Quantitative Assessment of Immunohistochemistry Laboratory Performance by Measuring Analytic Response Curves and Limits of Detection

Seshi R. Sompuram, PhD; Kodela Vani, MS; Anika K. Schaedle, BS; Anuradha Balasubramanian, PhD; Steven A. Bogen, MD, PhD

Context.—Numerous studies highlight interlaboratory performance variability in diagnostic immunohistochemistry (IHC) testing. Despite substantial improvements over the years, the inability to quantitatively and objectively assess immunostain sensitivity complicates interlaboratory standardization.

Objective.—To quantitatively and objectively assess the sensitivity of the immunohistochemical stains for human epidermal growth factor receptor type 2 (HER2), estrogen receptor (ER), and progesterone receptor (PR) across IHC laboratories in a proficiency testing format. We measure sensitivity with parameters that are new to the field of diagnostic IHC: analytic response curves and limits of detection.

Design.—Thirty-nine diagnostic IHC laboratories stained a set of 3 slides, one each for HER2, ER, and PR. Each slide incorporated a positive tissue section and IHControls at 5 different concentrations. The IHControls comprise cell-sized clear microbeads coated with defined concentrations of analyte (HER2, ER, and/or PR). The laboratories identified the limits of detection and then mailed the slides for quantitative assessment.

Results.—Each commercial immunostain demonstrated a characteristic analytic response curve, reflecting strong reproducibility among IHC laboratories using the same automation and reagents prepared per current Good Manufacturing Practices. However, when comparing different commercial vendors (using different reagents), the data reveal up to 100-fold differences in analytic sensitivity. For proficiency testing purposes, quantitative assessment using analytic response curves was superior to subjective interpretation of limits of detection.

Conclusions.—Assessment of IHC laboratory performance by quantitative measurement of analytic response curves is a powerful, objective tool for identifying outlier IHC laboratories. It uniquely evaluates immunostain performance across a range of defined analyte concentrations.

(Arch Pathol Lab Med. 2018;142:851–862; doi: 10.5858/arpa.2017-0330-OA)
reckoning is that it does not provide for objective feedback, verifying the actual navigational position. Dead reckoning does not as optimally account for the unexpected drift off-course. This is the type of feedback lacking in the field of diagnostic IHC.

For a typical IHC laboratory, generating an analytic response curve has not been feasible, until now. In other laboratory disciplines, this would be analogous to verifying calibration. Short of that, the average laboratory can verify immunostain performance by using slides mounted with cell lines and tissue arrays, such as are available through national proficiency testing (PT) surveys. However, cell lines and tissue arrays are mostly of unknown analyte concentrations. Laboratories do not measure analytic response curves as part of any proficiency test. Moreover, PT data is comprised of formalin-fixed, paraffin-embedded tissue blocks were obtained from the Biorepository of the Department of Pathology & Laboratory Medicine, Tufts Medical Center, Boston, Massachusetts, under an approved Institutional Review Board protocol. For ER and PR tissue controls, we mounted pathologic discard samples of normal uterine endometrium on slides. We measured only the stromal cells, as they express lower levels of ER and PR than the glandular epithelium. HER2 controls were composed of 2 different breast carcinomas, one expressing HER2 at a 3+ level and another expressing HER2 at a 1- level.

The slides for HER2, ER, and PR were prepared by marking each with a study number, slide serial number, and the analyte initials (eg, HER2, ER, or PR), as shown in Figure 1. On the premounted tissue slides, IHControls were spotted in 2 rows, 1 µL per spot. One row was above the tissue sample (toward the label) and the other below. For all but 1 IHControl, the top and bottom rows used different manufactured lots with slightly different analytic concentrations. Different lots were used to test the reproducibility of our manufacturing process. There were no noticeable differences between the two. Since the 2 rows are not exactly identical, they are not replicates. We picked 1 row (bottom) for the graphs shown in this article. For the lowest concentration IHControl (140 molecules per microbead) we used the same lot. We did not prepare 2 separate lots at that (lowest) concentration. The IHControl concentrations are shown in the Table.

The IHControls are compatible with all of the commercial heat-induced epitope retrieval (HER) protocols we evaluated except one. The IHControls microbeads will sometimes detach from the slide when heated to boiling in a strongly alkaline buffer. This was a problem only for the Dako PT-Link HIER buffer (Dako/Agilent Corp, Carpinteria, California), such as used for the ER/PR PharmDx kit (Dako/Agilent). Therefore, for laboratories using the PT-Link, an additional brief (15 minute) fixation step was added. For slides destined for use with the high pH Dako PT-Link, the IHControls spots were covered with 5 µL of formalin and incubated at room temperature for 15 minutes, rinsed with distilled water, and air-dried. This additional fixation step hardened the matrix, ensuring that the microbeads remained affixed to the slides. This additional fixation step affected the analytic response curves that include PR 636 and 1294.

Mailing the Test Slides to Participant Hospitals

Three separate test slides (HER2, ER, and PR), each with their respective tissue control and 2 rows of IHControls spots (as described above), were placed in plastic slide mailers and shipped to participating hospitals either by overnight express mail (if beyond the Boston area) or by local courier (within the Boston vicinity). To prevent freezing of the slides during shipping, the slide mailer was placed inside an insulated foam box, sandwiched with
The antigen retrieval (HIER) methods also varied among IHC laboratories. One laboratory did not perform HIER for HER2 because they used a formalin substitute. HIER was not warranted in that context.

### Determination of Limit of Detection

Pathologists at participating hospitals were asked to examine each of the 5 IHControls concentrations per row and identify the lowest analyte concentration that yields a clearly discernable stain. We called this the *limit of detection*. An explanation and examples were provided with an IHControls Evaluation Guide that was included in the mailing. The participants provided their evaluation of the LOD for each immunostain.

### Image and Data Acquisition and Analysis

Quantitative image analysis was performed in-house by using the returned stained slides. Images were acquired with a Zeiss Axioskop microscope fitted with a Spot Imaging Solutions Insight Gigabit CCD camera (Diagnostic Instruments Inc, Sterling Heights, Michigan). Briefly, IHControls quantification was performed by using a custom image analysis algorithm embedded in MatLab (The MathWorks Inc, Natick, Massachusetts), as previously described.\(^1\) The algorithm segments the images so as to identify microbeads, extracts the brown color by using color vector analysis, and then quantifies the mean pixel intensity along the microbead rim. Color intensity of the stained test microbead is expressed as a ratio to the color intensity of a color standard microbead. Using an internal color standard promoted reproducibility in measurement (by normalizing variability in optical settings). HER2 tissue stain quantification was performed by using the Immunomembrane plugin for Image J, as previously described.\(^2\)

## IHControl Product Concentrations (Molecules/Microbead)*

| Spot \(^b\) | Epitope | Concentration |
|-----------|---------|--------------|
| 1         | ER:1D5  | 1 103 921    |
|           | 6F11    | 924 478      |
|           | SP1     | 703 460      |
|           | PR: 636 | 800 503      |
|           | 1294/1E2| 1 379 113    |
|           | 16      | 1 089 808    |
|           | HER2: HercepTest/4B5/CB11/1E3 | 1 086 658 |
|           | SP3     | 1 023 265    |
| 2         | ER:1D5  | 934 651      |
|           | 6F11    | 703 460      |
|           | SP1     | 876 853      |
|           | PR: 636 | 1 096 864    |
|           | 1294/1E2| 1 365 001    |
|           | 16      | 1 146 258    |
|           | HER2: HercepTest/4B5/CB11/1E3 | 1 051 490 |
|           | SP3     | 876 853      |
| 3         | All     | 77 913       |
| 4         | All     | 91 494       |
| 5         | All     | 8187         |
| 6         | All     | 8357         |
| 7         | All     | 928          |
| 8         | All     | 956          |
| 9         | All     | 1 40         |
| 10        | All     | 1 40         |

Abbreviations: ER, estrogen receptor; HER2, human epidermal growth factor receptor type 2; PR, progesterone receptor.

* Analyte concentration is expressed in MEF, molecules of equivalent fluorochrome, as described in Materials and Methods.

\(^{\text{b}}\) Spots 1 and 6; 2 and 7; 3 and 8; and 4 and 9 are different manufactured lots of the same product and have slightly different concentrations. Spots 5 and 10 are the same lot.

## Participating Hospitals

Thirty-nine hospitals, scattered across the United States, participated in the study. The hospitals were located in 19 different states, the most common ones being Massachusetts (11), California (5), Illinois (3), New York (2), Rhode Island (2), Ohio (2), and Pennsylvania (2). Each hospital was asked to stain the test slides in exactly the same way as they would their patient slides.

Numerous automated immunostainers were represented, including the Autostainer Link 48 and Omni (Dako Corp/Agilent), Bond III (Leica Corp, Buffalo Grove, Illinois), Benchmark XT and Benchmark Ultra (Ventana Medical Corp/Roche, Tucson, Arizona), and Labvision Autostainer 360 (ThermoScientific, Waltham, Massachusetts). One laboratory stained IHC slides manually. Seven IHC laboratories used more than 1 instrument vendor, one for ER and PR and another for HER2.

The antigen retrieval (HIER) methods also varied among IHC laboratories. For ER and PR, 33 of 39 laboratories used the HIER method associated with their instrument. For HER2, 9 laboratories used a heated water bath or steamer, 1 used a pressure cooker, and the remainder used the HIER method associated with their instrument. One laboratory did not perform HIER for HER2 because they used a formalin substitute. HIER was not warranted in that context.

### Determination of Limit of Detection

Pathologists at participating hospitals were asked to examine each of the 5 IHControls concentrations per row and identify the lowest analyte concentration that yields a clearly discernable stain. We called this the *limit of detection*. An explanation and examples were provided with an IHControls Evaluation Guide that was included in the mailing. The participants provided their evaluation of the LOD for each immunostain.

### Image and Data Acquisition and Analysis

Quantitative image analysis was performed in-house by using the returned stained slides. Images were acquired with a Zeiss Axioskop microscope fitted with a Spot Imaging Solutions Insight Gigabit CCD camera (Diagnostic Instruments Inc, Sterling Heights, Michigan). Briefly, IHControls quantification was performed by using a custom image analysis algorithm embedded in MatLab (The MathWorks Inc, Natick, Massachusetts), as previously described.\(^1\) The algorithm segments the images so as to identify microbeads, extracts the brown color by using color vector analysis, and then quantifies the mean pixel intensity along the microbead rim. Color intensity of the stained test microbead is expressed as a ratio to the color intensity of a color standard microbead. Using an internal color standard promoted reproducibility in measurement (by normalizing variability in optical settings). HER2 tissue stain intensity was evaluated by using the Immunomembrane plugin for Image J, as previously described.\(^2\)

Stain intensity in this program is quantified on a relative 0 to 10 scale. Estrogen receptor/PR stain intensity was evaluated by using Image Pro Premier (Media Cybernetics, Rockville, Maryland), as previously described.\(^3\) Estro-
gen receptor/PR stain intensity is quantified as mean pixel intensity, on a relative 0 to 255 scale.

**Statistical Analysis**

For experiments performed in-house (Figure 1), each data point represents the mean ± standard deviation (SD) of triplicate slides. Each slide bears an IHControl spot containing approximately 5000 analyte-coated microbeads. To quantify a single IHControl spot, we sampled 3 different microscopic areas of the same spot (>40 objective lens). This is analogous to sampling 3 fields of a patient’s breast carcinoma for assessment of HER2 or ER/PR. From these 3 fields, we calculated the mean stain intensity per spot (slide). Each slide had 2 rows of IHControls. As previously described (under Preparation of Survey Tool in Materials and Methods), the IHControls in the 2 rows were from different manufactured lots and have slightly different concentrations. Therefore, the 2 rows were not treated as replicates of each other. For all of the data, the bottom row (further from the slide label) was used for the graphs shown in this article. For survey data, each data point represents the mean ± SD from a single slide, as is typically performed for PT.

**Validation of the Survey Tool**

Before initiating this survey, we performed a series of experiments designed to compare immunostaining of IHControls to tissue samples. In particular, we wished to understand how each responds when an immunostain is not performed properly. The purpose of this validation effort was to determine if outliers are equally detected by both types of test samples (IHControls and tissue sections). If deviations are detected by stain intensity of IHControls, does that directly translate to problems with the staining of patient samples? To perform this evaluation, a series of immunostains for HER2, ER, and PR were performed at a wide range of primary antibody dilutions, as previously described. Increasing dilutions simulated worsening degradation of the primary antibody. Each slide had both a tissue section and IHControls across a range of analyte concentrations. The stain intensities of the tissue samples and IHControls were quantified and compared to one another.

**RESULTS**

**Survey Tool Description**

Each participating laboratory received 3 slides, one each for HER2, ER, and PR. Figure 1 depicts a schematic representation of an exemplary ER survey slide. This is not an exact image. Slides did not have “high” or “low” designations or numbers; these are added for clarity. The tissue section is bounded both above and below with small (1 µL) spots of IHControls. The color of the spots has been enhanced so as to be plainly visible. Also, the 2 rows of IHControls are not exact duplicates, as different manufactured lots were used for 4 of the 5 spots in the 2 different rows (to evaluate manufacturing reproducibility). The exact concentrations of HER2, ER, or PR IHControls are listed in the Table. Each spot holds approximately 5000 microbeads, some bearing the relevant formalin-fixed analyte and others an antigenically irrelevant negative control. Laboratory staff judged the LOD and mailed back their slides for image quantification. Additional details on the survey are described in Materials and Methods.

**Survey Tool Validation Data**

If IHControls and tissue samples are functionally interchangeable in evaluating immunostain performance, then the 2 types of samples (tissue sections and IHControls) will produce comparable stain intensities. If the immunostain is performed correctly, both types of samples will be strongly stained. Conversely, if the immunostain is defective, both should be weakly stained or unstained. If we graph stain intensity of tissue sections (x-axis) versus stain intensity of IHControls (y-axis), then a linear relationship passing through the origin (where y = x = 0) would show interchangeability of the 2 types of samples. The dashed black lines in Figure 2 represent this theoretical equivalence.

Figure 2 shows 3 such graphs evaluating equivalence of IHControls and tissue sections, one each for HER2 (Figure 2, A), ER (Figure 2, B), and PR (Figure 2, C). The slides for this initial validation study were like those in Figure 1, but stained in-house with varying dilutions of primary antibodies. The collection of slides in this validation study included some with normal immunostaining, others with mild or moderate diminution of stain intensity, and yet others with no stain at all. Immunostain intensity depended on the primary antibody dilution. The complete spectrum of stain intensity was represented (Materials and Methods).

Figure 2, A, illustrates the stain intensity relationship for HER2 immunostaining of IHControls versus a 3+ HER2 breast carcinoma. In this particular example, the IHControls and the tissue section were both stained with the HercepTest kit (Dako/Agilent). Figure 2, A, shows that the stain intensity relationship depends on which IHControl concentration is used. The IHControl bearing 77 913 molecules per microbead (Figure 2, A; green line) yields a regression line that most closely approximates the dashed line of equivalence. This regression line shows that strong staining on the tissue sample equates to strong staining on the 77 913 molecules IHControl. Similarly, weak staining of the tissue sample equates to weak staining of the 77 913 molecules IHControl. The other IHControls do not match the tissue sample as closely.

Figure 2, B, shows a series of regression lines for the SP1 ER immunostain, comparing the stain intensity of a section of normal endometrial stroma to a group of IHControls at different ER concentrations. There is no perfect match. Endometrial stromal cells were scored (rather than the epithelial cells associated with the glands) because they express lower levels of ER and PR. Figure 2, B, suggests that the best match to this particular tissue sample may be with an IHControl having an ER concentration between 1331 and 8187 molecules per microbead.

Figure 2, C, illustrates the relationship with the PR 1E2 immunostain. The regression lines suggest that the best match to the tissue sample will be with an IHControl that has a PR concentration between 77 913 and 1 379 113 molecules per microbead.

We conclude that there is a close match between the IHControls and tissue sample for HER2 (Figure 2, A), but not for ER (Figure 2, B) and PR (Figure 2, C). However, the match appears entirely dependent on analyte concentration. There is nothing special about the analyte concentration of the tissue sample that we randomly selected. More important, the graphs of Figure 2, A through C, demonstrate the importance of analyte concentration when evaluating immunostain performance. This is in accord with a recent publication. For this reason, we believe that the availability of a broad range of known concentrations, as are available with IHControls, is advantageous. Testing over a range of analyte concentrations can offer insights not otherwise possible with tissue samples of unknown concentration.

**HER2 Consensus Performance**

Figure 3, A, shows the HER2 analytic response curves for HercepTest and 4B5 immunostain. The 4B5 curve is shifted...
approximately 1 log to the right of the HercepTest curve. These consensus curves are also almost identical to previously published analytic response curves.9 The previous report was not a survey; analytic response curves were generated from a single instrument (1 for each commercial vendor). The analytic response curves for other HER2 immunostains, using monoclonal antibodies CB11, SP3, and EP3, are not shown because the study did not include at least 5 participating laboratories.

Figure 3, B, shows the analytic response curves of the 5 participating laboratories that comprise the HercepTest consensus curve from Figure 3, A. It is noteworthy that laboratories 1 through 3 and 5 have almost identical analytic response curves. Laboratory 4 is different. This led us to examine the details of their protocol more closely. Like HercepTest, laboratory 4 reported using a polyclonal HER2 antibody from Dako/Agilent. Consequently, we initially concluded that laboratory 4 was using HercepTest. However, upon further questioning, we learned that the laboratory is using a laboratory-developed test that is similar to HercepTest. The laboratory uses a polyclonal HER2 primary antibody from the same vendor as HercepTest (Dako/Agilent) but couples it with a different vendor’s instrument and brand of immunodetection reagents. The fact that the 4 HercepTest analytic response curves are so close to one another demonstrates the high reproducibility associated with commercial US Food and Drug Administration (FDA)–cleared immunostains. Even though we later learned that laboratory 4 is not using HercepTest, we kept laboratory 4 in this group to illustrate the point. Apart from a few outliers, as will be described, the analytic response curves for commercial FDA-cleared in vitro diagnostic (IVD)–labeled immunostains had characteristic signatures.

HER2 Survey Outliers

From 30 diagnostic IHC laboratories that were evaluated for HER2 immunostaining, 3 showed unacceptably weak levels of immunostaining. Figure 4 exemplifies 2 quantitative methods for illustrating outlier IHC laboratory performance. Figure 4, A, illustrates the case of an outlier having an equivalent peer group, while Figure 4, B, shows an example that lacks an identical peer group. Figure 4, A, shows the same consensus analytic response curve for the 4B5 immunostain (red line) as in Figure 3, A. The dotted lines represent the 2SD confidence limits for the group of 19 participating laboratories using the 4B5 immunostain. All laboratories using the 4B5 monoclonal antibody were included in forming the 2SD limits. Figure 4, A, shows that laboratory 10 is an outlier, having noticeably weaker staining. Laboratory 10 demonstrated very weak staining when using the IHControl at 77 913 molecules per microbead. Therefore, it represents laboratory 10’s approximate threshold of detection. By comparison, the tissue section mounted on the same slide showed no staining (Figure 5, A). The absence of tissue immunostaining (Figure 5, A) indicates that the HER2 concentration in the 3+ tissue sample is below laboratory 10’s LOD (77 913 molecules).

Figure 4, B, illustrates an outlier with no corresponding peer group. Laboratory 27 used a laboratory-developed test with reagents and an instrument from different vendors. Laboratory 27’s analytic response curve shows no staining of the IHControls except at the highest concentration.
(1 086 658 molecules per microbead). The tissue section also showed no staining except at 1 edge, which had weak staining (not shown). In the absence of a peer group, we propose that the only reasonable option is to provide information (to the laboratory) on all of the other analytic response curves.

**Figure 3.** HER2 immunostains analytic response curves. Immunostain intensity (y-axis) is graphed as a function of HER2 concentration (molecules per microbead, x-axis). Immunostain intensity is enumerated as a ratio of test microbead intensity to the intensity of the color standard microbead (as per Materials and Methods). The consensus HercepTest and 4B5 immunostains from 5 and 19 diagnostic IHC laboratories (respectively) are graphed in (A) (left). Each point is the mean and standard deviation for the group of laboratories using that method. B, (right) Illustrates the analytic response curves for each of the 5 laboratories included in the HercepTest consensus curve in (A). Laboratory 4 uses a laboratory-developed test that is similar to HercepTest. Abbreviations: HER2, human epidermal growth factor receptor type 2; MEF, molecules of equivalent fluorochrome.

**Figure 4.** Two different methods of illustrating outlier laboratory performance. A, (left) Shows a well-established peer group (red line), comprising the 19 laboratories using the 4B5 immunostain. The dotted lines represent the upper and lower 2SD limits of that group. The green line represents the analytic response curve for laboratory 10, which is outside those limits. For laboratories without a peer group (B), laboratory 27 is compared to the analytic response curves of both peer groups without the delineation of specific bounds. These curves are exactly the same as shown in Figure 3, A. Laboratory 27 is an outlier from both of the peer groups. Abbreviations: HER2, human epidermal growth factor receptor type 2; MEF, molecules of equivalent fluorochrome.

**HER2 Limit of Detection Data**

The quantitative data previously described (Figures 3 and 4) were from image quantification measurements taken directly from the slide. This required mailing the slides back to the central laboratory. A potential alternative that avoids the need for mailing slides is for each participating...
laboratory to subjectively identify their own LOD. The LOD is the lowest concentration \textit{IHControl} that demonstrates clearly discernable staining. Figure 6 shows the LOD data for the HercepTest (Figure 6, A) and 4B5 (Figure 6, B) immunostains. The HER2 concentration is plotted along the x-axis. The percentage of laboratories reporting a particular LOD is plotted along the y-axis. The LOD data of Figure 6 mirror the analytic response curves for those same immunostains (Figure 3, A). Like the analytic response curve data shown in Figure 3, the LOD for the 4B5 immunostain is shifted to the right relative to HercepTest. In other words, the LOD data also show approximately a log difference in sensitivity between the 2 HER2 immunostains. The LOD data did not accurately identify unacceptable immunostaining performance. The 3 aforementioned diagnostic IHC laboratory outliers reported a similar LOD as the others. The staff at those 3 outlier laboratories were able to see enough stain, even if it was quite weak, as to correctly identify the same LOD as the others. On the other hand, a different IHC laboratory reported an LOD that was out-of-range when compared to the consensus. That laboratory is shown in Figure 6, B, as the single laboratory with an LOD to the far right, at approximately 10^6 molecules per microbead. These findings highlight the problem of postanalytic (interpretive) variability when trying to measure analytic performance.

**Progesterone Receptor Survey Data**

Figure 7 shows the analytic response curves for 3 commercial PR immunostains from 34 diagnostic IHC laboratories. The 1E2 immunostain is shifted to the right by approximately 1 log. The graph also shows that the 636 and 16 immunostains have similar analytic response curves. These sensitivity data closely align with our previously published in-house studies. Of the 34 laboratories, 3 had such low immunostaining intensity that they would be considered to have failed the assessment. Their analytic response curves are not shown for space considerations. These 3 laboratories were not the same laboratories that had problems with the HER2 survey.

The LOD data for the PR immunostains are shown in Figure 8, A through C. These data mirror the findings from Figure 7. The data demonstrate a similar LOD for PR 636 and 16. The fact that PR 1E2 immunostain has a higher LOD does not necessarily mean it is less sensitive than the others for reasons described in the Discussion.

Of the 3 aforementioned outlier PR laboratories, 2 would have been detected through the laboratories’ LOD measurement. These 2 are included in the group shown in Figure 8, C, in the vertical bar positioned to the far right, at the highest PR \textit{IHControl} concentration (1 379 113 molecules per microbead). In addition, 3 other diagnostic IHC laboratories are included in that group, despite the fact that objective quantitative measurement showed them to have adequate staining. This is another example of postanalytic (interpretive) variability. For each of these 3 laboratories with acceptable performance, the next lower concentration (77 913 PR molecules per microbead) demonstrated significantly lower but still detectable staining.

**Estrogen Receptor Survey Data**

Figure 9 depicts the analytic response curves for the SP1 and 6F11 immunostains from 34 laboratories. Other ER immunostains are not shown because there were not at least 5 participating laboratories on which to form a consensus. Like other HER2 and PR immunostains, the ER immunostains have characteristic analytic response curves. Figure 9 shows that the SP1 immunostain is more sensitive, by almost 2 logs. The 6F11 data for the highest \textit{IHControl} concentration are not shown because of a technical
limitation; there is background staining with the 6F11 immunostain at the highest IHControl concentration. This limitation was previously described.9

The corresponding ER LOD data are shown in Figure 10. The participants’ subjective assessment of sensitivity mirrored the image quantification data, albeit to a lesser magnitude. Of the 39 laboratories (34 shown in Figure 9 plus 5 others using different immunostains), 4 are outliers as based on the analytic response curves. Of these 4, two would have failed as based on the laboratory-reported LOD assessment. These 2 are represented in the farmost right vertical bars in Figure 10, A and B. There were no IHC laboratories with acceptable immunostaining (based on analytic response curves) that failed as based on the reported LOD.

DISCUSSION

The broad goal of this study is to evaluate a new tool (IHControls) for improving intralaboratory consistency and interlaboratory comparability in diagnostic IHC testing. A unique aspect of the tool is its calibration to specific analyte concentrations, in units of molecules (MEF) per microbead. We manufactured IHControls to have analyte concentrations that are spaced at approximately every log interval, facilitating the measurement of analytic response curves and LODs for each commercial HER2, ER, and PR immunostain. Analytic response curves are a step toward converting a 0 to 3+ IHC readout, which is not traceable to any objective standard, to an actual molecular concentration. In this study, we evaluated this new method of immunostain performance assessment in a PT survey format. For most participating laboratories, the feedback provided to participating laboratories through the survey was reassuring. In several instances, however, the feedback was instrumental in making the laboratory staff aware that their immunostains were notably weaker than those of their laboratory peers.

From an evaluation of 39 laboratories using a variety of different instruments and methods, we found some striking disparities. Analyte concentrations that were strongly positive on some immunostains were unstained on others. As exemplified through the LOD data (Figures 6, 8, and 10), the thresholds of detection vary among commercial immunostaining kits. This is potentially important for patient samples expressing analyte levels above the threshold of detection for one vendor’s immunostain but below that of a second.

It is important, however, to be cautious in making conclusions about these immunostains’ relative sensitivities. Our method of generating analytic response curves provides for valid comparisons of IHC laboratories using the same immunostain. For example, diagnostic IHC laboratories using the 4B5 HER2 immunostain can be compared to one another. However, the data do not prove superior sensitivity for one commercial immunostain over another. This is because we have not yet evaluated commutability of measurement. The term commutability refers to the ability of a reference or control material to yield similar results regardless of the assay method. In other laboratory disciplines that routinely use manufactured reference materials for PT, laboratories using different methods often obtain different measurements. Sometimes these differences can be large. Commutability is typically evaluated for

Figure 6. Limits of detection for the HER2 immunostains HercepTest (A) and 4B5 (B). These data are from the same laboratories from which analytic response curves were measured (Figure 3, A). The limit of detection analyses for SP3 and 1E3 HER2 antibodies are not included because there were not at least 5 laboratories per group. Abbreviations: HER2, human epidermal growth factor receptor type 2; LOD, limit of detection; MEF, molecules of equivalent fluorochrome.

Figure 7. Analytic response curves for PR immunostains 636, 16, and 1E2. Immunostain intensity (y-axis) is graphed as a function of PR peptide concentration (molecules per microbead, x-axis). The limit of detection analysis for PR 1294 is not included because there were not at least 5 laboratories per group. Each point is the mean and standard deviation for the group of laboratories. Abbreviations: MEF, molecules of equivalent fluorochrome; PR, progesterone receptor.
quantitative analytes using calibrated reference materials, which do not exist in the field of diagnostic IHC.

To our knowledge, Figures 3, A; 7; and 9 are the first published consensus analytic response curves for major commercial HER2, PR, and ER immunostains. Previously, McCabe et al19 and Welsh et al20 demonstrated the measurement of analytic response curves for laboratory-developed HER2 and ER immunostains, using calibrated cell lines that the investigators characterized. The relative unavailability of precisely calibrated cell lines like these, for broad use, with concentrations spanning the biologic range of expression, has limited their applicability for diagnostic IHC laboratories.

With the already described caveat of commutability, our data show the following sensitivity differences of a magnitude equal to or greater than 1 log: HER2 HercepTest greater than 4B5; PR 16 and 636 greater than 1E2; and ER SP1 greater than 6F11. The high sensitivity of ER monoclonal antibody SP1 is in agreement with the higher affinity of that antibody21 and previous comparative assessments using patient samples.22,23 Another study6 confirmed stronger SP1 immunostaining (relative to 1DS) even though the difference was deemed diagnostically insignificant. Therefore, our finding that SP1 has a greater analytic sensitivity, by almost 2 concentration logs, is no surprise. For PR, variability in sensitivity among different immunostains has been described but not specifically for the PR antibodies 1E2, 636, and 1624; quite the contrary. Kornaga et al25 described a similar percentage of positive PR samples with these 3 particular PR immunostains. There was no consistent pattern suggesting greater sensitivity by one of the PR immunostains. Our finding that HercepTest has a greater analytic sensitivity relative to the 4B5 immunostain is also at odds with prior literature. The analytic sensitivity difference is not corroborated by several previously published reports comparing 4B5 and Hercep-Test.26-28 Prior comparative studies of HER2 immunostains using patient tissue samples evaluated concordance to each other and to fluorescence in situ hybridization amplification data.

As already mentioned, the differences we observe in analytic sensitivity may be due to limitations associated with commutability. Since diagnostic IHC is not a quantitative field, commutability has never before been relevant. For HER2, we felt that this explanation is less likely because both the 4B5 and HercepTest antibodies are immunoreactive with the same carboxy-terminal peptide of HER2.29 In forming these analytic response curves comparing HercepTest and 4B5, both were evaluated with the same IHControl, with the same peptide, on the same microbeads. The fact that HercepTest is a polyclonal antibody, resulting in multiple binding events, could explain a higher sensitivity. Differences in the analytic response curves cannot be due to preanalytic variables such as length of tissue fixation, cold ischemic time, dehydration, or embedding in paraffin. None of those preanalytic variables apply in this context.

There is another potential explanation for why our observed differences in HER2 and PR analytic sensitivities are at odds with prior published reports. Those prior reports were correlative studies using random patient samples of unknown analyte concentrations. They may not be optimally suited to detect a 1-log concentration difference in analytic sensitivity. In studies comparing random patient samples by 2 or more immunostains, the differences in analytic sensitivity will only show up on positive samples with analyte concentrations above the threshold of detection of one immunostain and below the threshold of detection for the other. The number of samples in that group may be small. Also, heterogeneity of cell expression within a single tumor can further confound the analysis.30,31 In such instances, discrepant test results on serial sections of the

**Figure 8.** Limits of detection for the PR 636 (A), 16 (B), and 1E2 (C) immunostains. These data were provided by the same laboratories from which analytic response curves were measured (Figure 7). Abbreviations: LOD, limit of detection; MEF, molecules of equivalent fluorochrome; PR, progesterone receptor.
same sample, measured by 2 different immunostains, can arise because of differences in the serial sections. Finally, there is the potential for discrepancies due to postanalytic interpretive variability.

These differences in analytic sensitivity among commercial immunostains raise an additional question. If IHControls were to be used in a PT context, which response curve would be considered the standard for acceptable performance? As a start, we suggest that diagnostic IHC laboratories can be compared to their peers using the same immunostain. This is a common PT protocol in other laboratory disciplines. For example, laboratories using HercepTest can be evaluated in comparison to other IHC laboratories using HercepTest. The same would be applicable for the 4B5, CB11, and SP3 immunostains. For each HER2 (or ER, PR) immunostain, a distinct consensus response curve can be generated, even if those curves are different from one another. No single particular immunostain need be considered “correct,” at least without more data on the subject of commutability. This method of PT grading is similar to that used in most traditional PT surveys, where nonpatient materials are used to compare laboratories using the same method.

Our findings with laboratory 4 particularly impressed upon us the power of quantitative review using analytic response curves. The analytic response curves of commercial immunostains are so characteristic that it caused us to realize that we had initially misclassified laboratory 4 as using HercepTest (Figure 3, B), based on the information the laboratory provided us regarding their method. The objectivity and precision associated with the measurement of an analytic response curve can be a powerful tool in assessing diagnostic IHC laboratory performance.

These characteristic analytic response curves are not solely a function of the primary antibody. The published literature contains numerous descriptions of certain types of primary antibodies being more or less sensitive. For example, there are numerous articles focusing on the relative merits of rabbit versus murine primary antibodies. Despite the importance of primary antibodies, other factors also affect the analytic response curves. The shape of each curve is determined by a combination of influences, including antibody affinities, concentrations, incubation times, wash steps, temperature, the degree of amplification associated with the detection reagents, and antigen retrieval. The remarkable reproducibility of the analytic response curves (from many different laboratories) for FDA-cleared IVD tests is a testament to the high level of reproducibility in current Good Manufacturing Practices reagent manufacturing and automated instrument processing. Our findings also highlight the risks associated with laboratory-developed tests. Without the means to directly measure analytic sensitivity, it is difficult to assess analytic performance.

The study data also demonstrate that measuring analytic response curves is a superior assessment of analytic performance than the subjective assessment of LOD. In this study, interpretive variability was sufficiently large as to obscure true differences in immunostaining. There were several instances of laboratories with subpar stain intensity that still identified the same LOD as other laboratories. This

---

Figure 9. Analytic response curves for ER immunostains SP1 and 6F11. Immunostain intensity (y-axis) is graphed as a function of ER peptide concentration (molecules per microbead, x-axis). The ER 1D5/2-123 antibody is not included because there were not at least 5 laboratories per group. Each point is the mean and standard deviation for the group of laboratories. Abbreviations: ER, estrogen receptor; MEF, molecules of equivalent fluorochrome.

Figure 10. Limits of detection for the ER immunostains SP1 (A) and 6F11 (B). These data were provided by the same laboratories from which analytic response curves were measured (Figure 9). Abbreviations: ER, estrogen receptor; MEF, molecules of equivalent fluorochrome.
suggests that such laboratories may be accustomed to their lower level of stain intensity and adapt their interpretations accordingly.

Our survey protocol incorporating central review of slides is similar to the UK NEQAS or NordicQC IHC PT program formats. With central review, the most common cause of failure is weak immunostaining.\(^3\)\(^4\) According to the cumulative NordicQC experience (as of 2014), approximately 41% of slides are judged as optimal, 30% are acceptable, 18% are borderline, and 11% are unacceptable.\(^3\)\(^4\) In comparison, our data show a similar percentage of laboratories, approximately 10%, failing because of poor immunostain intensity. However, we scored laboratories’ immunostaining in a different manner. Rather than subjectively evaluate immunostain intensity, we used statistical methods of comparison. Such methods are commonly associated with PT for quantitative analytes, such as a serum glucose. We selected the mean ± 2SD as “passing.” This decision was arbitrary and can be modified. Since the mean ± 2SD by definition includes more than 95% of laboratories, only a few laboratories are outside the 2SD limits. Moreover, we drew those acceptability lines after including all of the laboratories using a particular method, including obvious outliers. In practice, a PT program might exclude certain outliers before establishing the acceptability boundaries.

From a logistical standpoint, the use of IHControls simplifies image quantification. We found it is easier to quantify stain intensity of IHControls than tissue sections. The microbeads are homogeneous, without the complexity and heterogeneity of cell types associated with tissue sections. Consequently, no special histologic training is required in selecting the microscopic fields to be photographed. Moreover, if the photograph is taken properly, the high contrast and perfectly round shape of the microbeads facilitate rapid and reliable image segmentation during analysis.

A potential limitation of using IHControls for PT is that they are formaldehyde fixed. Therefore, antigen retrieval is usually required for optimal stain intensity. Laboratories using an alternative nonformalin fixative, and therefore not performing antigen retrieval, will not match their peer group using an alternative nonformalin fixative, and therefore not usually required for optimal stain intensity. Laboratories

Financial support for this project was provided by the National Institutes of Health, through grant numbers R44CA183203 and R44CA213476 (National Cancer Institute, both funded to SAB) and grant number UL1TR001064 (National Center for Advancing Translational Sciences). We are grateful to Ron Zeheb, PhD, for his ongoing advice during the course of this project; Doroit Bogen, MS, for administrative support; to the many laboratory staff and pathologists who participated in this study; and to Gary Horowitz, MD, and Barbarajean Magnani, MD, PhD, for their helpful critique of the manuscript.

References

1. Paxton A. Rebooting IHC diagnostics for companion diagnostics. C A P Today. January 2016;30(1).
2. Paxton A. PD-L1, other targeted therapies await more standardized IHC. C A P Today. February 2016;30(1).
3. Vyberg M, Nielsen S. Proficiency testing in immunohistochemistry: experiences from Nordic Immunohistochemical Quality Control (NordiQC). Virchows Arch. 2016;468(1):19–29.
4. Nielsen S. External quality assessment for immunohistochemistry: experiences from NordicQC. Biotech Histochem. 2015;90(3):331–340.
5. Nunes C, Rocha R, Reis-Filho J, et al. Comparative analysis of six different antibodies against Her2 including the novel rabbit monoclonal antibody (SP3) and chromogenic in situ hybridisation in breast carcinomas. J Clin Pathol. 2008;61(8):934–938.
6. Brock J, Hornick J, Richardson A, Dillon D, Lester S. A comparison of estrogen receptor SP1 and 1D5 monoclonal antibodies in routine clinical use reveals similar staining results. Am J Clin Pathol. 2009;132(3):396–401.
7. Kornaga E, Klimowicz A, Guggisberg N, et al. A systematic comparison of three commercial estrogen receptor assays in a single clinical outcome breast cancer cohort. Mod Pathol. 2016;29(8):799–209.
8. Cho E, Srivastava A, Park K, et al. Comparison of four immunohistochemical tests and FISH for measuring HER2 expression in gastric carcinomas. Pathology. 2012;44(3):216–220.
9. Vani K, Sompuram S, Schaudle A, Balasubramanian A, Bogen S. Analytic response curves of clinical breast cancer IHC tests. J Histochem Cytochem. 2017;65(5):273–283.
10. Rhodes A, Jasani B, Cottier J, et al. A formalin-fixed, paraffin-processed cell line standard for quality control of immunohistochemical assay of HER2/neu expression in breast cancer. Am J Clin Pathol. 2002;117(1):81–89.
11. Rhodes A. Developing a cell line standard for HER2/neu. Cancer Biomark. 2005;1(4–5):229–232.

Quantitative Assessment of IHC Test Performance—Sompuram et al
12. Xiao Y, Gao X, Maragh S, Telford W, Tona A. Cell lines as candidate reference materials for quality control of ERBB2 amplification and expression assays in breast cancer. Clin Chem. 2009;55(7):1307–1315.

13. Rhodes A, Jasani B, Balaton A, et al. Study of interlaboratory reliability and reproducibility of estrogen and progesterone receptor assays in Europe: documentation of poor reliability and identification of insufficient microwave antigen retrieval time as a major contributory element of unreliable assays. Anat Pathol. 2001;115(1):44–58.

14. 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.

15. Sompuram S, Vani K, Tracey B, Kamstock D, Bogen S. Standardizing immunohistochemistry: a new reference control for detecting staining problems. J Histochem Cytochem. 2015;63(9):681–690.

16. Vani K, Sompuram S, Naher S, Goldsmith J, Fulton R, Bogen S. Levey-Jennings analysis uncovers unsuspected causes of immunohistochemistry stain variability. Appl Immunohistochem Mol Morphol. 2016;24(10):688–694.

17. Vani K, Sompuram S, Schaedle A, et al. The importance of epitope density in selecting a positive IHC control. J Histochem Cytochem. 2017;65(8):463–477.

18. Tuominen V, Tolonen T, Isola J. Immunomembrane: a publicly available web application for digital image analysis of HER2 immunohistochemistry. Histopathology. 2012;60(5):758–767.

19. McCabe A, Dilled-Filhart M, Camp R, Rimm D. Automated quantitative analysis (AQUA) of in situ protein expression, antibody concentration, and prognosis. J Natl Cancer Inst. 2005;97(24):1808–1815.

20. Welsh A, Moeder C, Kumar S, et al. Standardization of estrogen receptor measurement in breast cancer suggests false-negative results are a function of threshold intensity rather than percentage of positive cells. J Clin Oncol. 2011;29(2):2978–2984.

21. Huang Z, Zhu W, Szekeres G, Xia H. Development of new rabbit monoclonal antibody to estrogen receptor: immunohistochemical assessment on formalin-fixed, paraffin-embedded tissue sections. Appl Immunohistochem Mol Morphol. 2005;13(1):91–95.

22. Bogina G, Zamboni G, Sapino A, et al. Comparison of anti-estrogen receptor antibodies SP1, 6F11, and 1D5 in breast cancer. Am J Clin Pathol. 2012;138(5):697–702.

23. Cheang M, Treaba D, Speers C, et al. Immunohistochemical detection using the new rabbit monoclonal antibody SP1 of estrogen receptor in breast cancer is superior to mouse monoclonal antibody 1D5 in predicting survival. J Clin Oncol. 2006;24(36):5637–5644.

24. Press M, Spaulding B, Goshen S, et al. Comparison of different antibodies for detection of progesterone receptor in breast cancer. Steroids. 2002;67(9):799–813.

25. Kornaga E, Klimowicz A, Guggisberg N, et al. Evaluation of three commercial progesterone receptor assays in a single tamoxifen-treated breast cancer cohort. Mod Pathol. 2016;29(12):1492–1500.

26. Ascoli S, Maletta F, Verdun di Cantogno L, et al. Approaching heterogeneity of human epidermal growth factor receptor 2 in surgical specimens of gastric cancer. Hum Pathol. 2012;43(11):2070–2079.

27. Tabakman EJ, Borg ST, Hooijer GK, et al. Determining sensitivity and specificity of HER2 testing in breast cancer using a tissue micro-array approach. Breast Cancer Res. 2012;14:R93–R105.

28. Nunes C, Rocha R, Buzelin M, et al. False positivity in HER2 testing of breast cancer: novel paths for approaching an old dilemma. J Clin Pathol. 2013;66(11):946–950.

29. Schrohl A-S, Pedersen H, Jensen S, Nielsen S, Brunner N. Human epidermal growth factor receptor 2 (HER2) immunoreactivity: specificity of three pharmacodiagnostic antibodies. Histopathology. 2011;59(5):975–983.

30. Vance G, Barry T, Bloom K, et al. Genetic heterogeneity in HER2 testing in breast cancer: panel summary and guidelines. Arch Pathol Lab Med. 2009;133(4):611–612.

31. Perelman E, Press M, Dueck A, et al. Immunohistochemistry and fluorescence in situ hybridization assessment of HER2 in clinical trials of adjuvant therapy for breast cancer (NCCTG N9831, BCIRG 006, BCIRG 005). Breast Cancer Res Treat. 2013;138(1):99–108.

32. Vani K, Sompuram S, Fitzgibbon P, et al. National HER2 proficiency test results using standardized quantitative controls: characterization of laboratory failures. Arch Pathol Lab Med. 2008;132(2):211–216.