Clinical Chemistry 48:6
919–927 (2002)

Laboratory Management

Setting Performance Goals and Evaluating Total Analytical Error for Diagnostic Assays

Jan S. Krouwer

Background: **Total analytical error has been a useful metric both to assess laboratory assay quality and to set goals. It is often estimated by combining imprecision (SD) and average bias in the equation: total analytical error = bias + 1.65 × imprecision. This indirect estimation model (referred to as the simple combination model) leads to different estimates of total analytical error than that of a direct estimation method (referred to as the distribution-of-differences method) or of simulation.**
Methods: **A review of the literature was undertaken to reconcile the different estimation approaches.**
Results: **The simple combination model can underestimate total analytical error by neglecting random interference bias and by not properly treating other error sources such as linear drift and outliers. A simulation method to estimate total analytical error is outlined, based on the estimation and combination of total analytical error source distributions. Goals for each total analytical error source can be established by allocation of the total analytical error goal. Typically, the allocation is cost-based and uses the probability of combinations of error sources. The distribution-of-differences method, simple combination model, and simulation method to evaluate total analytical error are compared. Outlier results can profoundly influence quality, but their rates are seldom reported.**
Conclusions: **Total analytical error should be estimated either directly by the distribution-of-differences method or by simulation. A systems engineering approach that uses allocation of the total analytical error goal into error source goals provides a cost-effective approach to meeting total analytical error. Because outliers can cause serious laboratory error, the inclusion of outlier rate estimates from large studies (e.g., those conducted by manufacturers) would be helpful in assessing assay quality.**

Krouwer Consulting, 26 Parks Dr., Sherborn, MA 01770. Fax 508-647-9380; e-mail jan.krouwer@attbi.com.
Received November 2, 2001; accepted March 4, 2002.

Performance goals for laboratory testing have been discussed for many years, with perhaps the best known starting point being Tonks *(1)*. A collection of ideas on performance goal strategies has been published from a recent conference *(2)*. Many of these efforts to develop goals have produced valuable insights in understanding quality requirements for laboratory tests. For example, Klee et al. *(3)* showed that lot-to-lot reagent bias, as one error source, could adversely affect patient treatment.

Performance goals for laboratory testing have most often been developed for total analytical error and for imprecision (SD) and bias. A total analytical error goal requires that the combination of errors from all sources is within some acceptable limit. From a clinician's standpoint, this is the most useful goal, because an incorrect laboratory result, regardless of which component(s) of total analytical error has caused it, is harmful. A total analytical error goal also enables a simple and cost-effective assessment of the suitability of a particular assay because there is only one error source to estimate.

On the other hand, manufacturers are interested in total analytical error sources because knowledge of these error sources and their subsequent correction are the only way to reduce total analytical error and hence improve quality. Laboratories have a position between manufacturers and clinicians. They do not have the resources (and often the proprietary knowledge required) to perform the extensive studies carried out by manufacturers, but they are responsible in part for the quality of assay results and thus must be knowledgeable in total analytical error as well as its sources.

Currently, most total analytical error performance goals are not provided directly *(4, 5)*. Rather, the total analytical error goal is constructed from a combination of a bias goal and an imprecision goal (Eq. 1).

$$\text{Total analytical error} = \text{Bias} + 1.65 \times \text{imprecision} \quad (1)$$

This model is intuitively appealing for its simplicity because it would seem that bias and imprecision (e.g., systematic and random error) cover all possible error sources.

This report will show that the model represented by Eq. 1 (hereafter referred to as the simple combination model) can underestimate total analytical error because some possible error sources are either absent from the model or not treated properly. Additional error sources that may be used to establish a more complete model of total analytical error will be discussed. An alternative method to estimate total analytical error, referred to as the distribution-of-differences method and which does not require modeling at all, will be discussed and contrasted with the simple combination model. A goal allocation method that is commonly used in systems engineering will be reviewed, where it is a common task to allocate an overall system goal into a series of component goals. Finally, the role of outliers will be considered.

Additional Error Sources Can Help Establish a More Complete Model for Total Analytical Error

The bias represented in Eq. 1 represents average bias at a particular concentration and is commonly estimated from a method comparison experiment and represented by a regression equation. Imprecision is the random error in the method under evaluation and is estimated by measuring replicate samples from the same patient.

Three additional error source types are discussed as examples of error sources that are neglected, improperly handled in Eq. 1, or difficult to incorporate. These additional error sources can affect the estimates of imprecision and average bias as well as total analytical error.

random biases attributable to interferences in patient samples: a neglected error source

A patient sample that is being analyzed contains not only the analyte of interest, but also a unique mixture of thousands of other chemical substances. If assays were completely specific, the presence of these additional substances would be of no consequence. However, most assays, including immunoassays *(6)*, suffer from some degree of nonspecificity. This means that each patient sample will possibly exhibit a bias unique to that patient's mixture of substances that exhibit nonspecific reactions in the assay. Examples of this bias, shown in Fig. 1, demonstrate that some patient samples that are assayed repeatedly and compared with a reference method will consistently produce values that are on one side of a regression line, whereas samples from a different patient specimen will fall on the opposite side *(7)*. In a method comparison experiment, this random bias will inflate the standard error of the estimate ($S_{y|x}$) and contribute to the average bias by influencing the regression coefficients.

Lawton et al. *(8)* represent this interference bias as a random error. Krouwer *(9)*, using actual data from a cholesterol evaluation, showed that failing to account for this error can underestimate total analytical error. The random interference bias attributable to nonspecific effects in patient samples is different from the random error attributable to repeatedly assaying the same patient sample. The combination of bias and imprecision in Eq. 1 does not account for the effect of random interference bias and will thus underestimate total analytical error unless random interference bias is zero. See *Appendix A* for a mathematical explanation.

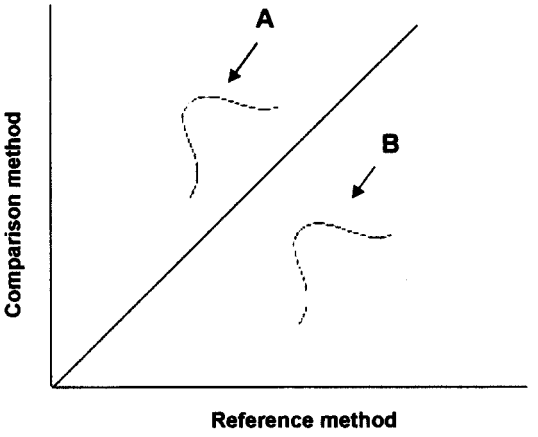

Fig. 1. Patient samples *A* and *B* have reproducible and different biases from the reference.
Each *curve* represents replicated samples from patient A or B.

effect of linear drift on imprecision and average bias: an example of an incorrectly handled error source

Linear drift, if present, is an example of another error source that is not correctly accounted for by Eq. 1. Consider a protocol for estimating imprecision in an assay that exhibits a positive linear drift. Krouwer *(10)* has shown, based on work by Haeckel and Schneider *(11)*, that the observed imprecision will actually be a combination of pure random error and bias, according to Eq. 2.

$$s_a = (s_p^2 + b^2)^{1/2} \quad (2)$$

where

s_a = the imprecision observed
s_p = the true random error (imprecision)
b = the average bias attributable to linear drift

The amount of bias observed will depend on the protocol. It will be smallest for a protocol that samples consecutive duplicate specimens and largest for a protocol that samples the first and last specimens in a calibration run. A common protocol of running 10 consecutive replicate specimens will exhibit an intermediate amount of bias.

Consider the effect of drift on bias estimation from a method comparison. The concept of average bias implies that for any sample assayed, the test result should be equal to random error plus the regression equation (i.e., bias is explained as a proportional plus a constant difference from the reference result). With drift present, this is not true. Samples assayed early in a calibration run will

have a reproducibly different average bias than samples assayed later in the run (see Eq. 3).

One can estimate linear drift from a suitable protocol using a multiple regression model such as Eq. 3.

$$Y = b_0 + b_1X + b_2t + e \quad (3)$$

where

Y = the observed result
b_0 = the estimated intercept coefficient (constant error)
b_1 = the estimated slope coefficient (proportional error)
b_2 = the estimated linear drift coefficient (linear drift error)
e = the estimated random error (pure random error)
X = the reference result
t = time of assay

Another interpretation of this example is that the model implied by Eq. 1 (e.g., Eq. 3 without the drift term) contains less knowledge about the true state of the process than Eq. 3. Goldschmidt and Krouwer *(12)* showed an example where the proportional bias was incorrectly estimated when a protocol was used that did not have a drift term in the model. An illustration, using a simple regression equation, of the different states of knowledge provided by random and systematic error is shown in Table 1.

TREATMENT OF OUTLIERS

In a method comparison experiment used to estimate average bias, it is standard and accepted practice *(13)* to remove outliers, should they occur. It makes sense to remove these outlier samples when assaying a small number of samples, otherwise the parameters estimated will not be representative. The problem is that there is no mechanism for these outliers to play any role in the simple combination method. They simply disappear from the analysis, although they will still be present in real life. When the distribution-of-differences method (below) is used, there is no basis for removing outliers, nor is there anything wrong (from an estimation sense) in a skewed distribution of differences.

A More Complete List of Total Analytical Error Sources

Additional error sources and the relationship among all of these error sources are shown in the cause and effect diagram in Fig. 2. In this diagram, each error source box is caused by all of the boxes connected below it. The darker boxes represent the estimation methods used in the simple combination model. Errors not listed in Fig. 2 may also be present and can be proposed based on knowledge of the assay technology. For a nearly perfect assay, all systematic effects will be negligible, and the assay will exhibit only imprecision. Each error source box in Fig. 2 is explained below. The first two error sources are the error sources in Eq. 1.

Table 1. Characteristics of systematic and random error.

	Systematic effect	Random effect
Error source is	Known (deterministic)	Unknown (probabilistic)
Y explained by[a]	$\beta_0 + \beta_1 X$	$+\varepsilon$
Y is	Known for each sample	Known for the average sample

[a] Assuming that the model is correct.

APPARENT RANDOM ERROR

Apparent random error is the imprecision estimated from protocols where replicates of a sample are assayed. If there are no systematic biases present, apparent random error and pure random error will be equal.

AVERAGE BIAS

Average bias is the method used by the simple combination method to estimate all systematic error. It estimates the slope (to convert the regression equation to a bias equation, 1 is subtracted from the slope) and intercept of a regression equation. The slope and intercept represent proportional and constant error, respectively. If there are other systematic errors present, the average bias will be incorrect. For example, if an assay is not linear at the upper end of the assay range, the slope and intercept of the regression equation will only partially express the average bias at the upper end of the range.

PURE RANDOM ERROR

Pure random error (not shown in Fig. 2) is the apparent random error with systematic error removed. This removal can be achieved by a suitable multifactor protocol *(9)* whereby pure random error is the residual error term from the model, as in Eq. 3. If systematic effects exist and are not removed, the apparent random error (e.g., that calculated without a multifactor protocol) will be greater than pure random error.

RANDOM ERROR

Random error refers to the collection of error sources whose effect is modeled as samples from a probability distribution. One can sometimes model the same error source as either random or systematic, such as discussed below for interferences.

SYSTEMATIC ERROR

Systematic error is the collection of error sources whose effect is modeled by an equation that describes the effect of the error sources for every sample.

PROTOCOL-INDEPENDENT BIAS

Protocol-independent bias refers to a collection of error sources that are largely independent of the protocol used to estimate them. Here, the protocol refers to every aspect of the assay, e.g., the sample order, reagent lot, and calibration sequence. Protocol-independent refers to the fact that the protocol usually is not a factor in the

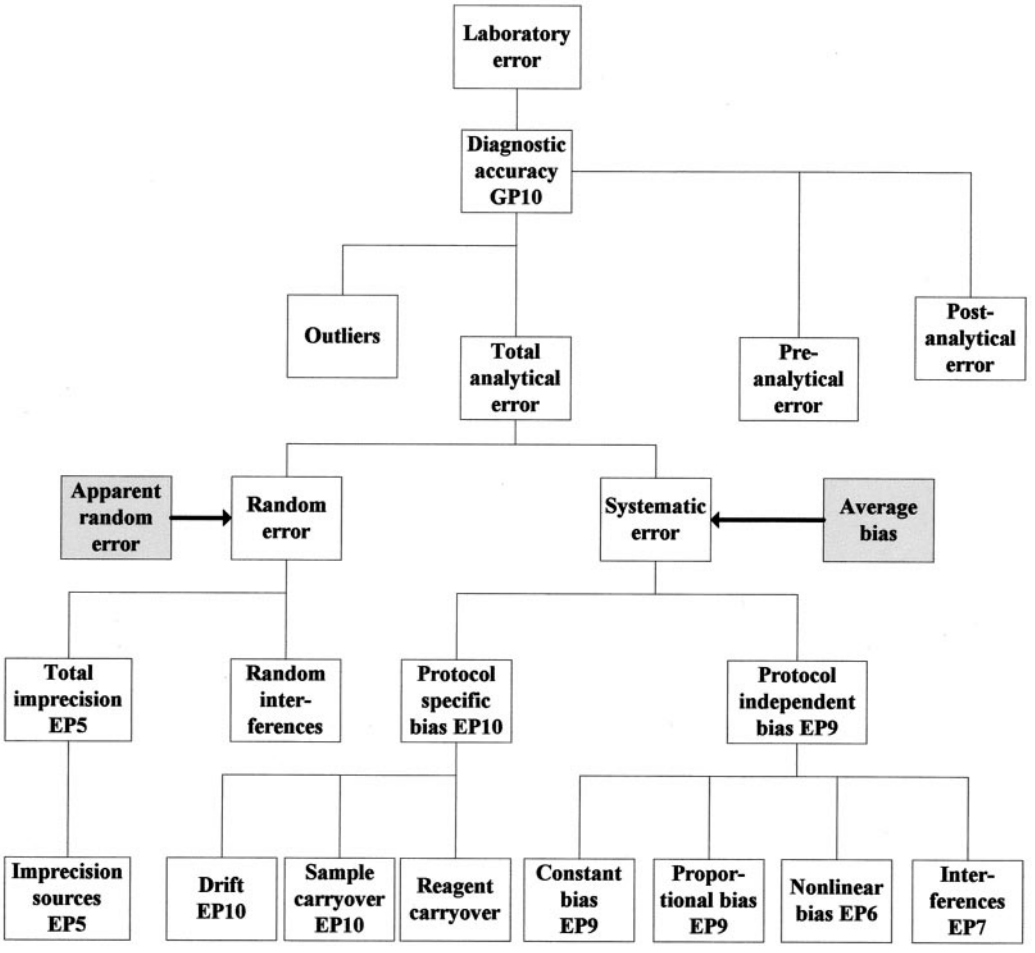

Fig. 2. Cause and effect diagram of laboratory error sources.

EP or *GP numbers* refer to NCCLS guidelines that provide methods to estimate those particular error sources. The *shaded boxes* are the error sources used in the simple combination method.

magnitude of the error source. For example, if an assay is inherently nonlinear and the nonlinearity is not corrected by software, then one can always expect this nonlinearity to be present. However, in some cases, nonlinearity may not be independent of the protocol. Nonlinearity can be caused by instability in a reagent, in which case the magnitude of the error source may depend on the reagent lot and its age.

PROTOCOL-DEPENDENT BIAS

Protocol-dependent bias refers to a collection of error sources that are largely dependent on the protocol used to estimate them. For example, linear drift depends not only on an instability in the assay response, but also on the sample order (e.g., the time of assay since the last calibration). Thus, the tenth sample assayed always has 10 times as much linear drift as the first sample assayed; hence the protocol is always involved in the bias equation. In addition, the extent of drift may vary from run to run. Usually the magnitude of the drift in a specific run cannot be predicted and is modeled by sampling the drift run magnitude from a suitable probability distribution. Thus, linear drift has a random as well as a systematic component. As an equation:

$$\text{Drift error} = \text{Drift run} \times \text{sample order} \qquad (4)$$

where drift run is the amount of drift effect present in a specific run and is a random effect (a random bias), and sample order gives the time since the last calibration that the sample was assayed and is a systematic effect (the protocol-dependent bias)

RANDOM INTERFERENCE BIAS

For each patient sample, a seemingly random bias component, additional to pure random error and caused by nonspecificity of the assay and the presence of interfering substances, may exist. For an assay with perfect specificity, the random interferences term would be zero. This error source can be estimated from a method comparison experiment *(8)*. One can test for the presence of random

interference bias by ANOVA by comparing $S_{y|x}$ to the imprecision estimated from replicates. The actual substance(s) causing the interference does not need to be known for a random interference term to be estimated.

SPECIFIC INTERFERENCE BIAS

Specific interference bias is error caused by nonspecificity of the assay attributable to the presence of a specific substance. This error is measured by interference experiments *(14)*. Manufacturers often test large numbers of potentially interfering substances. It is conceptually possible to estimate the effect of every possible interfering substance in an assay as well as to determine the concentration of each interfering substance in each patient sample assayed. Were this to be done, this error source would be completely deterministic. Because this is impractical (one cannot even be sure that one has thought of all possible interference candidates), interferences are also modeled as a random error source in estimations of total analytical error.

NONLINEAR BIAS

Nonlinear bias is bias that cannot be represented by a proportional relationship between the test and reference assay concentrations. This bias can be estimated from a method comparison experiment with higher order polynomial terms in the regression equation *(15)*. The high-dose hook effect in immunoassays is an example of nonlinear bias.

DRIFT

Drift is an error that is related to the time of assay since the last calibration. Drift may be linear or nonlinear and can be estimated by multifactor protocols or by protocols that specifically account for time of assay.

SAMPLE CARRYOVER

Sample carryover is an error attributable to the contamination of the current sample with the previous sample. Sample carryover errors are important only if the concentration differences of the two samples are reasonably large. Sample carryover may be estimated by multifactor protocols or by protocols that specifically account for the possibility of sample contamination, such as assaying a high-concentration sample followed by a series of low-concentration samples.

REAGENT CARRYOVER

Reagent carryover is an error in random access analyzers whereby the current assay is contaminated by reagent from the previous assay. Reagent carryover errors are important only when the contamination causes an effect, such as when an aspartate aminotransferase reagent precedes a lactate dehydrogenase reagent (lactate dehydrogenase is often part of the formulation of an aspartate aminotransferase reagent). Reagent carryover is estimated from protocols that take into account these combinations.

REAGENT/CALIBRATOR LOT EFFECTS

The presence of something that is different in a new calibrator or reagent compared with the previous calibrator or reagent can cause lot effects. For example, a calibrator with an erroneously assigned value will cause a bias in every value assayed with that calibrator lot. These error sources are often difficult to assess from protocols because sufficient different lots are often unavailable. In cases where there are enough samples, these error sources can be treated as random imprecision components. Manufacturers can assess these error sources for reagent lots from factorial studies in which different reagents lots are made with appropriate concentrations of reagent constituents to simulate manufacturing variances. Effects of calibration lot errors can often by estimated by mathematical simulation.

Accounting for All Terms in an Expanded Model

To estimate total analytical error by a more complete model, one must first decide which of the above error sources (or additional sources not in the above list) require estimation. Using a suitable protocol, one must not only estimate the magnitude of the error source, but also its distribution. This is required because the error source might not be constant. For example, drift might vary from one run to the next. Sampling a sufficient number of runs will provide an estimate of the distribution of drift coefficients.

Given the distribution of each error source, it is possible to create a simulation model (e.g., with software) that samples each error source from its distribution (which may not be a gaussian distribution) and combines all errors to arrive at the total analytical error *(16)*. To test the accuracy of the simulation, one can compare total analytical error estimated from the simulation with total analytical error estimated directly from a method comparison experiment.

Typically the detailed equation for this model will be quite complicated, with every possible effect having its own term, although in principle, the model will simply be an expansion of Eq. 1.

Alternative Method to Estimate Total Analytical Error: The Distribution-of-Differences Method

The distribution-of-differences method does not rely on a model at all, but simply estimates percentiles from the ordered distribution of differences collected from a method comparison experiment. The percentiles are estimated by parametric *(17)* or nonparametric methods *(18)*. Tolerance intervals can also be calculated *(19)*. The only requirement for this estimation is that samples assayed are representative; this requirement exists for the simple combination model as well.

Comparison of Approaches to Estimation of Total Analytical Error

DISTRIBUTION-OF-DIFFERENCES METHOD

Clearly, the main advantage of the distribution-of-differences method is its simplicity. There is no model to create or complicated calculations to perform. Note that the distribution-of-differences method does not preclude estimation of individual error sources, which many manufacturers or some laboratories may require. This is simply a separate activity and not required to estimate total analytical error.

The following example illustrates a benefit of the distribution-of-differences method compared with the simple combination model.

Consider a blood gas laboratory that is evaluating a lactate assay over a 2-week period. The laboratory has three analyzers; each analyzer receives a new electrode once a week and is calibrated every 30 min. This means that the variables instrument, electrode, and calibration are all potential error sources. The only way that the simple combination model can accommodate these error sources is to consider them as random error sources in an ANOVA model to estimate the imprecision term. For most laboratories, the correct formulation of the ANOVA model will be a challenging task. In the distribution-of-differences method, no ANOVA model is needed. For this or for any evaluation, one always simply computes all differences.

A disadvantage of the distribution-of-differences method is that the "differences" may not be solely attributable to the candidate method. Nevertheless, it is important to predict the outcome of switching an assay from the current assay (likely to not be a reference assay) to a new assay. Estimation of differences, whether they are attributable to the candidate or the comparison method, is nevertheless important because it is these differences that clinicians will observe.

The problem of determining which method is causing the difference is equally true for the simple combination method when the error source is attributable to bias. However, imprecision is treated differently in the two estimation methods. In the distribution-of-differences method, a difference is the bias between methods plus the imprecision of each method. Of course, laboratories are interested in knowing whether candidate methods are better; therefore, to ascribe as much error as possible to the comparison method, one should use a reference method to minimize bias in the comparison method and run replicate comparison method specimens to minimize imprecision in the comparison method. An ideal evaluation would be to run a three-way comparison consisting of the candidate, current, and reference methods.

Although the distribution-of-differences method does not provide an estimate of imprecision, laboratories will always evaluate separately the imprecision of a candidate assay to ensure that it will meet regulatory requirements.

FULL COMBINATION MODEL

The advantage of the full combination model is that, in addition to giving an estimate of total analytical error, it also provides detailed information about all error sources. The main disadvantage of this method is the large effort required both experimentally and with modeling to arrive at proper estimates.

SIMPLE COMBINATION MODEL

The main problem with the simple combination model, as described above, is that it often underestimates total analytical error. Moreover, because this method is also used to construct goals for total analytical error, these goals will be suspect as well. An example of this is the total analytical error goal suggested by the National Cholesterol Education Program, which uses the simple combination method *(20)*.

Detection vs Estimation

The three methods described above are for estimation of total analytical error and (except for the distribution-of-differences method) total analytical error sources. Given that an assay is in use, various quality-control strategies are useful in signaling results that are beyond prescribed quality limits attributable to various error sources *(21)*. These detection studies are used to optimize quality-control rules. Quality control monitors changes in assay performance as opposed to estimation, which is a one-time event or snapshot of assay performance obtained during an evaluation.

Note, however, that optimal quality control can never detect error attributable to random interference bias because quality-control samples contain the same matrix in every sample, unlike patients samples, which contain mixtures of different substances needed to detect random interference bias. This highlights the importance of determining the significance of random interference bias during a method evaluation and underscores the limitation of the simple combination method, which does not account for this error source.

ALLOCATING TOTAL ANALYTICAL ERROR GOALS

Assay performance goals allow evaluation results to be compared to a limit to determine whether an assay is acceptable. Goals are established by manufacturers (or laboratories) for several reasons:

- To satisfy regulatory requirements
- To meet commercial needs
- To meet medical needs

In addition, assays can be used in different ways, which may require different goals. For example, an assay that is used for diagnostic purposes is different from an assay that is used to monitor patients. In the latter case, serial measurements require that imprecision is the main parameter specified *(22)*. This section deals with medical need goals for assays that are used for diagnostic pur-

poses and assumes that a total analytical error goal has already been established.

Most assay performance goal setting in clinical chemistry has focused on setting goals for individual error sources *(2)*. Although most work has been devoted to goals for imprecision and bias, other error sources, such as reagent-to-reagent bias *(3)* and interference bias *(23)*, have been studied. These suggestions provide valuable insights into assay quality.

One limitation to the above goal-setting process, however, is that focusing on specifying a performance goal for an individual error source makes it is difficult to account for all other possible error sources, which is necessary to avoid specifying a performance goal for an individual error source that in practice causes the total analytical error goal to be exceeded. A solution to this problem is to create error source goals by allocation, using a systems engineering approach *(24)*.

Using reliability as an example, the systems engineering approach starts with the desired overall system reliability goal. One then estimates the reliability of each component from all subsystems and combines the individual estimates into an overall system reliability estimate. This is then compared with the goal. If the estimated reliability does not meet its goal, one must allocate the desired system reliability goal into goals for each component. Typically, the method used for this allocation is cost-based. The following example illustrates the systems engineering approach for an assay.

A Goal Allocation Example

Consider an assay with two error sources that are equal in magnitude for both imprecision and random interference bias. Assume that total analytical error was estimated both directly and by simulation and did not meet its goal. A manufacturer could allocate the total analytical error goal into goals for imprecision and random interference bias so that if these error source goals were met, the total analytical error goal would be met. Any combination of error reduction of components that leads to satisfying the total analytical error goal would work. To perform a cost-based allocation, a manufacturer would choose the least expensive design changes that would fulfill the total analytical error goal.

The above example could be further complicated by assuming that in addition to the above error sources, there were errors from five systematic biases (e.g., lot-to-lot reagent bias). It would almost always be a bad idea to allocate error equally among all of these error sources because dividing a total analytical error goal into seven equal parts would lead to each error goal being quite stringent. Moreover, the probability that all seven error sources occur simultaneously and that each at its maximum level would be extremely low; the pitfalls of such a "worst case" approach are shown in *Appendix B*. Thus, the allocation must take into account probability of occurrence.

Outliers Must Be Accounted for

Outlier results are those results that have unusually large deviations from an expected value. Outliers can cause medical errors because a large error in a laboratory result can cause an erroneous patient treatment decision *(25)*, but quantifying outlier rates is generally uncommon.

The use of a total analytical error goal does not solve the outlier issue, in spite of the word "total". The problem is that specifying total analytical error to mean that at least 95% of results are within an acceptable limit also means that up to 5% of results could be outside of this limit. Even with 99% limits, 1% of a large number of assay results is still a big number. Because laboratories can easily report 1 million results per year, if all results just met 99% acceptance limits, there would still be 10 000 results per year that were unacceptable according to total analytical error goals.

It would be naive to assume that a result just inside a total analytical error goal would be perfectly acceptable and that a result just outside this goal would cause a disaster. There is, rather, a continuum of quality. Thus, if all results outside the total analytical goal were nevertheless close to the goal, it is unlikely that these results would cause problems. This implies the use of another set of limits to define what "close to the goal" is. In addition to total analytical error limits, a wider set of limits could specify values that should never occur. Of course, one cannot test for the occurrence of "never"; however, if no outliers are found in a large sample size, one can guarantee that outlier rates can be no larger than a very small percentage.

Practically speaking, only manufacturers conduct studies of this magnitude. Although samples sizes are different for each assay, extremely large sample sizes (thousands) are common during product development, and the combination of results from all field trials often also produces a large sample size.

The most conservative way to estimate outlier frequency is to consider an outlier as a discrete event and use the binomial distribution *(26)*. It would be unwarranted to estimate potential outlier magnitudes and rates by simply calculating higher multiples from an estimated standard deviation. This is because there is no guarantee that an outlier comes from the same distribution that is used to calculate the standard deviation.

Typically, when a manufacturer finds a root cause for an outlier, either a design change for the assay is implemented, an algorithm is incorporated that prevents the result from being reported, or in some cases, and the least desirable, a caution is noted for the condition that could cause the outlier. Although development is a proprietary process, it would be helpful if manufacturers reported on the summary results of studies that estimate outlier rates.

CONCLUSIONS

Total analytical error is a useful metric for laboratory assay quality. The use of Eq. 1 to estimate total analytical

error is incorrect because it does not account for all potential error sources. Total analytical error can be estimated directly from a method comparison experiment. This estimate can be compared with a total analytical error goal. This simple approach can be used by both laboratories and manufacturers, with manufacturers using much larger sample sizes and sampling from all known potential error sources.

By studying the details of the assay process, one can enumerate various total analytical error sources. Different protocols are needed to estimate each total analytical error source. With knowledge of the distribution of error sources, a simulation model can be used to combine these sources to estimate total analytical error. The goal for total analytical error can be allocated into goals for each total analytical error source. Outlier rates must also be quantified.

I would like to acknowledge helpful discussions with several members of NCCLS subcommittee EP21 (Estimation of Total Analytical Error for Clinical Laboratory Methods).

References

1. Tonks D. A study of the accuracy and precision of clinical chemistry determinations in 170 Canadian laboratories. Clin Chem 1963;9:217–33.
2. Petersen PH, Fraser CG, Kallner A, Kenny D, eds. Strategies to set global analytical quality specifications in laboratory medicine. Scand J Clin Lab Invest 1999;59:475–585.
3. Klee GG, Schryver PG, Kisbeth RM. Analytic bias specifications based on the analysis of effects on performance of medical guidelines. Scand J Clin Lab Invest 1999;59:509–12.
4. Petersen PH, Stöckl D, Blaabjerg O, Pedersen B, Birkemose E, Thienpont L, et al. Graphical interpretation of analytical data from comparison of a field method with a reference method by use of difference plots. Clin Chem 1997;43:2039–46.
5. Petersen PH, Stöckl D, Westgard JO, Sandberg S, Linnet K, Thienpont L. Models for combining random and systematic errors. Assumptions and consequences for different models. Clin Chem Lab Med 2001;39:598–95.
6. Kricka LJ. Human anti-animal antibody interferences in immunological assays. Clin Chem 1999;45:942–56.
7. Kringle RO, Bogavich M. Statistical procedures. In: Burtis CA, Ashwood ER, eds. Tietz textbook of clinical chemistry, 3rd ed. Philadelphia: WB Saunders, 1999:265–309.
8. Lawton WH, Sylvester EA, Young-Ferraro BJ. Statistical comparison of multiple analytic procedures: application to clinical chemistry. Technometrics 1979;21:397–409.
9. Krouwer JS. Estimating total analytical error and its sources. Arch Pathol Lab Med 1992;116:726–31.
10. Krouwer JS. Multifactor protocol designs IV. How multifactor designs estimate the estimate of total error by accounting for protocol-specific biases. Clin Chem 1991;37:26–9.
11. Haeckel R, Schneider D. Detection of drift effects before calculating the standard deviation as a measure of analytical imprecision. J Clin Chem Clin Biochem 1983;21:491–7.
12. Goldschmidt HMJ, Krouwer JS. Multifactor experimental designs for evaluations. In: Haeckel R, ed. Evaluation methods in laboratory medicine. New York: VCH, 1993:71–86.
13. Kennedy JW, Carey RN, Coolen RB, Garber CC, Lee HT, Levine JB, et al. Method comparison and bias estimation using patient samples; approved guideline EP9-A. Wayne, PA: NCCLS, 1995.
14. Powers DM, Boyd JC, Glick MR, Miller WG. Interference testing in clinical chemistry; proposed guideline EP7-P. Wayne, PA: NCCLS, 1986.
15. Tholen DW, Kroll M, Krouwer JS, Lasky F, Happe TM, Caffo AL, et al. Evaluation of the linearity of quantitative analytical methods; proposed guideline EP6 P2, 2nd ed. Wayne, PA: NCCLS, 2001.
16. Aronsson T, de Verdier CH, Groth T. Factors influencing the quality of analytical methods—a systems analysis, with use of computer simulation. Clin Chem 1974;20:738–48.
17. Bland JM, Altman DG. Statistical agreement for assessing agreement between two methods for clinical measurement. Lancet 1986;1:307–10.
18. Krouwer JS, Monti KL A simple graphical method to evaluate laboratory assays. Eur J Clin Chem Biochem 1995;33:525–7.
19. Hahn GJ, Meeker WQ. Statistical intervals, a guide for practitioners. New York: Wiley, 1991:58,90.
20. National Heart, Lung, and Blood Institute. Recommendations regarding public screening for measuring blood cholesterol. NIH Publication 95-3045. Bethesda: National Heart, Lung, and Blood Institute, NIH, September 1995.
21. Westgard JO, Groth T, Aronsson T, Falk H, de Verdier CH. Performance characteristics of rules for internal quality control: probabilities for false rejection and error detection. Clin Chem 1977;23:1857–67.
22. Galen RS, Peters T Jr. Analytical goals and clinical relevance of laboratory procedures. In: Tietz N, ed. Textbook of clinical chemistry, 3rd ed. Philadelphia: WB Saunders, 1986:387–409.
23. Fuentes-Arderiu, Fraser C. Analytical goals for interference. Ann Clin Biochem 1991;28:393–5.
24. Blanchard BS, Fabrycky WJ. Systems engineering and analysis, 3rd ed. Prentice Hall international series in industrial and systems engineering. Upper Saddle River, NJ: Prentice Hall, 1998:729pp.
25. Sainato D. How labs can minimize the risks of false positive results. Clin Chem News 2001;27:1,6–8.
26. Hahn GJ, Meeker WQ. Statistical intervals, a guide for practitioners. New York: Wiley, 1991:103.
27. Mandel J. The statistical analysis of experimental data. New York: Dover, 1964:104–5.

Appendix A: Mathematical Relationship between Total Analytical Error and Imprecision (Reproducibility) and Patient Biases from Reference

Mandel *(27)* showed how a difference between a test and reference result can be described as a combination of random and systematic error (Eq. 1A).

$$TAE = (y - R) = (y - \mu) + (\mu - R) \qquad (1A)$$

Eq. 2A is an expansion of Eq. 1A to account for n replicates of each of m different specimens.

$$TAE = \sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij} - R_j) = \sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij} - \overline{\mu_i}) + \sum_{j=1}^{m}(\overline{\mu_j} - R_j) \qquad (2A)$$

where

TAE = total analytical error
y_{ij} = the ith observation from the jth sample of the new method
R_j = the reference method result for the jth sample
$\overline{\mu_i}$ = the mean of the jth sample of the new method

In Eq. 2A, the second double summation term is a measure of imprecision, and the last term represents the distribution of bias that is observed in each sample as seen in Fig. 1.

Appendix B: The Problem with the Worst Case Method of Allocating Goals

Consider an assay that has a total analytical error goal of ± 2 mg/dL and the only error sources comprise five independent, random biases, each with a normally distributed error source with zero mean and a SD of 0.2 mg/dL. A worst-case goal-setting method might work as follows. For each bias, the worst case might be designated as a 3 SD error = ± 0.6 mg/dL. Applied to all biases, this would equal an error of 5 × 0.6 = ± 3 mg/dL, which exceeds the goal of ± 2 mg/dL. Therefore, the SD goal for each bias would need to be reduced to 0.133 mg/dL because 0.133 × 3 = 0.4 and 0.4 × 5 = 2.0. This would also require that each random bias would have to be improved by 33%.

To see why this is a poor strategy, consider the likelihood of a result that occurs because each bias is ≥3 SD (each with the same sign) simultaneously. This is equal to $2 \times 0.003^5 = 4.86^{-11}\%$. On average, we would need to run >40 billion assays before seeing one such occurrence. Hence, allocation must take into account probability of occurrence.