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.

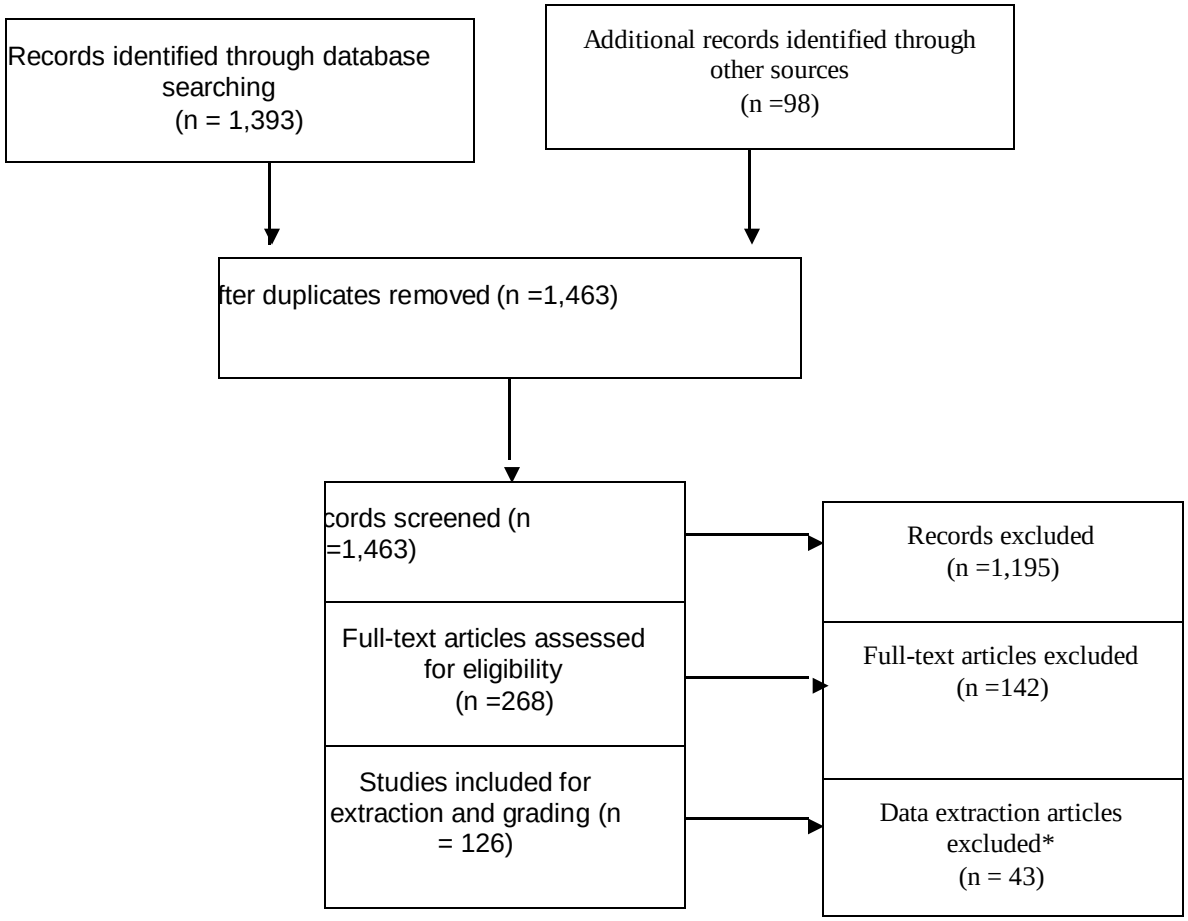
63. Ferguson J, Chamberlain P, Cramer HM, Wu HH. ER, PR, and Her2 immunocytochemistry on cell-transferred cytologic smears of primary and metastatic breast carcinomas: a comparison study with formalin-fixed cell blocks and surgical biopsies. *Diagn Cytopathol*. 2013;41(7):575-581.
64. Gong Y, Symmans 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