The History of Evaluation Criteria for CAP Surveys

William B. Hamlin

I summarize the evolution of target-value assignment methods and laboratory performance assessment (evaluation) within the College of American Pathologists Interlaboratory Survey Programs and review the importance of, and longstanding interest in, medical relevance criteria in setting analytical goals.

Indexing Terms: proficiency testing • analytical goals

The origin of proficiency testing as we know it today was the 1947 Belk and Sunderman study reported in the American Journal of Clinical Pathology (1). In the early 1960s, the College of American Pathologists (CAP) began to distribute unknown samples in several laboratory disciplines, in an effort to assess the state of the art with regard to interlaboratory performance. By 1960, the CAP had initiated its subscription “survey” programs in multiple disciplines; and by the mid-1960s, the CAP Laboratory Accreditation Program required participation in interlaboratory surveys as a condition of accreditation. This interlaboratory survey activity has steadily expanded to include virtually all clinical laboratory subspecialty disciplines and now consists of >100 different individual surveys, subscribed to by >20,000 laboratories in several countries besides the US and Canada.

Any discussion of the evolution of the evaluation criteria for these interlaboratory survey programs necessarily involves consideration of two related components: (a) value assignment (i.e., what is “truth”?), and (b) performance assessment (i.e., how close to “truth” are the reported results?).

With respect to value assignment, the CAP has progressed through a series of techniques. In the early years, the basis for “truth” was a single referee, or “Olympian” laboratory that was presumed to be “expert.” It rapidly became apparent that Olympians were neither necessarily “right” nor the sole possessors of “truth,” and groups of highly reputable referee laboratory facilities were used to establish target values for the survey samples. Again, problems with this method were soon recognized, and it became apparent that with this system bias was not only possible but, in fact, likely.

In the early to mid-1960s, it was recognized that, when data were available in sufficiently large volumes, participants’ results could be used for determining the “true” value of interlaboratory survey samples, and the technique of using the “all participant—all method mean” for value assignment was initiated. This method for value assignment was verified during the 1970s by several studies reported by Gilbert (2), which compared “all participant—all method mean” values derived from the results submitted by many laboratories with analytical values obtained by the National Bureau of Standards, using Definitive Methods. These studies clearly demonstrated the validity of the “all participant—all method mean” technique for value assignment.

No system, however, is flawless. Not all analytes, and not all laboratory disciplines, lend themselves to such a method of value assignment. Additionally, technological advancements have progressed at a phenomenal rate. The rapid development of new instrumentation and new analytical methods and the ability to measure new analytes have contributed to the difficulty in assigning values to survey samples.

It has become increasingly apparent that, for survey samples, some instrument/method systems produce results that differ from the “all participant—all method mean” but nevertheless can be demonstrated to produce accurate results for samples from patients. This phenomenon has more recently been attributed, in part, to a “matrix” effect of the manufactured sample material. The recognition of this reality resulted in the development of using “peer group” data to assign values for specific instrument/method systems and specific analytes. The results reported by laboratory groups of sufficient size who use the same instrument/method system can be used to assign values for survey samples for that peer group.

Today, the CAP uses all these techniques for assigning values to survey samples—all participant—all method means, peer group means, and referee or reference laboratory values for certain disciplines and (or) analytes.

With regard to evaluating results reported for analytes by individual laboratories, an analogous evolutionary process has occurred. Initially, the CAP did not attempt to “grade” results but rather compared the results reported by individual laboratories with those reported by all laboratories after a rather detailed statistical analysis of the data. It was thought that if a laboratory that was producing significantly deviant results could be provided the data to demonstrate that fact, and could be given the information in a format that indicated the level of accuracy of the overall laboratory population, that facility would be stimulated to take appropriate action to improve its performance. Such has in fact occurred for most laboratory participants.

With the introduction of the Medicare regulations in 1965, and the CLIA '67 regulations shortly thereafter, it became necessary to formally grade survey results. Primarily, this was precipitated by regulatory requirements designed to distinguish “good” laboratories from
"bad" laboratories, i.e., those with "acceptable quality" from those with "unacceptable quality." The CAP attempted to develop the required grading system in a manner that would help laboratories improve their performance.

The CAP approach was to designate reported analyte results as "acceptable," "needs improvement," or "unacceptable." For numeric results, this evaluation was based on the degree of deviation from the assigned value for that analyte. Results within ±2 SD were considered acceptable; those that fell outside the ±2 SD range, but were within ±3 SD, were designated as needs improvement; and finally, reported values that were outside the ±3 SD range were considered unacceptable.

For nonnumeric results, such as those in microbiology, blood banking, and microscopy, various other mechanisms were used to determine what was acceptable, unacceptable, etc. It was in these areas that medical relevance or medical usefulness was first utilized as part of the basis for evaluation or grading. For example, failure to isolate and identify a significant bacterial pathogen might be unacceptable, whereas successful isolation of the pathogen but inability to identify the organism at the species level might be considered as needs improvement, depending on the clinical importance of species identification.

In the early 1980s, it was determined that the evaluation system needed to be enhanced for some analytes. Assessment of submitted survey results through the use of fixed-limits criteria was introduced. Two situations were occurring that required attention. The ubiquitous ability to measure some analytes very accurately produced mean ±3 SD ranges that were so narrow that it was possible for a participant laboratory to report results outside that range and be evaluated as unacceptable, even though the reported results were within a medically useful and reasonably acceptable range. Conversely, for certain analytes, the mean ±3 SD range remained excessively broad and the laboratory population appeared to show no improvement in accuracy toward a level more attuned to medical needs. In both instances, the use of fixed-limits evaluation criteria consistent with medical decision-making and within the technical capabilities of the instrument/method systems being used was considered to be a more desirable way to assess performance and stimulate improvement.

These evaluation mechanisms, with some exceptions for peer group problems and nonnumeric results, continue to be used today. Notably, the CAP has always been willing to make a decision to not grade a specific analyte in a specific survey sample when the situation warrants such action. Survey samples are not specimens from humans; they do not always behave like human specimens in all analytical situations. The manufacture and distribution of survey samples can result in unexpected vagaries that must be taken into account with any evaluation system.

Throughout the history of CAP interlaboratory survey development, there has been a continuing recognition that any system for evaluating survey results reported by laboratories should include medical usefulness criteria as a major consideration. This need has been the principal subject of numerous CAP consensus conferences convened over the past 15–20 years to evaluate analytical goals (3–7).

The necessity for incorporating medical usefulness criteria into interlaboratory comparison evaluation systems and the difficulty in doing so are best described in a paper by Batsakis, published >10 years ago (8). All of his points are no less valid today than they were then.

Today, the laboratory community finds itself facing the implementation of CLIA ’88, which, through regulatory dictate, forces the use of interlaboratory surveys (i.e., proficiency testing) for purposes for which they were neither designed nor well-suited, and without appropriate attention to medical usefulness as an evaluation criterion.

The use of interlaboratory survey samples as the principal means of identifying "good" or "bad" results in the absence of defined analytical goals based on clinical utility (medical usefulness) cannot succeed. Boutilier, in 1975, while at the Center for Disease Control, made the following statement in a discussion of proficiency testing: "It is a tool which, when combined with other parts of a total program, can be used for laboratory improvement. It has been shown to be quite ineffective when used alone" (9).

The purpose of laboratory results is to provide physicians with information that can be used for optimum patient care; it is not necessarily simply to demonstrate accuracy. The ultimate analytical goal for clinical laboratories must be to achieve medical usefulness.

We hope that this conference will serve as a catalyst to refocus and accelerate the effort to incorporate medical usefulness criteria both into analytical goals and into the evaluation criteria that are used to assess clinical laboratory performance.

References
Discussion

Bradley Copeland: One comment and one question: Because I had extensive experience with the development and implementation of CAP surveys in the 1960s, I am sure that participants were never evaluated on the basis of comparison with a single laboratory's results. We chose multiple university-level laboratories and used the overall average and standard deviation.

My question has to do with matrix effects. Is the College of American Pathologists going to study and revise the survey material so that they can be confident that matrix effects are no longer a problem?

William Hamlin: I'm not sure that anyone has yet clearly defined what matrix effects are. Certainly, the College is interested in trying to define matrix effects and why they appear. Whether a sample can be produced, manufactured, and distributed without any matrix effects is a question for which I do not have the answer, but I doubt that it can be done. On the other hand, it is practically impossible for a large program to use fresh frozen samples; the distribution problems are overwhelming. That possibility, however, is being looked at in some depth. When I mentioned "Olympian" laboratories, I was looking at activities in the late 1940s and early 1950s, not as late as 1960. That method was rapidly abandoned. The only reason to use peer groups in evaluating proficiency testing or survey samples is when the problem arises where a specific instrument/method system can be demonstrated to produce valid results on patients' samples but, for whatever reasons, cannot produce the same degree of accuracy with survey samples. Even if you don't know for certain that the problem is due to matrix effects, there is almost no way to fairly deal with that participant except through peer group evaluation.

Carl Garber: Dr. Fraser, I was intrigued earlier by your comment that, for some analytes, performance was so much better than medical need. We might reflect a little on industries outside of the medical field and how they view this "so much better than need" situation. For example, Motorola looked at customer need vs their process performance capability, and set a goal of a factor of 6σ. On the other hand, we get excited when we see something that is 3σ, and say that it's too good. I challenge us in this field to start thinking forward into the 21st century and see where we come out relative to quality-improvement programs in the world marketplace—6σ, or target plus or minus six standard deviations.

Perhaps the reason clinicians do not understand variability is because the laboratory provides confusing data. Perhaps they don't understand variability because the analytical variability muddies the water relative to the clinical variability. If our performance were so outstanding (around the 6σ level), we could go back to the clinician and say, "We are very confident that the variability you see in your medical report comes from the patient." But right now, everything is so equivalent that we can't specifically interpret variability in certain laboratory results. Dr. Hamlin, what do you think about 6σ?

William Hamlin: Let me respond by specifically citing an example. At one time, the participants' results reported for potassium in this country in the CAP surveys had become so tight that when the means plus or minus several standard deviations were computed, the acceptable range was ±0.1 or 0.2 mmol/L. That meant that we were failing laboratories who were reporting potassium at ±0.3 mmol/L. In terms of grading and government oversight, this is not fair because that level is perfectly reasonable for dealing with patients or clinical problems. This kind of example prompted the use of fixed criteria. So, I don't see anything wrong with evaluating such things as long as we are evaluating them for medical usefulness purposes. However, that's not what proficiency testing or interlaboratory comparisons are being used for any more, and that is a fact.

Basil Doumas: Concerning fixed limits and the use of medical criteria: Which approach to setting them was used or is being used? Is it an objective approach or just a gut feeling based on experience? I agree that if you comply strictly with the standard deviation definition, the acceptable performance limits for some of the tests, e.g., urea nitrogen, will be too narrow.

William Hamlin: I was not present when those decisions were made. Some people in this room were present at most of the meetings where the fixed-limit criteria were decided upon. Perhaps Dr. Batsakis can enlighten us; I suspect that it was empirical.

John Batsakis: Semi-empirical. We had to select from the three venues of testing. We selected the diagnosis, e.g., hypercalcemia vs normal calcemia. In such cases, we came up with a consensus of the pathologists around the table and fixed limits. Some analytes cannot be done that way. Tomorrow, perhaps, when we talk about thyroid testing, that will become evident.

I want to go back to Dr. Fraser and Dr. Hamlin. Both of you use "too stringent" and "too loose." I think both of you said (I'm paraphrasing), let's not waste too much time on the too-stringent limits. Shouldn't we concentrate on the analytes where limits are manifestly too loose. Is that a correct interpretation of your statements?

Callum Fraser: Some desirable performance standards based on biology are too stringent and cannot be achieved with current technology. For these, desirable performance standards should be viewed as targets that are worthy of attainment. Where the desirable performance standards are loose, then the goals are achieved; in theory, improvement is not required, but it would be very advantageous for both clinical and quality-management reasons.

The US is not unique in using fixed limits in proficiency testing programs. Germany and Australia use fixed limits. A comparison of the limits is quite inter-
testing: They are very different for some quantities, which shows they were derived mainly by empirical means; they are not very objective.

Perhaps part of the problem concerning interpretation of laboratory results is that we do not try hard enough to give clinicians the detail that allows them to infer the total reliability of test results. For example, it has been proposed that clinicians should know the SD of tests. How many clinicians actually document in their laboratory handbook the data on biological variation? How many show tables of significant differences vs probability? How many flag reports of serial results when changes are significant, highly significant, or very highly significant? How many teach medical students about objective numerical interpretation? How many publish or try to publish data on test-result variability in the widely read medical journals?

Peter Wilding: Individuals who practiced clinical chemistry in the 1950s will remember assaying for sodium and potassium by titration. The error in those assays undoubtedly influenced the concept of biological variation held at that time. Therefore, Dr. Fraser's suggestion that biological variation should be the basis of quality goals is flawed. The 1976 Aspen Conference identified, or qualified, "achievable" and "desirable" precision for various analytes. What is interesting, is that the North American diagnostics industry used these figures or categories extensively because they clearly identified the priorities for method improvement. Where the desirable precision was not achievable, at least there was a clear target for an analytical goal. Today we need a mechanism that identifies the top priorities for method improvement and that indicates where targets can be relaxed to achieve point-of-care utility.

Jack Levine: I wonder why programs don't look at all the medical requirements, all the medical limits, and start developing criteria for the extreme situations that can occur in a clinical laboratory. I think this is where we miss the mark. A lot of our current proficiency surveys focus on a rather narrow analytical window. Instruments now have broad analytical ranges that can meet the measurement requirements for almost any condition an acutely ill patient can present to the laboratory. I don't believe your proficiency programs really address meeting medical requirements for the extremely ill patient.

William Hamlin: Regarding the distribution of proficiency testing samples or surveys, the College distributes proficiency samples and surveys in all of the ranges of health and disease for which we can reliably manufacture the samples. That is true for all of the surveys I am aware of, including microbiology, clinical chemistry, therapeutic drug monitoring, forensic toxicology, etc. The problem with that, from a survey program perspective, in much of the discussion today, is how can we know, particularly with a manufactured sample, what the biological variation may or may not be for a given laboratory's patient population? It may be very different from one environment to another. If one is dealing with a facility in which hospitalized, elderly patients with debilitating or terminal diseases are dealt with almost exclusively, the problems involved are different from those of the laboratory for which the bulk of the samples are coming from upper-middle-class, private physicians' patients, who are basically being screened to determine if the physicians can find something for which they can prescribe medicine or order more tests. Variations of those extremes are almost unlimited. I doubt that it is practical or possible for a program distributing proficiency testing samples to cover all possibilities. I think the comments made about the criteria are what this conference is supposed to address. Perhaps some of the evaluation criteria used for College's PT programs may be inappropriate. If so, then as soon as this Forum is over, and all of you have defined how it can be done, using clinical criteria and medical usefulness criteria, tell us how to do it, and we will incorporate those criteria into the College survey programs.

Neal Dawson: Being a general internist, I come to this discussion from a very different set of circumstances. I have a question about your concern with increasing precision. Have people looked at issues such as the marginal cost-effectiveness of increasing precision? At least within the realm of other medical technologies, looking at that ratio has become essential in view of the current emphasis on total health-care costs. Certainly this is an important thing to consider. When you are talking about effectiveness, the effectiveness would generally relate to some clinical outcome: at least, as a clinician, that is how I would view it. A lot of the issues about passing the various criteria for performance of tests appear to be problems with the criteria themselves rather than other sorts of issues. Perhaps the criteria are inappropriate in light of the clinical implications or the actual goals in testing. Second, the term 'empirical' has been used by several individuals, but how they are using it is not how I think of its meaning; so I would like a clarification as to what people mean by empirical. Do people mean intuitive or something different?

Callum Fraser: "Empirical" means based on observation, not on numerical fact or theory.

Eugene Harris: I think we ought to look at the approaches taken by Europeans to resolve the question of medically relevant analytical goals. Callum Fraser of the UK and Per Hyltoft Petersen of Denmark use biological variation as a general tool, whereas Torgny Groth and Carl de Verdiere of Sweden develop a specific mathematical model for each medical problem. In all cases, however, these and other European workers are using objective criteria to arrive at analytical goals. They might be looking at different populations, but they all have an objective, scientific philosophy. I think this is the way to go. When we in the US talk about defining medical usefulness by asking the opinions of individual clinicians, or even groups of clinicians, we are not using scientific criteria to resolve our problem, because medicine is primarily a subjective art to achieve certain results in patients. And we will inevitably get widely varying, anecdotal advice that is bound to change as
biochemists and other scientists teach clinicians more about physiology and disease. In my view, if we continue to ask practicing physicians for off-the-cuff answers to what is basically a scientific question, we will never get solid, reliable results that will satisfy governmental agencies or clinical chemists themselves.

Mario Werner: I would like to challenge the concept that greater precision is forever desirable. The central point in this meeting rather seems to be that we can say, "No, we have a medically definable target for precision, where we can stop further improvement and expense." I have no way of evaluating the relevance of error reduction at the Motorola Corporation as an example without knowing the details necessary for comparing this model with a model for medical practice. The issue is, to what specific purpose and under what specific circumstances do you gear your efforts? In the case of analytical precision in the clinical laboratory, medical relevance provides these relevant benchmarks, and my bias is that today's laboratory in many respects meets medical needs. Consider, for instance, acute myocardial infarction. The World Health Organization has said that this diagnosis can be based on enzyme assays in some cases. Consequently, one can argue that extremely precise assays of indicator enzymes are required because patient management will rest exclusively on these findings. However, one need also know that, among cardiologists, it is an accepted number that only half of the admissions to the intensive cardiac care unit for acute myocardial infarction are expected to be correct. Indeed, if more than half have acute myocardial infarction, you are considered to be underadmitting; if under half have myocardial infarction, you are overadmitting.

Meningitis provides another example. The essential differentiation here is between a viral and a bacterial etiology. Clearly, the laboratory provides crucial information for that decision. However, in actual practice, almost always when meningitis is diagnosed, antibiotic therapy will be instituted regardless of test outcome.

From these examples, it becomes obvious that, in defining required analytical precision, the full circumstances in which the data or information will be used must be considered.