Far too many laboratories consider proficiency testing just a necessary evil, little more than periodic pass–fail exercises we perform solely to meet regulatory requirements. In addition, too many of us belittle point-of-care (POC) testing as a passing fad, a technology so inferior to what we use in our own laboratories that it hardly warrants our attention. Clearly, a report that combines these 2 topics, such as the one in this issue of Clinical Chemistry (1), runs the risk of commanding little attention. That would be a very unfortunate mistake, because it has important lessons for all of us who practice laboratory medicine in our efforts to improve patient care.

As Stavelin et al. (1) suggest, POC testing represents an important and growing segment of laboratory medicine. Many important clinical decisions are based on these technologies. Even if the CVs of POC testing, and indeed their accuracies, are not as good as central-laboratory techniques, POC testing modalities can sometimes have a more positive impact on certain aspects of healthcare. A POC creatinine test that predicts a patient’s risk for contrast-induced nephropathy and is carried out in 10 min in the emergency room prior to radiologic imaging trumps (as long as it is “good enough”) a central-laboratory test that takes 30 min (or longer) to reach the caregivers (2). Even in nonurgent situations, the availability of such POC tests can be extremely valuable in preventing contrast-induced nephropathy (3).

Among the more common, and more important, decisions that clinicians make on the basis of laboratory test results are those related to achieving therapeutic international normalized ratios (INRs) in the management of oral anticoagulants (specifically, vitamin K antagonists such as warfarin). In general, data indicate that POC devices can perform reasonably well analytically (4, 5). In addition, the use of POC devices in ambulatory settings can enhance patient care by allowing immediate feedback between physicians and patients (6, 7). Even more recently, the use of POC devices by patients themselves, analogous to how diabetic patients use glucose meters, has been shown to be effective (8, 9).

But how do we know, in an ongoing way, whether POC devices are actually performing well (or poorly)? As is the case for the central laboratory, proficiency testing (PT) can shed some light. Doing measurements on blinded samples and then comparing them with the results of others should reveal, at a minimum, how close one’s results are to those of one’s peers. Although participating in PT may not be an absolute requirement for POC devices, many authorities, if not most, believe that PT in this setting is especially important (10).

Even for central-laboratory techniques, traditional PT suffers from “matrix effects,” in that samples used for testing often react differently from native patient samples. Therefore, comparisons must be made only to peer groups, rather than to the “true value” (11). What if the peer group as a whole is wrong (12)? The problem of matrix effects becomes even more complicated for POC devices, in which the sample typically used for patient testing is fresh whole blood, a sample type especially difficult to simulate in PT programs.

In their report, Stavelin et al. (1) propose a novel method for performing PT for INRs with POC devices. In brief, for each of 3 different POC methods, they asked a relatively small number (approximately 20) of “expert” centers to measure 5 patient samples with their POC device in parallel with their central-laboratory method. They then used these data to calculate a mean percentage bias for each POC method. At the same time, these centers and all the other participants in the program (approximately 1500) performed PT testing with standard, noncommutable PT material. The medians of the values from the “expert” centers became the POC method–specific target values. Stavelin et al. then plotted each participant’s performance on each (noncommutable) challenge on a graph in which the x axis represented the percent bias established by the “expert” laboratories for the POC method itself compared with the central laboratory method and the y axis represented the percent bias of the participant’s value compared with the peer group target value. The authors used this graph to define a set of regions representing different levels of performance.

---

1 Department of Pathology, Beth Israel Deaconess Medical Center, Boston, MA.
* Address correspondence to the author at: Beth Israel Deaconess Medical Center, 330 Brookline Ave., Boston, MA 02215. Fax 617-667-4533; e-mail ghorowitz@bidmc.harvard.edu.
Received November 19, 2012; accepted November 20, 2012.
Previously published online at DOI: 10.1373/clinchem.2012.197574
2 Nonstandard abbreviations: POC, point-of-care (testing); INR, international normalized ratio; PT, proficiency testing.
For example, a participant reporting a result close to the target value for its peer group, via a method that itself has a small bias compared with the central laboratory method, would fall into a region of very good performance, whereas a participant using that same method and obtaining a result deviating >15% from the target value would fall into a different region, indicating that the result was much less acceptable. More interestingly, a participant using a POC method that has a large bias compared with the central laboratory method but produces a result close to the peer group target would fall into a region that indicated, rightfully, a problem: Although they were performing their method well, their results with patient samples were almost certainly biased. Perhaps most interesting, however, is that a participant that uses a POC method with a large negative bias compared with the central-laboratory method but obtains a result far above the peer group target value would be in a region indicating a very serious problem. Although that participant might be reporting accurate results for patient samples, it would be the result of 2 serious but compensating errors.

One might quibble with the exact percentages the authors chose to use for defining their regions on each axis, but these percentages are clearly easy to modify according to clinical needs and analytical capabilities. What is more important is that one axis represents a realistic assessment of how close the POC method itself is to the “true value.” Independently, the second axis measures competency on the basis of fixed percentages of deviation from the target. A participant does not “benefit” from using a method that is imprecise, nor is a participant “penalized” for using a precise method, as is the case with many traditional PT-grading methods (11). Participants who perform a method well but obtain results that are not accurate can discern that easily from the graphical display. One hopes that manufacturers of such methods would have a strong incentive to improve the accuracy of their products; otherwise, participants can decide to change methods. What matters in this assessment is how close one’s result is to an accurate patient value.

A good PT program should be able to help establish the causes of poor performance, among which are problems inherent to certain methods, performance issues with particular lot numbers of reagents, and technical errors made by end users. In this study, the problems uncovered had little to do with the end users; they were ascribed to one of the methods themselves. In other words, it will do little good, at least in this case, to admonish users to do better (or to use peer group grading for PT for that matter). The manufacturer needs to improve the method itself.

Also of importance is that this proposal for PT need not be restricted to POC methods. Clearly, it can also be used for PT performed by central laboratories. It may be a reasonable and potentially more cost-effective alternative to the use of minimally processed actual human samples, a practice that has been introduced for selected measurands (e.g., 25-hydroxyvitamin D, testosterone, estradiol) in which noncommutable materials have led to particularly misleading conclusions (11, 13, 14).

The method as presented in this report has some (minor) limitations. First, the authors point out that they assumed that each method’s proportional bias with noncommutable material is comparable to its bias with commutable material. A second limitation is that the authors decided to adjust INRs from each central-laboratory method to “a typical Norwegian hospital method”; the adjustments were based on commutable control materials included with the surveys and were minimal. This limitation reflects less on the POC methods than on the lack of standardization of INRs, even with central-laboratory methods, a problem seen whenever a true reference method does not exist. Third, it is not clear that the bias, which is expressed in the report as a percentage, is constant across all concentrations of the measurand (15). That may be true for INRs (no data are shown), but even if it is not, the assumption is probably valid over the narrow range of INRs of particular interest.

This novel idea deserves careful consideration from all of us. If others can confirm these results and those for other measurands, I would strongly recommend that this methodology be incorporated as much as possible in all PT, because it offers a way for all of us to ensure that the values we routinely report for patients are accurate. If users were to demand, or if regulators were to require, a system of this type, the costs (which should not be too large) could be spread out among all users.

Our physicians and our patients deserve no less than to know that, no matter where their testing is done, it is accurate and “fit for purpose.”
Stock Ownership: None declared.
Research Funding: None declared.
Expert Testimony: None declared.
Patents: None declared.
Other Remuneration: G.L. Horowitz, AACC and College of American Pathologists.

References