CHEST®

Official publication of the American College of Chest Physicians

Patterns of Dissimilarities Among Instrument Models in Measuring Po_2, Pco_2, and pH in Blood Gas Laboratories

James E. Hansen and Richard Casaburi

Chest 1998;113;780-787
DOI 10.1378/chest.113.3.780

Patterns of Dissimilarities Among Instrument Models in Measuring Po_2, Pco_2, and pH in Blood Gas Laboratories*

James E. Hansen, MD, FCCP; and Richard Casaburi, PhD, MD, FCCP

Study objectives: **To ascertain the degree of dissimilarities among blood gas and pH analyzer models of the same and different manufacturers in measurement of Po_2, Pco_2, and pH using fluorocarbon containing emulsion (FCE) proficiency testing material.**

Design: **Statistically and graphically analyze data from six recent proficiency testing surveys for the 20 more frequently used models of analyzers.**

Setting and participants: **Over a 2-year period, approximately 900 participants from blood gas laboratories in the United States analyzed similar ampules from each of 30 lots.**

Measurements and results: **Both graphic and statistical comparisons were used to demonstrate differences between manufacturers. For each of the four major manufacturers, comparisons revealed statistically significant differences not only for Po_2, but also for Pco_2 and pH. Additionally, comparison models within each of the three manufacturers (those with multiple models and >15 instruments per model represented) disclosed statistically significant dissimilarities among models for each analyte in 115 of 153 model pairings. Previously reported tonometered blood differences among analyzer models for Po_2 are qualitatively similar to the differences found in these same models in this FCE study. Model differences are important in research studies and may be clinically important in deciding abnormality, selecting oxygen therapy, or the treatment of patients with respiratory failure or severe respiratory alkalosis.**

Conclusions: **To minimize the likelihood of misleading clinicians, laboratory directors should consider the degree of dissimilarity among blood gas analyzer models in current use and when changing instrumentation.** ***(CHEST 1998; 113:780-87)***

Key words: acid base equilibrium; ANOVA; blood gas analysis; carbon dioxide; oxygen therapy; quality control; respiratory failure

Abbreviations: AIM=all instrument mean; ANOVA=analysis of variance; D=statistically dissimilar; AVL=AVL Scientific Corporation; COR=Corning, Chiron Diagnostics Corporation; FCE=fluorocarbon containing emulsion; IL=Instrumentation Laboratories; N=statistically not dissimilar; RAD=Radiometer America, Inc.; V=statistically very dissimilar

It has been evident for more than a decade that different instrument models can yield model dependent results for Po_2, Pco_2, and pH.[1-11] This has been demonstrated using aqueous or fluorocarbon containing emulsion (FCE) proficiency testing materials as well as aliquots of tonometered blood. Logically, instruments that are technologically similar should reveal near-identical results for Po_2, Pco_2, and pH intensities along their entire range, from low to high values, whereas differences in electrodes, calibration, sample flow, or signal processing might cause two instrument models to give differing results.[8,12] We sought to determine the extent of these differences among the 20 most commonly used analyzers, whether made by the same or different manufacturers, by analyzing FCE proficiency testing data employing a wide range of analyte intensities from >900 instruments. This analysis leads to the conclusion that measurements made by blood gas analyzer models differ not only between manufacturers, but also within manufacturers.

Materials and Methods

Data Acquisition and Tabulation

Every 4 months, the American Thoracic Society-California Thoracic Society Proficiency Testing Survey sends out ampules of

*From the Department of Medicine, UCLA School of Medicine, Division of Respiratory and Critical Care Physiology and Medicine, Harbor-UCLA Medical Center, Torrance, Calif.

Manuscript received May 6, 1997; revision accepted September 10.

Reprint requests: James E. Hanson, MD, FCCP, 1692 Morse Drive, San Pedro, CA 90732

identical composition from five lots of FCE proficiency testing material[13] to approximately 400 laboratories with >900 enrolled instruments. Within a month, data are received from each enrolled instrument. From these responses, the mean and SD for each lot for each model are determined. Results are reported back to enrollees and, as necessary, to governmental agencies. We performed a retrospective analysis of these data for six successive periods from late 1994 to mid 1996, in which all instrument mean (AIM) lot values ranged from about 40 to 180 mm Hg for Po_2, 20 to 75 mm Hg for Pco_2, and 7.15 to 7.60 for pH units (U).

We tabulated the mean values for Po_2, Pco_2, and pH values from each of the 30 lots for each of the manufacturers with the largest number of participating instruments (AVL Scientific Corporation, [AVL], Roswell, Ga; Corning, Chiron Diagnostics Corp [COR], Medfield, Mass; Instrumentation Laboratory [IL], Lexington, Mass; and Radiometer America Inc [RAD], Westlake, Ohio) and calculated their positive or negative deviations from AIM values. To ascertain if the analyses differed among manufacturers, we prepared three graphs, one for each analyte, with AIMs of each lot on the abscissa and the means of the lot deviations from the AIMs for each of the four manufacturers on the ordinate. Three of the four manufacturers had several models represented by ≥15 instruments in the database: COR with 7 models; IL with 6; and RAD with 6. To determine whether analyses differed among models made by the same manufacturer, after calculating model deviations from AIM values, we prepared nine additional graphs, one for each analyte for each of these three manufacturers, with AIMs of each lot on the abscissa and lot deviations from the AIMs for each of their models on the ordinate.

Statistical and Graphic Analyses

We performed statistical analyses to seek differences among manufacturers and among models made by a specific manufacturer. Three analyses of variance (ANOVAs) were performed (one for each analyte) to determine whether there were systematic differences among manufacturers in their deviations from the AIM for each of the 30 lots. Nine additional ANOVAs were used (three analytes times three manufacturers) to compare the 30 lot values of six or seven models of each of the three manufacturers with their own manufacturer mean lot value. Scheffé tests were used to define the probability of significant differences between specific manufacturers or models. Manufacturers or models were graded as follows: V="very dissimilar" if they differed with $p<0.001$; D="dissimilar" if they differed with $p<0.05$ to >0.001; and N="not dissimilar" if they differed with $p>0.05$.

Review of the graphs suggested that for some of the nine latter ANOVAs, some of the N grades were not justified because the model patterns along the full range of intensities were visually very dissimilar. For example, model A might give consistently lower values than model B at low intensities and consistently higher values than model B at high intensities. Because of such visually apparent "crossover" patterns, we calculated four more ANOVAs, each including all six or seven models of that manufacturer, but using only the upper half of the range of values (including 15 rather than 30 lots).

Next, the 20 most commonly used models from the three 1996 FCE surveys were selected to compare the actual deviations for low, medium, and high intensities of Po_2, Pco_2, and pH to demonstrate the actual differences among models observed. First we tabulated the range of differences for 7 models; then we counted the number of model pairs for all 20 models that exceeded the following arbitrary limits: 4 mm Hg, 6 mm Hg, and 6% for Po_2; 3 mm Hg and 3% for Pco_2; and 0.030 U for pH at these same intensities.

Finally, the FCE data were compared with tonometered blood data of Scuderi et al,[11] who reported the differences among four blood gas analyzers (AVL995, COR178, IL1312, and RAD330) by measuring tonometered blood at 16 levels of Po_2 (their Table 3). They did not report Pco_2 or pH values. From their data, we selected the 11 blood Po_2 levels that we could reasonably match with 11 survey FCE lot levels (47 vs 43.4, 56 vs 57.4, 66 vs 65.1, 75 vs 78.1, 85 vs 86.4, 94 vs 94.4, 104 vs 108, 113 vs 111.5, 122 vs 123.1, 142 vs 152.3, and 189 vs 175.5 mm Hg). The deviations in mm Hg of each of these four analyzer models at each level of tonometered blood (their data) were paired with FCE deviations (our data) of the same models (44 pairs) and regression analysis was performed to compare the model biases of blood and FCE.

Results

Comparisons Between Manufacturers

The three graphs comparing manufacturers with each other for each analyte (Figs 1-3) visually demonstrate the differences between manufacturers along the complete range of intensities. Table 1 indicates the degree of the statistical dissimilarity between manufacturers for each analyte using the three traditional ANOVAs. For the 18 paired analyte comparisons, only the AVL and RAD mean Po_2 values are not graded V. In Figure 1, however, even for this comparison, their visual dissimilarity can be seen, because a "crossover" pattern occurs with AVL deviations consistently higher than RAD values at Po_2 intensities below 80 mm Hg, and consistently lower than RAD deviations at Po_2 intensities above 110 mm Hg. Seeing this pattern, we performed an unscheduled ANOVA using the 10 lots with Po_2 values below 79 mm Hg. Scheffé tests showed that all four manufacturers produced V results for these 10 lots, including the RAD and AVL comparison.

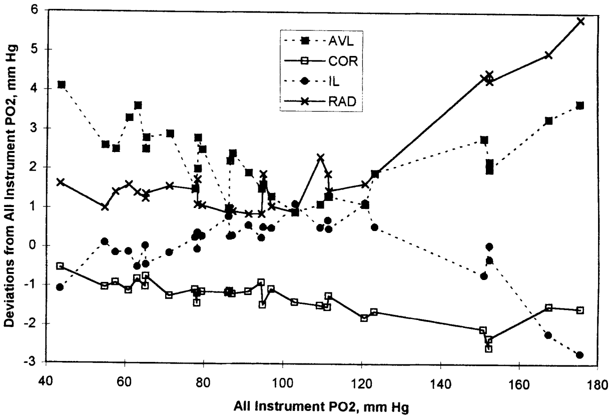

FIGURE 1. Po_2 measures of FCE by blood gas analyzers of four manufacturers are compared. Mean values are derived from approximately 75 AVL analyzers, 360 COR analyzers, 300 IL analyzers, and 180 RAD analyzers for each of 30 lots. Deviations from each lot AIM values are on the ordinate. Each pattern is different by ANOVA. See the first paragraph of the "Results" section regarding the statistical confirmation of these differences.

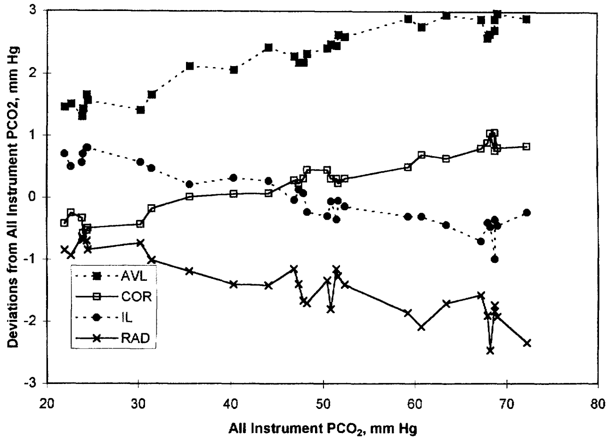

FIGURE 2. PCO_2 measures of FCE by blood gas analyzers of four manufacturers are compared. Values are derived from the same lots and analyzers as in Figure 1. Differences between manufacturers range from 2.5 mm Hg at low PCO_2 intensities to 5.5 mm Hg at high PCO_2 intensities. Each pattern is different visually and statistically.

For PCO_2 comparisons (Fig 2), AVL deviations are consistently positive, RAD deviations are consistently negative, and the COR and IL patterns cross. For pH comparisons (Fig 3), the AVL and IL deviations are consistently negative and different while the COR values are consistently slightly more positive than the RAD values. Thus, differences between each manufacturer for all analytes are evident both graphically and statistically.

The nine graphs from three manufacturers were studied carefully. To demonstrate the importance of graph analyses, three simplified graphs (with only four instruments per graph for clarity) were selected (Figs 4-6). They illustrate dissimilarities among the models of each manufacturer for each of these three analytes.

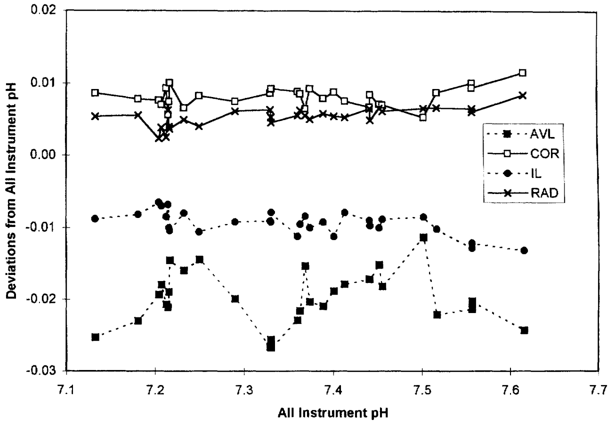

FIGURE 3. pH measures of FCE by blood gas analyzers of four manufacturers are compared. Values are derived from the same lots and analyzers as in Figure 1. The range of differences between manufacturers is approximately 0.03 U over the entire pH range. Each manufacturer's mean values are quite consistently related to each other. Each pattern is different visually and statistically, even for the COR and RAD models. Contrast this consistency with the COR 238 and 278 model patterns shown in Figure 6.

Table 1—*Dissimilarity Between Blood Gas Analyzers by Manufacturer**

	COR	IL	RAD
AVL	VVV	VVV	NVV
COR		VVV	VVV
IL			VVV

*Analyses are for 30 lots of FCE proficiency testing materials analyzed by approximately 75, 360, 300, and 180 instruments for AVL, COR, IL, and RAD models, respectively. By initial ANOVA, differences between manufacturers are graded as follows: V="very dissimilar" if they differ by $p<0.001$; D="dissimilar" if they differ by $p<0.05$ to >0.001; and N="not dissimilar" if they differ by $p>0.05$. In each series of three letter grades, the first letter denotes PO_2, the second letter denotes PCO_2, and the third letter denotes pH comparisons between manufacturers. The single "N" grade found in the initial ANOVA comparing PO_2 for AVL vs RAD should be replaced by a "V" grade because of the "crossover" pattern seen in Figure 1 and a subsequent ANOVA confined to lower PO_2 intensities.

Comparisons Within Manufacturers

For the three manufacturers, statistical comparisons were made comparing each model with other models of that manufacturer. Nine ANOVAs were performed, each revealing statistically significant differences. The subsequent Scheffé tests yielded 53 N grades, 15 D grades, and 85 V grades. Thus, the mean values for nearly two thirds of the paired comparisons revealed a statistical model dissimilar-

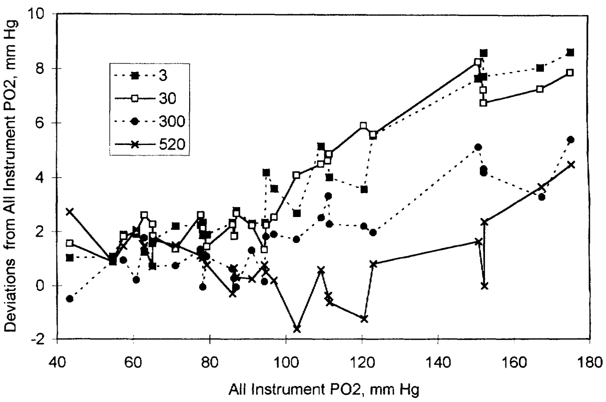

FIGURE 4. Four RAD models compared for PO_2. Fifteen to 41 analyzers of each model determine each deviation from the AIM (representing >900 instruments) values for 30 lots of FCE proficiency testing materials for PO_2 values from 43 to 175 mm Hg. The 520 model diverges from the relatively similar 3 and 30 models and dissimilar 300 model as PO_2 intensity increases above 90 mm Hg.

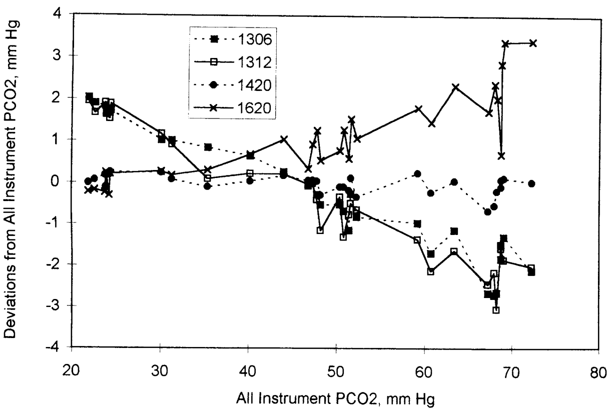

FIGURE 5. Four IL models compared for $P\text{CO}_2$. Some models demonstrate a "crossover" pattern. Thirty to 100 analyzers of each model determine each deviation from the AIM (representing >900 instruments) values for 30 lots of FCE for $P\text{CO}_2$ range of 22 to 72 mm Hg. The 1306 and 1312 models are visually similar, but the 1420 model is clearly different. However, this difference is not statistically significant using data from the full range of $P\text{CO}_2$ values for the ANOVA, because the *mean* deviations for the 1306 and 1312 models are only 0.2 to 0.4 mm Hg, respectively, different from the 1420 models. When the ANOVA is restricted to the upper half of $P\text{CO}_2$ intensities, it confirms that the 1420 model is V from the 1306 and 1312 models, with mean deviations of 3.1 and 3.3 mm Hg, respectively, from the 1420. The 1620 model is visually different and V from the other three models.

ity. After adding the results from the four additional ANOVAs (used because of visually apparent near mid-range graphic crossing), there were 38 N grades, 18 D grades, and 97 V grades, indicating an even greater incidence of dissimilarity than recognized by the earlier ANOVAs. Examples of the necessity for adding graphic review to the initial ANOVAs are given in Figs 4-6 and their legends.

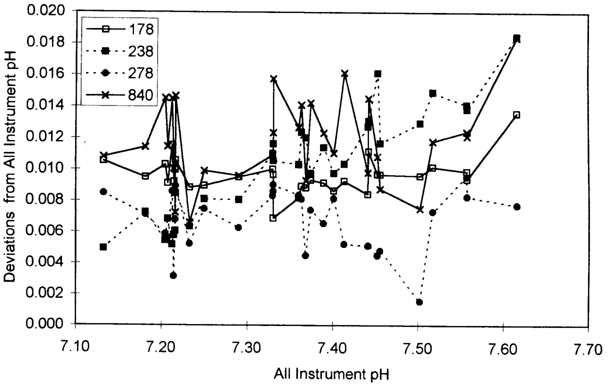

FIGURE 6. Four COR models compared for pH. Eighteen to 170 analyzers of each model determine each deviation from the AIM (representing >900 instruments) values for 30 lots of FCE for pH range of 7.132 to 7.616. The 278 model values are visually lower than the other three models by <0.01 U on average, but are V from them. The 178 and 840 models are minimally dissimilar. The 238 model values visually "crossover" the 178 and 840 values. Using the full range of values for ANOVA, the 238 does not statistically differ from the 178 and 840 models. However, using pH values above 7.40 for ANOVA, the 238 model is V from the 178 model.

Table 2 shows statistical comparisons among models of the same manufacturer. For the seven COR models, only the 278 model is not dissimilar from the 288 model for all three analytes. The 170 and 178 models are not dissimilar for $P\text{CO}_2$ and pH, nor are the 238 and 840 models. Otherwise all models differ from each other for at least two analytes. Comparing all COR models for $P\text{O}_2$, average differences between models are <1 mm Hg for seven comparisons, 1 to 3 mm Hg for six comparisons, 3 to 5 mm Hg for seven comparisons, and 5 to 6 mm Hg for one comparison. For the full range of $P\text{CO}_2$, average differences between models are <1 mm Hg for 19 comparisons and 1 to 2 mm Hg for 2 comparisons; while at $P\text{CO}_2$ values above 40 mm Hg, average differences are <1 mm Hg for 11 comparisons, 1 to 2 mm Hg for 8 comparisons, and 2 to 3 mm Hg for 2 comparisons. For pH, average differences between models are <0.001 U for 4 comparisons, 0.001 to 0.003 U for 10 comparisons, 0.003 to 0.005 for 5 comparisons, and 0.005 to 0.006 for 2 comparisons.

Statistically, none of the six IL models measure all three analytes without statistically significant differences. The closest matches are the 1304, 1306, and 1312 models that are not dissimilar for two analytes. The 1400 and 1420 models are not dissimilar for $P\text{O}_2$ and $P\text{CO}_2$. The 1620 models differs from the other five models for all three analytes. Comparing all IL models for $P\text{O}_2$, average differences between models are <1 mm Hg for 11 comparisons and 1 to 3 mm Hg for 4 comparisons. For the full range of $P\text{CO}_2$, average differences between models are <1 mm Hg for 10 comparisons and 1 to 2 mm Hg for 5 comparisons; while at $P\text{CO}_2$ values >40 mm Hg, average differences are <1 mm Hg for 6 comparisons, 1 to 3 mm Hg for 7 comparisons, and 3 to 4 mm Hg for 3 comparisons. For pH, average differences between models are <0.001 U for three comparisons, 0.001 to 0.003 U for one comparison, 0.003 to 0.005 U for four comparisons, 0.005 to 0.007 U for five comparisons, and 0.007 to 0.014 U for two comparisons.

Statistical comparison of the six RAD models shows that the 30 and 3 models are not dissimilar for any analytes. The same is true of the 300 and 330. The latter models are also not dissimilar from the 500 model for $P\text{O}_2$ and $P\text{CO}_2$. The 500 and 520 models are not dissimilar for $P\text{CO}_2$ and pH. The two models are statistically dissimilar for $P\text{O}_2$ only above 80 mm Hg, where the average $P\text{O}_2$ differences are <1.5 mm Hg. Comparing all RAD models for full range $P\text{O}_2$, average differences between models are <1 mm Hg for six comparisons and 1 to 3 mm Hg for nine comparisons; while for $P\text{O}_2$ levels >90 mm Hg,

Table 2—*Dissimilarity Between Models of Blood Gas Analyzers by Manufacturer**

	Chiron Corning Models					
	178	238	278	280	288	840
170	DNN	VDD	VDV	VVV	VDD	VND
178		VNV	VVV	VVV	VVN	VND
238			VVV	VVV	VVV	VNN
278				DVN	NNN	VVV
280					DVD	VVV
288						VVV

	Instrumentation Laboratory Models				
	1306	1312	1400	1420	1620
1304	DNN	NNV	VDV	VVN	VVV
1306		NNV	VDV	VVN	VVV
1312			VDV	VVV	VVV
1400				NNV	VVV
1420					VVV

	Radiometer Models				
	30	300	330	500	520
3	NNN	VVD	VVN	VVV	VVV
30		VVN	VVN	VVV	VVV
300			NNN	NNV	VND
330				NNV	VDV
500					VNN

*Model pairs are graded as in Table 1 footnotes. In each series of three letter grades, the first letter denotes Po_2, the second letter denotes Pco_2, and the third letter denotes pH comparisons between models of that manufacturer.

differences were <1 mm Hg for four comparisons, 1 to 3 mm Hg for eight comparisons, and 3 to 5 mm Hg for three comparisons. For Pco_2, average differences between models are <1 mm Hg for 13 comparisons and 1 to 2 mm Hg for 2 comparisons. For pH, average differences are <0.001 U for four comparisons, 0.001 to 0.003 U for four comparisons, 0.003 to 0.005 U for two comparisons, and 0.005 to 0.007 U for five comparisons.

Table 3 endeavors to illustrate representative variations that might be expected among models for each of three analytes. It can be noted that the range of values (*ie,* the difference between the model with the highest mean value and the model with the lowest mean value) obtained by these seven models at low, intermediate, and high intensities averaged 8.8 mm Hg for Po_2, 4.4 mm Hg for Pco_2, and 0.033 U for pH.

Table 3—*Deviations From AIM Values for Seven Models for Three Lots of Proficiency Testing Materials**

		Models							
	AIM	AVL995	COR178	COR280	IL1312	IL1620	RAD30	RAD520	Range†
Po_2, mm Hg	43.4	+4.3	−4.1	+1.0	+0.4	−2.4	+1.6	+2.7	8.4
Po_2, mm Hg	78.6	+2.6	−4.1	+0.4	+1.0	−1.3	+2.7	+0.4	6.8
Po_2, mm Hg	167.5	+3.5	−3.8	+1.2	−3.5	−2.2	+7.3	+3.7	11.1
Average range									8.8
Pco_2, mm Hg	22.6	+2.7	−0.5	−0.9	+1.7	−0.2	−0.4	−0.9	3.6
Pco_2, mm Hg	47.5	+1.6	+0.9	−1.1	−0.4	+1.0	−2.6	−1.2	4.2
Pco_2, mm Hg	67.9	+2.0	+2.0	−1.1	−2.2	+2.4	−2.9	−2.0	5.3
Average range									4.4
pH, U	7.207	−0.018	+0.009	+0.004	−0.002	−0.014	+0.008	+0.002	0.026
pH, U	7.377	−0.019	+0.010	+0.005	−0.005	−0.015	+0.009	+0.003	0.034
pH, U	7.616	−0.025	+0.013	+0.010	−0.006	−0.020	+0.008	+0.008	0.038
Average range									0.033

*One lot from each period in 1996.

†Difference between the models with the highest and lowest values.

Table 4 shows the number of model pairs out of the possible 190 model pairs that differed by arbitrary limits for each analyte at each level. For $\mathrm{Po_2}$, 168 of 570 model pairs differed by ≥4 mm Hg; 66 of 570 model pairs differed by ≥6 mm Hg; and 92 of 570 model pairs differed by ≥6.0%. For $\mathrm{Pco_2}$, 86 of 570 model pairs differed by ≥3 mm Hg, while 123 of 570 model pairs differed by ≥3.0%. For pH, 37 of 570 model pairs differed by ≥0.030 U.

Comparing the tonometered blood[11] and FCE biases for $\mathrm{Po_2}$ measurements, the maximum range of model biases at 11 levels for blood was 9 mm Hg and for FCE was 9.5 mm Hg. The slope (bias) of the FCE deviations regressed against blood deviations was 0.81, indicating that, on average, each 0.81 mm Hg of difference in blood measures between models might be associated with a 1.0 mm Hg difference in FCE measures between models. The correlation coefficient was 0.629 with $p<0.001$. Therefore, it is likely that our detection of differences among models when measuring FCE indicates that the differences among models when measuring blood exists as well.

Discussion

The clinician expects that laboratory analyses of clinical specimens will be accurate and reproducible. This study explores one source of analytic error in blood gas analysis: systematic differences among the results provided by commercially manufactured models of blood gas analyzers. In analyzing FCE proficiency testing material in a large number of instruments over a wide range of analyte values, we detected highly significant differences not only among manufacturers but also among models of a given manufacturer.

Our initial analyses were directed toward detecting differences among manufacturers (Figs 1-3). Having found appreciable and consistent differences among manufacturers, we expected to find only small differences among models of the same manufacturer. This often was not the case.

Usefulness of Graphic Displays

Our initial analysis strategy involved seeking differences in the mean values produced by a number of instruments of a given model across a range of analyte values utilizing ANOVA. However, we found this strategy failed to detect many appreciable differences. Specifically, when two models exhibit a "crossover" pattern of mean model values as a function of the analyte values, ANOVA may fail to detect a significant difference, because consistently positive differences over half of the analyte range cancel out consistently negative differences over the other half. Therefore, we supplemented our statistical analyses with graphic analyses.

When a crossover pattern was visualized graphically, the portion of the data sets where the divergences were seen (*eg*, the upper or lower portion) was used for a second ANOVA. This often statistically confirmed the visually apparent differences between models. This analysis strategy increased the incidence of statistically significant model differences from 65% (100/153) to 75% (115/153). Although we have presented mean differences from the full range of values in the "Results" section, it is clear that such comparisons may sometimes minimize real model differences in one portion of the analyte range, as noted in Table 3 and seen in Figs 4-6.

Limitations and Advantages

Do data analyses based on FCE ampules measured in a proficiency testing survey validly reflect

Table 4—*Number of Pairs of Models Differing by Specific Amounts**

Measure	Mean Lot Value				
		Minimal difference:	6 mm Hg	4 mm Hg	6%
$\mathrm{Po_2}$, mm Hg	43.4		10	41	62
$\mathrm{Po_2}$, mm Hg	78.6		5	31	18
$\mathrm{Po_2}$, mm Hg	167.5		51	94	12
		Minimal difference:	3 mm Hg		6%
$\mathrm{Pco_2}$, mm Hg	22.6		9		65
$\mathrm{Pco_2}$, mm Hg	47.5		3		10
$\mathrm{Pco_2}$, mm Hg	67.9		74		48
		Minimal difference:	0.030 U		
pH, U	7.207		5		
pH, U	7.377		5		
pH, U	7.616		27		

*Twenty models assessed; 190 comparisons at each level.

model differences for blood that might be reflected in clinical practice? Currently no proficiency testing material available for shipment is equivalent to freshly tonometered fresh human blood (see below). Another possible disadvantage is that the measurements are made by hundreds of technicians with differing experience and training using hundreds of different instruments. As previously noted,[14] laboratories with better quality control, more equipment, more frequent analyses, and dedicated personnel are likely to have less imprecision and less inaccuracy in their analyses. Advantages of using this database are as follows: (1) it is unlikely that the distribution of any one model of analyzer is concentrated at high or low altitudes or "better" or "poorer" laboratories; (2) the differences between technician practices tend to cancel out when a large number of technicians are used; (3) a large amount of data can be collected and analyzed uniformly; (4) the inherent differences between models tend to be more evident when large numbers of analyses are made; (5); it is unlikely that any single site would have this diversity of operating analyzer models available at any one time; and (6) the infectious problems associated with the handling of large quantities of blood are avoided.[15] Certainly, the finding of consistent deviations between models using any single type of proficiency testing material is strong evidence that the models actually differ in some way from each other.

Relevance to Blood Measurements

With minor quantitative changes, we believe our findings are relevant to clinical and research blood analyses. Because aqueous and FCE proficiency testing materials differ from fresh human blood in viscosity, oxygen capacity and content, oxygen half-saturation pressure of hemoglobin values, and temperature dependence,[13,16] they can be expected to differ from blood, but FCE has shown much less variability than aqueous materials in measuring Po_2.[4,8,17] Two studies using fresh tonometered blood at several Po_2 intensities found minimal model differences between a few instruments.[9,18] In contrast, four studies using fresh human blood demonstrated major model differences in measuring Po_2 at several intensities.[5,10,11,17] Our comparison of tonometered blood data from the four instruments used by Scuderi et al[11] with FCE data from approximately 150 instruments of the same models discloses that blood deviations averaged 81% of the FCE deviations we detected in the same models at 11 Po_2 levels. If we combine data from two of our own studies,[10,17] using multiple models and seven levels of Po_2 intensity, blood deviations averaged 77% of FCE deviations between Po_2 of 42 and 92 mm Hg and 100% for Po_2 values of 145 mm Hg. Thus, one can estimate that fresh human blood Po_2 deviations below the hyperoxic range are likely to be roughly four-fifths of those for FCE.

For Pco_2 and pH comparisons, the differences between proficiency testing materials and fresh human blood have been studied less but are probably smaller, because both aqueous and FCE materials are stable and have similar buffering capacities. Blood is not a primary standard nor a good quality control material for pH, due to its inconvenience, infectivity, and variations in buffering capacity. One study[7] using only two models of analyzers over a wide range of Pco_2 showed blood deviations that were about 60% of those of ampules of several quality control materials. In contrast, another study[10] using several analyzer models showed that blood deviations slightly exceeded FCE deviations for four Pco_2 intensities from 22 to 65 mm Hg. Because Pco_2 electrodes are modified pH electrodes, one would suspect that the pH deviations for blood are reasonably similar to those of either FCE or buffered aqueous materials. Because proficiency testing ampules can be manufactured in huge quantities and are stable over long periods, either aqueous or FCE materials seem preferable to blood and acceptable for comparing Pco_2 and pH values between instruments and models. In the absence of other data, it seems reasonable to conclude that for Pco_2 and pH measurement, FCE instrument and model differences are likely to be between 1⅓ and ⅔ of those for blood.

The finding that models of the same manufacturer usually differ significantly from each other for two or more analytes (Table 2) initially was surprising to us. However, these model differences are likely due to continuing improvements in analyzer geometry, calibration and flushing techniques, temperature control, electronic signal modification, and other unknown factors. Manufacturers can be expected to continue to upgrade and introduce new instrument models, in order to decrease inaccuracy, decrease imprecision, decrease sample size, decrease instrument and technician errors, increase speed of analysis, and improve ease of quality control. Such improvements may have increased the quality and ease of blood gas and pH measurements, but currently significant differences between models and between manufacturers continue.

Clinical Importance of These Findings

How important are these model differences in research or clinical practice?[19] The difference between models often exceeds twice the standard deviation (SD) of a single model, and especially the SD of a single instrument. (The average SDs for the 20 models at all participating laboratories at the three analyte levels are 2.5 mm Hg for Po_2, 1.3 mm Hg for Pco_2, and 0.008 for pH; historically SDs are even less for individual instruments.[10,11]) In research, it is obviously unwise for an investigator to shift between models without first ascertaining their comparability. In clinical practice, the answer to relevance depends on whether one is distinguishing between normality and abnormality, defining the degree of abnormality, or making changes in therapy. Expressed on an absolute basis by referring to Table 4, numerically higher measures are most likely and middle measures are least likely to be affected by model differences; expressed on a percentage basis, both low Po_2 and Pco_2 values are most likely to be affected by model differences. It would be comforting, but unwise, to ignore these model differences. Differences in or unrecognized alterations in laboratory instrumentation could importantly influence clinical decisions regarding the following: (1) the presence of degree of impaired oxygenation in disability evaluation; (2) the necessity for oxygen supplementation; and (3) the management of respiratory failure or severe respiratory alkalosis.

These manufacturer and model differences suggest that laboratory directors should consider the similarities and differences between models when reporting data and also when retiring older instruments and adding new instruments to their laboratories. When doing so, the laboratory director can find specific information on probable model differences by examining the complete reports of their recent proficiency testing surveys. Otherwise clinicians may be misled by the values they receive in the same institution from different instruments.[19]

ACKNOWLEDGMENT: The writers thank Lauren Malonis, Dr. Robert L. Jensen, and other American Thoracic Society and California Thoracic Society committee members and staff for providing access to the database of the American Thoracic Society-California Thoracic Society Blood Gas Proficiency Testing Survey.

References

1 Rej R, Vanderlinde RE. Proficiency testing in acid-base analyses: an interlaboratory evaluation. Clin Chim Acta 1973; 49:161-67

2 Itano M. CAP blood gas survey—first year's experience. Am J Clin Pathol 1980; 74:535-41

3 Clausen JL, Hansen JE, Misuraca L, et al. Interlaboratory comparisons of blood gas measurements [abstract]. Am Rev Respir Dis 1981; 123(Suppl):104

4 Hansen JE, Clausen JL, Mohler JG, et al. Blood gas proficiency testing materials: a multilaboratory comparison of aqueous solution and a fluorocarbon-containing emulsion. Clin Chem 1982; 28:1818-20

5 Sutt-Corbett B, Fonzi C. Instrumental biases in blood gas analysis of tonometered whole blood. Clin Chem 1982; 28:550-52

6 Itano M. CAP blood gas survey—1991 and 1982. Am J Clin Pathol 1983; 80(suppl):554-62

7 Burnett D, Henfrey RD, Woods TF, et al. Regional quality assessment of pH and blood gas analyzers. Ann Clin Biochem 1986; 23:26-36

8 Hansen JE, Clausen JL, Levy SE, et al. Proficiency testing materials for pH and blood gases: the California Thoracic Society experience. Chest 1986; 89:214-17

9 Van Kessel AL, Eichhorn JH, Clausen JL, et al. Interinstrument comparison of blood gas analyzers and assessment of tonometry using fresh heparinized whole human blood. Chest 1987; 92:418-22

10 Hansen JE, Jensen RL, Casaburi R, et al. Comparison of blood gas analyzer biases in measuring tonometered blood and a fluorocarbon-containing proficiency-testing material. Am Rev Respir Dis 1989; 140:403-09

11 Scuderi PE, MacGregor DA, Bowton DL, et al. Performance characteristics and interanalyzer variability of Po_2 measurements using tonometered human blood. Am Rev Respir Dis 1993; 147:1354-59

12 Holbek CC. The Radiometer ABL300 blood gas analyzer. J Clin Monit 1989; 5:4-16

13 Feil MC, Cormier AD, Legg KD. Perfluorocarbon emulsions as pH/blood-gas controls. Clin Chem 1982; 28:2187-88

14 Hansen JE. Participant responses to blood gas proficiency testing reports. Chest 1992; 101:1240-44

15 Protection of laboratory workers from infectious disease transmitted by blood, body fluids, and tissue. 2nd ed. Tentative guideline, publication M29-T2. Villanova, Pa: National Committee for Clinical Laboratory Standards, 1991

16 Ong ST, David D, Snow M, et al. Effects of variations in room temperature on measured values of blood gas quality-control materials. Clin Chem 1983; 29:502-05

17 Hansen JE, Feil MC. Blood gas quality control materials compared to tonometered blood in examining for interinstrument bias in Po_2. Chest 1988; 94:49-54

18 Hansen JE, Stone ME, Ong ST, et al. Evaluation of blood gas quality control and proficiency testing materials by tonometry. Am Rev Respir Dis 1982; 125:480-83

19 Clausen JL, Murray KM. Clinical applications of arterial blood gases: how much accuracy do we need? J Med Technol 1985; 2:19-21