Upper-echelon QA through Accuracy-Based Programs

Anne Paxton

June 2015—HbA1c, creatinine, testosterone, vitamin D, lipids, and maybe albumin. If you know what the common thread is among these analytes, then you may be familiar with the CAP's Accuracy-Based Programs and their evolution over the past couple of decades.

While a few providers around the world offer accuracy-based surveys, the College's Accuracy-Based Programs are by far the largest. As the CAP Accuracy-Based Testing Committee looks back on the progress of this set of Surveys and considers new biomarkers to include, it has no trouble showing that the Surveys have been effective in raising laboratories' awareness of accuracy, improving standardization of tests, and moving manufacturers, correspondingly, to improve their assays.

The current list of Accuracy-Based Programs includes: Accuracy-Based Lipids (ABL), Accuracy-Based Testosterone and Estradiol (ABS), Accuracy-Based Vitamin D (ABVD), Accuracy-Based Urine (ABU), Hemoglobin A1c-3 Challenge (GH2), Hemoglobin A1c-5 Challenge (GH5), Hemoglobin A1c CVL (LN15), and Creatinine Accuracy CVL (LN24).

While proficiency tests serve as a valuable check on the accuracy and reliability of laboratories' testing, a laboratory's results are not compared against a gold standard in most such tests. Rather, proficiency testing works on the basis of peer group ratings. "Your individual lab result is compared to the average of all other participants that used the same method," explains Greg Miller, PhD, a member of the CAP's Accuracy-Based Testing Committee and professor of pathology and director of clinical chemistry at Virginia Commonwealth University.

"If you have the same value as everybody else using the same method, you can be confident you're using the method the way it was intended. It's a good check that the lab has implemented a method and is using it correctly, but it does not tell whether a particular method itself has a bias against a correct value from a reference method or that all methods give the same results for patient samples."

Dr. Miller

Biases in proficiency testing results can be caused by matrix effects—the effects of all components of the sample other than the analyte of interest—on the measurement of the analyte. Matrix effects can stem from modifications made to a proficiency testing material during its preparation.

"Generally speaking, the matrix-related bias is a property of a method type," Dr. Miller says. "So if you score an individual against the average of everybody using the same method, you can ignore the matrix-related bias because it affects all users of the same method in the same way."

While peer group grading has limitations, there's a practical basis for conducting proficiency testing that way. The reason was that the materials in the proficiency tests were generally stabilized or processed material that could be produced in large quantity, says Accuracy-Based Testing Committee member John H. Eckfeldt, MD, PhD, professor of laboratory medicine and pathology at the University of Minnesota.

"They could test dozens of potential analytes in chemistry. But by adding various things or stabilizing the material, they often introduced what's called non-commutability. The material no longer behaves like a patient sample in the

measurement sample. So biases across peer group results are seen as an artifact of non-commutability. That's the concern—and it's left something of a cloud over whether actual patient sample results are accurate or not," Dr. Eckfeldt says.

"That's where the Accuracy-Based Programs come in to fill the gap," Dr. Miller says. "They provide a way to assess the bias between different procedures in a reliable manner."

Much of the laboratory world takes it for granted—mistakenly—that bias has been systematically rooted out of the testing process. "We find, even today, that a number of international guidelines and clinical textbooks are predicated on the assumption that all methods are producing comparable results," says Accuracy-Based Testing Committee member Darryl Erik Palmer-Toy, MD, PhD, medical director of Kaiser Permanente Regional Reference Laboratories in North Hollywood, Calif. "So it's important that we, at the very least, harmonize our test results to make different methods look the same, or better still, produce accurate results."

Dr. Eckfeldt

Such harmonization is necessary for proper clinical care, Dr. Eckfeldt notes. "For any test where specific clinical decision points exist, for example based on a large research study, there really has to be a concern about the absolute accuracy of your result in the laboratory. Because if you have a method-specific bias while you have a single clinical target, you really can't appropriately manage patients if different laboratory measurement procedures give different quantitative results."

For example, one of the early analytes for which accuracy-based Surveys became important was cholesterol. In the 1980s, national guidelines were issued, saying that if your cholesterol was greater than 200 you were at an increased risk of cardiovascular disease, Dr. Miller says.

"At the time, cholesterol wasn't standardized, so there was a huge effort to standardize the tests, and that was probably the most significant initial understanding of the problem with non-commutable samples and the inability to use them to assess actual performance of a test using patient samples. Out of that learning experience evolved the awareness that PT samples needed to be improved if we were going to really understand the relationship among different measurement procedures for patient samples. And CAP was one of the key leaders in the early decision to develop approaches to address that need."

The CAP's fresh frozen serum studies of the mid- to late 1990s were one result, Dr. Miller says, and the information collected then became the basis for the College's commutable frozen serum, which was a supplement to the regular PT programs. That serum was offered in vials to customers for several years, with a table of reference method targets, to allow them to check their calibration status as needed. The program was discontinued when the supply of serum was exhausted.

"Historically, questions have arisen about the validity of some of our PT results, given that the materials we use are often not human samples; they're chemically modified samples and reconstituted to resemble serum," Dr. Palmer-Toy explains. "We have seen some evidence that they don't always behave like clinical patient samples. So how do we get around that? One notion was we could make these pools of selected plasma or serum which are human source and do behave much more as a clinical sample would. But even with those, there is some manipulation that goes on. So the creation of ideal samples still remains a challenge."

Another test standardization effort was launched after the national Diabetes Control and Complications Trial, conducted from 1983 to 1993, showed that tight glycemic control was important for diabetes, Dr. Eckfeldt says. "So hemoglobin A1c targets were established for 'good' control of diabetes, and now for diagnosis of diabetes, and they were not method-specific clinical targets but universal ones regardless of the measurement procedure being used, so it became extremely important to have comparability among the hemoglobin A1c measurement procedures." The CAP's resulting glycohemoglobin Survey was probably the first sustained accuracy-based survey in the U.S., he says.

Similar but more complex, because of the biological sources of reagents used in the measurement process, was development of the INR in an effort to improve interlaboratory comparability by using INR to guide warfarin anticoagulation, rather than the prothrombin time in seconds that varied widely from lab to lab, Dr. Eckfeldt says.

Building on those earlier understandings and research, the CAP's Accuracy-Based Programs evolved so that today they include a number of different Surveys designed to allow labs to determine they are getting accurate values, Dr. Miller says. But equally important is the feedback that the Accuracy-Based Program results provide to manufacturers about their testing methods.

"The Surveys provide information that allows the manufacturers to recognize that their method is not giving results that agree with others, and it provides the evidence the laboratory medicine profession needs to develop a standardization program," Dr. Miller says. "In a number of cases, manufacturers have gone back and revisited their calibration traceability processes and adjusted them to ensure they conform to available reference systems."

For example, Dr. Miller explains, the fresh frozen serum studies allowed the field to recognize that creatinine needed to be standardized so eGFR could be calculated and reported, and based on that information, the program to standardize creatinine was launched. The LN24 accuracy-based Survey for creatinine was developed some 10 years ago, "before we called them accuracy-based Surveys," says Dr. Eckfeldt, who was then chair of the National Kidney Disease Education Program Laboratory Working Group. That particular Survey and the standardization of creatinine have had significant implications in kidney disease diagnosis, he says.

Dr. Palmer-Toy agrees that creatinine is the biggest recent success story in harmonization. "Until the last several years, we haven't had consistent results by method." But HbA1c presents a somewhat different kind of success story, he points out. "That's a test where we 'agreed to be wrong.' We agreed we would favor harmonization over accuracy per se in our measurement of HbA1c. What we commonly report is actually off by a couple percent of hemoglobin. Yet because the assay has become so prevalent in clinical care, there was a reasonable clinical decision to say we're going to relate this to a reference method but will continue to report our results."

"Even though the results are not entirely accurate, they won't cause too much disruption of clinical care. So that's the compromise that was struck, and it's one where harmonization and clinical expediency trumped accuracy," Dr. Palmer-Toy says. While accuracy is extremely important, he stresses it is not the only worthwhile goal. "We can't be ideologues about accuracy."

The most important outcome of a standardization or harmonization program, Dr. Miller adds, is for all routine methods to give equivalent results for an analyte so that clinical guidelines for patient care decisions will be applied uniformly.

The addition of the testosterone and estradiol accuracy Survey a few years ago took the Accuracy-Based Programs in a new direction. As the CAP has noted in a recent practice guidance, there has been a dramatic increase in recent years in testosterone testing in middle-aged and elderly males, partly the result of widespread direct-to-consumer marketing of testosterone supplementation via patch.

As this trend was developing, quality problems began to surface. "What happened is that at a conference in 2010 and in a consensus statement published in 2013, the Endocrine Society expressed huge concern that many of the immunoassays were cross-reacting with other substances in clinical samples and reporting values that were

erroneously high," Dr. Eckfeldt says. There was a resulting push to improve the accuracy of the clinical lab measurement procedures. "But it became clear that the materials being used in the CAP Survey at the time had some commutability problems. Although the Endocrine Society immediately jumped on how different some of the CAP Survey method-specific means were, you couldn't really tell how good or bad the clinical assays' comparability was because of non-commutability issues," Dr. Eckfeldt says.

Testosterone presents challenges on several fronts, Dr. Palmer-Toy says. "Some of the early publications on inaccuracy of testosterone results were somewhat sensationalized. They faulted the immunoassays for their inability to correlate with mass spectrometry results. Yet at the same time, mass spec results were not terribly well standardized. I think mass spec was a worthy tool, but as our experience with vitamin D has shown, it's taken many years to bring mass spec results into good alignment. We continue to make progress, but we certainly need to challenge all of our methods."

One particular weakness of testosterone assays is that they are not reliable in measuring low values of serum testosterone in women or children—for example, for tumor assessment. "The assays are really designed to measure testosterone in men, which is at a much higher level, and they're designed to measure testosterone deficiency in men, which is something that can be treated. For women, the immunoassays just don't have adequate sensitivity," Dr. Miller says.

Testing of women and children for testosterone increases the need for better accuracy. "That's one of the more challenging diagnoses and one where interference from cross-reacting substances like dehydroepiandrosterone sulfate is an important consideration," Dr. Palmer-Toy says. "In regards to older men, there are open questions on what levels of testosterone really require treatment, if any. And even the form of testosterone that should be measured is in debate."

Total testosterone is what is commonly measured, he says. "In fact, however, what is biologically active is not necessarily the total testosterone but the fraction that actually interacts with receptors in your cells. There is debate whether 'free' testosterone or a slightly different 'bioavailable' pool is more relevant."

In short, it's not just a matter of standardization when it comes to testosterone, Dr. Palmer-Toy points out. "It's also a matter of whether we can determine the best ways to use results and which form of testosterone we should be measuring."

Another survey target that has presented technical challenges to the Accuracy-Based Programs along the way is vitamin D, Dr. Miller says. For example, there are two versions of vitamin D, and one—D2—is synthetic and not normally found in food. "But people are going to be treated with the synthetic version, and if the assays don't measure that, then you don't know if you're proceeding with the proper doses."

Harmonization of results has improved since the CAP accuracy-based Survey in vitamin D was introduced, Dr. Eckfeldt believes. "The CAP Survey showed that certain measurement procedures had serious problems in quantifying vitamin D, and particularly patients getting vitamin D2 rather than D3 supplementation. And I think it's made pretty clear that several of the immunoassays needed improvement and clarification on what they were really measuring. That accuracy-based Survey also allowed labs that were doing liquid chromatography mass spec measurements to check on how comparable they were to the reference procedures."

The Centers for Disease Control and Prevention, Dr. Eckfeldt notes, has its own vitamin D standardization program that is far more complex and expensive than the CAP Survey—costing $3,000 to $9,000 per year per subscriber depending on the number of samples and mailings per year—but it is primarily geared toward manufacturers of 25-hydroxy vitamin D IVD measurement procedures and research labs doing work in the vitamin D field. "The CAP Survey gives labs a reasonable sense of how accurate their results are, and the manufacturers of certain immunoassays have improved them based on the results from the first few rounds of the vitamin D accuracy-based Survey."

A lingering issue is how to interpret 25-hydroxy vitamin D results. "There is a well-established threshold for vitamin

D deficiency," Dr. Palmer-Toy says. "There is less agreement when it comes to the optimal levels. That's a concept that is somewhat contentious."

The number of participants in the Accuracy-Based Programs has been stable over the years, Dr. Miller says. As of May 2015, enrollment stood at 221 for lipids; 107 for testosterone and estradiol; 466 for vitamin D; 74 for urine; 1,001, 2,529, and 1,055 for the three HbA1c Surveys; and 406 for creatinine.

The labs that sign on to the Accuracy-Based Programs are willing to go the extra mile to ensure quality of their results and to enhance the care of the patients who use their services, Dr. Palmer-Toy says. "They are subjecting themselves to additional scrutiny by doing so. And I think it's commendable." But, he notes, some laboratories might not be entirely confident of their ability to meet such a challenge. And that might reflect their willingness or unwillingness to subscribe to the accuracy-based Surveys.

A number of European countries also have accuracy-based survey programs, some of them more extensive than what the College offers, Dr. Miller says. For example, in the U.K. and the Netherlands, most proficiency testing survey samples are accuracy-based rather than measurement procedure peer-group–based.

"They're able to do this because they have much smaller volume requirements than the College does," Dr. Miller points out. "With only about 160 labs in the Netherlands, the national survey program only needs to prepare a modest quantity of materials, whereas for the College, with 8,000 participants in its proficiency testing program, it's not economically possible to prepare those quantities of commutable serum to be used. That's why the College has gone down the pathway of these more specialty voluntary Surveys through its Accuracy-Based Programs."

One of the limitations in making an accuracy-based Survey is you can't make it measure dozens of different analytes at multiple concentrations, Dr. Eckfeldt says. "The general chemistry Survey—the proficiency test that about 8,000 labs subscribe to—has 60 or 70 analytes that are measured. To get varying concentrations of all those different analytes and enzymes, they put in all kinds of stuff. That makes it potentially non-commutable for general use. So large Surveys that have many different analytes being measured are probably never going to be Accuracy-Based Programs."

One complication of that, Dr. Eckfeldt notes, is that some patients are being seen in different clinic and hospital settings—say, because they are living in the Sunbelt in winter and the northern states in the summertime—and their test results cannot be compared. "If you don't have comparable results, it's very hard to track how the patient is responding or progressing."

The CMS may have to do some rethinking, in his view, on how well the CLIA regulations work in judging labs. "I just think peer group grading may give labs a false sense of security sometimes, and they may be turning out patient results that are not harmonized with what other labs are reporting."

A recent study he conducted looked at performance (in the CAP 2014 CYS Survey) of certain cystatin C measurements using actual patient sample pools from people with kidney disease rather than normal volunteers (Eckfeldt JH, et al. *Arch Pathol Lab Med.* Epub ahead of print April 17, 2015. doi: 10.5858/arpa.2014-0427-CP). "It showed that the assays differed by as much as plus or minus 20 percent, which leads to huge variations in the computed estimated GFR. But that degree of variation was always sort of masked before that, because there was an assumption that the material used for the Survey was in fact non-commutable. This recent study pretty clearly showed some assays' calibrations need adjustment if people are ever going to use cystatin C successfully for estimating GFR," Dr. Eckfeldt says.

In the course of the study, he found that getting samples from patients with kidney disease was quite a challenge. "It depends on whether you can artificially make an accuracy-based Survey or whether the abnormal value samples are easy to come by." By contrast, finding samples for a Survey of glycohemoglobin is relatively straightforward. "You can find diabetics with all degrees of control over their diabetes, so it's pretty easy to get samples that range from normal up to 10 or 11 percent HbA1c." Samples for other analyte Surveys, however, can

be more difficult to get.

The cost of preparing accuracy-based Surveys is significantly higher than for preparing conventional proficiency testing. “That’s one of the barriers,” Dr. Miller says. The cost per participant is relatively similar. What makes the difference is that an accuracy-based Survey only has a few analytes and sometimes only one, so the number of samples is not sufficient to satisfy the CLIA requirements, which specify 15 challenges in a year. Some of the accuracy-based Surveys are circulated only twice a year, some of them only once a year. “So they’re intended to fulfill a different need in laboratory medicine, and that is to assess the accuracy or the trueness of measurement systems.”

Serum albumin is the next analyte that may be added to the Accuracy-Based Programs list, although a formal decision to go forward with it has not been made because the CAP still needs to determine the definitive reference method to be used. “We do offer serum albumin, cortisol, TSH, and sex hormone-binding globulin in the Accuracy-Based Testosterone and Estradiol [ABS] Survey,” says Sharon Burr, MBA, MT(ASCP), senior technical manager of CAP’s Proficiency Testing Program. “The Accuracy-Based Urine [ABU] Survey also offers albumin, but all of these analytes are for harmonization purposes only and are not graded against a reference measurement procedure target.”

Dr. Palmer-Toy, who is advocating for inclusion of serum albumin in an accuracy-based Survey, says albumin was added as a challenge in one of the linearity Surveys. Albumin, he points out, is perhaps one of the first clinical chemistry analytes ever recognized in medical science. “We’ve been measuring this for centuries, to some degree, and of course it plays an important part in renal function. It’s something that is regularly measured in most diabetics, and is also an important marker of malnutrition when serum albumin is low.”

“In addition, it’s an important part of the quality matrix by which the care of dialysis patients is assessed. Hundreds of laboratories measure serum albumin, but we have a disparity in the most common methods used.” Based on the linearity challenge that the ABS program conducted, there is about a 10 percent spread in results across methods used just with the labs that were surveyed.

In gauging the value of accuracy-based Surveys to their own quality assurance efforts, laboratories should be aware that improving accuracy has many benefits, Dr. Eckfeldt says. “By using accuracy-based Surveys, labs will have a far better sense of how they compare to reference measurement procedures’ established target values, and therefore they can tell their clinicians that they’re providing accurate results they can use in clinical trials or research-based cut points with confidence,” he says.

Dr. Miller views the Accuracy-Based Programs as an essential component in the progression of laboratory medicine. “The driver for the College’s Accuracy-Based Programs and for lab standardization efforts is the recognition that there are analytes people are using to make clinical decisions for which the lab methods are not adequately standardized. So the Surveys are just part of a bigger picture of trying to improve the overall quality of lab testing to keep patient needs at the fore.”

“It’s important to participate in these Surveys so that the laboratory medicine profession gains useful information about the performance of different methods. Then, when the performance is inadequate, steps can be taken to improve it.” The College is strongly committed to the Accuracy-Based Programs, he adds. “CAP has definitely stepped up to the plate to recognize the importance of accuracy-based Surveys and is providing the resources to address some of these more challenging analytes.
[hr]

Anne Paxton is a writer in Seattle.