Proficiency Evaluation of Clinical Chemistry Laboratories


The problem of proficiency evaluation of clinical chemistry laboratories is considered in terms of the accuracy, precision, and allowable limits of error of 10 common analytical procedures, and in terms of the proficiency of individual analysts and the occurrence of gross laboratory mistakes. The concept of the operational chart is developed, and its use is illustrated. Operational charts summarize basic quality-control information in a clear, concise, goal-oriented perspective which is useful for evaluating the performance of individual laboratories and for comparing groups of laboratories. The charts aid in identifying major problems relating to the reliability and clinical usefulness of laboratory results, and should facilitate managerial decision-making processes.

Additional Keyphrases allowable laboratory error • analytical accuracy, precision • health care and laboratory performance • homeostatic limits • individual variability • laboratory mistakes • operational chart • operational line • proficiency evaluation • quality control • reliability of clinical laboratories

Clinical chemistry increasingly is being extended from its traditional role in the evaluation of hospitalized or acutely ill patients to a health surveillance function (1–5). Multiphasic screening is now an accepted approach to the identification of atypical individuals within a population (4), and prospective biochemical profiling is a conceptually valid approach to the early detection of subtle deviations from an individual's established norm (5, 6). However, the realization of the full health benefits of these modern applications of clinical laboratory science requires excellence in laboratory performance with a high degree of interlaboratory comparability of results. Hence, the question of the long-term reliability of laboratory results is becoming of critical importance (7).

Several years ago a quality-control laboratory was established at Ohio State University Hospital as a subdivision of the Clinical Chemistry Laboratory. In practice, this laboratory functions as an external Proficiency Laboratory, and has provided the opportunity to monitor the performance of the Clinical Chemistry Laboratory and to obtain objective evidence of its reliability. In this report we will present some fundamental principles of laboratory proficiency evaluation, and illustrate some uses of the evaluation program in identifying analytical problems, in comparing the performance of the Chemistry Laboratory with that of other laboratories, and in detecting laboratory mistakes.

The purpose of this report is to emphasize both the need and the broad opportunity for improved performance by clinical chemistry laboratories.

Methods

Brief Description of the Laboratory Evaluation Program

The evaluation program is based primarily on the reference sample procedure introduced by Levey and Jennings (8) and further described by Copeland (9) and Sax et al. (10). Commercial serum reference specimens are prepared daily in the Proficiency Laboratory, assigned fictitious patient names and numbers, and introduced as blanks at random times into the mainstream of clinical specimens flowing through the Clinical Chemistry Laboratory. About 30 analytical procedures are monitored routinely at two or more concentrations.

The determined values and analyst identification numbers are recorded daily, results exceeding established limits are noted, and the causes of excessive deviations are investigated promptly. The data for each reference specimen are analyzed monthly, quarterly, semi-annually, and annually in terms of the mean, standard deviation, and relative standard deviation (coefficient of variation). The results are interpreted in various ways as described below.

From the Division of Clinical Chemistry, Department of Pathology, The Ohio State University, 410 W. 10th Ave., Columbus, Ohio 43210.

1 Postdoctoral Fellow in Clinical Chemistry. Present address: Biomedical Services Inc., 8301 Arlington Blvd., Fairfax, Va. 22030.

Received June 28, 1971; accepted Nov. 22, 1971.
In addition to this external evaluation program with blind reference specimens, the analysts in the chemistry laboratory include known commercial reference specimens and laboratory pool with each set of clinical specimens analyzed. Because of the extensive use of these internal controls, the external program has been limited to about 2 to 3 specimens daily per analytical procedure. In a previous publication the use of blind (external) and known (internal) reference specimens were compared and appreciable analytical bias were demonstrated with the internal controls (11). The present report is concerned only with external evaluation by the Proficiency Laboratory.

Principles of Laboratory Evaluation: A Conceptual Framework for the Interpretation of Results

The “operational line” and evaluation of accuracy. The objective in any routine analytical procedure is to obtain consistently accurate and reproducible results (18), and the performance of a laboratory should be evaluated initially in terms of the accuracy and precision of its determinations. The role of accuracy as a primary component of laboratory proficiency is illustrated in Figure 1, in which the determined values of reference specimens are graphed against the control values. If the laboratory were absolutely accurate, and if the control values were “true” values, then the determined values should correspond exactly with the control values, and the analytical results should fall on line 1. Certain kinds of systematic errors, such as faulty standardization, or use of an inappropriate factor in calculations, should affect the slope of the line, illustrated by line 2. Other kinds of errors, such as an inadequate blank correction, might yield results along line 3. Still other errors, such as those caused by deteriorating reagents or faulty operation of “AutoAnalyzer” components, might cause results to fall along line 4.

Every laboratory, in every analytical procedure, does in fact operate along some line such as illustrated in Figure 1. Ideally, every laboratory, in every procedure, should operate along the line of slope of 1.0. A laboratory evaluation program should determine the operational line for each procedure in order to evaluate accuracy and to facilitate the maintenance of an operational line of slope 1.0.

The operational line can be determined by preparing proportionate mixtures of high and low reference specimens as suggested in Table 1, graphing the results of analyses of these solutions against the known values, as illustrated in Figure 2, and drawing the best line through the determined values. (This dilution technique establishes a known linear relationship between the reference specimens, and the analytical results should verify this relationship, as well as the actual concentrations of analyte in the reference specimens). Alternatively, high and low concentration reference specimens could be analyzed repeatedly and the mean values calculated. The line connecting the mean values, on a graph of determined vs. control values, should then approximate the laboratory’s mean operational line.

---

**Table 1. Preparation of Proportionate Mixtures of High and Low Reference Specimens**

<table>
<thead>
<tr>
<th>Tube no.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low specimen A</td>
<td>5.0</td>
<td>3.0</td>
<td>2.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>High specimen B</td>
<td>0.0</td>
<td>1.0</td>
<td>2.0</td>
<td>3.0</td>
<td>4.0</td>
<td>5.0</td>
</tr>
<tr>
<td>Total volume</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
</tr>
</tbody>
</table>

Percent

| Low specimen A | 100 | 80 | 60 | 40 | 20 | 0 |
| High specimen B | 0   | 20 | 40 | 60 | 80 | 100 |

---

**Fig. 1. Illustration of the concept of the laboratory operational line**

Ideally, each laboratory should operate along a line of slope 1.0 (line 1), but systematic proportional (line 2), absolute (line 3), or combined (line 4) errors may cause the actual operational line to differ from the ideal.

**Fig. 2. Determination of the operational line by using proportionate mixtures of low (A) and high (B) reference specimens**
The operational line can be characterized in terms of its intercept, "a," on the ordinate, and its slope. For example, the characteristics of the lines of Figure 1 are summarized in Table 2. As implied above, these characteristics of the operational line should provide important clues to the possible nature of the principal errors affecting a procedure, and expedite their correction.

**Deviation about the operational line and evaluation of precision.** In all analytical procedures there are two principal classes of errors: *absolute errors*, and proportional or *percentage errors*. In addition, errors may occur systematically or randomly. *Systematic* errors affect accuracy. For example, in Figure 1, lines 2, 3, and 4 illustrate the effect on accuracy of, respectively, a systematic percentage error, a systematic absolute error, and combined systematic percentage and absolute errors. *Random* errors affect precision. Random percentage errors such as those caused by variations in the accuracy of measuring samples, or run-to-run variation in a standard curve or color factor, should affect the slope of the operational line and should cause results to be distributed within certain limits as illustrated by the dotted lines of Figure 3; i.e., the magnitude of the error increases as the concentration increases. With this type of error, the standard deviation (measured in concentration units) increases as the concentration increases, but the coefficient of variation (measured as a percentage) tends to be relatively constant.

Random absolute errors, such as those caused by *variations* in a blank correction or variable turbidity of solutions, should tend to be constant at all concentrations of analyte, hence should affect the intercept, "a," of the operational line, and cause results to be distributed within certain limits, as illustrated by the dashed lines of Figure 3. With such errors the standard deviation of reference specimens should be constant at all concentrations, while the coefficient of variation should decrease as concentration increases.

Most analytic methods are subject to both kinds of random error, but in a given method one kind may predominate. In evaluating laboratory performance and identifying sources of error, it should be helpful to know how precision normally varies with concentration in each procedure, and whether the normal relationship has changed.

Every laboratory, in every procedure, does operate within certain limits of precision. Proficiency evaluation should determine these *operational limits* for each procedure, to identify errors as they develop, and thus facilitate the maintenance of precision within specified limits. Precision is evaluated by repeated random analyses of reference specimens at high and low concentrations and calculation of the standard deviation and coefficient of variation. From such calculations the limits of variability about the operational line can be drawn (Figure 4).

**The limits of allowable laboratory error.** The evaluation of the accuracy and precision of a laboratory is meaningful only in relation to established limits of acceptable error. The maximum limits of total analytical error can be specified on both theoretical and practical grounds. Theoretically, the rationales of clinical chemistry stem largely from the concept of homeostasis, according to which physiologic

---

**Table 2. Characteristics of Operational Lines of Figure 1**

<table>
<thead>
<tr>
<th>Line no.</th>
<th>Ordinate intercept</th>
<th>Slope</th>
<th>Slope error(^*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1.00</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1.37</td>
<td>+37%</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>1.00</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>a'</td>
<td>0.75</td>
<td>-25%</td>
</tr>
</tbody>
</table>

* Slope error (%) = (slope of operational line - 1.00) \(\times\) 100.

---

**Fig. 3.** Illustration of the concept of laboratory precision

Various kinds of random errors cause results to be distributed about the operational line within absolute (dashed lines) or proportional (dotted lines) limits

**Fig. 4.** Estimation of the operational line and limits of variability

Reference specimens are analyzed repeatedly by the laboratory and the mean values \(\pm 2 \text{ SD}\) calculated. The line through the mean values is the apparent operational line, and the lines through \(\pm 2 \text{ SD}\) limits indicate the limits of analytic variability.

---

224 CLINICAL CHEMISTRY, Vol. 18, No. 3, 1972
parameters are normally maintained relatively constant, under standard conditions, by various complex regulatory mechanisms (13, 14). The concept of biochemical homeostasis is expressed symbolically as:

\[ \sum \frac{\partial A}{\partial t} = 0 \]  

(1)

That is, the changes in concentration of some parameter, \(A\), over a period of time, \(t\), normally tend, within certain limits, to average zero (Figure 5). The limits of variability of \(A\) are fixed by the nature of the physiologic system and the adequacy of the regulatory mechanisms affecting the concentration of \(A\), and are characteristic of the standard physiologic steady state.

If changes from an individual's established norm are to be detected, analytical error must be no greater than the individual's normal homeostatic limits. The normal limits of individual variability for a few substances have occasionally been reported (15–17) and recently these limits were carefully determined for 15 parameters (18). For illustrative purposes for this report, we determined the variability of one apparently healthy individual by analyzing eight specimens drawn weekly over a period of two months. Each specimen was analyzed on the day it was drawn and then stored frozen. Later, to eliminate the effect of day-to-day and run-to-run analytic variability, all specimens were analyzed together in a single run. No significant deterioration was noted, and a range of values similar to that reported by Cotlove et al. (18) was observed. (The limits thus obtained by us are shown in Figure 8 (a–g)).

The limits of allowable error have also been defined in practical terms: The normal range for many parameters is defined as \(\pm 2 \, \text{SD}\) of the mean of values observed in an apparently healthy population. If it is assumed that the variability of a healthy individual is about \(\pm 1 \, \text{SD}\) of the normal population range, then laboratory error, defined as \(\pm 2 \, \text{SD}\) of the range of analytical values of a reference specimen having the normal mean concentration, should be less than these limits. That is to say,

\[
\text{laboratory error (±2 SD, analytical range)} \leq \text{individual variation (±1 SD, normal population range)}
\]  

(2)

This relationship is illustrated in Figure 6. From Equation 2 it follows that the laboratory error should be less than \(\pm 1/4\) the normal population range and, expressed as percent of the normal mean value:

\[
\text{acceptable laboratory error} \leq \pm \frac{1/4 \, \text{normal range}}{\text{normal mean}} \cdot 100
\]  

(3)

Equation 3 is Tonks' formula, which has been used to establish arbitrary limits of acceptable performance in laboratory surveys (19). In specific applications the formula requires modifications. For example, the physiologic control of electrolyte concentrations is normally very close, and it is difficult to maintain, over an extended period of time, a laboratory error less than \(\pm 1/4\) the normal range. However, this degree of precision is probably not necessary to detect clinically significant changes from normal. At the other extreme, the normal range for cholesterol is quite large and an analytical error that is less than \(\pm 1/4\) the normal range can easily be maintained.

Despite these limitations, the definition of acceptable error in terms of the normal range provides a clinically useful guideline for interpreting results of many procedures: if the laboratory consistently performs within these limits it is capable of discriminating between low and high values within the normal range (Figure 6), and should be capable of detecting significant changes from an individual's established norm.
**Detection of laboratory mistakes.** Detection of laboratory mistakes is an important but neglected aspect of Proficiency Evaluation. In a system as complex as a busy clinical chemistry laboratory there are many opportunities for mistakes to be made. Factors such as design of the laboratory, managerial decisions concerning the quality of personnel employed, the analytical and specimen-handling procedures used, the training and assignment of clerical and technical personnel—all these could influence the frequency with which mistakes are made, and the overall quality of performance of the clinical laboratory.

Mistakes usually occur when laboratory personnel deviate from established rules of procedure. For the present purpose, a mistake is defined as an instance in which the reporting of a clearly erroneous analytical result was found to be due to such human failure, rather than to a random analytical error. By systematically documenting mistakes as they occur and identifying their causes, steps may be taken to minimize or prevent their recurrence.

The structure of a clinical chemistry laboratory as a specimen-processing system is illustrated in Figure 7. In the proficiency evaluation program, reference specimens are randomly interspersed with clinical specimens. When the reported value for a reference specimen is beyond established limits, or otherwise questionable, the cause of the aberrant result is investigated and a report filed. These reports are later analyzed and an estimate is made of the principal causes and frequencies of mistakes \( f_n \), Figure 7) in various sections of the laboratory.

**Analytical Methods**

The methods used in the procedures discussed in this report are listed in Table 3. In survey comparisons of our laboratory with other laboratories, only those laboratories that use similar methodologies have been included.

**Results and Discussion**

**Typical Operational Charts**

The performance of the chemistry laboratory in 10 common determinations is summarized in Figure 8 \( a-j \). In each chart the laboratory's determined values are graphed against the reference or control values. The control values used are those stated by the manufacturer of the reference specimens. For the 10 procedures considered, the reference specimens were "Versatol," "Versatol A," and "Versatol A-Alternate" and, except for total protein the control values are based on a weighed-in amount of analyte. The ideal operational line of slope 1.00 is shown on each chart (solid line). The observed range of analytical variability (i.e., \( \pm 2 \sigma \) of blind reference specimens) over a six-month period is indicated by the dashed lines. The actual mean operational line of the laboratory is not shown (because it would tend to obscure other details of the charts), but it lies midway between the dashed limit lines. Also shown is the normal population range \( (N) \), the apparent homoeostatic limits of an individual \( (I) \), and the limits of acceptable performance given in various interlaboratory surveys. The control values used in plotting the survey data were the mean values of the referee laboratories used in each survey.
Fig. 8. (a-j). (In each case, abscissa is control value, ordinate the determined value). Operational charts of the Ohio State University Hospital Clinical Chemistry Laboratory for 10 common determinations.

Each chart shows the theoretical operational line of slope 1.0 (solid line), the range of analytical variation over a six-month period (dashed lines), the normal population range (N), the homeostatic limits of a normal individual (I), and the range of 'acceptable performance' reported in various laboratory surveys (C: College of American PATHOLOGISTS, O: Ohio Department of Health, W: Wisconsin Department of Health). See text for further details.
The Accuracy of Analytical Procedures

Sodium and chloride determinations (Figures 8a and 8b). The determinations of sodium and chloride are analytically exacting, because of the relatively high concentrations in serum and the narrow normal ranges. Nevertheless, a mean operational line of intercept 0 and slope 1.00 is maintained for the chloride procedure, which indicates that there is good agreement between the aqueous standards used by the Chemistry Laboratory and the stated values of the weighed-in, serum-based reference specimens used by the Proficiency Laboratory.

Sodium values determined by the chemistry laboratory are slightly lower than the weighed-in values, and the mean operational line has an intercept of 0 and a slope of 0.993. Although this discrepancy is observed consistently, the difference is small, and considered a minor problem.

Potassium determination (Figure 8c). There are constant discrepancies between potassium values reported by the Chemistry Laboratory and the stated values of weighed-in reference specimens: high potassium values determined by the laboratory are lower than the reference values, while low potassium values are higher than the reference values. These differences result in an apparent intercept error as well as a slope error. These errors are small and tend to cancel one another within the concentration range of most clinical determinations, so their possible net effect on the quality of results is small.

We believe the apparent error in the potassium determination to be related to interference by sodium in the flame photometric procedure, and to the use of standard solutions that have different proportions of sodium and potassium than the reference specimens (Dr. Vir D. Anand, personal communication). This problem will be discussed in detail in a later report.

Urea-N, creatinine, uric acid, and glucose (Figures 8 d-g). These four determinations are performed on a four-channel AutoAnalyzer (Technicon Corp., Tarrytown, N. Y. 10591). The urea-nitrogen and uric acid procedures appear to be quite accurate (i.e., they have mean operational lines of slope 1.00). However, creatinine values determined by the laboratory are consistently about 3% higher than the weighed-in values of reference specimens, while glucose values are about 4% lower.

The reasons for these differences are unknown. The chemistry laboratory is standardized with commercially available aqueous standards, while the serum-based reference specimens have a weighed-in amount of analyte. We have analyzed certified ASCP standard glucose solutions in parallel with reference specimens, and the results for laboratory standards agree with those for the certified standards, while the reference specimens have the characteristic lower values. The failure of the AutoAnalyzer glucose method to recover the weighed-in amount of glucose is also evident in laboratory surveys. For example, in a recent Ohio Department of Health survey involving six referee laboratories and about 50 participating laboratories using the AutoAnalyzer method, the mean determined values were 3% less than the weighed-in values. Thus, there is evidence from many laboratories that the values based on a weighed-in amount of glucose in reference specimens may be inconsistent with values actually determined in clinical laboratories.

Total protein procedure (Figure 8h). Total protein determination by the biuret procedure is technically simple and analytically sound, but presents several problems. As illustrated in Figure 8h, the mean operational line has a small negative intercept, and a slope of about 1.08 (i.e., the determined values are about 8% higher than the stated values of reference specimens). The normal range determined in our laboratory (and presumably used in clinical interpretations of results) is also higher than the values usually cited in the literature (Table 4).

The high determined values are possibly related to the fact that our total protein determination is standardized with a bovine serum albumin preparation (29, 30), while the serum-based reference

<p>| Table 4. Some Normal Ranges Reported for Total Serum Protein |
|-------------------|------------------|-------------------|-------------------|</p>
<table>
<thead>
<tr>
<th>Range, g/100 ml</th>
<th>Reference no.</th>
<th>Range, g/100 ml</th>
<th>Reference no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.8-8.8</td>
<td>our values</td>
<td>6.1-7.9</td>
<td>(40)</td>
</tr>
<tr>
<td>6.6-8.4</td>
<td>(38)</td>
<td>6.0-7.8</td>
<td>(36)</td>
</tr>
<tr>
<td>6.0-8.6</td>
<td>(34)</td>
<td>6.2-7.7</td>
<td>(41)</td>
</tr>
<tr>
<td>6.2-8.5</td>
<td>(35)</td>
<td>5.7-7.7</td>
<td>(41)</td>
</tr>
<tr>
<td>6.0-8.3</td>
<td>(36)</td>
<td>6.5-7.6</td>
<td>(43)</td>
</tr>
<tr>
<td>6.0-8.2</td>
<td>(37)</td>
<td>6.5-7.5</td>
<td>(44)</td>
</tr>
<tr>
<td>6.5-8.0</td>
<td>(38)</td>
<td>6.0-6.9</td>
<td>(45)</td>
</tr>
<tr>
<td>6.0-8.0</td>
<td>(39)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Plasma protein, corrected for fibrinogen.
specimens contain a heterogeneous mixture of human serum proteins. The protein contents of both the standards and reference specimens are calculated by multiplying the protein nitrogen content by 6.25; i.e., it is assumed that these proteins have exactly 16.0% nitrogen. However, a factor of 6.54 may provide a better estimate of total protein in normal, but not dysproteinemic serum (31, 32), while a factor of 6.37 may yield a more accurate estimate of bovine serum albumin (32).

The results of a comparison of bovine albumin standard and a variety of human serum reference specimens is given in Figure 9. Dilutions of the reference specimens and of a standard bovine albumin solution obtained from the American Society of Clinical Pathologists were analyzed by the biuret procedure, and the determined values were graphed against the stated values of the various reference specimens (Figure 9). The determined values of the ASCP bovine albumin fell on the operational line of slope 1.00, indicating an excellent correlation with the laboratory standard. However, all determined values of the human serum reference specimens fell on a line having a slope of 1.08. Thus, the stated protein content of a human serum-based reference specimen, which is calculated from the protein nitrogen content, is apparently inappropriate when the total protein procedure is standardized with bovine serum albumin.4

The total protein determination by the biuret reaction is an example of a relatively simple analytical procedure that can be performed quite reproducibly in any one laboratory. However, if groups of laboratories are to obtain comparable results there must be some agreement to use comparable standards, standardization technique, and analytical procedure.

Cholesterol procedure (Figure 8j). There is a major discrepancy between values for cholesterol as determined by the clinical laboratory and the manufacturer's stated values for reference specimens (“Serachol,” “Versatol A-Alternate,” and a 1:1 mixture of these preparations). The discrepancy appears as an apparent intercept error of 11 mg/100 ml and a slope error of +4%. The laboratory uses standards that have been checked on many occasions with National Bureau of Standards cholesterol and found to be accurate. The apparent error is the result of our use of a method that is subject to nonspecific interference by bilirubin; the ideal operational line is attained with reference specimens having a low bilirubin concentration. Modification of the cholesterol procedure to eliminate interference by bilirubin is under investigation.

Bilirubin (Figure 8j). The bilirubin procedure appears to have a slope error of about −7%, if the weighed-in values of reference specimens are assumed to be correct. Most of this difference could reasonably be attributed to differences in standard bilirubin preparations, because the molar absorptivity of such preparations may vary by 5% (46). However, in this case the difference is actually due to a rigorous standardization procedure that did not take into account the deterioration of bilirubin during the routine processing of specimens.

The bilirubin procedure is standardized by the procedure recommended by the Joint Bilirubin Committee (47). Standard solutions were prepared in the dark and were protected from light throughout the critical stages of the analytical procedure. From the results of analyses of these solutions a standard table relating the absorbance of azobilirubin to serum bilirubin concentration was prepared.

Figure 8j shows that blind reference specimens, processed through the chemistry laboratory in the same manner as clinical specimens, consistently have low determined values. However, when the same reference specimens are analyzed immediately after preparation, the stated values and determined values correspond very well. It is well known that bilirubin is light-sensitive and deteriorates quite rapidly when exposed to direct

---

4 In addition to the problem of standardization, a minor systematic error is present: when the standard table relating biuret absorbancy to protein concentration was prepared, the standard albumin solution was diluted with sodium chloride, while in the analysis of clinical specimens samples were diluted with a sodium sulfate precipitant solution. In conjunction with the study summarized in Figure 9, the ASCP albumin solutions were also diluted with the sulfate solution, and the results fell along the dotted line (Figure 9). Apparently this small difference in procedure introduced a systematic absolute error of about 0.1 g/100 ml.
sunlight, and at a significant rate under ordinary conditions of laboratory lighting (48). Although the Chemistry Laboratory takes specific precautions to minimize bilirubin deterioration in clinical specimens, apparently a 7% loss still occurs.

**Nationwide Comparisons of Laboratories**

_Determination of the operational lines of laboratories._ In interlaboratory surveys, referee laboratories are sometimes asked to analyze several specimens covering a range of concentrations. From the results of these analyses the laboratories’ operational lines may be compared. For example, in a Wisconsin State Department of Public Health survey seven laboratories situated in various parts of the United States served as referees for the uric acid determination by the AutoAnalyzer method. Results obtained by four of these laboratories are shown in Figure 10a, in which the determined values are plotted against the mean values of all laboratories. The best line through each set of determined values is the apparent operational line of the laboratory. Evidently, Laboratory No. 1 was operating with a small absolute or intercept error (in comparison to the other laboratories), No. 2 was operating along the mean operational line of all laboratories, and No. 3 had a small slope error. The data from Laboratory No. 4 do not fall on a straight line, and the best line through the points suggests an appreciable slope error. The remaining three laboratories had operational lines intermediate between those of laboratories No. 1 and 3. Because of the uncertainty concerning the data from Laboratory No. 4, all of its results probably should have been excluded from the calculation of the mean reference values.

The differences in operational lines in Figure 10a are very similar to the day-to-day variations observed in our chemistry laboratory. For example, in Figure 10b lines 1 and 2, respectively, are representative “high” and “low” operational lines observed in our laboratory during a one-month period. Line 3 was obtained on a busy day when work pressures delayed the regular servicing of the AutoAnalyzer. When Figures 10a and 10b are compared, it is evident that laboratories that are widely separated geographically can be standardized to a common mean operational line, but the solution to the problem of consistently obtaining the same operational line in all laboratories depends primarily on learning how to stabilize the line within a single laboratory.

The results shown in Figures 10a and b demonstrate that it is practicable to determine the operational lines of laboratories. This determination would be especially worthwhile to facilitate the selection of referee laboratories for surveys, and to standardize laboratories that help to establish assay values of commercial reference specimens.

_Survey comparisons of laboratories._ In surveys such as the College of American Pathologists’ “Quality Evaluation Program,” specimens are sent to participating laboratories several times a year. The results are analyzed statistically by the College in terms of the mean value and “limits of acceptable performance.” The limits are based primarily on the ±2 sd of the reported results, after outlying values have been excluded. The Wisconsin and Ohio Departments of Health surveys are conducted and analyzed in a similar manner.

In these kinds of surveys, each reported result presumably furnishes a point on the laboratory’s operational line. The mean value then represents a point on the mean operational line of all laboratories, and the limits represent the range of variation of the operational lines.

The limits found in various recent surveys are shown in Figures 8a–j. In the automated determinations of sodium, chloride, potassium, uric acid, and glucose, the participating laboratories (i.e., 400 to 700 laboratories in the CAP surveys) have similar operational lines, and the range of variation of the line is only slightly greater than the range in referee laboratories, or in our laboratory over a six-month period. Apparently, further
improvement in these determinations mainly depends on a decrease in the variability of the operational line in each laboratory.

In the automated urea nitrogen determination, a similar degree of proficiency is observed in the concentration range of greatest clinical interest, although the variability is greater at higher concentrations.

In contrast to these results, the range of values observed in the automated creatinine determination (Figure 8e) is quite large. Laboratories apparently have difficulty in standardizing on a common operational line and, before performance will be improved, the reasons for this difficulty must be found.

The data for total protein, cholesterol, and bilirubin (Figures 8 h-j) are for manual procedures. In each of these procedures the survey limits are two or more times the limits observed in our laboratory over a six-month period. The wide range of values observed in the bilirubin determination (Figure 8j) is especially noteworthy. For example, it is not clear from data supplied to participants in the CAP survey whether the specimen having a mean value of 4.5 mg/100 ml actually had a concentration of about 6.2 mg/100 ml and the wide range of values is due to different degrees of specimen deterioration, or whether the actual value was about 4.5 mg/100 ml and the wide range is due to differences in the standardization of the laboratories. As we and other referee laboratories found about 4.5 mg/100 ml, the latter possibility is the more likely. Thus, the survey limits represent a variation in the slopes of the operational lines of ±38%. This range of variation cannot reasonably be considered “acceptable performance” in a quantitative analytical procedure. Improvement of laboratory performance apparently is dependent on developing a standardization technique that can be successfully applied in all laboratories.

Accuracy and precision of manual procedures also depend on the proficiency of the individuals performing the analyses. The reliability of results is related to the analyst’s recognition of which procedural steps critically affect the determination, and his skill in performing each step. Analysts in our chemistry laboratory rotate from station to station each week or two. Thus, over a period of several months, a dozen or so analysts may have been responsible for a single procedure. Figure 11 compares the mean values and standard deviations observed during a six-month period, during which each of 12 analysts performed the total protein determination 10 or more times. There are obvious differences in the quality of results reported by these individuals. Technologists No. 1 and 7 appear to have been well standardized and obtained highly reproducible results. Technologist No. 3 obtained fairly reproducible, but elevated results. Technologists 10, 11, and 12 obtained highly variable results, although they did maintain the laboratory’s mean value.

Thus, a laboratory’s performance in manual procedures such as the total protein, cholesterol, and bilirubin determinations depends not only on correct standardization, but also on the proficiency of the analysts performing the tests. Improved performance probably involves careful instruction of the analysts in the subtleties affecting each determination.

Analysis of Laboratory Mistakes

During 1969 and 1970 a total of about 5200 external (blind) reference specimens were processed by the Clinical Chemistry Laboratory. Several tests were requested on each specimen, and 242 “incorrect” analytical values were detected. (An incorrect value is a result that was reported to the Proficiency Laboratory, appeared questionable, and was found by subsequent investigation to have been erroneous because of some mistake, rather than random analytical error). The 242 incorrect values were due to 191 mistakes (i.e., some mistakes affected several reference specimens). Eleven of these mistakes occurred in the Proficiency Laboratory and 180 in the Chemistry Laboratory. Thus, mistakes affecting one or more clinical specimens apparently occurred in the Chemistry Laboratory at the rate of 3.46 mistakes per 100 specimens (180/5200 × 100 = 3.46). This figure is probably an underestimate, since it is unlikely that all mistakes were noted. However, it should also be noted that the frequency of mistakes per test would be much lower as several tests were performed on each specimen.

Some kinds of mistakes that occurred are listed in Table 5, and their frequencies of occurrence are given in Table 6. These mistakes have varying
Table 5. Types of Laboratory Mistakes

(1) Specimen mix-up:
(a) Specimens labeled with wrong accession numbers in the clerical area.
(b) Sera transferred to mislabeled tubes in the specimen preparation area.
(c) When a specimen was removed from the AutoAnalyzer sampler wheel to insert an emergency specimen, an improper cup number was recorded and all specimens on the wheel were assigned false values.
(d) Analytical tubes interchanged during pipetting of specimens, or by misplacing cuvets in the Gilford spectrophotometer 4-position cuvet holders.

(2) Incorrect chart readings:
(a) Incorrect reading of AutoAnalyzer peak.
(b) Incorrect read-off from standard curve.
(c) Read-off from standard curve assigned to wrong specimen.
(d) Read-off from wrong standard curve (e.g., reading LDH values from SGOT or SGPT chart, or reading sodium from the potassium peak, or vice versa).

(3) Dilution and calculation errors:
(a) Analyst forgot to correct results for dilution.
(b) Samples diluted by first-shift technologist were analyzed by a second-shift technologist, who was not informed of the prior dilution.
(c) A newly employed analyst thought a "1 to 2" dilution meant one volume of serum and two volumes of diluent, rather than one volume of each.

(4) Reagent and standard solutions:
(a) Distilled water, rather than buffer, was used to prepare a reagent.
(b) pH meter standardized with wrong buffer.
(c) Reagent contaminated.
(d) New reagent used without checking against old reagent; baseline correction changed and all specimens had elevated values.
(e) Out-dated substrate or standard solutions used.

(5) Instrument problems:
(a) Slow clock used for a timed reaction.
(b) Recorder not properly warmed up; blank reading unstable.
(c) Broken balance was used to weigh out standards.
(d) AutoAnalyzer: dirty membrane; clot in line; manifold lines worn; manifold lines improperly installed.

(6) Other:
(a) Specimens left at room temperature by first-shift technologist to be analyzed by second-shift technologist; second-shift technologist did not report for work and specimens were not analyzed until the next day.
(b) Analyst calculated results mentally, rather than drawing a standard curve or calculating a factor; the results were grossly incorrect.
(c) Initial computer print-out was incorrect and subsequent corrected print-out was ignored.

Table 6. Frequency of Laboratory Mistakes

<table>
<thead>
<tr>
<th>Type of Mistake</th>
<th>Rate of occurrence (mistakes per 100 specimens)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Specimen mix-up</td>
<td>0.89</td>
</tr>
<tr>
<td>(a) clerical area</td>
<td>(0.35)</td>
</tr>
<tr>
<td>(b) specimen preparation area</td>
<td>(0.12)</td>
</tr>
<tr>
<td>(c) analytical areas</td>
<td>(0.42)</td>
</tr>
<tr>
<td>1. manual area</td>
<td>(0.13)</td>
</tr>
<tr>
<td>2. AutoAnalyzer area</td>
<td>(0.13)</td>
</tr>
<tr>
<td>3. enzyme area</td>
<td>(0.16)</td>
</tr>
<tr>
<td>(2) Incorrect chart readings</td>
<td>0.66</td>
</tr>
<tr>
<td>(3) Dilution and calculation</td>
<td>0.60</td>
</tr>
<tr>
<td>(4) Poor reagent or standard solutions</td>
<td>0.75</td>
</tr>
<tr>
<td>(5) Other, or unexplained</td>
<td>0.56</td>
</tr>
<tr>
<td>(6) Mistakes in Proficiency Laboratory</td>
<td>0.19</td>
</tr>
<tr>
<td>Total:</td>
<td>3.65</td>
</tr>
</tbody>
</table>

degrees of significance for the overall reliability of the laboratory, and for patient care. If inappropriate accession numbers are assigned in the clerical or specimen preparation areas, the result is serious because two specimens are involved and all analytical results reported for these specimens will be incorrect. A specimen mix-up in an analytical area also involves two specimens, but should ordinarily affect only one of the analyses performed on these specimens. Incorrect readings and mistakes in dilutions or calculations usually affect only a single result on a single specimen, although sometimes a series of specimens may be affected.

Altogether, specimen mix-ups, incorrect readings, and mistakes in dilutions or calculations apparently affected about 2% of all specimens passing through the laboratory (Table 6). Because it is virtually impossible to detect each mistake (only those involving reference specimens can be detected) it is essential that every effort be made to prevent their occurrence.

Although the number of affected specimens is alarmingly high, it seems realistic when one considers the nature of the specimen-processing system. Each specimen and associated paperwork travels a tortuous course through the laboratory, passing through perhaps eight or more stations, or hazards, at each of which there is a certain probability of a mistake occurring (Figure 12). For example, Table 6 shows that specific laboratory areas—such as the Proficiency Laboratory, specimen preparation area, and the manual, automated, and enzyme analytical areas—each have similar frequencies of specimen mix-ups: 0.12–0.19%. Hence, the total probability that a specimen will be affected by a mistake may well approach 1 to 2%, as observed. Obviously, the de-
Class C analyst contributed as many mistakes as 30 Class A analysts.

Conclusions

The Reliability of Clinical Laboratory Results

Analytical considerations. The concepts of the operational chart and allowable limits of error (Figures 1–6) are consistent with fundamental principles of analytical and physiological chemistry, and provide an appropriate goal-oriented perspective for assessing the accuracy, precision, and probable clinical usefulness of laboratory results. This particular perspective is useful because the ideal goals (an operational line with a slope of 1.0 and an intercept at 0.0) are shown in relation to limits of acceptable error (±1/4 the normal range, or individual variability). The operational chart facilitates comparison, over a wide range of

---

Class C analyst contributed as many mistakes as 30 Class A analysts.

Conclusions

The Reliability of Clinical Laboratory Results

Analytical considerations. The concepts of the operational chart and allowable limits of error (Figures 1–6) are consistent with fundamental principles of analytical and physiological chemistry, and provide an appropriate goal-oriented perspective for assessing the accuracy, precision, and probable clinical usefulness of laboratory results. This particular perspective is useful because the ideal goals (an operational line with a slope of 1.0 and an intercept at 0.0) are shown in relation to limits of acceptable error (±1/4 the normal range, or individual variability). The operational chart facilitates comparison, over a wide range of

---

Class C analyst contributed as many mistakes as 30 Class A analysts.

Conclusions

The Reliability of Clinical Laboratory Results

Analytical considerations. The concepts of the operational chart and allowable limits of error (Figures 1–6) are consistent with fundamental principles of analytical and physiological chemistry, and provide an appropriate goal-oriented perspective for assessing the accuracy, precision, and probable clinical usefulness of laboratory results. This particular perspective is useful because the ideal goals (an operational line with a slope of 1.0 and an intercept at 0.0) are shown in relation to limits of acceptable error (±1/4 the normal range, or individual variability). The operational chart facilitates comparison, over a wide range of
concentrations, of actual laboratory performance with specific performance goals, and helps to identify the nature and magnitude of particular analytical problems. These concepts can be broadly applied by single laboratories attempting to systematically improve their performance, by groups of laboratories attempting to obtain analytically comparable results, and by agencies that conduct periodic surveys of laboratories.

Our present evaluation program makes use of commercially available control or reference materials that are designed only to simulate normal or abnormal clinical specimens. These materials are not entirely satisfactory, because of the uncertain validity of some of the manufacturer's stated values, occasional inconsistencies noted when changing lot numbers, the lack of a known quantitative relationship between high and low concentration specimens, and the redundancy of information that is necessarily obtained when normal and abnormal specimens are used to monitor large numbers of tests. Other materials, specifically designed for the evaluation of accuracy and precision in accordance with the concepts developed earlier in this report, would be much more useful.

Nevertheless, the present consideration of ten common analytical procedures provides useful insight into the current status of the analytical reliability of clinical laboratories. For example, Figure 8f shows that the range of values observed in surveys of large numbers of laboratories performing the automated uric acid procedure is only slightly greater than the range of variation of the operational lines of referee laboratories (Figure 10a) or of a single laboratory (Figure 10b). In this and some other automated procedures (Figures 8a,b,c, and g) the range of observed values approaches the range of individual variability, or limits of error necessary to obtain maximally useful clinical results. Thus, large numbers of laboratories using similar methodologies can obtain quite comparable results. In other automated procedures (Figure 8d,e) and in manual procedures (Figure 8h–i), the range of values observed in groups of laboratories is much greater than the variability of one laboratory, and exceeds the limits of acceptable error; in these procedures, laboratories do not produce comparable results and their usefulness is thereby diminished.

Effect of laboratory mistakes. Clerical mistakes undoubtedly occur in all laboratories. If the data of Table 6 are typical, about 2% of all specimens processed may be assigned erroneous results because of a specimen mix-up or an erroneous calculation at some stage of processing and analysis. However, this figure might vary widely among laboratories, depending on the design of the laboratory, the workload, and the type of personnel employed. The results shown in Figure 14 suggest that Class C analysts may be prone to make mistakes that most analysts (Class A) either avoid making or recognize and correct. Regardless of its apparent accuracy and precision, it is doubtful that a laboratory staffed primarily by Class C analysts could produce reliable results.

The percentage of erroneous values owing to analytical mistakes also probably varies considerably among laboratories. These kinds of mistakes can affect large numbers of specimens unless they are corrected promptly. The effect of analytical mistakes on the reliability of a laboratory will depend on the willingness of laboratory personnel to recognize that mistakes can and do occur, and on their ability to recognize and correct their mistakes. According to Table 6, analytical mistakes may occur with a frequency of about 1%, in a laboratory staffed predominantly with Class A and Class B analysts. A vigorous quality-control program is essential if the adverse effects of these kinds of mistakes are to be minimized. It seems reasonable to suppose that laboratories that have minimal quality-control programs, or are inattentive to changes in control specimens, are unlikely to detect analytical mistakes and may produce an inordinate number of unreliable results.

Improvement of Laboratory Performance

A clinical chemistry laboratory is not unlike a complex, multifaceted instrument in the modern physician's armamentarium of diagnostic aids.
The attainment of high-quality performance and efficient use of this instrument must be a joint endeavor of clinical chemists, of manufacturers of laboratory equipment, reagents, and standards, and of the physician-users.

The clinical chemist must calibrate each facet of the instrument to the required degree of accuracy, adjust its sensitivity so that it performs within the required limits of precision, and organize the total laboratory system, so far as possible, to preclude the possibility of mistakes.

To this end, manufacturers of equipment should design their products for reliable performance in routine use, provide clear instructions for maintenance and trouble-shooting, and offer prompt, dependable service when needed. Manufacturers of reagents should maintain rigid standards of manufacturing quality control, and certify the analytical accuracy of standards and control or reference materials in practical terms consistent with accepted laboratory practice.

In this broad perspective, it would seem desirable for clinical chemists to calibrate their "laboratory-instrument" with common standards, to use a minimum number of analytically reliable methodologies, and to undertake a program of systematic improvement of laboratory performance. Physicians are aided in interpreting results if uniform normal ranges, homeostatic limits, and the magnitude of significant physiologic and pathologic deviations from normality are reliably determined and clearly specified. And to assure consistently reliable results, high-quality programs of technical training and continuing education should be encouraged to ensure an adequate supply of competent analysts.

In normal operation these complex "instruments" require regular maintenance and servicing. The calibrations must be checked frequently, faulty components must be identified and readjusted or replaced, and new components must be checked for proper function. Proficiency evaluation and routine quality control are powerful tools for monitoring total laboratory performance, to detect significant deterioration in performance, to identify major problem areas, and to obtain objective evidence of the reliability of the clinical laboratory.

We thank Drs. J. C. Geer and C. R. Macpherson for their advice and encouragement, Miss Penny Nichols for her technical assistance, and Mrs. Kay Cook for preparation of the manuscript. Also, we acknowledge the cooperation, support, and many helpful suggestions of the Clinical Chemistry Laboratory supervisors, medical technologists, and clerks. We are especially indebted to Miss Rachel Earp, who several years ago initiated the program of systematic documentation of mistakes as a training device.

---

1 For the 10 determinations discussed in this report there are at least 33 distinctly different methodologies in common use.

References

20. AutoAnalyzer technic N-20b, Technicon Corp., Tarrytown, N.Y.