Laboratory Medicine in the Era of Disruptive Technology
LMCE 2017 & KSLM 58th Annual Meeting
October 18-20, 2017
Grand Walkerhill Seoul, Korea
www.lmce-kslm.org

EP15, 3rd Edition – Improving on a CLSI performance evaluation guideline

A. Paul Durham
APD Consulting, Culver City, CA, USA

EP15 is a consensus-based guideline from the Clinical and Laboratory Standards Institute (CLSI.org), now in its third edition. Released in September 2014, EP15-A3 represents a substantial refinement of the previous, April 2006 edition. CLSI's EP-series includes several guidelines devoted to establishing and/or verifying relevant performance characteristics of measuring procedures ("assays"). Developed by volunteers representing the principal stakeholders, they aim to formulate testing strategies mutually acceptable to IVD manufacturers, clinical laboratories, and the regulatory/accrediting agencies entrusted with their oversight.

EP15 addresses verifying the accuracy ("trueness" in ISO terminology) and precision of quantitative assays — especially commercial assays already validated and cleared for marketing. The guideline is intended for use primarily when a clinical laboratory first introduces such an assay into its repertoire for patient sample testing and reporting. At this juncture, the laboratory faces regulatory — in the US: CLIA-based — demands for verifying precision, accuracy, and more (viz., the assay's "reportable range" and reference intervals, matters addressed in other CLSI guidelines). These require the laboratory to demonstrate, experimentally, performance *consistent* with certain expectations for the assay.

For imprecision, this has always been construed by EP15 as expectations *set by the manufacturer*: "claims" encapsulated in the assay's labeling as a table summarizing statistics from a suitably definitive precision study. (As yet, the package insert is not required to include confidence intervals or a precision profile.) The study typically conforms to EP05's single-site design and analysis, which involves testing samples representing a spectrum of concentrations on at least twenty days, two runs per day, two replicates per run — a "20 by 2 by 2" design — using analysis of variance (ANOVA) to extract sample-specific estimates for repeatability and "within-laboratory" precision — formerly designated within-run and "total" precision, respectively.

To demonstrate *consistency* with these claims, EP15-A3 requires measurements for at least two samples on five days, one run per day, five replicates per run — a "5 by 5" design, associated with far simpler data processing (one-way ANOVA, rather than two-way). This small study is *comparable* to the manufacturer's in the sense of yielding estimates for the same two precision types.

As for bias, the measure of "trueness" (accuracy in the narrow sense), the assay's labeling may contain no claims, explicit or implied, which the laboratory is in a position to verify. The package insert nearly always summarizes a method comparison study; all too often, however, the assay which served therein as the comparative (reference) method is no longer available to the laboratory. Some inserts may report on a *recovery* study, assessing bias relative to one or more samples with known values — reference preparations — but this is common for only a very few measurands. Another frequent obstacle is superficial reporting in the package insert that renders extracting manufacturer estimates of bias and their uncertainty problematic or impossible. Moreover, even when the laboratory can revisit such studies, there may be little value in doing so beyond meeting a regulatory demand.

As an acceptable and often more informative approach to "verifying accuracy", EP15-A3, like its predecessor, encourages laboratories to perform a recovery study using samples with value assignments based on peer group data from proficiency testing (PT) or interlaboratory QC programs, data representing measurements obtained in actual practice by a substantial number of clinical laboratories. In assessing their uncertainty, EP15 measures the "effective study size" (degrees of freedom) associated with these value assignments conservatively, in terms of the number of participating laboratories, rather than the total number of measurements. If successful, this approach serves to demonstrate performance — in the laboratory seeking to introduce the assay — consistent with relevant expectations set for it, not by the manufacturer's original validation studies, but by representative end-user laboratories already making use of the assay.

For the recovery study, EP15-A3 recommends the same experiment as for the precision verification study: a "5 by 5" design involving five days, five replicates per daily run, and at least two samples. Clearly there are efficiencies associated with combining these studies, not only in benchwork, where overhead is limited to just one set of runs, but also in data processing, which is the same for both, up to a point. Using the same samples yields still greater savings, but may entail compromise. For example, when assessing imprecision, native patient samples are inherently preferable to typical QC materials; but if the latter are suitable for internal QC, they should behave in a similar manner with respect to precision.

Making do with just two samples amounts, at best, to "spot checking" for precision and accuracy. If the goal is to check performance throughout most of the assay's stated measuring interval, one could argue (though the guideline doesn't) for processing however many samples may be needed to construct imprecision and bias profiles, namely, at least 3 or 4 samples, likely more, with suitably spaced concentrations. Such an approach cannot be dismissed as unduly burdensome if the same samples are processed for both imprecision and bias in the same "5 by 5" study.

Because the guideline has so little to say about patient sample method comparison studies, one might think that it no longer supports this approach to verifying bias claims. On the contrary, and in agreement with the previous edition, EP15-A3 states that, when applicable, this approach may be preferable if the comparison is of interest (Section 3.2, 3rd bullet). However, to avoid conflict or redundancy, EP15-A3's document development committee made a decision to defer, in all matters associated with bias estimation via patient sample method comparisons, to EP09-A3, the guideline dedicated to this topic and addressed to both manufacturers and laboratories. (EP09-A3, incidentally, requires 40 samples when verifying a manufacturer's bias claims, twice the number required in EP15-A2.)

This means that, as of the third edition, EP15 is no longer self-contained: laboratories must consult EP09-A3 for one of the acceptable approaches to bias verification. They should also have access to EP05-A3, the guideline for establishing and validating an assay's precision characteristics. As it happens, all three of these guidelines (EP15, EP09, EP05) were undergoing revision at about the same time; some participants worked on more than one committee, helping to ensure appropriate coordination.

Although addressed primarily to manufacturers, EP05-A3 includes a tutorial (Section 1.5) on basic precision concepts, pitched at a level that should make sense to any laboratory personnel with a practical understanding of internal QC; it should therefore be instructive for the intended readership

of EP15. Conversely, the EP15-A3 computations are relevant to EP05-A3, now that the latter, in response to regulatory pressure, includes guidance on reproducibility (i.e., multi-site precision) studies. Of the two principal formats for such a study, one amounts to three or more laboratories implementing an EP15-A3 "5 by 5" study on a common set of samples. The salient role played by the "5 by 5" in EP05-A3 underscores the reasonableness of EP15-A3's independent adoption of it for precision verification.

Also worth noting: EP15's relevance and intended readership have expanded considerably, since the bias estimation procedure in the latest edition's chapter on recovery studies, abstracting from considerations of experimental size, is as applicable to establishing or validating such estimates, a task for manufacturers and developers, as it is to verifying or checking them. EP15-A3 is thus far the only CLSI document with extensive guidance on recovery studies.

EP15-A3 now addresses data integrity issues in detail; and it includes a formal test for "outliers" (the Grubbs' Test), with an extensive table making it relevant in other contexts.

The guideline devotes considerable attention to *interpreting* the multi-sample precision verification and recovery studies, surveying possible failure modes from a "quality goals" perspective based on measurand-specific total allowable error (TEa) limits and their decomposition into limits for imprecision and bias. The manufacturer's precision claims require inspection for consistency with allowable precision limits; while the laboratory's estimates in a recovery study may be deemed acceptable, even when they differ significantly from the value assignments, providing the differences are within allowable limits for bias.

EP15-A3 draws attention to a major difficulty, passed over without comment in previous editions, namely the problem of determining from a package insert's precision table the claims applicable at the mean values obtained in a precision verification study, which will nearly always differ from any concentrations listed in the manufacturer's table. EP15-A3 suggests various *ad hoc* stratagems — interpolation, nearest neighbor, local smoothing, etc. — for coping with this intelligently, but the determination inevitably introduces considerable subjectivity at this juncture.

The problem traces to what remains the principal shortcoming of guidelines for establishing an assay's precision, including its low-end precision. EP05 stops short at constructing a table of concentration-specific precision estimates, when it should also require (a) fitting to these statistics "precision profiles" characterizing repeatability and within-laboratory precision as continuous functions of concentration across all or most of the stated measuring interval, and (b) incorporating algebraic and/or graphical representations of these profiles in the package insert. For more than a decade, this topic has been on the agenda for EP05 committees, but it remains unresolved, despite the emergence of sound precision profiling algorithms and readily available implementations.

The EP15-A3 precision verification and recovery studies both rely on classic ANOVA-based statistical techniques yielding not just estimates of imprecision but also information on their uncertainty — information essential to assessing *consistency* of the laboratory's measurement results with the corresponding precision claims and value assignments.

The guideline implements what are essentially confidence interval calculations in the precision verification analysis. The laboratory's estimate for a sample, even if higher than the manufacturer's

claim, is deemed consistent with that claim unless the estimate also exceeds what the latest edition calls the UVL (upper verification limit), where "the UVL represents the upper 95th percentile expected for imprecision estimates obtained in an experiment similar in size and design to the user's precision verification study when the claim is correct." (Section 2.3.6)

Manufacturers are expected to have access to statistical know-how and software; but laboratories, especially small laboratories, must all too often cope on their own. Accordingly, EP15-A3 gently spells out the computations involved in one-way random-effects analysis of variance for potentially unbalanced data sets — i.e., even when there may be missing measurement results: neither EP15-A2 nor EP05-A2 addressed this. (The corresponding two- and three-way computations called for in EP05-A3 were considered too complex to spell out in that guideline!)

A prominent feature of EP15-A3 is the disappearance of worksheets, reflecting a recognition that most laboratories would opt nowadays for tabulating and manipulating numeric data in a spreadsheet environment, rather than on paper. Modern spreadsheets typically include a facility for generating one-way ANOVA summaries from a suitable tabulation of the raw data. EP15-A3 recommends taking advantage of this to by-pass some tedious number crunching. Subsequent calculations are either quite simple or implemented via look-up tables with the flexibility to provide for studies involving 5 to 7 days, up to six samples, etc. Hence, for both analyses (precision, recovery), laboratories should no longer have to perform difficult calculations manually, such as evaluating the treacherously complex Satterthwaite formula, unless they conduct an atypically large or small experiment.

As guiding principles, CLSI document development committees try to ensure that the recommended studies are rigorous, relevant, informative and clearly delineated, while shielding manufacturers and laboratories so far as possible from unduly burdensome benchwork, excessive material costs or delays, error-prone computational drudgery, and arcane technicalities.

During the committee's first teleconference, for example, one participant proposed increasing the minimum number of replicates in the precision study from three to five, to achieve better repeatability estimates. Others argued for increasing the number of days instead — requiring at minimum, say, a "7 x 3" study design — to improve the reliability of estimates for within-laboratory precision, the clinically more important precision type, even for assays with relatively large between-day (or between-run) sources of variability. The first proposal was soon accepted, on the grounds that for many, if not most assays, it would improve both precision estimates, with a relatively modest increase in benchwork and no additional delay in completing the study. Thus did EP15-A3's "5 x 5" design emerge as a tradeoff between competing considerations.