multitissue block (MTB)?
The terms are not always used consistently and TMAs and MTBs are not necessarily different. TMA often refers to a tissue block constructed using a commercially available instrument that results in uniform cores while MTBs may be assembled manually; these are sometimes referred to as “sausage blocks” or “spring rolls.”

If we temporarily move our laboratory while the existing one is being remodeled, do we have to revalidate all assays after both moves?
A complete revalidation of all assays is not required when equipment is moved, but a limited assessment of a selection of assays is recommended following laboratory relocation. In this situation, re-assessment of assay performance would apply to both moves. After each move, the laboratory medical director should select a group of assays that encompass different clinical uses (i.e., predictive and non-predictive markers, pathogen-specific markers, etc) and immunolocalizations (i.e., nuclear, cytoplasmic and membranous) and compare results of testing after the move with the results of testing done previously. The number of validation tissues tested should be determined by the director.

Does the guideline address validation of research use only (RUO) antibodies?
Not specifically, but the principles of analytic validation described in the guideline apply to all antibodies that may be used in patient testing.

Could you give some advice on the interpretation of the following terminology for IHC tests?:

1. Accuracy/Precision (Repeat measurement of samples at various concentrations or activities)
2. Sensitivity (Lower limit of detection)
3. Specificity
4. Reportable Range (Analytic Measurement Range)
5. CLIA requirements to determine test performance specifications apply to all lab tests including all IHC assays, but the nature of these assays is such that some of them aren’t relevant. For instance, reportable range and reference intervals are generally not applicable to tests that are reported qualitatively or semi-quantitatively, which represents most IHC tests.

    - With respect to determining accuracy, precision, analytical sensitivity and analytical specificity, CLIA distinguishes between FDA approved and laboratory-developed tests (LDTs). For FDA-approved test kits, laboratories must demonstrate performance characteristics that are comparable to those established by the manufacturer (often called “verification”). Manufacturers may provide users with directions and/or materials for this verification. By contrast, laboratories must “establish” their own performance specifications for LDTs. For IHC assays, accuracy, analytic sensitivity and specificity are determined by analytic assay validation, which is theoretically done by testing a validation tissue set against a gold standard. Since the majority of IHC tests do not have a "gold standard" referent test, analytic sensitivity and specificity are determined by measuring positive and negative concordances with an appropriate comparator. This may be another validated IHC assay (i.e., different clone), testing done in another lab with a

validated assay, a different test (e.g., ISH), or even clinical outcome if you have the resources. For most laboratories and tests, it's some combination of the first two.

- In our literature review we could not find strong evidence to say how IHC assay precision (inter-run and inter-operator) should be measured. Until stronger evidence is available, the laboratory director must determine the extent to which these performance specifications are established based on the method, testing conditions and personnel performing the test.

Aren't commercially available antibodies already validated for clinical use by manufacturers?
The guideline applies to analytic validation of assays, not antibodies. An antibody marketed as an FDA Class I in vitro diagnostic device may be produced following current good manufacturing practices and with documentation of specificity, but if the laboratory's assay is improperly designed or is not performed correctly (e.g. incorrect antibody dilution, inadequate antigen retrieval, wrong buffer, incorrect scoring system used), the test results will be incorrect. For antibodies marketed as "analyte specific reagents," the laboratory performing the test must establish the performance characteristics of the clinical assay.

Does the guideline apply to validation of controls?
No. The guideline applies to assays, not antibodies or controls.

Can negative internal cells be used as a negative tissue test or do the negative validation samples need to be separate tissue samples?
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. When validating a new antibody lot with one positive and one negative case, for example, a single control slide that contains both antigen-positive and antigen-negative cells might be sufficient.

Does the guideline apply to assays that have been in use in the laboratory for many years or do they only apply to newly introduced assays?
The guideline applies to all assays used on patient specimens. CLIA requires laboratories to verify the performance characteristics of all assays before issuing results on patient specimens. Thus, even if an assay has been in use, if there is no documentation that validation was ever done, the laboratory may not be compliant with federal law and could be subject to citation by an accrediting agency.

Do we have to revalidate every existing assay to provide the number of cases recommended?
Revalidation of existing assays would not be expected if a previous validation was performed, but the Medical Director should determine if the previous validation was sufficient.

Must all tissues from a validation set be acquired by and processed in the laboratory validating the IHC panel?
No. This would be ideal but is not possible for many laboratories, especially reference laboratories, and may be impossible for some rare antigens.

How long must laboratories do validations on all the antibodies they currently use?
For each assay, initial validation is done once and not repeated unless the assay is changed. Validation records should be retained indefinitely to demonstrate to future inspectors that it was done.

Some laboratories use microwave fixation to decrease processing time. How does this reduced fixation time influence IHC results?
This specific issue was not addressed in the guideline, but because any change to a procedure

can introduce variation in test results, assays done on microwave fixed tissues should be compared to routinely fixed and processed specimens to determine if IHC results are affected.

Is a single daily control slide sufficient for validation?
No. Daily quality control is essential to ensure the assay has not changed and continues to perform as expected, but this is not a substitute for initial assay validation.

REFERENCES:

1. Fitzgibbons PL, Bradley LA, Fatheree LA, et al. Principles of analytic validation of immunohistochemical assays: guideline from the College of American Pathologists Pathology and Laboratory Quality Center. *Arch Pathol Lab Med.* 2014;138(11):1432-1443.

Principles of Analytic Validation of Clinical Immunohistochemistry Assays

Jeffrey D. Goldsmith, MD, Patrick L. Fitzgibbons, MD,† and Paul E. Swanson, MD‡*

Abstract: All assays performed in anatomic and clinical pathology laboratories must be validated before they are placed into clinical service. This review summarizes strategies for validation of clinical immunohistochemistry assays, and is chiefly based on the recently released guideline released by The College of American Pathologists.

Key Words: clinical immunohistochemistry, analytic validation, practice guideline

(*Adv Anat Pathol* 2015;22:384–387)

In the current practice of anatomic pathology, immunohistochemistry (IHC) is a critical ancillary test that aids in the accurate diagnosis of a host of neoplastic and non-neoplastic conditions. In addition, IHC is being increasingly used to predict response to therapy and screen for inherited diseases. In the last decades of the 20th century, IHC assays were being developed that could be reproducibly performed on paraffin-embedded, formalin-fixed tissues; these methods were developed as "home-brew" assays, more appropriately termed "laboratory developed tests." As such, assay conditions often varied significantly between laboratories. As IHC became more widespread and its use expanded to industry, detection methods became more standardized. However, as many preanalytic factors may affect the results of IHC tests, assay conditions still may vary significantly between laboratories.

Many IHC laboratories continue to use laboratory developed tests; as preanalytic factors may significantly affect assay results, robust and standardized analytic validation before use on patient samples is required, particularly for those assays with quantitative results or for IHC tests that predict responsiveness to specific therapies. Indeed, analytic validation of all clinical laboratory tests, including IHC, is required by the Clinical Laboratory Improvement Amendments of 1988.[1] Despite both this regulatory mandate and the common sense notion that quality testing is predicated on carefully validated methodology, up to 28% of surveyed IHC laboratories did not have a written procedure for initial assay validation at the time a recent interlaboratory practice survey.[2] The same survey noted that laboratories in compliance with CLIA'88 validation requirements nonetheless followed strikingly variable IHC assay analytic validation practices. To address these challenges to the uniformity and quality of diagnostic IHC, the College of American Pathologists (CAP) convened a panel of experts in 2012 with the charge of creating an evidence-based guideline that would serve as a standard for analytic validation of IHC assays. The resulting recommendations were published in 2014.[3]

With these introductory comments in mind, we herein review the relevant concepts behind analytic validation with particular focus on analytic validation of IHC assays. The authors of this review served on the expert panel that created above-mentioned guidelines; however, this article has been created without input from the CAP.

GENERAL CONSIDERATIONS

The United States Food and Drug Administration defines "validation" as "confirmation by examination and provision of objective evidence that the particular requirements for a specific intended use can be consistently fulfilled."[4] In other words, analytic validation is a process that confirms that a test has the expected level of sensitivity, specificity, and reproducibility for its intended use. In the context of the clinical pathology laboratory, validation is achieved by comparing the test's result with a known gold standard. However, a vast majority of IHC assays do not have a gold standard referent test that can be feasibly obtained by most laboratories. As such, most laboratories must compare their results to comparators that are not considered gold standards in the strict sense. Such comparators tend to fall in the following 4 categories.

(1) Morphology and expected results according to the medical literature: This comparator is frequently used when new assays are being initiated. Typically, the medical director of the laboratory performs a review of the literature pertinent to the new assay. From those data, a set of validation cases is chosen, typically from the laboratory archives from cases fixed and processed in the same manner as those that will be run on patient samples.

(2) Previous results from a previously validated assay from the same laboratory: This method is often used if the assay conditions change to such an extent that merits

From the *Department of Pathology, Beth Israel Deaconess Medical Center, Children's Hospital Boston, and Harvard Medical School, Boston, MA; †Department of Pathology, St. Jude Medical Center, Fullerton, CA; and ‡Department of Pathology and Laboratory Medicine, University of Calgary and Foothills Medical Centre, Calgary, AB, Canada.

In various combinations, authors have received honoraria and travel expense reimbursement for speaking at the 2011 (PS), 2013 (PF, JG), 2014 (PF, JG), and 2015 (PF, JG) College of American Pathologists Annual Meetings regarding The Guidelines cited in this article. J.D.G. and P.E.S. spoke at the 2014 annual meeting of the American Society of Clinical Pathology regarding these guidelines. J.D.G. also spoke at the 2014 annual meetings of the American Society of Cytopathology and the United States and Canadian Academy of Pathology. In addition, all authors were on the committee that produced the aforementioned guidelines and received reimbursement from the College of American Pathologists for expenses incurred as a result of committee participation.

Reprints: Jeffrey D. Goldsmith, MD, Department of Pathology, Beth Israel Deaconess Medical Center, 330 Brookline Ave., Boston, MA 02215 (e-mail: jgoldsmi@bidmc.harvard.edu).

some sort of revalidation (see below). For example, if a manufacturer discontinues a primary antibody and it is replaced with a different primary antibody clone, this change is considered a fundamental modification to the assay that requires complete revalidation. In this circumstance, the use of results obtained from previously validated assays from the same laboratory as a comparator would be a reasonable approach.

(3) Another laboratory's results from the same validation set using a previously validated assay: This method is particularly useful for assays that are difficult to validate. In this situation, interlaboratory comparison allows the laboratory to directly compare results from a previously validated assay on the same tissues.

(4) Previously validated results from a sufficiently validated nonimmunohistochemical assay: As noted above, this comparator applies to very few assays, but is often the most robust validation method. Examples include chromogenic or fluorescent in situ hybridization (CISH/FISH) for Her2-neu as applied to Her2 IHC, flow cytometric analyses for markers such as CD3, CD20, and other common hematopoietic analytes, and mutation testing for the *BRAF* V600E mutation as compared with mutation specific b-raf IHC.

CONCORDANCE AND SIZE OF VALIDATION SET

The desired level of concordance between the new assay and the comparator is tightly related to the size of the validation set. This is due to the fact that both of these parameters have a hand in determining the confidence interval for a particular level of concordance. The confidence interval, generally set at 95%, is the statistical value that determines the level of confidence that the observed concordance level reflects the true performance of the test. Thus, as the size of the validation set increases, the level of confidence that the observed concordance is the true value increases. For an example, see Table 1; this table shows that the 95% confidence intervals are smaller and overall confidence levels higher with a validation set composed of 40 cases compared with 20 case validation set. Thus, as a general rule, a larger validation set is desirable, whenever possible. Of course, larger validation sets can be difficult to obtain, especially in smaller laboratories.

The size of the validation set should also be dictated by the intended clinical use of the assay. The clinical use of IHC assays fall into 2 general groups. The first are markers that are interpreted in the context of the morphologic findings and are typically used as ancillary stains for diagnosis (eg, cytokeratin 7, cytokeratin 20, TTF-1, GATA-3, etc.). The second group of stains includes those that are interpreted without regard to the histologic context; many of these markers give predictive information about the sensitivity of a tumor to various treatments (eg, Her2 IHC on breast carcinoma, b-raf mutation-specific IHC on melanoma). Markers that are used for histologic diagnosis and are interpreted in a histologic context have less direct clinical impact than predictive markers that result in an actionable result that is independent of the morphologic context. Thus, the size of the validation set for a predictive marker should be larger than that prepared for a diagnostic marker. The expansion of the size of the validation set for predictive markers increases the confidence that the observed concordance level truly reflects the desired level of concordance. As such, the CAP Guideline mandates that the size of the validation set for predictive markers should contain at least 40 challenges, whereas nonpredictive/diagnostic markers should have at least 20 challenges. Depending on the resources available, expansion of the size beyond the prescribed amount of the validation set is optimal and would add additional assurance that the assay will behave as expected.

In theory, the level of aggregated positive and negative concordance between the new test and the comparator should be 100%. However, this is not practically obtainable due to a number of factors including, but not limited to, preanalytic factors, intratumoral heterogeneity of analyte expression and the quality of the originally validated method or comparator set. For these reasons, the Guideline set the desired level of concordance at 90%; this was chiefly based on evidence from concordance data between Her2 IHC and Her2-neu FISH, in which concordance levels higher than 90% were not feasible for a majority of laboratories.[5–7]

COMPOSITION OF THE VALIDATION SET

As a general rule, the composition of the validation set should reflect the intended clinical use of the assay. Not only should relevant positive cases be included, but also judicious inclusion of cases that show lack of expression of the analyte of interest should be part of the validation set. For example, TTF-1 is a transcription factor that is often used as an ancillary test in the workup of metastatic carcinoma of unknown origin. It is expressed in a majority of small cell carcinomas of the lung, most primary pulmonary adenocarcinomas, and many types of primary epithelial tumors of the thyroid gland. Inclusion of tumor types that are known to be positive for TTF-1 should be part of the validation set. In addition, tumor types that are known to be TTF-1 negative and are in the histologic differential diagnosis of either metastatic pulmonary adenocarcinoma and metastatic thyroid carcinoma should be included. Such examples of clinically relevant TTF-1 negative carcinomas might include ductal carcinoma of the breast, colorectal carcinoma, and pancreatic ductal adenocarcinoma. Inclusion of clinically relevant cases in the validation set adds additional assurance that the validation accurately reflects the performance of the assay when performed on patient samples.

Occasionally, assays are used in more than 1 clinical context. In this circumstance, it would be wise to tailor the validation set to reflect all potential clinical uses. For example, CD30 is a marker that is often used to diagnose

TABLE 1. Comparison of Concordance Rates and 95% Confidence Intervals for Validation Sets Composed of 20 and 40 Tissues

No. Validation Tissues	Concordance for 0 Discordant Cases	Concordance for 1 Discordant Case	Concordance for 2 Discordant Cases
20	100% (81%-100%)	95% (75%-100%)	90% (69%-98%)
40	100% (90%-100%)	97.5% (86%-100%)	95% (83%-99%)

Hodgkin lymphoma and various germ cell tumors, such as embryonal carcinoma. In this circumstance, the validation set should include cases of Hodgkin lymphoma, in which Reed-Sternberg cells are the expected positive cells, and cases of embryonal carcinoma. Relevant negative cases include nodular lymphocyte-predominant Hodgkin lymphoma and CD30-negative primary mediastinal diffuse large B-cell lymphoma, which are known mimics of Hodgkin lymphoma and are CD30 negative. In addition, expected CD30-negative cases in the differential diagnosis of embryonal carcinoma should be included in the validation set that might include seminoma and yolk sac tumor.

FORMAT OF VALIDATION CHALLENGES

Classically, validation is achieved by applying single tissue sections on slides analogous to the practice on patient samples. More recently, tissue microarrays have been used as a more efficient and cost-effective method of displaying multiple challenges on a single microscopic slide.[8–12] Tissue microarrays are usually an acceptable method of validation. However, caution should be exercised with assays that are known to show significant heterogeneity of staining. Examples of this include bcl-6 staining in normal tonsillar tissue. Bcl-6 expression is limited to germinal center cells; as such, a tissue microarray would not be an effective method. Similarly, CD15 and CD30 validation using classic Hodgkin lymphomas should not be performed using tissue microarrays, as the CD15-positive and CD30-positive Reed-Sternberg cells are very often heterogenously distributed within lesional tissue in this tumor type.

PREANALYTIC CONSIDERATIONS

Once tissues become devitalized at the time of biopsy or resection, they are fixed, processed, and prepared for microscopic diagnosis. This process can differ between laboratories and, in fact, may vary within a particular laboratory depending on the specimen type. These variations in tissue processing and handling may have dramatic effects on IHC results. For example, for some antibodies, acidic decalcification solutions can change the avidity of the primary antibody for its epitope(s).[13] Although it is impossible to control for all possible preanalytic factors during validation, attention to major causes of preanalytic variation should be taken into account. Some of the major preanalytic factors that may impact results include fixative type and preparation method (ie, cytologic preparations vs. formalin-fixed, paraffin-embedded tissues).[14,15]

If tissues fixed in alternative fixatives or tissues exposed to decalcifying solutions are to be used for IHC, efforts should be made to ensure that the results are clinically valid. A reasonable approach would be to validate a subset of assays that are often run on decalcified samples. Examples of such assays might include cytokeratins, CD45, S-100, and estrogen receptor.

Similarly, if IHC is run on cytologic preparations, including smears, cytospins, cell blocks, and ThinPrep preparations (or core samples submitted with aspirate fluid or other preparation to the cytology laboratory in CytoLyt or other nonformalin solutions), reasonable efforts should be made to assure that these assays perform adequately before they are used on patient samples. The selection of markers tested and number of cases included in these separate validation studies must be determined by the laboratory medical director.

REVALIDATION AFTER CHANGES TO ASSAY CONDITIONS

Once initial assay validation is successfully completed and a test is placed in clinical service, it is common for assay conditions to change. When that occurs, some sort of revalidation is needed to assure that the assays perform as expected. In general, changes to assay conditions fall into 3 categories. The first, and perhaps most straightforward, is a change to the antibody clone. As different antibody clones target different epitope(s), changes in antibody clone are considered a fundamental change to the assay. In this circumstance, full analytic revalidation is required.

The second category includes modifications to assay conditions that are common to all assays in the laboratory. Examples include changes to detection chemistry, water supply, antigen retrieval solution(s), and tissue processing equipment. When such changes occur, it is not necessary to fully revalidate all assays affected by the change. Instead, it is reasonable to choose a representative sample of assays run in the laboratory and compare cases prepared with the modified assay conditions with examples representative of original conditions. If the subset of modified assays performs as expected, it would be reasonable to assume that the remaining assays will perform adequately. If, however, significant changes to the assay conditions are necessary to achieve expected results, more extensive revalidation may be necessary.

The final set of condition changes that merit revalidation are changes that apply to single assays. Examples of this might include changes to antibody lot, primary antibody dilution, primary antibody incubation time, and change of primary antibody vendor using the same clone. Of these changes, a new antibody lot (same clone) often results in minimal perturbation of the assay. As such, verification of continued expected assay results is achieved by running 1 known positive and 1 known negative case. It may be judicious to include a third case that shows a low-positive reaction as an additional indication of appropriate assay performance. Changes to primary antibody dilution, incubation time, and vendor are more substantive changes to the assay. In these circumstances, it is reasonable to run 2 known positive and 2 known negative cases to assure continued assay performance; again, it may be wise to run a fifth, low positive, case to assure appropriate assay sensitivity.

CONCLUSIONS

IHC is a critical ancillary test in the modern anatomic pathology laboratory that often has significant impact on patient care. To be assured of accurate results, robust analytic validation must be performed on all assays before their use on clinical samples. This review summarizes best practices for analytic validation for IHC assays and outlines an approach for revalidation necessitated by changes to assay conditions after successful completion of initial validation procedures.

REFERENCES

1. US Department of Health and Human Services. Clinical laboratory improvement amendments of 1988: Final Rule. *Fed Regist*. 1992;57:7001–7186.

2. 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;137:19–25.
3. Fitzgibbons PL, Bradley LA, Fatheree LA, et al. Principles of analytic validation of immunohistochemical assays: Guideline from the College of American Pathologists Pathology and Laboratory Quality Center. *Arch Pathol Lab Med*. 2014;138: 1432–1443.
4. Institute CLS. *Quality Assurance for Design Control and Implementation of Immunohistochemistry Assays: Approved Guideline (CLSI Document I/LA28-A2)*, 2nd ed. Wayne, PA: Clinical and Laboratory Standards Institute; 2011.
5. Mayr D, Heim S, Werhan C, et al. 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; 454:241–248.
6. Rhodes A, Jasani B, Anderson E, et al. 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;118:408–417.
7. van der Vegt B, de Bock GH, Bart J, et al. Validation of the 4B5 rabbit monoclonal antibody in determining Her2/neu status in breast cancer. *Mod Pathol*. 2009;22:879–886.
8. Fons G, Hasibuan SM, van der Velden J, et al. Validation of tissue microarray technology in endometrioid cancer of the endometrium. *J Clin Pathol*. 2007;60:500–503.
9. Gulbahce HE, Gamez R, Dvorak L, et al. Concordance between tissue microarray and whole-section estrogen receptor expression and intratumoral heterogeneity. *Appl Immunohistochem Mol Morphol*. 2012;20:340–343.
10. Kwon MJ, Nam ES, Cho SJ, et al. Comparison of tissue microarray and full section in immunohistochemistry of gastrointestinal stromal tumors. *Pathol Int*. 2009;59:851–856.
11. Thomson TA, Zhou C, Chu C, et al. Tissue microarray for routine analysis of breast biomarkers in the clinical laboratory. *Am J Clin Pathol*. 2009;132:899–905.
12. Dessauvagie BF, Thomas C, Robinson C, et al. Validation of mitosis counting by automated phosphohistone H3 (PHH3) digital image analysis in a breast carcinoma tissue microarray. *Pathology*. 2015;47:329–334.
13. Athanasou NA, Quinn J, Heryet A, et al. Effect of decalcification agents on immunoreactivity of cellular antigens. *J Clin Pathol*. 1987;40:874–878.
14. Hanley KZ, Birdsong GG, Cohen C, et al. 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;117:279–288.
15. Nishimura R, Aogi K, Yamamoto T, et al. Usefulness of liquid-based cytology in hormone receptor analysis of breast cancer specimens. *Virchows Arch*. 2011;458:153–158.