Proficiency Testing for Blood-Gas Quality Control

Collene J. Delaney, Elizabeth Teng Leary, Vidmantas A. Raisys, and Margaret A. Kenny

We have conducted a voluntary, community blood-gas proficiency testing program, with use of tonometered human blood, for 32 analyzers located in 16 laboratories. Instruments initially showed inaccuracies as large as $-30.8 \text{ to } +17.3\%$ for $p_{O_2}$ and $-14.0 \text{ to } +42.9\%$ for $p_{CO_2}$, but inaccuracy and imprecision decreased in most laboratories during the program. For a typical 15-week period, mean group precision (CV) was 4.3 to 5.1\% for $p_{O_2}$ from 6.92 to 33.3 Pa (52 to 250 mmHg), and 4.0 to 6.9\% for $p_{CO_2}$ from 2.0 to 6.8 Pa (15 to 51 mmHg). This program can detect increasing imprecision or inaccuracy caused by analyzer deterioration, and can identify interlaboratory or interinstrument bias and problems not detected by participant quality-control programs. Participants have used the proficiency information in discussing data quality with clinicians, promoting internal control and maintenance programs, and justifying instrument purchases. We believe that proficiency-testing documentation of variability in blood-gas analysis may help to establish realistic patient-care protocols.

Additional Keyphrases: analytical error • interlaboratory performance

Five years ago we began a proficiency-testing program to monitor the analytical validity and medical reliability of blood-gas analysis in two University of Washington hospitals. We were then in the midst of a rapid expansion period, which eventually resulted in an 18-fold increase in the number of blood-gas analyses. Many more physicians were ordering the tests, and many more patients who needed pulmonary, renal, and cardiac assessment were routinely being monitored. The number of tests per patient rose from a few to as many as 250.

Some physicians served at both institutions, so we felt that uniform quality of blood-gas testing would help them standardize their expectations and responses. We had purchased two new blood-gas analyzers, but had no solid documentation of their long-term precision and accuracy. Without such criteria for achievable performance, we could not isolate analyst-related from instrument-related errors, nor could we justify modifications in our existing quality-control program. Furthermore, we wished to provide the technologists who used the instruments with an additional index of malfunction, in hopes of detecting problems before they deteriorated into crises. We projected that blood-gas proficiency testing would be a partial answer to these problems.

It soon became evident that other laboratories in the metropolitan Seattle area were experiencing the same uncertainties that we were. Hoping that they might derive the same projected benefits of proficiency testing, we gradually expanded our program to include 32 blood-gas analyzers in 16 laboratories.

We report here the results of this endeavor. We know of no similar program. The spectrum of testing environments represented is comprehensive, ranging from service to research laboratories in federal, state, county, private, and commercially managed hospital situations. Thus, the proficiency testing program now presents a unique opportunity to study the quality of blood-gas analysis in a group that we believe reflects current laboratory practice on a broader scale.

From our experience, from retrospective study of our records, and from analysis of a questionnaire recently completed by the participants, we have drawn information presented here regarding (a) the feasibility of local proficiency testing with tonometered blood; (b) the "state of the art" of blood-gas testing, as reflected by each participating laboratory's first performance in the program; (c) the sensitivity of the testing system to an

\[\text{Department of Laboratory Medicine, University of Washington School of Medicine, Seattle, Wash. 98195.}\]
\[1\text{ Address correspondence to this author at the Department of Laboratory Medicine, SB-10, University of Washington, Seattle, Wash. 98195.}\]

\[\text{Received June 7, 1976; accepted July 13, 1976.}\]
individual laboratory's quality control problems, including instrument failure, instrument design, and staff changes; (d) the impact of the program on the performance of the participants, as shown by current data; and (e) the use that the participants have made of performance-analysis data. Related studies on internal quality control of blood-gas analysis, tonometry, and pH calibration will be reported elsewhere.

Materials and Methods

Program Protocol

The program's format is diagrammed in Figure 1. At the University Hospital, pooled human blood is tonometered to specific \( O_2 \) and \( CO_2 \) tensions, as described below. We then send three specimens by taxi, courier, shuttle bus, or ferry to participating laboratories. The participants analyze the bloods and telephone their results and the time of analysis to the reference laboratory. The technologist receiving the telephoned information records it and describes to the caller (a) the results obtained in other laboratories, (b) the results obtained in the reference laboratory, and (c) the theoretical value of the specimen. If the participants or the reference laboratory personnel believe a serious discrepancy exists between the results, they may discuss possible explanations with reference laboratory staff or faculty or otherwise make use of limited problem-solving services available from that resource. Finally, the testing data are evaluated, summarized, and distributed to the participants at monthly or bimonthly intervals.

Specimens

Day-old whole blood collected (ethylenediaminetetraacetic acid anticoagulant) for routine hematology evaluation is pooled. We exclude specimens with high leukocyte count, with lipemia, hemolysis, or abnormal hematocrit, or from patients suspected of having hepatitis or cancer. We then determine the leukocyte count of the pool. We use an IL 137 tonometer (Instrumentation Laboratory, Inc., Lexington, Mass. 02173) to equilibrate 6-ml aliquots of blood. Gases (Airco Industrial Gases, Murray Hill, N. J. 07974) for tonometry and for calibration of the blood-gas instruments in the reference laboratory are assayed in triplicate by the method of Scholander (1). The three common gas mixtures are (a) low gas, equilibrating at 6.65 kPa (50 mmHg) \( P_{O_2} \) and 2 kPa (15 mmHg) \( P_{CO_2} \); (b) medium gas, equilibrating at 20 kPa (150 mmHg) \( P_{O_2} \) and 4 kPa (30 mmHg) \( P_{CO_2} \); and (c) high gas, equilibrating at 33.3 kPa (250 mmHg) \( P_{O_2} \) and 6.65 kPa (50 mmHg) \( P_{CO_2} \) (The exact tensions vary from tank to tank, however.) Tonometry is performed according to our own method (Teng Leary, E., Delaney, C. J., and Kenny, M. A., Use of tonometered blood for internal blood gas quality control, text in preparation for publication), and the specimens are transported to the participating laboratories in ice baths.

Reference Laboratory

The University Hospital Clinical Chemistry Laboratory, base of the program, was the "reference laboratory" in this study. It is a service laboratory in which 2000 blood-gas specimens per month are analyzed by any of 24 rotating and four non-rotating certified medical technologists. Four analyzers are currently used: one IL 213, two IL 313s (modified) (Instrumentation Laboratory, Inc.), and one Radiometer ABL-1 (The London Co., Cleveland, Ohio 44145). The quality-control program shown in Table 1 regularly includes items 1-5. It may include items 6-9 for problem solving or research purposes.

Analytical protocol includes one-point calibration of the instruments before every analysis, two-point calibration at least once every 8 h, and daily monitoring of measuring temperatures. Electrolyte solutions and membranes of all electrodes are changed each week. Weekly maintenance of the IL instruments also includes changing pump windings and air filters, cleaning sample

---

Table 1. Internal Quality-Control Measures Used in the Reference Laboratory

<table>
<thead>
<tr>
<th>Measure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Controlled quality of the specimen</td>
<td>Assay of control sera</td>
</tr>
<tr>
<td>2. Assay of tonometered blood daily at three tensions</td>
<td>Assay of room air-equilibrated water</td>
</tr>
<tr>
<td>3. Assay on at least two instruments</td>
<td>Assay of room air and gas mixtures</td>
</tr>
<tr>
<td>4. Regular preventive maintenance</td>
<td>Analysis of blood bicarbonate by other methodology</td>
</tr>
<tr>
<td>5. Documentation of all maintenance, repair and problem-solving</td>
<td></td>
</tr>
<tr>
<td>6. Assay of room air-equilibrated water</td>
<td></td>
</tr>
<tr>
<td>7. Assay of room air and gas mixtures</td>
<td></td>
</tr>
<tr>
<td>8. Analysis of blood bicarbonate by other methodology</td>
<td></td>
</tr>
<tr>
<td>9. Assay of control sera</td>
<td></td>
</tr>
</tbody>
</table>

* Routine measures

---

2 In the remainder of this presentation, the more familiar non-SI unit, mmHg, is used. 1 mmHg = 133 Pa = 133 N/m².
Table 2. Description of Participating Laboratories and Their Current Blood Gas Analyzers

<table>
<thead>
<tr>
<th>Laboratory orientation</th>
<th>Hospital size (beds)</th>
<th>Gases, * per week</th>
<th>Instruments</th>
<th>Proficiency tests per mo.</th>
<th>Staff size</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Chemistry (reference laboratory)</td>
<td>273</td>
<td>480</td>
<td>IL 213</td>
<td>20</td>
<td>28</td>
</tr>
<tr>
<td>B. Pulmonary function</td>
<td>273</td>
<td>15</td>
<td>Radiometer BMS3 Mk2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>C. Anesthesiology (R)*</td>
<td>273</td>
<td>5</td>
<td>Radiometer BMS3 Mk2</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>D. Anesthesiology (R)*</td>
<td>273</td>
<td>3</td>
<td>IL 113</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>E. Surgery</td>
<td>273</td>
<td>80</td>
<td>Radiometer BMS3 Mk2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>F. Chemistry</td>
<td>350</td>
<td>200</td>
<td>IL 313</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>G. Respiratory therapy</td>
<td>300</td>
<td>100</td>
<td>IL 113</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>H. Pulmonary function</td>
<td>293</td>
<td>7</td>
<td>Radiometer BMS3 Mk2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>I. Chemistry</td>
<td>281</td>
<td>350</td>
<td>IL 313</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>J. Chemistry</td>
<td>225</td>
<td>60</td>
<td>IL 213</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>K. Chemistry</td>
<td>200</td>
<td>300</td>
<td>IL 113 (2)</td>
<td>4</td>
<td>34</td>
</tr>
<tr>
<td>L. Chemistry</td>
<td>198</td>
<td>40</td>
<td>IL 313 (2)</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>M. Chemistry</td>
<td>173</td>
<td>73</td>
<td>Radiometer ABL 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N. Chemistry</td>
<td>160</td>
<td>25</td>
<td>Corning 160°</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>O. Chemistry</td>
<td>108</td>
<td>34</td>
<td>Corning 165°</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P. Chemistry</td>
<td>80</td>
<td>8</td>
<td>Corning 185°</td>
<td>4</td>
<td>12</td>
</tr>
</tbody>
</table>

* Each gas analysis included pH, pO₂, and pCO₂
* Indicates research rather than patient service work
* Corning Scientific Instruments, Medfield, Mass. 02052

chambers and, in the IL 313s, cleaning leadex switches (stainless-steel replacement fabrications). Monthly maintenance includes changing transmission tubing and fluid diodes, cleaning potentiometers, baths, and hydration chambers, and replacing seals as needed. All instruments are cleaned with dilute hypochlorite solution much more frequently than recommended by manufacturers. The ABL-1 may be backflushed as often as once per 15 specimens.

Proficiency-testing specimens are analyzed with standard technique. In special cases, proficiency-testing specimens may be analyzed from 30 min to 8 h after equilibration to parallel expected delay in a participating laboratory, but most are analyzed within an hour.

Participating Laboratories

Table 2 describes the laboratory orientation, hospital size, workload, manufacturer, and model of instruments currently in use, and number of staff performing blood gas analyses in the participating laboratories. The 188 analysts operating the equipment include laboratory assistants, technicians, technologists, therapists, and physicians. No laboratory engages more than three operators at once; most use only one at a time, although that individual often operates several instruments simultaneously. Eighty-six percent of these individuals rotate to other assignments, some spending as little as one week out of 16 performing blood-gas analysis. None of the chemistry laboratories have discrete personnel for blood-gas analysis. However, supervisors of blood-gas analysis in four chemistry subsections have limited rotations. At most participant laboratories, depth of technical expertise, i.e., maintenance and problem-solving ability, is retained by one or two individuals—technicians, technologists, or therapists—but never by physicians. Physicians perform analysis in two laboratories only. The weekly workload of all program participants combined is 1780 specimens. However, no laboratory has a constant workload and the figures from individual locations may vary by 30–40% from one month to the next. Program participation ranges from 2 to 30 sets of tonometered specimens per month, with most receiving four per month. Specimens are transported up to 25 miles away, with delivery times ranging from 5 min to 3.5 h.
Data and Records

When specimens are prepared, we record the barometric pressure, the calculated gas tensions adjusted to the day's pressure (theoretical gas tension), the duration and temperature of tonometry, the time of specimen preparation, and the name of the operator. When participating laboratories report their performance by telephone, we record their results, the time of their analysis, and any special or additional comments that may be pertinent.

Written performance reports are prepared as illustrated in Figure 2. Using either a computer or a programmable calculator, we normalize the telephoned results to standard barometric pressure, thus eliminating the effects of day-to-day pressure variation. (All data cited in the text are corrected to standard barometric pressure.) The data are then processed to obtain a tally of the number of samples, the mean, standard deviation, coefficient of variation, relative error, and a 95% confidence range for the group as a whole and for the individual instruments being monitored. Because some instruments are monitored daily and others are monitored weekly, the data reduction is usually done at 5–7 week intervals to enhance the statistical significance. Confidential summary reports containing all laboratory performance information are mailed to supervisors in participating laboratories.

Performance in the reference laboratory, which may use more than one set of tonometered specimens per day, is monitored daily as a quality-control check of both the blood-gas analysis and the tonometry technique.

Problem-Solving Services

Participants may consult with the reference laboratory staff, as previously indicated. The reference laboratory does not initiate problem solving except in special circumstances. In addition to verbal consultation, the reference laboratory may direct participants to appropriate commercial repair services, or loan them temperature monitoring devices, spare calibrated gases, or even tonometers.

Results

First-Time Participation

When laboratories initially enrolled in the program, we assessed the accuracy of their instruments by comparing their first results with the tonometered specimen's theoretical value and with the reference laboratory's usual performance. (The tonometry technique adopted to achieve maximum precision with the IL 137 precludes absolute accuracy for the highest O2 tension; hence, comparison of these values is made with reference laboratory performance.) “Reference laboratory performance” is defined as the average mean deviation from theoretical ±2 SD, based on data gathered from our four instruments over 89 consecutive days.

Figure 3 presents first-day performance of 19 instruments in 13 laboratories. Participants joined the program at various times during a four-year period, so the range shown for each gas is the range spanned by the theoretical tonometry values during that period. Both linear and nonlinear instrument performances were observed. The deviation range from −78 to +19 mmHg for high $p_{O_2}$ analysis is remarkable. The relative error ranges calculated from these data are: 0.0 to +42.9% for low $p_{CO_2}$, −10.7 to +20.0% for medium $p_{CO_2}$, −14.0 to +13.5% for high $p_{CO_2}$, −27.5 to +17.3% for low $p_{O_2}$, −19.5 to +6.4% for medium $p_{O_2}$ and −30.8 to +7.5% for high $p_{O_2}$. When compared with the reference laboratory mean, the high $p_{O_2}$ relative error range is −22.6 to +21.2%. Some laboratories having more than one instrument had good agreement between instruments for one analyte but not for both (e.g., laboratory E).

Sixty-five percent of the first-day assays fell within the “achievable” limits defined by reference laboratory performance (shaded areas). $p_{O_2}$ assays were as frequently within acceptable limits as were $p_{CO_2}$ assays.
Performance data gathered from each instrument during the second week of program participation demonstrated most improvement in $p_{CO_2}$ analysis. Three additional instruments matched reference laboratory performance on the low $p_{CO_2}$, three more on the medium $p_{CO_2}$, five more on the high $p_{CO_2}$, five more on the low $p_{O_2}$, and two more on the high $p_{O_2}$. Overall, 81% of the data returned were within the "achievable" performance limits. These accuracy changes resulted from analyzer temperature-setting adjustments, electrode maintenance, electrode replacement, or instrument repair. The instrument in laboratory C, which initially assayed the high $p_{O_2}$ specimen at 63 mmHg below its theoretical value, gave a result 111 mmHg below theoretical in the second week. This and one other instrument could not be rendered more accurate by their operators until significant electrical repair, requiring two to three additional weeks, had been done. Increased accuracy of these instruments was observed in the second month of their program participation.

Long-term Participation

To assess the impact of long-term proficiency testing, we calculated the mean deviation from theoretical ±2 SD for 15 consecutive weeks for the 21 instruments participating at least once a week (Figure 4). The briefest period of participation prior to this interval of data collection was six weeks. If an instrument was tested five times a week during the period, every fifth value was used, so $n = 15$ for each instrument. The reference laboratory instruments are represented by code numbers 18–21. For ease of interpretation, the theoretical gas tensions have been normalized to the single value shown.

The data in Figure 4 indicate that the accuracy of most of the instruments compared well with that of the reference instruments. The relative errors ranged from +6.7 to +47.7% at 15 mmHg $p_{CO_2}$, from −3.3 to +3.3% at 30 mmHg $p_{CO_2}$, from −3.9 to +5.9% at 51 mmHg $p_{CO_2}$, from −3.8 to +13.5% at 52 mmHg $p_{O_2}$, from −14.1 to +0.7% at 149 mmHg $p_{O_2}$, and from −19.2 to −6.0% at 250 mmHg $p_{O_2}$. The average deviations from theoretical for the whole group at the six tensions, in the same sequence, were: +1.8, +0.4, −0.2, +3.0, −7.8, and −29.4 mmHg.

Instrument 9 had a notable bias during the period represented in Figure 4. If the limits set by reference laboratory performance shown in Figure 3 (shaded areas) are superimposed on Figure 4 data, the mean deviation of instrument 9 from theoretical falls outside of those reference limits for $p_{CO_2}$ at 15 mmHg, $p_{O_2}$ at 149 mmHg, and $p_{O_2}$ at 250 mmHg. Instrument 10, the other analyzer at the same locale as instrument 9, agreed closely with instrument 9. However, it was only for $p_{O_2}$ at 250 mmHg that the mean of instrument 10 lay beyond the 2 SD limits of the reference laboratory performance.

Analysis of these data for between-week precision showed surprising group consistency. The average CV values are: 6.9% at 15 mmHg $p_{CO_2}$, 4.2% at 30 mmHg $p_{CO_2}$, 4.0% at 51 mmHg $p_{CO_2}$, 4.4% at 52 mmHg $p_{O_2}$, 4.3% at 149 mmHg $p_{O_2}$, and 5.1% at 250 mmHg $p_{O_2}$. The most precise individual performance was the 1.6% CV of instrument 17 ($p_{CO_2}$ of 51 mmHg); the least precise was the 14% CV of instrument 14 ($p_{CO_2}$ of 15). Instrument 14 had the least precision at all three $p_{CO_2}$ tensions. Instrument 13, located in the same laboratory as 14, was considerably more precise except for a $p_{O_2}$ of 52 mmHg. Results from instrument 12 ranged from 177 to 258 mmHg for the theoretical value of 250 mmHg $p_{O_2}$. This extreme imprecision, which occurred within a one-month period, is not consistent with the instrument’s usual performance (ordinarily only 1 or 2% less precise than the group average).

The data of Figure 4 are representative of group performance (mean and range) and of interlaboratory bias. Specific instrument performance varies, however, in response to changing stability of electronic components and transducers. For this reason, some interinstrument bias may be obscured when evaluated in a limited time period. If expected bias is not taken into account, however, some sensitivity of proficiency testing is lost. Both points are illustrated by the data for instruments 1 through 4 (Figure 4), which are located in a single laboratory.

Instruments 1 and 2, identical models, usually detected $p_{O_2}$ values closer to the theoretical value, whereas instruments 3 and 4 are design-limited and generated values 4 to 6% lower on the same specimens. During the period illustrated, instruments 1, 3, and 4 performed predictably. However, at $p_{O_2}$ of 250 mmHg, instrument 2 showed a mean markedly lower than that of instrument 1 but similar to the means of instruments 3 and 4. Examination of any other 15-week period of the preceding year shows that both instruments 1 and 2 had mean deviations from theoretical comparable to that of instrument 1 in Figure 4, and that the means of both instruments always agreed within 5 mmHg at $p_{O_2}$ of 250 mmHg. Historical perspective or performance expectations provided by proficiency testing are more sensi-
tive indicators of malfunction than group performance when accuracy is changing.

We also examined the high $p_{CO_2}$ performance data from Figure 4 as a possible index of tonometry quality, since scatter of data is a possible reflection of proficiency specimen preparation. Data from instruments 14, 15, and 16 were eliminated entirely because these instruments had two to six results that were more than 4 mmHg from their mean tension for the 15-week period. Operators confirmed malfunction of these instruments during the interval. Of the remaining instruments, 15 analyzed all specimens within 4 mmHg of their mean. Four analyzers generated one value more than 4 mmHg from their mean for this period. Investigation revealed that in one instance a second analyzer in a laboratory accurately assayed a proficiency specimen and subsequently confirmed malfunction of the first instrument. In another, the membrane was replaced on the $CO_2$ electrode, the specimen re-assayed, and a more appropriate result subsequently obtained. Two other random inaccuracies were reported from a laboratory where both instruments tested use a single heating bath. Operators remained ambiguous whether the heating bath or the tonometry caused the aberrant results. Thus, only two of these 270 data points at high $p_{CO_2}$ tension may reflect tonometry quality rather than instrument performance. Precision of tonometry and instrument performance shown in Figure 4 were typical for the entire program.

While overall, long-term interlaboratory and inter-instrument differences remained relatively stable, the proficiency-testing program was able to identify random inaccuracies, progressive loss of accuracy or precision, and periodic imprecision. Figure 5 illustrates that the total analytical environment, not just instrument design limitations, accounts for such changes. Laboratories A and B used the same model of instrument, both of about the same age. Laboratory A had recently joined the program at week 1 and was having a significant, unrecognized precision problem. More stringent electrode maintenance and education programs were instituted, and by week 18 the oscillations had decreased to those expected of an IL 313 analyzer. The analyzer in labora-

Table 3. Coefficient of Variation Data as Indicators of Progressive Instrument Failure

<table>
<thead>
<tr>
<th></th>
<th>IL 313 Instruments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td>1972</td>
<td></td>
</tr>
<tr>
<td>January to mid-March</td>
<td>2.7</td>
</tr>
<tr>
<td>Mid-March to June 1</td>
<td>4.1</td>
</tr>
<tr>
<td>June 1 to August 1</td>
<td>5.9</td>
</tr>
<tr>
<td>August 1 to October 1</td>
<td>13.5</td>
</tr>
<tr>
<td>October 1 to November 1</td>
<td>3.0</td>
</tr>
</tbody>
</table>

a Blood tonometered in the range of 230–250 mmHg PO2
b Instrument repair followed

Figure 6 illustrates the sensitivity of proficiency testing data to analytical technique in a participating
laboratory. (For simplicity, we show only two of the six tension levels measured.) The data for weeks 1–6 were obtained by an experienced analyst, who was then replaced by a person less familiar with specimen-handling techniques and instrument calibration. The initial data conformed reasonably well with the reference laboratory performance limits (indicated by the gray zone). The staff change coincided with an immediate and significant accuracy loss. During week 10 (point C), reference laboratory personnel worked with this technologist to refresh her skills. By week 12, proficiency performance was once again within expectations.

Usefulness of Proficiency Testing

We recently gathered information from the 16 laboratories participating in the program in order to evaluate the relevance of their quality-control programs and preventive maintenance systems to their instrument performance. No correlations could be drawn, but the information provided data about the “state of the art” and the usefulness of proficiency testing. This is summarized below.

Internal control programs. Three laboratories do not routinely apply any internal quality control; they use controls only when an analyst perceives a problem. Two laboratories rely solely on the proficiency-testing program to identify problems (both participate weekly). The remaining laboratories routinely use a variety of techniques: four use blood-gas analyzer calibration tanks assayed by the reference laboratory, and six have their own tonometers; some use tonometered blood daily for control purposes and others use it only for problem solving.

Maintenance programs. Four laboratories have no regular preventive maintenance program. The laboratory with the most complete regular preventive maintenance program had the most precise long-term performance data. The laboratory with complete access to the reference laboratory’s instrument repair service did not modify discrepant performance more rapidly than other laboratories that were without access to this technical expertise.

Evaluation of proficiency data. Seven of the laboratories evaluate their performance in relation to the theoretical gas content of proficiency specimens, and the other nine evaluate theirs in relation to the reference laboratory performance. Four laboratories would stop assays of patients’ specimens on an instrument that shows a 10% relative error in analysis of a proficiency specimen; all others would continue to process patient materials but would work full time to identify and correct the possible error. (Most participating laboratories receive very few patient specimens that have $PCO_2$ tension values as low as 15 mmHg or $PO_2$ tensions as high as 250 mmHg.) Most laboratories expect less than 5% CV on proficiency specimens. One laboratory has a 10% CV limit and one has no limit. Again, most laboratories would continue to assay patient specimens while they resolved the possible problems, but the same four laboratories would stop patient testing on the instrument in question.

Use of program data. No laboratory changed staffing in response to proficiency testing results. Three retired inaccurate or imprecise blood-gas analyzers, and eight purchased new instruments, basing their selection in part on performance data available from the program. Five changed or enlarged their preventive maintenance programs to improve performance, and seven started, modified, or enlarged their quality-control programs. Eleven used proficiency testing data in discussing the quality of blood-gas analysis with physicians who were ordering blood-gas determinations.

Program withdrawal. Two laboratories have withdrawn from the program. One of them has an internal quality-control program, including tonometry; the other has no internal quality-control program. Neither reported results back to the reference laboratory routinely. The performance of these laboratories did not change during the period of participation (three and six months, respectively).

Discussion

The deviations in first-day performance data (Figure 3) surprised the laboratories involved—no one had expressed any reservations about instrument function or operator skill before joining the program. Initial results were mostly nonlinear with respect to both theoretical values and reference laboratory performance. For example, instrument K tested 11.8% lower, 1.9% higher, and 9.7% lower than the reference laboratory performance at low, medium, and high oxygen tensions, respectively. Quality-control mechanisms, where used, were not effective in detecting such difficulties over the ranges surveyed. Only two laboratories had been routinely validating analytical quality for $PO_2$ tensions above 150 mmHg.

The magnitude and frequency of errors reported here are consistent with the findings of other investigators, whether they had employed tonometered blood, tonometered serum, or commercial control media. Dowd et al. (2), surveying seven analyzers in five laboratories within a single hospital, observed a deviation from the reference instrument performance of 9 mmHg at 35 mmHg $PCO_2$ and 16 mmHg at 83 mmHg $PO_2$. Weisbrot et al. (3) cited discrepancies as large as 16 mmHg for both gas tensions within the physiological range, in two instruments in different laboratories within the same hospital.

Rej and Vanderlinde (4) report a 20.4% incidence of aberrant participant $PCO_2$ tensions, defined as more than 3 SD beyond their reference laboratory performance limits. In our study, 26% of the $PCO_2$ tensions were more than 3 SD beyond our limits (Figure 3). These limits, however, are much narrower than Rej and Vanderlinde allowed. Their CV values for reference laboratory performance ranged from 3.2 to 7.6% (mean 5.1%) within the $PCO_2$ tension range of 34–53 mmHg.

---

3 One- and two-point calibrations and the frequency of calibration are considered operational, not quality-control mechanisms; neither is implied in this discussion.
The mean CV values for 89 consecutive days for the reference laboratory instruments shown in Figure 3 are 3.3% at 30 mmHg $p_{CO_2}$ and 3.0% at 49 mmHg $p_{CO_2}$. For the period during which the data in Figure 4 were collected, the corresponding reference CV values are smaller yet, 2.5 and 2.4%, respectively. If the mean reference precision limit of the Rej and Vanderlinde study (5.1% CV) were applied to our Figure 3, 10.5% of the $p_{CO_2}$ analyses would be aberrant among our first-time participants. Although one might argue that our narrower reference criteria are unrealistic for a larger group, we believe that they reflect achievable precision.

Disagreement between instruments in the same location is another aspect of the initial participation data (Figure 3) that deserves attention. Brantigan (5) has illustrated a 12-mmHg discrepancy at 25 mmHg $p_{CO_2}$ for two analyzers in a single laboratory. We report, in Figure 3, maximum discrepancies of 10 mmHg at 50 mmHg $p_{CO_2}$ and 13 mmHg at 252 mmHg $p_{CO_2}$. We do not feel that the comparatively good agreement among instruments in the five sets of multiple blood gas analyzers at the same locations illustrated here should be extrapolated as a reflection of common practice on a broader scale.

Although instrument comparison is a recommended quality-control mechanism (6), few of the 188 analyzers in our program, especially rotating staff, are familiar with such recommendations in the literature. However, since 1972, the reference laboratory has made strenuous efforts toward standardizing the performance of several models of gas analyzers from a single manufacturer. Our instrument comparisons made on patient test results and for problem detection became well known to local representatives of the instrument manufacturers, to local instrument servicing organizations, and to laboratory-management personnel. From these direct and indirect contacts, much local publicity has been given to this quality-control practice of instrument comparison and the application of internal tonometry. Consequently, these mechanisms probably are applied more frequently in the Seattle area than elsewhere in the country.

Changes observed after two weeks of participation in the program indicate that performance improvement can be generated in response to proficiency-testing information. However, the actual percentage and degree of improvement quoted here must be interpreted in the context of long-term precision capability within the participating laboratories. Nevertheless, because most of the changes have been sustained to some degree, we conclude that part of the value of proficiency-testing resides in the establishment of external "target" performances. The arbitrariness with which these targets are defined— theoretical, group average, or reference laboratory performance—has not mitigated their effectiveness.

It is important to note that two of the 19 analyzers in this study needed major electronic repair to improve proficiency performance. In addition, we selected the cases of progressive electronic deterioration (Table 3) and cited subsequent electronic deterioration of instrument B in Figure 5 to illustrate the beneficial applications of proficiency testing and to draw attention to unsuspected instrument malfunction as a source of imprecision and inaccuracy. So much attention had been focused on specimen instability or transducer-associated artifacts (7) that analysts and manufacturers both appeared not to respond in a timely manner to the possibility of electronic deterioration in the instruments. The data cited show that it took nearly eight weeks to solve the problems, during which time patient-sample testing was cautiously continued. In each case, most of the known transducer- and hydraulics-related sources of error were thoroughly investigated and eliminated as possibilities before suitable expertise was applied to detect or resolve the electronic malfunction. Again, our program participants and their experiences may have been atypical, but the long-term data do not corroborate the statement made by a manufacturer's representative, and quoted by Good (8), that "about 98 percent of all blood gas problems are due to technique or operator deficiencies rather than instrument failures. . . . Of the remaining 2 percent of the complaints, perhaps 1.99 percent are electrode problems. . . . A miniscule fraction are electronic in nature."

Analysis of our long-term program data confirmed the importance of operator and environmental impact on the quality of blood-gas analysis, if instruments were functional (Figure 5). When testing was done weekly, we found the performance of a specific model of instrument in an individual laboratory changed with time toward a level of performance similar to that of instruments of the same model in other laboratories (see Figure 5). Thus, most IL 313 instruments in the program now have comparable accuracy and precision. The Radiometer ABL-1 instruments likewise have comparable performance, but it is different from that of the IL 313. Testing less often than weekly did not appear to improve performance; at least, instruments so tested did not match their comparable group performance. Because sensitivity of proficiency testing is limited by frequency of testing, analytical precision within the participating laboratory, and precision of our tonometry, we now actively discourage less than weekly participation. The examples shown in Figure 6 and Table 3 illustrate that weekly testing has enough sensitivity to signal some types of performance change.

Credibility of the proficiency information among participants seemed to vary, as indicated by their degree of willingness to compare themselves to group, reference laboratory, or ideal performance, by the limits in accuracy they employed, and by the response they made to unanticipated performance. Participants also attached different importance to performance at each gas tension, depending on their patient populations. For

---

example, inaccuracy at low $p_{\text{CO}_2}$ elicited no response from some operators but was of great concern to pulmonary function-oriented analysts. Because newer participants were more suspicious of the quality of the tonometry and specimen handling, they sometimes delayed troubleshooting until a second set of proficiency samples had confirmed the existence of an analytical problem. Participants who had been in the program longer were more familiar with the quality of the tonometry and were better able to diagnose where their problems lay.

The incidence of spurious proficiency specimens distributed in the total program is unknown. It would appear to be low. We have followed our tonometry quality by monitoring long-term group precision but have encouraged participants to communicate directly with our tonometry operators if they were in doubt about the quality of a specimen. Because tonometry is subject to a variety of errors (7, and Teng Leary, E., Delaney, C. J., and Kenny, M. A., Use of tonometered blood for internal gas quality control, text in preparation) we recommend caution in applying our experience with it to other proficiency-testing situations.

Rej and Vanderlinde (4) reported a direct correlation between weekly workload and quality of $p_{\text{CO}_2}$ analysis: the higher the workload, the better the performance. We found no such correlation. Either our population was too small or an interlaboratory similarity had developed during the program that obscured any initial relationships between proficiency and workload, staff size, or laboratory orientation.

Finally, the proficiency testing data in Figure 4 may be useful to participating institutions in considering what changes in $p_{\text{O}_2}$ and $p_{\text{CO}_2}$ tensions have medical significance. A current “worst case” of between-week precision of instrument 14 at 51 mmHg $p_{\text{CO}_2}$ shows 1 SD = ±4.7 mmHg. This precision would seem insufficient to distinguish arterial from venous blood (6 mmHg difference), to assess “normalcy” in arterial blood (10 mmHg range), or to follow modest patient status compensations (4-6 mmHg) over any protracted period.

We believe that the willingness of participants to voluntarily retire inaccurate analyzers and buy new ones, to change control programs, and to initiate tonometry reflects their professional concern with the medical usefulness of their blood-gas data. Combining this introspective medical assessment with proficiency statistics on the relative capabilities of the other analysts in the program has helped several supervisors motivate their staff to apply quality control and preventive maintenance programs. There has been no external compulsion from the reference laboratory in this regard. Indeed, the absolutely voluntary nature of the program is at times a detriment to gathering complete program statistics. Some analysts are reluctant to report grossly aberrant results, even though reference laboratory personnel have tried to stress that a malfunction should be expected as a spontaneous occurrence in any instrument. The program is not advertised nor is participation solicited, so its steady growth can be taken as a measure of its adjudged value to the participants.

No participating laboratory has indicated that it uses performance or medical usefulness criteria from the literature. Perhaps these criteria are not practical measures in terms of achievability or medical value. Adams et al. (9) reported a 2% CV for $p_{\text{CO}_2}$ at physiological tensions. Weisbrot et al. (3) found between-day averages of 6.8 and 9.7% CV at tensions of 20—76 mmHg $p_{\text{CO}_2}$ for the two instruments they studied. The worst instance noted was 12.6% at 22 mmHg, and the best was 5.8% at 59 mmHg. They suggested a medically useful precision limit of 3.75% CV at a mean tension of 40 mmHg $p_{\text{CO}_2}$. They derived this target from the requirement that such a limit should assume a 95% likelihood of being able to detect the arterial/venous difference accurately—a difference of 6 mmHg $p_{\text{CO}_2}$.

Twelve of the 21 instruments shown in our Figure 4 met this criterion if we extrapolate from the 51 mmHg $p_{\text{CO}_2}$ data. The CV values of five other instruments are between 3.8 and 4.8%. Thus, a high proportion of $p_{\text{CO}_2}$ tension data from our participating laboratories has suitable medical usefulness.

Projected medical usefulness criteria for $p_{\text{O}_2}$ take into account precision only. Adams et al. (9) quote an expected standard deviation of 1 mmHg for $p_{\text{O}_2}$ tensions between 0 and 150 mmHg. Our reference laboratory could not meet this limit if it is applied to between-day performance. Weisbrot et al. (3) observed CV values of 5.0 and 5.6% in their study when tensions were between 0 and 106 mmHg. Their best $p_{\text{O}_2}$ tension precision was 3.0% at 106 mmHg, and their worst was 7.4% at 96 mmHg. They projected a medically useful precision of 6% CV at 40 mmHg. However, we question the breadth of these limits. If they were used, hemoglobin saturation data (interpolated from a standard $p_{\text{O}_2}$ saturation plot) would range from 65—85% if the $p_{\text{O}_2}$ is 40 mmHg ±2 SD (35.2—44.8 mmHg). Such a spread would at least obscure assessment of oxygenation adequacy and efficiency in some cases.

Our proficiency data indicate that narrower limits can be achieved. Extrapolating from the $p_{\text{O}_2}$ tension of 52 mmHg in Figure 4, 19 of 21 instruments meet Weisbrot's limits; 16 could meet a 5% CV limit. After the scheduled retirement of two less-precise instruments later this year, these figures are expected to improve further.

Consideration of medically useful performance limits for $p_{\text{O}_2}$ analysis should also include accuracy criteria. Although many patients are managed in response to defined incremental $p_{\text{O}_2}$ tension changes, others are managed when the tension exceeds a fixed limit (e.g., 80 or 90 mmHg for neonates). Definition of such a limit in any given institution may be safer for the patient if done in the context of achievable relative error and achievable precision as delineated by proficiency testing.

Moreover, subsequent blood-gas analyzers installed in any single laboratory should generate data that are consistent with the clinical limits being imposed. If a
new instrument’s bias will cause artifactually lower or higher assessment of \( pO_2 \) values, then laboratorians need to define the discrepancies. Proficiency testing over a broad range of \( pO_2 \) tensions can help define these interinstrument biases.

In conclusion, several limitations of our proficiency testing program deserve attention. First, the proficiency specimens are not prepared with absolute accuracy. Second, the blood pool changes daily, which occasionally results in poor stability and lower results from laboratories that delay analysis. Third, participant performance data inevitably include tonometry imprecision and inaccuracy. Fourth, we have gathered all proficiency data on weekdays, when supervision and staffing are maximum in all laboratories; except in the research environments, it may not be appropriate to infer that observations made under what should be optimal conditions adequately reflect total performance. Fifth, the reporting of all proficiency-testing data rests with the participants. Specimen volumes are large enough that several aliquots may be analyzed. Because gas tensions are predictable, an operator may detect an aberrant performance in a test, resolve its cause, and subsequently reanalyze the specimen. Whether either or both results are telephoned to the reference laboratory for our records depends on the operator. We encourage the use of proficiency specimens for problem solving as strongly as we endorse the value of proficiency testing as a true reflection of performance.

Despite these shortcomings, we conclude that our local blood-gas proficiency testing program is practical and useful. It provides all of the benefits originally projected: (a) statistics for definition of actual performance, (b) data for interinstrument comparisons, and (c) an additional index for problem detection.

We thank the participating laboratories and the postdoctoral fellows who have been involved in the program. This work was supported in part by NIGMS Training Grant in Clinical Chemistry ST01-GM00776 and by NIGMS Research Fellowship 4P22-GM01872.

References