# It’s here: whole slide imaging validation

## Anne Ford

May 2013—For the past four years, a group of pathologists has been diligently considering one question—Exactly how should whole slide imaging be validated?—all the while knowing that some laboratories consider WSI validation an unnecessary undertaking.

“The biggest argument I’ve heard is: ‘Why should we validate these instruments? We don’t validate our microscopes. It seems to be overkill,’” says Alexis B. Carter, MD, a member of the expert panel that created the CAP’s guideline titled “Validating Whole Slide Imaging for Diagnostic Purposes in Pathology,” published online May 1 in the *Archives of Pathology & Laboratory Medicine*.

Dr. Carter, assistant professor of pathology and laboratory medicine at Emory University School of Medicine in Atlanta, obviously disagrees. So does Liron Pantanowitz, MD, the leader of the panel that wrote the 50-page, 55-footnote document, which represents the first standard guideline regarding validation of WSI for diagnostic use.

“It’s just another instrument that we need to make sure is safe, like any other device,” says Dr. Pantanowitz, associate director of the pathology informatics division in the Department of Pathology at the University of Pittsburgh Medical Center.

Still, he adds, “Not everyone supports the fact that there should be validation.”

So is the working group expecting resistance to the guideline? “I’m anticipating some flak,” Dr. Carter says. “But what I’m hoping this guideline will do is show people the medical evidence behind these recommendations. People who are thinking that whole slide imaging is no different from a microscope aren’t aware of the literature, and hopefully this guideline will help educate them about that.”

The literature to which she refers: 767 international publications, of which the panel considered 27 strong enough to be subjected to data extraction and review by an independent methodologist. Twenty-three of those publications, along with comments from the public and consensus from the expert panel, formed the basis of the guideline. Depending on the strength of the evidence behind it, each item in the guideline has been categorized from strongest to weakest as a “recommendation,” a “suggestion,” or an “expert consensus opinion.”

Dr. Pantanowitz

A summary of the guideline’s findings makes them sound relatively straightforward. “Validation of the entire WSI system, involving pathologists trained to use the system, should be performed in a manner that emulates the laboratory’s actual clinical environment,” the summary reads. “It is recommended that such a validation study include at least 60 routine cases per application, comparing intraobserver diagnostic concordance between digitized and glass slides viewed at least 2 weeks apart. It is important that the validation process confirm that all material present on a glass slide to be scanned is included in the digital image.”

Simple, no? No. Several of those points—the 60 cases, the intraobserver issue, even that two-week period—required a hefty amount of discussion and deliberation during the panel’s 19 meetings. And when draft recommendations were posted on the CAP Web site in July 2011, they drew more than 500 comments from 132

respondents, requiring further modifications.

Take the question of how many routine cases a validation study should include. “There were pathologists not on the panel who called me and said, ‘One case would be good enough,’” Dr. Pantanowitz recalls. “Well, just from a practical point of view, one case wouldn’t be sufficient to make sure this works in a laboratory.” Then, too, Dr. Carter says, “There were a number of people in the group who felt very strongly that the more cases that were used in the validation process, the safer the implementation was going to be.”

“Some people thought we should be doing hundreds of thousands of cases,” Dr. Pantanowitz confirms. “But thousands of cases? We’re trying to make it practical for people to use this. We’re not trying to get this FDA-cleared as a vendor.” (Speaking of the FDA, a quick aside: As of fall 2011, the agency considers WSI systems to be class III medical devices. “There hasn’t really been a final word on that,” Dr. Pantanowitz says.)

As a starting point, the panel suggested 100 cases, as “a number that is practical and easy enough for people to do and still provides some assurance of proper validation,” he says. However, comments on the draft recommendations did not strongly support that number. So the panel examined studies that had used the following average numbers of cases: 20, 60, and 200.

“When we looked at the literature that used an average of 20 cases, the concordance between this digital modality and glass was only 75 percent,” Dr. Pantanowitz says. “That’s not enough. Well, what about 200 cases? That turned out to yield 91 percent concordance. I don’t know why, but when we looked at 60 cases, that yielded the best concordance [95 percent]—I guess because you’re not overburdening people, but you’re still giving them sufficient cases.”

That said, that number applies to only limited applications, such as frozen sections for brain lesions. “If you’re going to use it for more than that, such as cytology or hematology, meaning smears or hematoxylin and eosin, including frozens and permanent sections, you’re going to have to do another 20 cases for each additional application,” Dr. Carter points out.

Dr. Henricks

As for intraobserver concordance, the aim of that suggestion is to “take out the variable of individual pathologist expertise,” explains panel member Walter H. Henricks, MD, medical director of the Center for Pathology Informatics, Cleveland Clinic. “It’s most important for an individual pathologist to make the same diagnosis on the same case whether he or she is using glass versus whole slide imaging. I might make a different call than someone else, but what’s important is that I’m using the same judgment regardless of how I’m looking at it.”

Unfortunately, “when we looked at the literature, no one had studied this,” Dr. Pantanowitz says, “so there was no real data to base this on,” though he adds that 86 percent of the commenters on the draft guideline agreed with the importance of establishing intraobserver concordance. Hence this element of the guideline was categorized as a suggestion rather than a recommendation.

Determining the recommended length of the washout period—that is, the length of time allowed to pass after a pathologist views a case or slide and before he or she reviews it using a different modality—proved tricky as well. Short washout periods can lead to bias, of course, as pathologists tend to remember especially interesting or difficult cases for at least some period of time. But there are problems with long washout periods, too. First, they can prove cumbersome for a laboratory. And second, diagnostic criteria can change over time, either because a particular pathologist becomes more skilled or because new criteria for certain diagnoses emerge.

Then, too, “Many studies don’t even report their washout periods,” Dr. Henricks points out, making it difficult for the panel to establish a recommended length of time. The studies that do report their washout periods tend to use periods of between one and three weeks.

“When we looked at studies with washout periods of less than one week,” Dr. Pantanowitz explains, “their accuracy wasn’t very good—about 70 to 75 percent. When we looked at those studies that waited more than six months, again, a lot of them didn’t include that data, but one of them showed concordance of 95 percent. But if we took just a two- or three-week period, we found that the accuracy was also around 95 percent. So why wait six months when you could achieve the same level of concordance in a two- or three-week period?” The panel originally recommended three weeks, changing it to two in response to comments it received on the draft guideline.

As for another potentially controversial question—Should digital and glass slides be evaluated in random or nonrandom order during the validation process?—either option is fine, the expert panel says. The guideline reads: “Our meta-analysis of selected articles showed no marked difference in concordance when comparing glass with digital slides viewed in random versus nonrandom allocation. Therefore, our panel felt that laboratories can decide to evaluate their cases in either random or nonrandom order (as to which is examined first and second) for a validation study.”

The CAP guideline aside, some holdouts remain skeptical of whole slide imaging in general. “There is some controversy about the ability of pathologists to interpret patient cases using digital images instead of microscopes,” says Thomas W. Bauer, MD, PhD, of the Department of Anatomic Pathology, Cleveland Clinic.

Dr. Bauer, who was not a member of the panel that created the CAP guideline, is the lead author of “Validation of whole slide imaging for primary diagnosis in surgical pathology,” a study published last month in the *Archives of Pathology & Laboratory Medicine* (137[4]:518–524).

Dr. Bauer

In his view, intraobserver variability is the key issue here. “Testing whether this technology [whole slide imaging] works does not have anything to do with competence,” he says. “It has to do with: Can I make a diagnosis just as well with this technology as with a microscope? It should not test if I get the answer right or wrong. What matters is that I get the same answer using both methods.”

To his frustration, “If you look at the literature, there are not many really good studies that document intraobserver variability using microscope slides alone,” he points out. “So in our study, we decided to directly compare intraobserver variability interpreting whole slide images with intraobserver variability interpreting microscope slides.”

The first question Dr. Bauer and his coauthors had to answer was: How many samples should they use so they can be reasonably confident that the two diagnostic methods are equivalent? With the help of an independent statistician who reviewed available literature, they determined the answer to be about 450.

The study used two primary pathologists—one who specialized mainly in orthopedic and gastrointestinal cases and a second general surgical pathologist in a community setting who reviewed a broader spectrum of cases—and a one-year washout period. “After obtaining IRB approval, microscope slides of consecutive cases interpreted by each pathologist were retrieved from the file by an independent case coordinator,” Dr. Bauer explains. “That coordinator generated working copies that would have been identical to what the pathologists saw when they first

saw the cases.

“The coordinator then distributed every other case back to the pathologist with the microscope slides, while alternate cases were scanned and distributed as whole slide images.” That way, each pathologist interpreted the exact same cases he or she had interpreted more than a year before, half using a microscope and half using digital imaging. “The idea was the pathologists would have exactly the same amount of information available as they did the first time.”

After the pathologists had recorded their diagnoses, other pathologists reviewed those diagnoses and marked them “concordant” or “possibly discordant.” “These were independent pathologists who are subspecialty experts in each individual area,” Dr. Bauer says. “So if there was a possible discrepancy in, say, a liver biopsy, then the pathologist in charge of the liver section would review not only the diagnoses but also the microscope slides. If there was a discrepancy, that referee pathologist would decide which diagnosis was actually better. This was necessary, because it was possible that the diagnosis made by reading the digital image could be better than the diagnosis made by reading the microscope slides. If so, that discrepancy should not count against digital imaging.”

In the end, the major discrepancy rate was determined to be about 1.6 percent for whole slide imaging and about one percent for microscope slides. “Those rates are not statistically different,” Dr. Bauer says. Or, as the study puts it: “. . . diagnostic review by WSI was not inferior to microscope slide review.”

“The major discrepancy rate for both diagnostic methods was favorable when compared to the literature,” Dr. Bauer says.

In addition to the study’s main conclusion, Dr. Bauer and his coauthors uncovered several other interesting findings related to whole slide imaging. First, “We learned that the color enhancement on the screen is not necessarily exactly what you see through the microscope,” he says. “It takes a little bit of practice to adjust to it, but with a little experience it is not a problem.”

Second, they found very few discrepancies with respect to benign versus malignant tumors. Instead, “The cases we had the most difficulty with on digital imaging were some of the subtle inflammatory lesions,” he says. “We learned that in certain cases where the pathologist knows he or she needs to look for individual inflammatory cells at high magnification, it’s a good idea to get a high magnification scan to begin with. The default scan magnification for the study was 20×. We learned that for some types of diagnoses, a 40× scan is better than a 20×.”

Because many of the cases in the study were not especially thorny ones, Dr. Bauer and his coauthors are in the midst of conducting a followup study that will apply the same methods to more difficult examples. “The results of that study, based on only cases we receive for consultation, are very good so far,” he reports.

Dr. Carter

He feels especially confident in the first study’s conclusion given that he and his coauthors used an additional method he has not encountered in other studies. “Many pathology cases are complicated—they consist of multiple parts,” he explains. “So, for example, we might get a prostate biopsy with six different needle specimens that were all taken at the same time. For a study like this, you might count each of those biopsies as completely independent, yielding $n = 6$. Or you might count it as $n = 1$, since the surgeon just makes one decision based on the outcome of the entire case. So we evaluated our outcome measures from both perspectives. We calculated our discrepancy rates as if you considered each part independently, and we also calculated them based on cases.”

By way of illustration, he suggests imagining a prostate biopsy with only one needle specimen. If that specimen were to be interpreted as Gleason score six by one method and Gleason score seven by another method, that's a major discrepancy. But if a prostate biopsy were to have six needle specimens, three of which showed similar Gleason score discrepancies, "that would only count as one major discrepancy for the whole case, not three," Dr. Bauer says. "We are not aware of previous studies addressing that kind of complexity, but we were trying to be as conservative as possible. Fortunately, the number of discrepancies was low, no matter how we calculated it."

All well and good. But, returning to the CAP guideline, what if a laboratory remains unconvinced that validation is necessary for whole slide imaging? Dr. Carter offers an analogy.

"Just like freezing tissue can introduce artifacts that have to be accounted for when making a diagnosis, creating a digital image from stained tissue sections on a slide can also introduce subtle yet important artifacts. Sometimes these artifacts can make a diagnosis easier, but they can also make it harder," she says. "Unlike with frozen sections, our training, knowledge, and competency in recognizing and accounting for these artifacts are in their infancy." This gap in knowledge will have to be addressed, she says, if the technology is to be used successfully in patient care.

*Anne Ford is a writer in Evanston, Ill.*