



**Article:**

Joely A Straseski, Hubert W Vesper, Thomas A Long, Gary L Horowitz.  
*Accuracy-Based Proficiency Testing Surveys for Testosterone Provide Important Information About Measurement Accuracy and Its Impact on Patient Care.*

Clin Chem 2026; 72(10): 1046–55. <https://doi.org/10.1093/clinchem/hvag068>

**Guests:** Dr. Joely Straseski is a Professor of Pathology at the University of Utah School of Medicine and the Head of Clinical Operations for Clinical Chemistry, Toxicology, and Biochemical Genetics. Dr. Gary Horowitz is a Professor of Pathology at Tufts University School of Medicine and Chief of Clinical Pathology at Tufts Medical Center.

Bob Barrett:

This is a podcast from *Clinical Chemistry*, a production of the Association for Diagnostics & Laboratory Medicine. I'm Bob Barrett. Clinical guidelines often recommend a decision threshold for the diagnosis of a clinical condition, but laboratorians know that commercially available assays don't all generate equivalent results.

Some are set a little higher and some lower. This assay bias can lead to very different applications of the guideline-defined cutoff, depending on which test method is used. Assays with high bias generate more values above the cutoff, while assays with a low bias will do the opposite. Clearly, assay standardization would address this problem, but there are lots of clinical labs throughout the world.

How can we assess accuracy across thousands of labs in an efficient way, ensuring uniform application of clinical guidelines regardless of where a patient's samples are tested? A research article in the October 2026 issue of *Clinical Chemistry* asks whether accuracy-based proficiency testing has a role to play. If commutable samples are used, can results from a small number of labs serve as a reasonable estimate of larger group performance?

Today, we are joined by the article's lead and senior authors. Dr. Joely Straseski is a Professor of Pathology at the University of Utah School of Medicine and the Head of Clinical Operations for Clinical Chemistry, Toxicology, and Biochemical Genetics. Dr. Gary Horowitz is a Professor of Pathology at Tufts University School of Medicine and Chief of Clinical Pathology at Tufts Medical Center. Both have served on the College of American Pathologists Accuracy-Based Programs Committee.

And Dr. Straseski, let's start with you. Your study highlights accuracy-based proficiency testing. How is that different from traditional proficiency testing?

Joely Straseski: Yeah, so traditional proficiency testing, or of course we usually refer to it as PT, that uses really very highly processed, they're non-commutable materials. So basically, they're artificial. And those are really helping laboratories directly compare their performance to peer labs that are using the same methods. But accuracy-based PT, that uses commutable specimens. So, they're actually human serum.

And they have reference values or target values assigned to them by a reference measurement procedure. So, that's going to allow true measurement of accuracy. It gives us a much more realistic picture of how an assay is performing when you're actually testing patient samples.

Bob Barrett: Well, thank you. Now, Dr. Horowitz, let's go to you. Isn't successful performance on traditional PT surveys sufficient for accreditation? Isn't that why it's required?

Gary Horowitz: Yes, that's true. PT is required for accreditation as it documents that labs are performing the assays well. But as Joely mentioned, traditional PT uses highly manipulated specimens in order to cover literally dozens of analytes in various concentrations for hundreds of laboratories. As a result, different methods or assays from different vendors can and do get different values for the exact same samples. That's why labs are graded by comparison to their colleagues using the same methods in hopes of correcting for those method-dependent variations.

As we move to national and international guidelines, for many analytes, rather than local reference intervals, labs have to ensure that the values they report on these analytes match the values that were used to establish those guidelines. Now, in truth, most of us assume that despite the differences observed on traditional proficiency testing, all the methods will get the same results for genuine patient specimens, and that they're accurate. Unfortunately, it turns out that this is often not the case. That's where accuracy-based surveys fit in.

Bob Barrett: Okay, could I get some more from both of you about accuracy-based surveys?

Joely Straseski: Yes, absolutely. Accuracy-based surveys are actually already available for several analytes, and that does allow assessment of accuracy. Good examples include hemoglobin A<sub>1c</sub> and cholesterol. We know these are frequently ordered tests. They have international guidelines, as we're all aware. As a result of these accuracy-based surveys, though, manufacturers and labs that develop their own tests, they can actually see how well their assays are actually working in the field.

Hemoglobin A<sub>1c</sub> is actually a really, really good example here. When this survey first switched to using commutable specimens, and it was many years ago now, the differences among the different methods, it was actually huge, disappointingly large.

Over time, manufacturers improved their methods to the point where virtually all the methods that are in use today, including point-of-care assays, I'll add, are actually quite accurate, and they can be used interchangeably for clinical purposes. Really, that's what's most important.

Gary Horowitz: You can imagine that it's not feasible for proficiency testing providers to manufacture accuracy-based surveys for every analyte. As we said earlier, we're talking about dozens of different analytes at many different concentrations, and huge amounts of material that have to be mailed to hundreds of laboratories. It would be very difficult to send accuracy-based surveys to all these labs.

Also, it wouldn't be affordable for most labs to purchase them if they were all commutable. What we wanted to do was to see whether results on the traditional surveys already in use could predict the results on the smaller accuracy-based surveys that use human specimens. For this study, we decided to focus on testosterone as an example. This is another assay with international guidelines, and the Endocrine Society recommends a threshold of 264 nanograms per deciliter for the diagnosis of male hypogonadism. We wanted to understand how accurate current testosterone methods are around that decision point, and how inaccuracies would affect patient management.

Bob Barrett: Tell us, how was your study designed?

Joely Straseski: We evaluated three commutable testosterone samples. What we did was one of those was included as what we called a wildcard specimen. That was added to one of the large, the traditional PT surveys from the College of American Pathologists, or CAP, as most of us would refer to it as. That included more than 1,400 labs. Then the other two samples were analyzed using a smaller accuracy-based CAP survey.

Then we compared performance among the different method groups, and we looked at potential effects on patient care as it related to that diagnostic threshold of 264 nanograms per deciliter that we just mentioned for testosterone.

Bob Barrett: What was the most important finding from that study?

Gary Horowitz: Well, if we're looking for some good news, I have it for us. The overall average performance across all the laboratories and all the peer groups looked excellent, truly excellent. The

average bias versus the reference method was under 2%. Yes, overall, under 2%, or roughly just plus or minus 5 nanograms per deciliter at 264 nanograms per deciliter. It's really amazing how good the average was. Unfortunately, even excellent performance, excellent overall performance on a proficiency testing survey, whether it's accuracy-based or traditional, can mask important differences among individual testing platforms or vendors.

What we saw was that individual method groups had mean biases, mean biases ranging from minus 16% to plus 15%. Let's put that into perspective. Consider aliquots from the same specimen whose true value is 264 nanograms per deciliter. Think about two aliquots being measured at two labs using different methods. The mean result could be as low as 222 or as high as 304 nanograms per deciliter, 40 units above or below the true result, the threshold for hypogonadism.

You can see that at a clinical decision point like this, management decisions could look quite different depending on the method used or equivalently, depending on which lab the specimen was sent to.

Bob Barrett: What did you learn from comparing results from both commutable and non-commutable specimens?

Joely Straseski: Well, we had hoped that we could predict performance in the commutable or the accuracy-based surveys by the performance that we saw on the larger traditional PT surveys. In other words, if a method group was 10% high on a traditional survey, would it be 10% high on the commutable survey? That was our basic hypothesis. However, several methods showed completely different bias patterns for the two different survey types. That finding really reinforced basically what's one of the more important messages here --

-- that traditional PT can only tell you whether your results agree with those from other labs using the same method, but it may not tell you whether or not the method itself is accurate when you're testing real patient samples.

Bob Barrett: Well, finally, doctors, what is the bottom line? What should laboratorians and clinicians take away from this work?

Gary Horowitz: The key message is that traditional PT surveys, the ones that are manipulated to get the different concentrations and all the different analytes, the one that most labs use, the one that's required for accreditation, those traditional PT surveys have limited value when it comes to ensuring accuracy. These surveys are necessary for accreditation and they can tell you how well your lab's results agree with your peers, but

you need to look at the performance of your peer group on accuracy-based surveys to truly assess accuracy.

Joely Straseski: Okay, so how to actually do that? For labs that are not enrolled in accuracy-based surveys themselves, what you could do is you could see whether your results from using traditional PT, do those results align well with the mean of your peer group? Don't just look at whether or not you pass the established criteria, but see whether or not your results are actually near the mean of the peer group. If they are, then you could point to the manufacturer. You could look back to your manufacturer and see if they have data that actually shows their performance on an accuracy-based survey.

If that performs well, if those results look good on the accuracy-based survey, it's reasonable to infer then that your lab's results are accurate. In order for any of this to work, though, widely used FDA-cleared assays need to be represented on accuracy-based surveys. We basically need that data. The goal should be really pretty similar to what we saw for hemoglobin A1C. We would want manufacturers, and perhaps also the FDA, they should be aiming toward ensuring accuracy of all methods, particularly those that have international guideline recommendations. And just to remind ourselves, at the end of the day, accuracy is really what's most important when we're testing actual patient samples.

Bob Barrett: That was Dr. Joely Straseski from the University of Utah in Salt Lake City, and Dr. Gary Horowitz from Tufts University in Boston, Massachusetts. They wrote a research article in the October 2026 issue of *Clinical Chemistry* using accuracy-based proficiency testing to evaluate assay standardization, and they've been our guests in this podcast on that topic. I'm Bob Barrett. Thanks for listening.