Although inference of the proper phase in this case is still obvious, with a chance of an erroneous phase assignment of only 0.64%, the $B$-score acts as a warning signal, alerting the user to the need for further investigation of the data.

With the rapid adoption of NGS in the clinical testing arena, our ability to collect raw data has far outpaced our capacity to analyze and interpret results. Although there have been numerous computational tools developed for these purposes, most of them have been designed for research purposes and lack the proper statistical underpinnings that are required at the level of clinical interpretation.

Phasing is a good example of an unmet clinical need. Accurate phasing of compound heterozygous alleles is an important and common problem for many clinical sequencing laboratories. Our method offers a simple yet robust solution that is capable of providing a statically relevant score to help laboratory personnel interpret haplotypes in the presence of experimental error. Our testing scenario showed complete concordance with much more laborious conventional phasing methods. The method is in principle applicable to any gene/gene locus and capable of phasing across significant genetic distances.

**Author Contributions:** All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.

**Authors’ Disclosures or Potential Conflicts of Interest:** No authors declared any potential conflicts of interest.

**References**

An advantage of this method is that the HRF and CRF values are allowed to adjust over time because they are updated using PRF in Eqs. (1) and (2). This compensates for slow process changes, such as degradation of internal standard (IS) stock solution concentration. Regarding more rapid changes, Pauwels et al. (1) point out that bad IS preparation contributed substantially to the variance obtained in their approach. With our strategy, the introduction of a bad lot of IS may initially produce erroneous, possibly unacceptable QC results if the change of IS concentration is large, but over a period of several runs the method will self-correct. If \( W \) is large (e.g., 0.9), it takes more time to adjust to new values of HRF and CRF. These errors can be minimized by replacing the IS solution or by other methods alluded to by both Olson et al. (2) and Pauwels et al. (1).

After trying several weighting factors (e.g., 0.5 and 0.9), we empirically found by subjective judgment that \( W = 0.75 \) provided a good compromise between stabilizing the process against random variations and the time to self-adjust after sudden unexpected systematic process changes. In both theory and practice, the dampening effect provided an improvement in the imprecision obtained (Table 1).

This algorithm presumes knowledge of a HRF and is therefore not self-starting. An initial HRF can be determined using a conventional multipoint calibration.

Analyzing many runs, we have found that patient sample batches may be unnecessarily rejected due purely to a bad calibration curve. Our strategy allows such batches to be processed and judged by acceptability of QC samples. We find use of the weighted mean RF can provide not only better imprecision but more accurate sample determinations as well. Traditional multipoint calibration can be performed every 6 months, to verify method linearity, as is standard for analytical measurement range evaluation. In some situations, such as adjustment of collision energy, it may be necessary to reestablish and reset the HRF. One can also adapt the method in Rule et al. (4) to certain nonlinear calibration schemes that can be characterized by a response factor.

CLSI document C43-A2 (5) suggests that historical calibration curves may be used if they are shown to be linear over time. At least 2 standards are suggested for each batch of samples with 1 standard at the threshold concentration. In our approach, we extract and analyze a sample at the limit of quantification before each run to verify acceptable method and instrument performance. Interestingly, in high-volume core laboratories it is common to use just 2 calibration standards to verify or make adjustments to a multipoint calibration. For example, 2 calibration standards may be evaluated just once every 7–28 days (Siemens, ADVIA Centaur package inserts).

<table>
<thead>
<tr>
<th>Table 1. Comparison of QC means, CVs, and bias, along with patient sample comparisons for 3 analytes by both conventional and alternative calibration approaches. a</th>
</tr>
</thead>
<tbody>
<tr>
<td>QC (n = 64)</td>
</tr>
<tr>
<td>Low</td>
</tr>
<tr>
<td>Conventional</td>
</tr>
<tr>
<td>Alternative</td>
</tr>
<tr>
<td>Bias, %</td>
</tr>
<tr>
<td>High</td>
</tr>
<tr>
<td>Conventional</td>
</tr>
<tr>
<td>Alternative</td>
</tr>
<tr>
<td>Bias, %</td>
</tr>
</tbody>
</table>

Deming regression

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Slope (95% CI)</th>
<th>Intercept (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>355</td>
<td>0.992 (0.989–0.995)</td>
<td>1.006 (1.003–1.008)</td>
</tr>
<tr>
<td></td>
<td>758</td>
<td>0.992 (0.989–0.995)</td>
<td>1.006 (1.003–1.008)</td>
</tr>
<tr>
<td></td>
<td>872</td>
<td>0.992 (0.989–0.995)</td>
<td>1.006 (1.003–1.008)</td>
</tr>
</tbody>
</table>

\[ 5 \mu \]

\[ 0.033 \]

\[ 0.175 \]

\[ 7.35 \]

*a Data are from a total of 32 runs acquired over a period of 2 weeks.

*b DHEA, dihydroepiandrosterone.

Author Contributions: All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.

Authors’ Disclosures or Potential Conflicts of Interest: Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:

Employment or Leadership: None declared.

Consultant or Advisory Role: None declared.

Stock Ownership: None declared.

Honoraria: None declared.

Research Funding: A.L. Rockwood, ARUP Laboratories.

Expert Testimony: None declared.


References


To the Editor:

We read with great interest the report of Stepman et al. (1) identifying analytical errors in commercial analytical systems by using an external quality assessment (EQA) scheme with single-donor specimens. The use of such materials provides an assessment of method accuracy that is generally regarded as the gold standard in such comparisons (2). Unfortunately, only a limited number of laboratories can typically participate in such assessments without resorting to pooled or processed proficiency materials. Analytical ranges are also usually limited using single-donor specimens unless these materials are supplemented with additions that potentially obviate the benefits of using such specimens.

Although intermethod bias cannot be definitively assessed with materials that have not been shown to be commutable for the examined analyte and analytical techniques, it is incorrect to assume a priori that pooled or spiked specimens, by their nature, are noncommutable and that method biases seen with such specimens are largely due to matrix effects. The data provided by Stepman et al. (1) allow for comparison of bias estimated with more commonly used materials in proficiency testing, and we undertook a review of bias estimated by traditional proficiency testing (PT) in light of the data reported in that paper.

Through the New York State Laboratory Reference System Proficiency Testing Program, we used pooled human serum, derived from plasma (Bioresource Technology), with analytes added as necessary to prepare specimens with clinically relevant concentrations for each test event. Specimens were sterile-filtered (0.22-μm pore size), aliquoted, and stored frozen at ≤−80 °C until they were shipped in the frozen state to 427 laboratories for analysis. Results for the analytes and instrumentation from participants matching the Stepman et al. study (1) are shown in Table 1. To approximate the analysis done in the original study (1), we stratified the data into low-, middle-, and high-concentration ranges for each analyte. Target values for each specimen and analyte were established with a robust statistical technique using all the participant data (3), and we calculated, for each assay system, the median bias from those target values within each category.

Comparison of the bias assessed by the 2 EQA schemes shows that processed proficiency fluids largely provide similar estimates of bias for 6 of the 8 analytes examined (Table 1). Excluding LDL and HDL cholesterol, 74 of 90 estimates of bias (82%) differed by less than the threshold criteria proposed by Stepman et al. (1): ≤4.5% for glucose, phosphate, and triglycerides; ≤4.0% for cholesterol, creatinine, and urate. Using a much stricter criterion of differences of ≤2% (likely within the measurement uncertainty of either study), we found concordance for 63% of the comparisons. Estimates of bias for LDL and HDL cholesterol in our study were substantially greater than those obtained with neat human serum. The lack of commutability of processed materials for lipoproteins and lipid measurements has been demonstrated previously (4), and this was not an unexpected finding. Nonetheless, results for cholesterol and triglycerides were largely within bias thresholds even though processed human serum has been previously shown to be noncommutable for these lipid analytes (4). The notable bias found in the Ortho procedure for phosphate (approximately 6%–15%) was nearly identical in both studies and followed the same degree of dependence on phosphate concentration (Table 1).

In addition to the type of specimens used in the 2 studies, there are other differences in design including the concentration range of analytes, number of participant laboratories, number of specimens distributed, and method for calculating bias that make the similarity in bias estimates all the more remarkable.

Although the comparison we performed here cannot substitute for a full commutability study, it provides insight into the utility of pooled and spiked materials in PT. Whereas there were instances in which the bias of the processed materials differed from those of the patient samples, the more common observation was agreement between the 2 EQA schemes for estimating intermethod bias for 6 analytes. These data suggest that accuracy-based PT with pooled and/or spiked specimens is achievable for at least