Detecting Errors in Blood-Gas Measurement by Analysis with Two Instruments

Louis F. Metzger, William B. Stauffer, Ann V. Krupinski, Richard P. Millman, George S. Cembrowski, and Allan I. Pack

We performed a two-stage prospective evaluation of the error detection capabilities of duplicate analysis of blood-gas specimens. In the first stage we analyzed 1601 specimens with a Corning Model 175 blood-gas analyzer as the test instrument and a Corning Model 178 analyzer as the reference instrument, and in the second stage we analyzed 1544 specimens with two Model 178 analyzers. In each stage the designated reference instrument underwent troubleshooting whenever an analytical error was detected; the test instrument underwent troubleshooting only when error conditions were indicated by means other than duplicate analysis. An error was considered to have occurred if the difference between the duplicate analyses exceeded 0.02 (for pH), 0.53 kPa, i.e., 4 mmHg (P$_{CO_2}$), or 7% (P$_O_2$). The number of specimens for which errors were detected was 97 (6.1%) in the first stage, 23 (1.5%) in the second. For each analyte more errors were detected with the Model 175 analyzer (of older design) than with the newer Model 178 analyzer. Furthermore, in certain periods associated with the use of particular electrodes there were very high error rates for individual analytes: 8% for P$_{CO_2}$; 18% for P$_O_2$. We conclude that duplicate analysis should be considered as a possible required standard for error detection.

Additional Keyphrases: quality control • pH • analytical-error assessment • economics of laboratory operation

The utility of duplicate analysis of blood-gas specimens for the detection of analytical errors is a subject of current debate. On the one hand, some consider it an essential part of quality control (1, 2), while others tend to de-emphasize its usefulness (3, 4). Part of the reason for this diversity of views is the lack of definitive data on the subject (5).

The value of dual-instrument analysis lies in its capability to detect intermittent ("random") errors. Continuous ("systematic") errors should be detected by the use of control specimens, which typically are analyzed every 8 h. The purpose of the study was to characterize the type (continuous vs intermittent), frequency, and magnitude of errors encountered in blood-gas analyses, and to determine whether duplicate analysis could be useful for detecting these errors. Because the results could depend on the specific instrument being tested, we performed prospective studies of the type and frequency of errors with two different blood-gas analyzers, one of a recent design and one older.

Methods

Samples

All blood specimens of at least 1.5 mL obtained from patients without contagious disease were accepted into the study and were simultaneously measured on two blood-gas analyzers. The volume requirement, which allows the four analyses enumerated below, precluded the use of specimens contained in 1-mL syringes and capillary tubes, specimens that exhibit greater variability than do the larger samples. Specimens from patients with contagious diseases were excluded to avoid unnecessary exposure of laboratory personnel.

All technologists participating in this study had received training in theoretical and practical aspects of blood-gas analysis and had performed well in this laboratory's annual proficiency tests.

Quality Control

For quality control of the analyzers we used "Confirm" quality-control reagent (Ciba-Corning Diagnostics Corp., Medfield, MA) and performed routine preventive maintenance, with immediate corrective maintenance in response to instrument-generated error signals.

Confirm is a tonometered aqueous buffer available in "Normal," "Alkalosis," and "Acidosis" formulations for verifying instrument performance throughout a range of pH, P$_O_2$, and P$_{CO_2}$ values. We used the multi-rule control protocol proposed for a three-level system by Westgard et al. (6).

All three formulations of Confirm were analyzed at least once on each 8-h shift. If the multi-rule control procedure was violated, the technologist checked the system, the first step being the re-analysis of the Confirm sample indicating "out of control" data; if the result was unchanged, the technologist investigated further, e.g., by removing bubbles or changing membranes or electrodes.

Study Protocol

The study consisted of two consecutive parts, each lasting approximately 10 weeks. In Part A, we analyzed each of 1601 specimens once with a Model 175 and once with a Model 178 Blood Gas Analyzer (both from Ciba-Corning), the former designated the "test instrument" and the latter the "reference instrument." Specimens were introduced into the analyzers without regard to order of injection. Both instruments were controlled as described above, but in addition, we verified the proper operation of the reference instrument if the duplicate (i.e., two-instrument) analysis protocol indicated a malfunction of that instrument (see below for full details). In contrast, if duplicate analysis indicated a malfunction of the test instrument, we did not initiate investigation of the test instrument until that malfunction was signaled by either the built-in error checks of the instrument or results for the control material.

In a preliminary study we checked for the presence of bias between the two models of analyzers. Only P$_{CO_2}$ assays show meaningful bias. This difference, which is expected (Ciba-Corning, personal communication), was eliminated during the study by applying a correcting equation to all Model 175 values for P$_{CO_2}$ before calculating the between-instrument differences. This equation was obtained from linear regression analysis of data for 1500 blood specimens analyzed in both the 175 and 178 analyzers: corrected P$_{CO_2}$ in kPa(Model 175) = 1.128 × measured P$_{CO_2}$ in kPa(Model 175) − 0.55.
In Part B of the protocol we analyzed 1544 specimens with two Model 178 blood-gas analyzers, the one designated the reference instrument being the same as the reference instrument in Part A of this study.

We compared the values obtained for each variable (pH, \(PCO_2\), \(PO_2\)) with the test and reference instruments for each blood specimen as follows:

\[
\Delta pH = \text{test pH} - \text{ref. pH}
\]

\[
\Delta PCO_2 (\text{kPa}) = \text{test } PCO_2 (\text{kPa}) - \text{ref. } PCO_2 (\text{kPa})
\]

\[
\Delta PO_2 (\%) = \frac{((\text{test } PO_2 - \text{ref. } PO_2) \times 100)}{\text{ref. } PO_2}
\]

If the difference between these instruments exceeded our chosen ranges of acceptability, based on those of Elser et al. (2)—pH ± 0.02, \(PCO_2\) ± 0.53 kPa, i.e., ± 4 mmHg, or \(PO_2\) within 7% of the \(PO_2\) as measured by the reference instrument—we considered the results to be in error. If such an error occurred, we followed the protocol shown in Figure 1. The first step of this protocol was to analyze the specimen again, once with each analyzer. If repeating the analysis on both instruments gave results within acceptable limits, we assumed that an intermittent error had occurred and took no further analytical steps. An error was defined as intermittent if, on re-analysis, the between-instrument difference was acceptable or if, on analysis of control material, no error was found. The intermittent error was assigned to the instrument that in the repeat analysis showed the greatest absolute change toward the average of the repeat results. If both instruments showed equal changes, we considered the situation ambiguous and did not assign an intermittent error to either instrument. The subsequent steps were as illustrated in the Figure. When the error was assigned only to the test instrument, troubleshooting was not performed, but rather, the study continued until (a) the error condition spontaneously disappeared; (b) the test instrument, through its own built-in error-detection system, signalled an error condition; or (c) a routine scheduled run of Confirm detected an out-of-control condition. When the Confirm data were out of limits on both instruments, the error was taken to be continuous (as opposed to intermittent) and was assigned to both instruments; in this event, the reference instrument was brought into control, but the test instrument was allowed to remain out of control until the error condition was signalled by other than discrepant between-instrument duplicate measurements.

Data Analysis

We kept complete records of all analyses, and recorded the serial numbers of the electrodes used throughout the study. For statistical analyses of these data, we used the BMDP statistical software programs (7).

Within-instrument differences. We calculated the within-instrument precision for blood specimens from the between-run differences obtained when duplicate analyses were done with the same instrument. We used only those data for which there was not a between-instrument error. The SD of the difference was calculated according to the equation: SD = \(\sqrt{\frac{\sum d^2}{2n}}\) (4). The paired t-test was used to assess whether for each analyte for each analyzer the mean difference between runs differed significantly from zero (\(p < 0.05\)).

Between-instrument differences. We calculated the between-instrument precision of analysis of blood, to evaluate whether our criteria for between-instrument errors were reasonable. For this we used the mean and SD of the difference between the instruments, after first eliminating outliers that exceeded three SDs from the initial mean. The paired t-test was used to assess whether differences between the two analyzers were significantly different from zero (\(p < 0.05\)).

We calculated the overall frequency of errors in specimens by dividing the number of specimens having errors in one or more analytes by the total number of specimens. To determine the frequency of errors in results for a given analyte, we divided the number of errors observed for that analyte by the total number of specimens. Using chi-square analysis (4), we compared both the total number of specimens with errors and the number of errors for each analyte in Part A with that in Part B.

We also examined the relation between the frequency of errors for each analyte and the specimen accession number, to see if errors were clustered into groups or dispersed randomly throughout the course of the study. For this, we divided our study into batches of 100 specimens and calculated the number of errors in each batch. We noticed clusters of errors that appeared to be related to specific electrodes, and therefore we used a chi-square analysis to compare the error rate for each electrode with that for all other electrodes considered together.

Clinical-correlation studies. Whenever we assigned an error to the test instrument, the results from the test and reference instruments were interpreted separately. We classified the acid-base status of each according to a modified version of the Goldberg nomogram (8). We called an error clinically significant if there was a difference in these acid-base classifications. For example, a \(PCO_2\) error that shifted the acid-base classification from metabolic alkalosis to acute respiratory alkalosis was called significant. For \(PO_2\), one of us (R.P.M.) determined whether the different \(PO_2\) values would have resulted in a change in patient management, such as a change in the prescribed concentration of inspired oxygen or a request for another specimen for blood-gas assay. We called the error clinically significant if such a change would have occurred.

Results

Within-instrument differences. We compared the means and SDs for Confirm for each of the analytes (pH, \(PCO_2\), \(PO_2\)) in each of the three formulations with the values obtained from all laboratories using the same lot of material and same model of analyzer, the latter data being provided by the manufacturer (Confirmation Statistical Analysis, Ciba-
Corning). We found our mean values to be within 2 SDs of the means of all laboratories and our SDs to be less than twice theirs.

The within-instrument precision for blood was similar for all instruments used in this study (see Table 1). Since the same Model 178 was used as the reference instrument in both parts A and B, the data obtained with this instrument were pooled. A small but statistically significant bias between the first and second runs was found for $p_{\text{CO}_2}$ on the test instrument of Part A and for $p_{\text{O}_2}$ on all instruments.

**Between-instrument differences.** Significant between-instrument differences in non-erroneous measurements indicate the presence of between-instrument biases that are extremely small (see Table 2).

The proportion of specimens for which duplicate analyses fell outside of the acceptable range was higher in Part A (6.1%) than in Part B (1.5%). This overall difference in the number of specimens with errors was significant ($p < 0.01$) by the chi-square test. The number of errors assigned to the five categories indicated in Figure 1 was for Part A: 1) 126, 2) 18, 3) 14, 4) 0, and 5) 0. For Part B the number of errors in each category was: 1) 35, 2) 0, 3) 1, 4) 0, and 5) 0.

For each of the analytes considered separately, there were significantly more errors in Part A than in Part B ($p < 0.01$) by the chi-square test.

These errors are shown in more detail in Figure 2a. We found 16 pH errors in Part A. Of the 10 assigned to the test instrument (error rate = 0.6%), eight were intermittent and two continuous. In Part B of the study no pH errors were found.

For $p_{\text{CO}_2}$, 76 errors were found in Part A (Figure 2b). Of these, 42 were assigned to the test instrument (error rate = 2.6%); 40 intermittent, 2 continuous. A total of 25 $p_{\text{CO}_2}$ errors were found in Part B. Of these, 14 were assigned to the test instrument (error rate = 0.9%): 13 intermittent, and 1 continuous.

When $p_{\text{O}_2}$ was examined, there were 84 errors in Part A (Figure 2c). Fifty-one were assigned to the test instrument (error rate = 3.2%): 41 intermittent, 10 continuous. A total of 11 $p_{\text{O}_2}$ errors were found in Part B. Nine of them, all intermittent, were assigned to the test instrument (error rate = 0.6%).

When error frequencies were related to specimen-accession number there were no peaks in the error rate for pH in either part of the study. For $p_{\text{CO}_2}$ (Figure 3, Panel I) there was an error peak of 8% between specimens number 201 and 300 in Part A. No error rates of 5% or over were observed in Part B.

In a similar manner, the distribution of $p_{\text{O}_2}$ errors relative to the specimen-accession number is shown in Figure 3, Panel II. In Part A there was an interval between specimens number 201 and 300 when there was a particularly large cluster of errors, resulting in an error rate of 18% for that interval. Furthermore, in the 1201 to 1400 range there is another peak, reaching a 5% error rate. In Part B, there was an error rate of 6% between specimens 1 and 100.

**Clinical-correlation studies.** The number of clinically significant errors in Part A were as follows: three of 10 pH errors, 26 of 42 $p_{\text{CO}_2}$ errors, and nine of the 51 $p_{\text{O}_2}$ errors. In all, 39 of the 103 errors (37%) attributed to the test instrument were clinically significant. In terms of the number of specimens for which there was an error for at least one analyte, 39 of the 97 specimens (40%) contained clinically significant errors. In Part B, five of the 14 errors in $p_{\text{CO}_2}$ and none of the nine errors in $p_{\text{O}_2}$ were clinically significant. That is, five of the 23 errors (22%) in Part B were clinically significant. For none of the specimens in Part B were there errors in results for more than one analyte, so the error frequency remains at 22% when considered in terms of specimens with one or more errors.

**Correlation between electrode use and error prevalence.** Two electrodes had a significantly greater number of errors than their cohorts. One was a $p_{\text{CO}_2}$ electrode from the test instrument in Part A ($p < 0.001$) that was in place during the high $p_{\text{CO}_2}$ error periods of Part A (specimens 201 to 300 and 1201 to 1400). The other was a $p_{\text{O}_2}$ electrode from the test instrument in Part B ($p < 0.005$) used during the high $p_{\text{O}_2}$ error period at the beginning of Part B (specimens 1 to 100). Three of four periods with error rates of 5% or greater were associated with the use of specific electrodes. Note, however, that the frequency of errors produced by these electrodes was variable, i.e., they were an intermittent source of error; they performed satisfactorily after the periods of high error prevalence.

### Table 1. Within-Instrument Differences: Blood-Gas Analyses

<table>
<thead>
<tr>
<th>Part A, Test Model (178)</th>
<th>Mean</th>
<th>SD</th>
<th>No. of specimens</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>0.0001</td>
<td>0.0021</td>
<td>129</td>
</tr>
<tr>
<td>$p_{\text{CO}_2}$ kPa</td>
<td>-0.05*</td>
<td>0.13</td>
<td>69</td>
</tr>
<tr>
<td>$p_{\text{O}_2}$ %</td>
<td>1.44*</td>
<td>1.76</td>
<td>73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part B, Test Model (178)</th>
<th>Mean</th>
<th>SD</th>
<th>No. of specimens</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>0.008</td>
<td>0.0034</td>
<td>34</td>
</tr>
<tr>
<td>$p_{\text{CO}_2}$ kPa</td>
<td>-0.03</td>
<td>0.07</td>
<td>9</td>
</tr>
<tr>
<td>$p_{\text{O}_2}$ %</td>
<td>0.85*</td>
<td>0.82</td>
<td>22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parts A &amp; B, Ref. Model (178)</th>
<th>Mean</th>
<th>SD</th>
<th>No. of specimens</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>0.0001</td>
<td>0.0022</td>
<td>143</td>
</tr>
<tr>
<td>$p_{\text{CO}_2}$ kPa</td>
<td>0.02*</td>
<td>0.14</td>
<td>82</td>
</tr>
<tr>
<td>$p_{\text{O}_2}$ %</td>
<td>1.48*</td>
<td>1.70</td>
<td>98</td>
</tr>
</tbody>
</table>

**Table 2. Between-Instrument Differences: Blood-Gas Analyses**

<table>
<thead>
<tr>
<th>Part A</th>
<th>No. of specimens</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>1530</td>
<td>-0.0012*</td>
<td>0.0043</td>
</tr>
<tr>
<td>$p_{\text{CO}_2}$ kPa</td>
<td>1518</td>
<td>0.02*</td>
<td>0.16</td>
</tr>
<tr>
<td>$p_{\text{O}_2}$ %</td>
<td>1541</td>
<td>0.48*</td>
<td>2.33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part B</th>
<th>No. of specimens</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>1465</td>
<td>0.0012*</td>
<td>0.0031</td>
</tr>
<tr>
<td>$p_{\text{CO}_2}$ kPa</td>
<td>1453</td>
<td>-0.06*</td>
<td>0.12</td>
</tr>
<tr>
<td>$p_{\text{O}_2}$ %</td>
<td>1437</td>
<td>-0.02</td>
<td>1.49</td>
</tr>
</tbody>
</table>
Discussion

In this study we addressed the question of how many errors in the measurement of blood-gas tensions and pH are detected by duplicate analysis and that were assigned to the test instrument. Intermittent errors are designated by solid lines, continuous errors by broken lines. The height of each line indicates the magnitude of one specific error.

Fig. 2. Distribution of errors in pH, pCO₂, and pO₂ that were detected by duplicate analysis and that were assigned to the test instrument. Intermittent errors are designated by solid lines, continuous errors by broken lines. The height of each line indicates the magnitude of one specific error.

Fig. 3. Time distribution of pCO₂ and pO₂ errors assigned to the test instrument throughout the study.

those aspects related to study design. Our laboratory procedure is typical of many laboratories in that we use three formulations of commercially prepared aqueous buffer as the quality-control material. We took particular care to designate one instrument the reference instrument and the other the test instrument. We did so because we intended to identify the true incidence of errors that would have occurred if duplicate-instrument analysis had not been performed. Thus troubleshooting was not performed on the test instrument, even if duplicate analysis suggested a malfunction, because this would have falsely decreased the number of errors detected by two-instrument analysis. Rather, we did troubleshooting only when a malfunction was signalled by other means, e.g., if the instrument was seen to be out of control when control material was analyzed or by the built-in error checks of the instrument. In the absence of duplicate analysis these are the only ways that would allow detection of instrument malfunction. In contrast, we felt it important to maintain optimal functioning of the reference instrument. Therefore this analyzer had troubleshooting performed whenever errors were detected, including those detected by duplicate-instrument analysis. Furthermore, we used the same Model 178 blood-gas analyzer as a reference in both Parts A and B, so that results in the two parts of the study could be compared.
We also had to decide, in planning the study, what would constitute an unacceptable difference between the duplicate analyses. Our criteria—0.02 pH units, ± 0.52 kPa, i.e., ± 4 mmHg for $p_{\text{CO}_2}$, ±7% for $p_{\text{O}_2}$—were based on previously suggested values (7). Comparison of the criteria established before the study with the within-instrument standard deviations that were actually found (Table 1) indicates that our limits were in fact very conservative. A large analytical error, exceeding three to four SDs, is required to exceed our limits; therefore it is unlikely that a result was called an error when in fact it was a valid measurement.

With this design we detected a relatively low overall prevalence of errors for each analyte, the highest error rates being for $p_{\text{CO}_2}$ (2.6%) and $p_{\text{O}_2}$ (3.2%), both in Part A of the study. Furthermore for all analytes, errors were more common with the older (Model 175) analyzer. Most errors were intermittent (89% of the total). Furthermore, the vast majority of these intermittent errors fell into category 1 of Figure 1, i.e., the error was not reproduced in the second analysis. This distribution emphasizes the fleeting nature of the intermittent error. Such errors will not be detected by means other than analysis of samples with two instruments.

Although the overall prevalence of errors was low there were, as mentioned, times during the study when the error rates were much higher. Error rates during these intervals can be disturbingly high, and their frequency may well impact on patient care. In considering this, however, it is important to consider not only the frequency of errors but also their magnitude. Thus one must define what magnitude of error to call clinically significant. This depends on the purpose of the analysis, e.g., monitoring changes in blood-gas tensions vs population screening. The former, which coincides with our purpose, requires greater accuracy and precision than does the latter (9). Rather than designating specific numerical limits to define the term "clinically significant," we chose a functional definition: if an error in pH or $p_{\text{CO}_2}$ resulted in a change in interpretation or if an error in measurement of $p_{\text{O}_2}$ would have resulted in a change in patient management, the error was called clinically significant. Thus, during periods of high error prevalence there may well be a high proportion of mistakes that adversely affect patient care.

We cannot definitely identify the cause of these periods of high error rate. We consider it likely, however, that they reflect malfunction of particular electrodes, because some electrodes had a significantly higher error rate than others (see Results). However, this does not represent definitive proof, because a high error rate owing to any cause would necessarily be attributed to some electrode. Moreover, these same electrodes were used during other parts of the study without producing unusually high error rates.

Although our study does not reveal the cause of the intermittent errors, it does indicate that even with modern analyzers and current quality-control techniques such errors do occur, as previously shown by Leary et al. (1). These authors reported that erroneous results were detected in 3.9% of specimens analyzed. However, our study design differed from theirs in two important respects. First, we designated one analyzer the test instrument so that a true prevalence of errors could be determined. More importantly, we analyzed not only the overall rate of errors but also how the rate varied during the study. Thus we identified periods of much higher error rate (up to 18%). This adds weight to the conclusion reached by Leary et al. that duplicate analysis is their "principal mechanism of random error control."

These data lead naturally to the question of whether concurrent analysis with two separate instruments should be made part of the routine quality-control procedure in the blood-gas laboratory. Before considering this it is prudent to examine whether our laboratory could be considered less well controlled than the typical laboratory. If this were true, then our experience might overestimate the error prevalence of others. We believe this not to be the case, for the following reasons. The technologists involved had all received in-house training and had passed annual written and practical examinations. Full-time technologists staff the day, evening, and night shifts, with part-time staff filling in for weekends, etc. This latter group was also extensively trained and met our examination requirements. Our analyzers are relatively modern; the Corning 175 being five years old and the 178 two years old. During the course of the study our analysis of quality-control material was similar to that for other similar instruments. The means of our analyses were within 2 SDs of the mean of all like instruments. Furthermore, our SDs were slightly smaller than the group deviations for all values for analyte, except for the high $p_{\text{CO}_2}$, for which our standard deviation was slightly larger. Thus the frequency of errors we experienced is not likely to be inflated in relation to other laboratories using similar equipment and similar control material. In fact, it may be that we experienced fewer errors than might occur in other laboratories, because we used a multi-rule quality-control program, which is potentially more sensitive in detecting errors than are other protocols. Furthermore, the frequency of errors noted in this study is probably lower than if we had included less-ideal samples, those with volumes less than 1.5 mL.

Because such duplicate analysis is the only procedure that will detect intermittent errors, it is a potentially important component of an optimal quality-control strategy. Obviously it will add to direct laboratory cost, but costs can be reviewed in a broader sense. In particular, what is the hidden cost of reporting an erroneous result? Will a decision negatively affecting patient care be made? Will a second specimen be drawn and analyzed? Nonetheless, in this time of concern about cost it seems inappropriate to simply add duplicate analysis to existing quality-control strategies. Rather, there is a need to consider the two major strategies—duplicate analysis and the use of quality-control material on a regular basis—and to develop an optimal combination of these that will lead to sufficient error detection at acceptable cost. Our study provides information as to the number of errors detected by dual-instrument analysis. We did not study the efficacy of duplicate analysis with a single instrument. However, Lunetsky et al. (10) evaluated this by computer simulation and showed unsatisfactory error-detection capabilities. A study similar to the present one is needed, to evaluate the utility of running quality-control material once per 8-h period, in a setting where duplicate analysis is and is not routinely performed. Then the results of the two studies could be used to objectively assess the efficacy of both types of quality-control strategies, and then to design an optimal quality-control protocol.

In conclusion, in the present study we have demonstrated that even in a laboratory with highly trained staff and with analyzers meeting quality-control standards, intermittent errors of clinical significance occur, at times frequently. Only duplicate-instrument analysis can detect intermittent errors, so it should be considered as a required routine.
procedure. However, there is a need to combine this in some fashion with other strategies to provide the optimal lowest cost approach to effective quality control in the blood-gas laboratory.

Many of our laboratory staff worked on this study. For this we thank all of the technologists involved in data collection and Lin Peters for the multiple retyping of this manuscript. Also, we are grateful to Robert F. Moran (Ciba-Corning) for many helpful suggestions.

References