Proficiency testing (PT), also known as external quality assessment (EQA), has earned a well-deserved position as one of the most important elements of a laboratory’s quality management program (1). By assessing its performance on blinded samples, a laboratory can determine to what extent its results agree with those of its peers or, ideally, with the reference method.

Unfortunately, as has been emphasized recently (1), the most frequently used PT materials are noncommutable, meaning that they do not behave exactly like real patient samples. Conclusions drawn from such materials are, of necessity, limited to understanding how well any given peer group performs rather than knowing whether the results on patient samples are, in fact, accurate (2). In the case of a method lacking a true peer group (including, for example, the overwhelming majority of mass spectrometry methods and other laboratory developed tests), even this limited type of comparison may not be possible. Additionally, when there are differences between peer groups comprising laboratories that use homogeneous systems (single manufacturer’s instruments, calibrators, and reagents), the differences are often ascribed to lack of commutability (matrix effects) of the materials, although the differences may in fact be real and reflected in patient samples as well (3, 4).

So-called accuracy-based surveys, which use commutable materials and reference method target values, are becoming more widespread and have generated several improvements (5) and interesting findings (6, 7). These surveys face some unique challenges. It is difficult to obtain sufficient materials to supply the number of laboratories participating in conventional surveys. It is also difficult to achieve the full range of concentrations one would like; manipulating commutable materials may render them noncommutable (for example, spiking analyte into them to increase concentrations or diluting them to decrease concentrations).

In addition, it can be extremely expensive to obtain reference method target values.

In this issue of Clinical Chemistry, Stepman et al. (8) meaningfully extend a novel approach some of them had previously reported (9). They collected 20 single-donor sera according to CLSI protocol 37-A, from which they made 1-mL aliquots, which were frozen at −70 °C until analysis. The authors arranged to have the aliquots analyzed singly for 8 common analytes by 63 individual laboratories. These laboratories were selected to have roughly 10 of each of 6 manufacturers’ homogeneous systems represented. On these samples for these analytes, one would expect agreement to be good, if not excellent. The results, however, indicate that there is much room for improvement. Importantly, because of the study’s design, the authors were able to detect individual sample problems and determine, among other things, individual laboratory imprecision vs that of its peer group and peer group agreement with target values.

The authors fully recognize some of the limitations of their study. First, use of the CLSI 37-A protocol to prepare samples did not in itself prove that the samples were commutable. Second, the number of laboratories included was, of necessity, relatively small. Third, the concentration ranges covered by these samples were small, leaving open the question of performance at other concentrations seen in real patient samples. Fourth, for some of the target values, the authors used the “all method trimmed mean” rather than the reference method. Fifth, the authors’ decisions on specific quality limits for individual analytes could be debated. Nonetheless, the insights their data provide are striking. If we cannot achieve agreement on analytes as simple and straightforward as these, is it reasonable to expect to achieve agreement on others (or on these at other concentrations)?

Some readers may argue that systematic differences between methods can be overcome by providing different reference intervals. There are several problems with this argument. For one thing, from a strictly scientific perspective, for the 8 analytes discussed in this paper as well as for many others, there are true values, which are the values we should report (10). At least as important, though, is that, in many cases (e.g., hemoglobin A₁c (5) and cholesterol (11)), physicians use decision limits promulgated by national or inter-

---

1 Beth Israel Deaconess Medical Center, Boston, MA.

* Address correspondence to this author at: Beth Israel Deaconess Medical Center, Yaminis 309, 330 Brookline Ave, Boston, MA 02215. E-mail ghorowit@bidmc.harvard.edu.

Received March 28, 2014; accepted March 31, 2014.

1 Nonstandard abbreviations: PT, proficiency testing; EQA, external quality assessment; MDRD, Modification of Diet in Renal Disease.
national guidelines rather than reference intervals established by individual laboratories, making it absolutely essential that we report accurate values.

Similarly, increasingly, physicians are expected or required to use laboratory values, or formulas that depend on them, for clinical care. For example, for patients on dialysis, estimates of prognosis, and sometimes even physician payment schedules (12) can be based on serum albumin measurements, often without specifying which method should be used, although differences between methods are well documented (10, 13). As another example, the Modification of Diet in Renal Disease (MDRD) and other equations for calculating estimated glomerular filtration rate depend on serum creatinine values. The widespread use of these equations requires that accurate creatinine values be used, a situation that did not prevail when the MDRD equation was first promoted (14).

No doubt, there will be some cases in which age, sex, and/or ethnicity differences persist even after methods are harmonized. In such cases, reliance on different reference intervals may be required (15). But we should work to minimize the frequency of these occurrences.

Finally, there may be cases, such as tumor markers or immunosuppressant drugs, where methods cannot be fully harmonized. Even in these cases, however, we can do a better job informing physicians of our limitations. For tumor markers, we should report the method used along with the result values, and we should assist in reestablishing baselines for patients when methods are changed (16). For immunosuppressants, we should ensure that methods are harmonized on at least the parent compounds, and we should inform physicians that values, even on the basis of the same principle and from the same manufacturer, may not be interchangeable (4).

In short, the contribution of Stepman et al. (8), as powerful as it is, will not solve all of the problems that we face in trying to achieve interchangeability and accuracy of clinical laboratory test results. However, to paraphrase Voltaire, let us not allow the perfect to become the enemy of the good. The fact that we may not be able to solve all the problems should not prevent us from solving many of them. The overwhelming majority of laboratory tests are generated with homogeneous systems. We can, and should, do more to ensure that end users who use these systems report accurate results.

It is very important to emphasize that Stepman et al. (8) propose that their method complement, not replace, conventional PT. By distributing a relatively small number of single donor samples to a relatively small number of laboratories representing the major homogeneous systems, one can determine whether those systems are accurate. One can then use conventional PT materials among large numbers of participants to assess whether individual laboratories are performing that method comparably to their peers. More concretely, assume that a group of 10 laboratories that use homogeneous system A is shown to obtain accurate results on commutable materials. Then laboratory X, which did not participate in the original survey using those commutable materials, but which uses that same homogeneous system A, can infer, with reasonable confidence, that if it gets values close to the mean for that system on the conventional survey, its results on patient samples are almost certainly accurate.

For now, we should celebrate the insights that Stepman et al. (8) have provided. As good as conventional PT may be, we can do better. We are indebted to these authors for shedding light on a problem we may have assumed we did not have and, more important, for providing a powerful tool to help us make things better.


