Bias and Precision of Pulse Oximeters and Arterial Oximeters*

Bruce C. Nickerson, M.D., F.C.C.P.; Catherine Sarkisian; and Kevin Tremper, M.D.

We evaluated measurements of oxygen saturation from four noninvasive pulse oximeters, and two conventional arterial oximeters. Simultaneous measurements were obtained on each instrument in three different occasions in five healthy subjects breathing gas with an FIO2 of 1.00, 0.50, 0.21, 0.17, 0.15, 0.13 and 0.11. Mean bias relative to the sum of oxy-, carboxy-, and methemoglobinbs ranged from −0.4 to −2.6 percent for the pulse oximeters and +1.8 to −4.2 percent for the conventional oximeters. Two pulse oximeters performed well at all saturations down to 65 percent, while the others showed either increasing or decreasing bias below 80 percent saturation. Precision was approximately 2 percent for all instruments except one conventional oximeter with a precision of 0.7 percent. In the clinically relevant range, the performance of the noninvasive pulse oximeters was similar to conventional oximeters using arterial samples.

The purpose of this study was to evaluate the performance of pulse oximeters using statistical techniques that would detect clinically important differences between measurements taken with different instruments. Previous studies* report the agreement between pulse oximeter readings and saturations measured on arterial blood in terms of a correlation coefficient for a least-squares linear regression by reporting a slope, intercept, and a correlation coefficient. Altman and Bland† have emphasized the pitfalls of linear regression analysis in evaluating a new technique of measurement. They note that devices with unacceptable properties such as always reading too low, nevertheless may have good correlation coefficients. They suggest that bias and precision (mean and standard deviation of the differences between measurements made with a new method and the accepted standard) are preferable to linear regression for evaluating a new technique.

Most data from previous studies are for saturations above 90 percent and very few saturations below 80 percent have been reported. Using regression analysis, previous studies found slopes significantly different from one and intercepts different from zero, indicating that pulse oximeter measurements were not identical to arterial saturations.

Reproducibility is also an important property in evaluating a new test.‡ Clinicians should know the range of values observed under identical circum-

*From the Department of Pulmonary Medicine, Children’s Hospital of Oakland, and Department of Anesthesia, University of California, Irvine.

†Supported by grants from the Novametrix Corporation, the Fund for Pediatric Care and Research, and the Northern California Sudden Infant Death Syndrome Foundation.


§Manuscript received May 23; revision accepted August 11.

Reprint requests: Dr. Nickerson, Pulmonary Function Laboratory, 747 52nd Street, Oakland 94609

Material and Methods

The Human Subjects Review Board of Children’s Hospital of Oakland gave approval for this study. Five 31- to 34-year-old nonsmoking men free from lung, heart, or blood disease gave informed consent. All completed the test and none complained of excessive discomfort during the procedure.

After local anesthesia, a 20-gauge radial arterial catheter was placed in the nondominant wrist for collection of arterial blood samples. Four different pulse oximeters (Biox/Ohmeda 3700, Boulder, CO, designated as A; Criticare CR501 Critical Care Systems, Inc., Milwaukee, WI, designated as B; Nellcor N-100, Hayward, CA, designated as C; Novametrix 500, Wallingford, CT, designated as D) were placed on the four fingers of the dominant hand. Each oximeter was used according to the manufacturer’s specifications with models available at the time of the study (March, 1986). Readings were taken only when all oxygen saturation readings were stable and the oximeters showed a strong pulse that agreed with a simultaneous ECG monitor.

The subjects were challenged with a succession of concentrations of inspired oxygen: 0.21, 0.17, 0.15, 0.13, 0.11, 0.15, 0.21, 0.50, 1.00 and 0.15. In order to establish equilibrium, five minutes of breathing each concentration was allowed before measurements were taken. The inspired and expired concentrations of oxygen were monitored with a mass spectrometer (McGaw RMS-3, San Marcos, CA).

At each FIO2, three arterial blood samples were taken at least 1 minute apart. The pulse oximeters were read as the arterial sample was being withdrawn. The blood samples were immediately placed on ice and analyzed within 10 minutes. We measured arterial blood gas tensions and calculated oxygen saturation, (Corning Medical Instruments Model 168, Ciba Corning Medical Instruments, Med-
hemoglobin, methemoglobin, and reduced hemoglobin were measured (Instrument Laboratories, Lexington, MA IL282). The IL282 is a five wave-length analyzer that is commonly used as a standard for oxygen saturation measurements.

We calculated bias as the mean of the differences between the oximeter readings and the sum of the oxy-, carboxy- and methemoglobin concentrations measured by the Instrument Laboratories IL282. This is because oximeters respond to these three types of hemoglobin as oxyhemoglobin. We defined bias so that it was positive when the test oximeter overestimated oxygen saturation. The precision was taken as the standard deviation of the bias values. We calculated the difference between the highest and lowest value of the three different samples taken at each FIO2 for each subject and each instrument. The mean of these differences was a crude measure of reproducibility for that FIO2 during a single session.

Another important property of a clinical device is its response time. To determine the response time, we recorded the oximeter readings on a strip chart and took arterial samples every minute for 10 minutes after switching the FIO2 from 1.00 to 0.15. All data are reported in units of percent saturation.

**RESULTS**

We obtained 165 arterial samples and simultaneous pulse oximetry measurements. The mean and standard deviation of pH measurements were 7.40 ± 0.03 and of the PaCO2 were 41.3 ± 5.50. The mean carboxyhemoglobin was 2.0 percent and the mean methemoglobin was 0.3 percent. Thus, an oximeter with a bias of -2.3 relative to the sum of oxy-, carboxy- and methemoglobin would closely approximate the oxyhemoglobin saturation reading of the IL282 device. In no case was the carboxyhemoglobin greater than 2.7 or the methemoglobin greater than 0.6.

Table 1 shows the bias of the saturation calculated from the arterial blood gas, the four pulse oximeters, and the two arterial oximeters for each range of saturations. Saturations calculated from the arterial blood gas measurements had a bias of -1.8, which approximated the measured oxyhemoglobin value of -2.3. The bias became more negative at saturation values below 80 percent.

Pulse oximeter A had a mean bias of -2.6 percent. It performed well at saturations above 80 percent. However, at saturations below 80 percent, it tended to underestimate saturation to an increasing degree.

Thus, the severity of episodes of desaturation was exaggerated. Oximeter B performed well at higher saturations, but overestimated saturations below 80 percent. Thus, it tended to underestimate the severity of desaturation episodes. Pulse oximeters C and D performed similarly. They had biases that were within 2 percent of the standard over the entire range studied. At low saturations, they neither over- nor underestimated saturation.

Arterial oximeter X had a mean bias of -4.2 percent. It tended to underestimate saturation relative to red hemoglobin at high saturations, but had less bias at lower values. Arterial oximeter Y was the only device with a positive bias so that it consistently overestimated oxygen saturation. Its bias remained within 2 percent of the standard over the range studied.

Table 2 shows the precision of the instruments studied. The saturation calculated from the arterial blood gas was relatively imprecise. Arterial oximeter Y was the most precise, while arterial oximeter X was one of the least precise instruments studied. The precision of the different pulse oximeters was similar.

Another approach to assessing clinical accuracy is the percentage of measurements within a given range around the standard. We found the percentage of measurements within 3 percent of the gold standard was 79 percent for pulse oximeter A, 91 percent for pulse oximeter B, 98 percent for pulse oximeters C and D, 43 percent for arterial oximeter X, and 96 percent for arterial oximeter Y and 86 percent for the saturation calculated from the arterial blood gas.

The reproducibility of all oximeters was 1 percent at FIO2 levels greater than 0.17, and was 2 percent at FIO2 below 0.17. The arterial oximeters, particularly the IL282 were slightly more reproducible than the pulse oximeters.

When we compared the response curves of the pulse oximeters and arterial oximeters, there were no significant differences. The pulse oximeters had the practical advantage that they could be read at the bedside, while the arterial sample had to be injected into the machine and analyzed before a result was available.

**Table 1—Mean Differences (Bias) between Oximeter Readings and IL282 Standard* **

<table>
<thead>
<tr>
<th>Bias</th>
<th>ABC</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>All values</td>
<td>-1.8</td>
<td>-2.6</td>
<td>-1.0</td>
<td>-0.4</td>
<td>-1.0</td>
<td>-4.2</td>
<td>+1.8</td>
</tr>
<tr>
<td>Sat &gt;90%</td>
<td>-1.6</td>
<td>-2.2</td>
<td>-2.1</td>
<td>-1.0</td>
<td>-1.0</td>
<td>-4.9</td>
<td>+2.0</td>
</tr>
<tr>
<td>Sat 80-90%</td>
<td>-1.7</td>
<td>-2.2</td>
<td>0.0</td>
<td>+0.7</td>
<td>-1.2</td>
<td>-3.2</td>
<td>+1.4</td>
</tr>
<tr>
<td>Sat 65-80%</td>
<td>-3.2</td>
<td>-4.9</td>
<td>+3.0</td>
<td>+1.5</td>
<td>-0.5</td>
<td>-1.8</td>
<td>+0.9</td>
</tr>
</tbody>
</table>

*Bias is taken as the mean of the differences between oximeter readings and the sum of the oxy-, carboxy-, and methemoglobin concentrations measured by the IL282, a five wave-length oximeter using an arterial sample. Positive bias means the test oximeter overestimates saturation. Units are percent saturation.

**Table 2—Precision (Standard Deviation of Difference from IL282) of Oxygen Saturation Measurements**

<table>
<thead>
<tr>
<th>Precision</th>
<th>ABC</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>All values</td>
<td>2.4</td>
<td>2.1</td>
<td>2.8</td>
<td>1.7</td>
<td>1.6</td>
<td>2.5</td>
<td>0.7</td>
</tr>
<tr>
<td>Sat &gt;90%</td>
<td>1.4</td>
<td>1.2</td>
<td>2.4</td>
<td>1.2</td>
<td>1.4</td>
<td>2.2</td>
<td>0.7</td>
</tr>
<tr>
<td>Sat 80-90%</td>
<td>3.0</td>
<td>2.7</td>
<td>1.7</td>
<td>1.8</td>
<td>2.0</td>
<td>2.3</td>
<td>0.5</td>
</tr>
<tr>
<td>Sat 65-80%</td>
<td>4.3</td>
<td>2.9</td>
<td>2.0</td>
<td>1.0</td>
<td>1.9</td>
<td>2.1</td>
<td>0.5</td>
</tr>
</tbody>
</table>

*Precision is taken as the standard deviation of the difference from the IL282 standard. Units are percent saturation.
DISCUSSION

Before accepting a new device in clinical practice, clinicians should be aware of the bias, precision, reproducibility, and limitations. This study confirms that noninvasive pulse oximeters perform adequately for clinical decision-making over the range of saturations likely to be encountered. Indeed, their performance compared favorably with that of conventional oximeters which use arterial blood samples. We found that caution should be exercised in interpreting pulse oximeter readings of less than 80 percent for certain devices and that different devices may not be equivalent. In patients with lung disease, this is not crucial since oxygen therapy is indicated whether the saturation is 75 or 78 percent. However, this is a significant problem in certain instances. For example, when evaluating patients with cyanotic heart disease, the calculated right-to-left shunt is highly dependent upon accurate measurement of arterial saturation in this range.

Our data on reproducibility in normal subjects suggest that changes of 1 percent in the range of saturations above 90 percent, and 2 percent in the range below 90 percent indicate a clinical change has occurred. This is important in evaluating the clinical significance of changes in therapy or of the patient’s status.

We found the response times were similar for the relatively slow changes in saturation which occurred when breathing subjects were switched from an \( \text{FIO}_2 \) of 1.0 to 0.15. The response time we measured included the alveolar equilibration time, circulation time and response time of the instrument. Thus, they reflect the patient as well as the instrument. Much more rapid changes of saturation may occur during apnea, especially in hypoxic patients or those with apnea beginning at low lung volumes. Also, a significant component of response time is the circulation time from lungs to the sampling site. Minor differences in response time may lead to large differences in lowest recorded saturation during a brief episode. Thus, sleep apnea studies reporting lowest saturation or scores based on number of episodes below a given saturation threshold should only be compared using identical oximeters used at identical sites.

We did not study the effects of other types of hemoglobin such as fetal hemoglobin, carboxyhemoglobin or sickle hemoglobin which might have an effect on bias. Jennis and Peabody found premature infants with more than 50 percent fetal hemoglobin decreased pulse oximeter readings relative to the IL282 device.

We are aware that manufacturers have made modifications and improvements in their monitors since this study was completed. Readers should use this article to understand the validity of methods for measuring oxygen saturation and not to decide which oximeter is best.

We conclude that the bias, precision and reproducibility of noninvasive pulse oximeters compares favorably with those characteristics of conventional oximeters using arterial samples.

ACKNOWLEDGMENT: The authors thank Drs. Marshall Klaus, Herman W. Lipow, Gabriella E. Molnar, Lily C. Kao and David Durand for their suggestions, and Joy Childers and Ceres C. Lark for preparation of the manuscript.

REFERENCES