LETTERS TO THE EDITOR

Statistical Methods for Comparing Methods of Measurement

To the Editor: We read with interest the recently published article by Fortuna and coworkers on the measurement of the fraction of exhaled nitric oxide (FE\textsubscript{NO}) using a hand-held analyzer (NIOX-MINO, Aerocrine, Solna, Sweden). This study falls within the literature comparing methods of measurement, or studies of agreement between methods. In this case, the methods compared are quantitative. One aim of this type of study is to evaluate the reliability of new measurement devices. To that end the authors compared readings taken with the NIOX-MINO device to those obtained with standard equipment considered to offer a reference, the N-6008 chemiluminescence analyzer (SIR, Madrid, Spain). They compared means with the Mann-Whitney test and calculated correlations between individual results with the Pearson method. The authors based their conclusion that the device is reliable on the finding of very good correlation between FE\textsubscript{NO} measurements made with the 2 devices (\(r=0.92, P=.001\)) and they suggest a range of reference values with an estimated upper limit set at 2 SDs above the mean.

This statistical method for evaluating agreement between quantitative measures has been shown to be inadequate and the interpretation of the Pearson correlation coefficient as a measure of agreement has also been shown to be erroneous. The lack of a difference between means only guarantees that 2 methods are based on the same value, not that all the values obtained are the same. Furthermore, in the particular case of this study, the authors compared means using the Mann-Whitney test, a nonparametric test appropriate when the means are from independent groups. Since the measurements in this study were paired, the authors should have used a nonparametric test for matched pairs, namely the Wilcoxon T test. Although other authors have also used only the Pearson correlation coefficient for the same purpose (References 2, 8, and 11 in the Table), this statistic only indicates strength of association between 2 variables (how close the data points for the 2 methods are to the regression line) and the linearity of their relationship, but not that the line necessarily bisects the set of data; even a perfect correlation cannot be considered synonymous with perfect agreement. Nor do
LETTERS TO THE EDITOR

Validation Studies, With the Statistical Method Used, for the Measurement of the Fraction of Exhaled Nitric Oxide

<table>
<thead>
<tr>
<th>Study</th>
<th>Pearson Correlation Coefficient</th>
<th>Intraclass Correlation Coefficient</th>
<th>Bland and Altman Plot</th>
<th>Bland and Altman Coefficient of Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.97 and 0.98</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>0.97 and 0.98</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>0.96</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>0.992</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>0.977</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>6, 7</td>
<td>0.94 and 0.96</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>0.99</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>9, 10</td>
<td>0.94-0.949</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>11</td>
<td>0.94-0.949</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

*List of validation studies:


*95% confidence interval of the correlation coefficient.

the authors give confidence intervals for the coefficient. We calculated the 95% confidence interval to be 0.83 to 0.96, implying that the values for the correlation coefficient could be less than the 0.9 or greater than the 0.9, which case the correlation would be classified as merely good. Other studies not mentioned by the authors report Pearson correlation coefficients much closer to 1 (Table).

Among the most frequently used procedures for determining agreement between quantitative measures are a) analytical methods involving the calculation of various measures of agreement such as the intraclass correlation coefficient or the concordance correlation coefficient of Lin, regression models (of Deming and of Passing-Bablock), analysis of variance of repeated measures, and structural equation modeling; and b) graphic methods such as the construction of a Bland and Altman plot4 (in which the interval of agreement is also calculated), the folded empirical cumulative distribution curve (or “mountain plot”), and the survival-agreement plot.5 Fortuna and coworkers6 applied none of these methods. A more exhaustive review of the literature (eg, at http://www.nioxmino.com/references.html) shows that several other studies have used these methods (Table). In fact, an important observation based on the Bland and Altman plots in the literature is that the relation between paired FE\textsubscript{NO} measurements between the devices compared is heterocedastic, which is to say, the higher the FE\textsubscript{NO} values, the greater the differences between measurements (References 6 and 7 in the Table). In addition, as mentioned above, the authors suggest a cutoff value of 2 SDs above the mean to be used for reference. Even if values measured fall within the limits the authors establish, this approach to estimating reference values can only be used if the data are normally distributed.6 However, they report no statistical test, either analytic or graphic, of normality. Other studies have performed log transformations in order to normalize data for a variable (References 5 and 10 in the Table).

Although Fortuna and colleagues7 describe the attractive features of the new FE\textsubscript{NO} measurement device (namely, that it is simple, fast, manageable, and inexpensive) and they note its usefulness for diagnosis and follow-up in asthma, they cannot assume from their analysis that the device gives measurements that agree with the reference method (reliability). Nor can they assume that the reference values they suggest are valid without providing more information about the distribution of the data.

Fortunately, a good study with insufficient analysis can always be re-analyzed. We believe that the above-mentioned information should be brought to light, so that the NIOX-MINO device can be considered validated in a Spanish population. There is a statistical package (MedCalc Software, Mariakerke, Belgium) that incorporates a procedure to compare methods directly. In this program, multiple measures of the aforementioned coefficients, regression models, and graphic methods can be implemented without the need for algebraic transformation of the data (whereas such ease of calculation is only possible for the intraclass correlation coefficient in the SPSS program). We encourage the authors to complete their analysis and publish the pertinent data for validating the device—showing its reliability—in a Spanish population.

José-Francisco Pascual-Lledó, Carmen Contreras Santos, and Carlos Martín Carrasco

Unidad de Neumología, Hospital La Inmaculada, Huércal-Overa, Almería, Spain