The following full text is a publisher’s version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/173083

Please be advised that this information was generated on 2019-04-16 and may be subject to change.
**Article 25fa pilot End User Agreement**

This publication is distributed under the terms of Article 25fa of the Dutch Copyright Act (Auteurswet) with explicit consent by the author. Dutch law entitles the maker of a short scientific work funded either wholly or partially by Dutch public funds to make that work publicly available for no consideration following a reasonable period of time after the work was first published, provided that clear reference is made to the source of the first publication of the work.

This publication is distributed under The Association of Universities in the Netherlands (VSNU) ‘Article 25fa implementation’ pilot project. In this pilot research outputs of researchers employed by Dutch Universities that comply with the legal requirements of Article 25fa of the Dutch Copyright Act are distributed online and free of cost or other barriers in institutional repositories. Research outputs are distributed six months after their first online publication in the original published version and with proper attribution to the source of the original publication.

You are permitted to download and use the publication for personal purposes. All rights remain with the author(s) and/or copyrights owner(s) of this work. Any use of the publication other than authorised under this licence or copyright law is prohibited.

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please contact the Library through email: copyright@ubn.ru.nl, or send a letter to:

University Library  
Radboud University  
Copyright Information Point  
PO Box 9100  
6500 HA Nijmegen

You will be contacted as soon as possible.
Evolution of Quality Assurance for Clinical Immunohistochemistry in the Era of Precision Medicine – Part 2: Immunohistochemistry Test Performance Characteristics

Emina E. Torlakovic, MD, PhD,*†† Carol C. Cheung, MD, PhD, JD,*§
Corrado D’Arrigo, MB, ChB, PhD, FRCPath,*‡‡‡ Manfred Dietel, MD, PhD,**
Glenn D. Francis, MBBS, FRCPath, MBA, FFSc (RCPA),††††††††§§§ C. Blake Gilks, MD,†††
Jacqueline A. Hall, PhD,‡‡‡‡‡‡ Jason L. Hornick, MD, PhD,### Merdoll Ibrahim, PhD***
Antonio Marchetti, MD, PhD,++++ Keith Miller, FIBMS,**** J. Han van Krieken, MD, PhD,++++
Soren Nielsen, BMS,#### Paul E. Swanson, MD,##### Mogens Vyberg, MD,######
Xiaoge Zhou, MD,####### Clive R. Taylor, MD,#### and
From the International Society for Immunohistochemistry and Molecular Morphology (ISIMM) and International Quality Network for Pathology (IQN Path)

Abstract: All laboratory tests have test performance characteristics (TPCs), whether or not they are explicitly known to the laboratorian or the pathologist. TPCs are thus also an integral characteristic of immunohistochemistry (IHC) tests and other in situ, cell-based molecular assays such as DNA or RNA in situ hybridization or aptamer-based testing. Because of their descriptive, in situ, cell-based nature, IHC tests have a limited repertoire of appropriate TPCs. Although only a few TPCs are relevant to IHC, proper selection of informative TPCs is nonetheless essential for the development of and adherence to appropriate quality assurance measures in the IHC laboratory. This paper describes the TPCs that are relevant to IHC testing and emphasizes the role of TPCs in the validation of IHC tests.

This is part 2 of the 4-part series “Evolution of Quality Assurance for Clinical Immunohistochemistry in the Era of Precision Medicine.”

Key Words: biomarkers, quality assurance, quality control, validation, immunohistochemistry, test performance characteristics

(Appl Immunohistochem Mol Morphol 2017;25:79–85)

Historically, immunohistochemistry (IHC) has for all practical purposes been considered a “special stain” similar to traditional histochemical preparations; how-
ever, it is more accurately described as a “cell-based in situ immunoassay” and resides firmly within a class of on-slide, in situ molecular tests that use one of a number of different probes (antibodies, RNAs, DNAs, aptamers or other synthetic molecules) and enable the identification of specific complementary molecular targets. Unlike histochemical tests, “on-slide molecular tests” use reagents that have high (or very high) sensitivity and specificity (in the nanomolar range) for individual targets. The great majority of in situ tests, including IHC tests, are developed in the individual laboratory and as such, they are best designated as “laboratory developed tests” (LDTs). Application of appropriate quality assurance measures and the proper validation of “in situ cell-based molecular tests” require a full understanding of which test performance characteristics (TPCs) apply and how validation of these tests is defined. In addition, it is necessary to adapt specific terminology relating to both quality assurance and validation that was initially developed for liquid-based assays. TPCs are essential for the characterization of LDTs. As we will show, TPCs as a group are an inherent component of LDT validation.

In this paper, we will define the TPCs that are relevant to in situ cell-based assays that have descriptive readouts primarily generated by human readers (ie, pathologists) ± assistive image analysis and will describe the role of TPCs in the validation of the IHC test. This work builds upon recently published recommendations and guidelines for the validation of antibodies, and molecular and IHC assays from the College of American Pathologists, recommendations for the standardization of controls from the Ad Hoc International Expert Committee, as well as research applications of biomarkers in preclinical and clinical trials for novel and emerging anticancer drugs.

**TPCs FOR IN SITU CELL-BASED ASSAYS**

Here, we adopt and define the following TPCs for IHC: (1) sensitivity and specificity, (2) reproducibility and robustness, (3) accuracy and precision, and (4) reportable range (Table 1).

**Sensitivity and Specificity**

An Ad Hoc Committee recently published a proposal for a new classification of sensitivity and specificity in IHC testing. The new concept, for the first time, allowed for observation of analytical sensitivity and specificity in a descriptive, nonlinear method such as IHC. The introduction of the concept of analytical sensitivity and specificity for descriptive tests, and its distinction from diagnostic sensitivity and specificity, and clinical sensitivity and specificity, resulted in a 3-tier classification of sensitivity and specificity, which sets the stage for building an appropriate validation framework for IHC. As will be shown technical validation is not possible in the absence of analytical sensitivity and analytical specificity. Diagnostic and clinical sensitivity and specificity do not apply to technical validation of IHC tests (even though a valid methodology based on proper technical validation is a fundamental requirement for clinical validation studies), but should be demonstrated/know before an IHC test is introduced into clinical practice. Therefore, they are also briefly discussed below.

**Analytical Sensitivity and Specificity**

Generally speaking, analytical sensitivity represents the smallest amount of substance in a sample that can accurately be measured (for quantitative assays) or detected (for qualitative assays) by an assay. In other words, analytical sensitivity determines the limit of detection (LOD). Analytical specificity refers to the ability of an assay to uniquely measure or detect one particular substance, rather than others. It is critical to recognize that for a descriptive method such as IHC, analytical sensitivity, and analytical specificity are also descriptive. A pertinent example is the demonstration of low molecular weight cytokeratin (LMW-CK) in hepatocytes (a known weak expressor of CK, thus suitable as a descriptive LOD cell type). If an IHC test for LMW-CK shows positivity in hepatocytes, then adequate analytical sensitivity has been demonstrated. LMW-CK is not expressed in benign

---

**TABLE 1. Test Performance Characteristic for Technical Validation of Immunohistochemistry (IHC) Assays**

<table>
<thead>
<tr>
<th>Test Performance Characteristic</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analytical sensitivity</td>
<td>Determines LOD* of an IHC test</td>
</tr>
<tr>
<td>Analytical specificity</td>
<td>Ability of IHC assay to detect one particular substance, rather than others</td>
</tr>
<tr>
<td>Preanalytical reproducibility/robustness</td>
<td>Ability of IHC assay to produce the same results with the same tissue sample for various preanalytical conditions</td>
</tr>
<tr>
<td>Analytical reproducibility/robustness</td>
<td>Ability of IHC assay to produce the same results with the same tissue sample for various analytical conditions</td>
</tr>
<tr>
<td>Readout accuracy</td>
<td>Readout agreement with a designated gold standard or designated true value for the readout</td>
</tr>
<tr>
<td>Readout precision</td>
<td>Readout reproducibility (intraobserver and interobserver reproducibility of the readout)</td>
</tr>
<tr>
<td>Reportable range</td>
<td>Clinically relevant range of values that can be described (or measured by image analysis) for a certain antigen that can be detected by IHC in specimens known to express the antigen with a provision that the specimens had undergone ideal/standard tissue processing</td>
</tr>
<tr>
<td>EQA accuracy for reportable range</td>
<td>Ability of IHC test to produce a correct result in EQA (PT or interlaboratory), where external IHC results act as the agreed upon actual (true) value</td>
</tr>
<tr>
<td>NIM accuracy for reportable range</td>
<td>Ability of IHC test to produce a correct result in NIM (PT, interlaboratory, or intralaboratory), where results of another (previously validated) method (eg, FISH, PCR, other) act as the agreed upon actual (true) value</td>
</tr>
</tbody>
</table>

*Limit of detection (LOD): the smallest amount of substance in a sample that can accurately be detected by an assay. EQA, external quality assurance; FISH, fluorescent in situ hybridization; NIM, non-IHC method; PCR, polymerase chain reaction; PT, proficiency testing.
lymphocytes, skeletal muscle, and many other tissues. If an IHC test for LMW-CK is reproducibly negative in known negative tissues, then adequate analytical specificity has been demonstrated.27 In the clinical laboratory, analytical sensitivity and specificity are traditionally determined against an analytical reference standard test (ie, a criterion standard or colloquially, a “gold standard”), but this is currently not the norm in IHC testing because for most IHC tests, gold standards are not yet defined. An Ad Hoc Expert Committee recently initiated development of descriptive gold standards for descriptive analytical sensitivity and specificity in IHC testing by establishing the concept of standardized controls for IHC tests, also referred to as IHC critical assay performance controls (iCAPCs).17 Although iCAPCs include relevant LOD tissue(s), analytical specificity is only narrowly addressed. To establish a more extended analytical specificity it is desirable to include a spectrum of benign human tissues in the evaluation of analytical specificity. Such tissues may be combined into a so-called “specificity tissue microarray (TMA)” (see part 4 of the Evolution Series).

Diagnostic Sensitivity and Specificity

Diagnostic sensitivity is the probability that a test outcome is positive in a diseased person (eg, for IHC tests this usually relates to a particular type of tumor), and diagnostic specificity is the probability that a test outcome is negative in a nondiseased person (eg, in a different type of tumor by IHC). In a clinical setting, 2 useful measures are positive predictive value and negative predictive value. Positive predictive value is defined as the probability that a person randomly chosen from the population who has tested positive actually has the disease; negative predictive value is the probability that a person who has tested negative does not have the disease. Clearly, these definitions assume that the values are calculated and are relevant in a defined diagnostic setting.28 Therefore, if sensitivity and specificity are defined in these terms, we refer to “diagnostic sensitivity and specificity”; this concept is applicable to IHC in the setting of diagnosing a disease or a condition. Diagnostic sensitivity and specificity of IHC tests are often published in the pathology literature (eg, sensitivity and specificity of DOG1 IHC assay in the diagnosis of gastrointestinal stromal tumor).29–31

Clinical Sensitivity and Specificity

To establish clinical sensitivity and specificity (eg, patient response to hormonal therapy for estrogen receptor-positive breast cancer), clinical trials are necessary; this requirement generally goes beyond the purview of an individual laboratory.32,33

Clinical utility is a concept that incorporates both diagnostic and clinical sensitivity and specificity. Diagnostic and clinical sensitivity and specificity may be known either from studies performed by the laboratory itself, or those reported in published literature or other reliable sources. For example, studies that document diagnostic, prognostic, and predictive relevance of IHC tests are frequently provided in the peer-reviewed literature.34–39 IHC tests that have not been diagnostically and clinically validated in published literature should not be introduced in a clinical IHC laboratory without that laboratory itself performing a diagnostic validation and/or clinical validation of its own. Tissue tools for demonstration of IHC analytical sensitivity and specificity are described in part 4 of the Evolution Series.

In summary, “clinical sensitivity and specificity” relate to the ability of an IHC test to predict clinical outcomes and clinical responses to therapeutic agents (prognostic and predictive); “diagnostic sensitivity and specificity” relate to the ability of an IHC test to diagnose a disease or biological state; “analytical sensitivity and specificity” relate to the ability of an IHC test to clearly demonstrate signal in an agreed upon “LOD” cell type, along with optimal signal-to-noise ratio in a fit-for-purpose manner.

Reproducibility and Robustness

Reproducibility and robustness were not specifically addressed by the recent College of American Pathologists guidelines for IHC validation.15 However, we would argue that because reproducibility is defined as the ability of a test to produce the same result for a given sample when repeatedly tested, and robustness relates to test reproducibility given some changes in test conditions, both are applicable to IHC validation.

Reproducibility Versus Precision

Reproducibility is the ability of a test to produce the same results in repeat testing within a laboratory (it may also be assessed across different laboratories as discussed below). It has been argued that reproducibility is to a qualitative test what “precision” is to a quantitative test.7,10,11,19,21,40 Therefore, for qualitative IHC assays, the term “reproducibility” is preferred and the term “precision” should not be used.

How to Measure Reproducibility?

Reproducibility is measured by the percentage of times the same results are obtained when a test is performed repeatedly on the same sample (specimen). Although this approach to measuring reproducibility is applicable to IHC testing, the phrase “the same results” needs to be defined for each test because IHC tests are descriptive in nature (eg, for CD31 testing, demonstration of strong staining of endothelial cells plus weak staining of mantle zone lymphocytes in 9 of 10 slides from the same paraffin block demonstrates 90% reproducibility). Intralaboratory reproducibility for qualitative testing was also recently referred to as “accordance.”41 The difference between the usual understanding of intralaboratory reproducibility and accordance is that reproducibility is expressed as a percentage of the same set of results (compared with a predefined standard, such as control materials), while accordance is measured by using paired samples without a predefined standard; it is a measure (also in percentage) of how often the paired samples produce the same results (positive vs. negative). Although
reproducibility can be easily accessed by monitoring results of “on-slide standardized (or well defined) controls” daily, to assess accordance, we would need to prepare at least 10 paired samples and compare their results for agreement in a separate evaluation.

Which Samples to Use to Demonstrate Reproducibility?

Proper selection of quality control samples is critical for demonstration of reproducibility in a descriptive test such as IHC, and such samples should optimally include LOD tissues/cell types. Therefore, iCAPCs or similar samples are well suited to be used for demonstration of reproducibility of qualitative IHC assays (see part 4 of the Evolution Series).

How Many Repeats of a Given Test are Necessary to Demonstrate Reproducibility?

The number of repeat tests that is sufficient to demonstrate reproducibility may be determined by regulatory agencies, professional bodies, or by a laboratory medical director. If standardized on-slide controls are used for IHC testing, then the reproducibility of the test is continually demonstrated and monitored with each IHC slide, because the obtained results are regularly inspected at the intralaboratory QC stage of IHC testing. This approach is endorsed because in clinical practice, slides that are deemed less than optimal must be recognized, investigated (with laboratory-initiated repeat testing if necessary), and properly documented. Whether a laboratory-identified suboptimal IHC slide is released to the pathologist who ordered the test is usually the decision of the medical director of the laboratory. A recent publication reported approximately 98% overall reproducibility of the IHC tests in a single large clinical laboratory.42

Robustness

The robustness of an IHC test (or other in situ cell-based assays) is defined as test reproducibility in the face of changes in various test conditions.43,44 These changes are often in preanalytical conditions, such as fixation time, decalcification, warm and cold ischemic time, among others. Reproducibility of IHC results for a clinically relevant range of preanalytical conditions defines preanalytical robustness. Similarly, reproducibility of IHC results for a clinically relevant range of analytical conditions (eg, different equipment, different operators, etc.) defines analytical robustness of the IHC test. Demonstration of preanalytical robustness is under the purview of the IHC laboratory, while analytical robustness is the responsibility of both the laboratory and industry (eg, manufacturers/suppliers of IHC reagents and equipment).

Many IHC laboratories do not establish robustness because robustness has not yet been mandated for IHC validation and the number of tissue samples for establishing preanalytical or analytical robustness in IHC validation has not yet been defined. Tissue tools for demonstration of IHC reproducibility and robustness are described in part 4 of the Evolution Series.

Accuracy and Precision

Accuracy is usually defined as the degree of closeness of a test measurement to the actual (true) or expected (previously measured) value; this approach to accuracy is not applicable to most IHC testing.43-45 Similar to polymerase chain reaction tests, IHC tests use amplification and are generally qualitative, and therefore cannot be analyzed based on exact quantitative measurements; this is in contrast to other, more quantitative methods that do not use variable amplification, where the output of the test more closely reflects actual protein content/ expression (eg, mass spectrometry, or even immunoassays that use directly labelled primary antibody, such as multicolor flow cytometry and direct immunofluorescence).10 However, accuracy may be applicable to IHC testing in defined circumstances as follows: (1) accuracy of the IHC protocol and (2) accuracy of the IHC readout.

(1) IHC protocol accuracy is sometimes inappropriately referred to as “concordance.” “Concordance” has also been considered as the equivalent of reproducibility when measured across different laboratories; namely the percentage chance that 2 identical test samples sent to different laboratories will both yield the same result (ie, both found positive or both found negative result).41 However, in this context most of the time one of the results is considered as the designated gold standard result (designated “true value”) and therefore, accuracy is a more appropriate term because concordance reflects agreement of the paired samples irrespective of the correctness of the results. In clinical IHC, the output (“stained” slide) is compared with a designated “gold standard,” where the designated gold standard result (designated “true value”) is provided by a reference laboratory (either through interlaboratory exchange or through a proficiency testing or PT program). In this context, the IHC protocol accuracy is best referred to as external quality assurance (EQA) accuracy.

Where the designated gold standard result (designated “true value”) is determined using a non-IHC based testing method, the IHC protocol accuracy is best referred to as non-IHC methodology (NIM) accuracy. In NIM accuracy, the ability of an IHC protocol to produce a correct result is assessed in comparison with another (non-IHC) test methodology for the same biological condition (eg, comparison of IHC test results with results obtained using fluorescent in situ hybridization or polymerase chain reaction testing).46 EQA accuracy may involve a reference laboratory that provides both IHC and fluorescent in situ hybridization results for a set of PT samples while testing for NIM accuracy may be performed by the same laboratory performing the validation (intralaboratory NIM accuracy) or by an external laboratory.

To assess IHC protocol accuracy, samples that contain no analyte, as well as samples with low, moderate, and high levels of analyte are usually included. Professional organizations are likely to
continue the practice of issuing recommendations regarding the number of samples for intralaboratory NIM accuracy, as well as the number of samples for interlaboratory IHC exchanges and/or proficiency testing that need to be included to enable meaningful calculations. 14,46–50 Although IHC protocol accuracy is also often referred to as “concordance,” when comparison is conducted with the agreed upon “true value” of the IHC test, or with a designated “true value” using another methodology, then “EQA accuracy” and “NIM accuracy” are the preferred terms. Concordance could be used when there is no reference to “correct results,” but only interlaboratory reproducibility is being assessed.

(2) There are many types of nonquantitative measurements that are applicable to the IHC readout, which produce different types of data (eg, binary data, categorical data, ordinal data, etc). 10 Where accuracy and precision apply only to the validation of the readout as a specific segment of the IHC protocol, they should be designated as “readout accuracy” and “readout precision.” Readout accuracy refers to readout agreement with a designated gold standard or designated true value, and readout precision refers to reproducibility of the readout per se (intraobserver and interobserver reproducibility of the readout). Readout accuracy and readout precision should be determined for type 2-IHC tests (class II and class III U.S. Food and Drug Administration IHC tests), and for type 1-IHC tests (class I) where the read out includes an element of quantification (eg, % cells with a defined qualitative cutoff point that distinguishes between a “negative” and “positive” result). Note that readout accuracy and readout precision are not the same (eg, image analysis is expected to have very high readout precision, but it may have very low readout accuracy if an inappropriately calibrated algorithm is used for the readout). The level of acceptability for readout accuracy and readout precision of various IHC tests may be set by regulatory bodies, professional organizations, or expert/subspecialty groups. Readout accuracy and readout precision are particularly relevant for biomarkers that are used for targeted therapy. 32

Reportable Range

For liquid-based quantitative laboratory tests, “analytical measurement range” is the term used by the College of American Pathologists to refer to the range of values that a quantitative method can measure for a sample, without dilution, concentration, or other pretreatment of the specimen before testing. Analytical measurement range is the same as “reportable range” in The Clinical Laboratory Improvement Amendments terminology. 31 When applied to IHC testing, which is a descriptive qualitative technique, the reportable range is usually considered to be the clinically relevant range of results that can be described (or measured by image analysis) for a given IHC target that can be detected in specimens known to express the antigen, with the provision that the specimens had undergone ideal tissue processing (eg, ischemic time <1 h, overnight fixation in formalin, and embedding in paraffin). For example, a given range may extend from rare cells positive, to all cells positive, etc. Depending on the IHC protocol conditions, the intensity of staining, above a threshold defined in the controls, may or may not be relevant. Examples include range of expression of estrogen receptor in breast cancer, range of expression of CD10 in follicular lymphoma, etc. Depending on the objective of the test, 3 to 10 samples may be well representative of the reportable range. Reportable range samples may be combined in the form of a TMA paraffin block (also known as “Index TMA”; see part 4 of the Evolution Series). Depending on the type and the number of cases included, accuracy TMA may also be suitable for demonstrating reportable range for most IHC assays (see part 4 of the Evolution Series).

Reportable range is also relevant for the proper selection of quality control samples. 17 Selection of quality control samples for daily controls, and other quality control samples must satisfy the minimum requirement of a negative control and 2 positive controls containing different concentrations of the target antigen, to be included in every run. Note that for modern automated IHC testing, each slide in fact represents a separate run and, hence, the on-slide control rule applies. 17,42 The expression levels should be chosen to verify assay performance for relevant analytic and clinical decision points. This usually means including a high-expressor positive control, near the upper limit of the reportable range, as well as a low-expressor positive control, near or at the lower LOD. 52

TPCs IN TEST VALIDATION

Demonstration of relevant TPCs is required for technical IHC assay validation. In this discussion it is assumed that primary antibodies and other reagents, as well as equipment that will be used in clinical IHC testing, have already been validated either by industry, or in published literature. Therefore, clinical IHC laboratories must perform additional technical validation of IHC protocols. Furthermore, the medical director (or designate) of the clinical IHC laboratory is responsible for validation of IHC readouts where applicable.

Technical validation of IHC testing is described in detail in part 3 of Evolution Series.

REFERENCES


