Med Intensiva. 2017;41(6):327–329

medicina intensiva

www.elsevier.es/medintensiva

EDITORIAL

Good science☆

Buena ciencia

V. Modesto i Alapont

Unidad de Cuidados Intensivos Pediátrica, Hospital Universitari i Politècnic La Fe, Valencia, Spain

In the present issue of MEDICINA INTENSIVA,[1] García-Soler et al., of the Pediatric Intensive Care Unit (PICU) of Hospital Regional Universitario Carlos Haya (Málaga, Spain), present an excellent study that aims to validate a transcutaneous system for measuring hemoglobin concentration in critically ill children at risk of bleeding. The methodology used is solid, the analysis rigorous, and the conclusions honest. In my opinion: very good science.

In the literature on diagnostic tests, the clinically relevant objective is usually to evaluate the accuracy of a measurement. Accuracy is a property established along a certain dimension: comparison of the diagnostic test with some other measurement taken to be the gold standard. Accuracy is measured using likelihood ratios or their decimal logarithms (the so-called ''weight of evidence'' invented by Alan Turing in 1940–1941 to break the coding of the Nazi Enigma cipher machines[2]).

However, we must note that concordance or reliability – a property established along another, different dimension – is a necessary prior condition for establishing accuracy. What we seek to evaluate on measuring the concordance or agreement of two measurements, or on studying the consistency or reproducibility of one same measurement repeated over time, is the ''discrepancy'' between the two measurements (or repetitions): the bias introduced when we taken one measure for the other. Only if the two measurements are concordant (*i.e.*, non-discrepant), and one of them is moreover the gold standard, can we subsequently assess the accuracy of the other as diagnostic method.

Analysis of the concordance between variables involving categorical data is based on the use of the (weighted) Cohen kappa index. However, in clinical research it is relatively common to have to assess the concordance between quantitative measurements. In this case, the Pearson correlation coefficient (R^2) is not of help, and it is a mistake to use it for this purpose.[3] Two measurements can have a very high and statistically significant correlation and yet be based on different scales. And what we seek to evaluate on measuring concordance is the ''identity'' between both – not only their capacity to vary through mutual influencing effects. The best way to measure this appears to be Lin's correlation and concordance coefficient (CCC).[4]

A very adequate alternative is that chosen by García-Soler and her team: calculation of the so-called intraclass correlation coefficient (ICC), which estimates the mean correlations among all the possible orders of the available pairs of observations. In order to understand intraclass correlation, let us suppose that all the observations of a variable are ordered into *m* groups (in this concrete case the two groups: transcutaneous hemoglobin measurement and laboratory hemoglobin measurement), each containing n observations. And let us also suppose (null hypothesis) that there are no reasons to expect differences in the mean level of the variable between the *m* groups. If there were such differences, the observations of the same group would tend to be positively correlated to each other, and their variability would be different from that of the observations of the other group. This correlation is what we know as intraclass correlation.

☆ Please cite this article as: Modesto i Alapont V. Buena ciencia. Med Intensiva. 2017;41:327–329.

E-mail address: vicent.modesto@gmail.com

The formula for calculation is based on a single-factor random effects analysis of variance model (ANOVA). The idea is that the total variability of the measurements can be broken down into two components: variability due to the differences between the different subjects (between-subject variance) and that due to the differences between the measurement methods of the variable for each subject (within-subject variance). The ICC, a parametric coefficient that can be regarded as the equivalent of the kappa statistic for continuous variables, is then calculated as the proportion represented by the between-subject variance with respect to the total variability.[5] Since ICC is a proportion, it takes values between 0 and 1: the coefficient is close to 1 if the observed variability is fundamentally attributable to the differences between the subjects and not to the differences between the measurement methods (or between the observers), and takes the value 0 in the opposite case. Although the interpretation is subjective, a consensus-based scale has been adopted for assessing the ICC as a measure of reproducibility: values under 0.4 indicate scant reproducibility, while values of 0.75 or higher indicate excellent reproducibility. Intermediate values are considered to indicate adequate reliability. The main limitation in the use of the ICC – apart from the limitations derived from non-compliance with the hypotheses for application of the ANOVA (normality, equality of variances and independence of errors) – is that it depends both on the range of variation of the measurement scale and on the number of measurement methods (or observers). Thus, for example, if a measurement presents very little variability, the ICC can be low, without this meaning that the method is scantly reliable.

In their analysis, García-Soler et al. also used the popular Bland–Altman plot, which detected the main weakness of the transcutaneous method for measuring hemoglobin. *A priori*, the authors established the tolerance limit for bias in the measurement (the maximum difference in the two hemoglobin concentration measurements in order for the new noninvasive method to be considered reliable) as ± 1 g/dl. This may seem an arbitrary decision, though in my opinion it is quite appropriate and is based on sound clinical reasoning: in the pediatric intensive care setting where the noninvasive transcutaneous method is intended to be used, a difference of 1 g/dl can cause us to modify the therapeutic approach. Ideally, the possible difference would fall below this margin. However, as commented by the authors in the Results section of their article, both the statistical analysis and the Bland–Altman plot revealed that ''the mean difference between the laboratory test and pulse-cooxymeter values was 0.66 ± 1.46 g/dl, with a median of 0.5 g/dl (interquartile range [IQR]: -0.2 to 1.4). The median difference in absolute values was 0.8 g/dl (IQR: 0.4–1.7)''. In other words, the tolerance limit was clearly exceeded. In fact, the 95% confidence interval (95%CI) of the mean population bias recorded in the Pediatric Intensive Care Unit of Málaga (assuming normality) was -2.2 to 52 g/dl: the value of ± 1 g/dl (tolerance limit) is a possible value at a confidence level of 95%. This bias is 10 times greater than that reported in a recent study on the subject,[6] though there are other publications that describe similar and even greater biases.[7–9]

Thus, the possible magnitude of the bias could invalidate the measurements and clearly affect the clinical relevance of the transcutaneous method. In fact, as pointed out by the investigators, their Table 1 indicates that ''60.9% of the noninvasive pulse-cooxymeter measurements exhibit a difference of ≤ 1 g/dl with respect to the value of the core laboratory analyzer'' – which means that 50% of the measurements fall outside the tolerance limit established *a priori* by the authors. Furthermore, with their multivariate models correctly adjusted according to the Akaike information criterion (AIC), Soler et al. found that precisely the ''perfusion index'' is a covariable that strongly influence concordance. This is clinically very relevant, since continuous hemoglobin monitoring to ensure the early detection of anemia in the pediatric setting is particularly crucial in patients at a high risk of bleeding (which is why sampling in this study involved such children). In this regard, the fact that precisely those patients with a greater risk of anemia (with hemodynamic instability secondary to incipient shock, and therefore with a high probability of presenting an altered perfusion index) are those in which the transcutaneous measurement loses reliability represents a strong handicap. In my view, it is in these clinical situations where the measurement must be most reliable.

This important limitation is one of the issues commented by the authors in the Discussion and cited in their conclusions. For this reason, and as I mentioned at the start, I think it is ''good science'' for a study to use the text not only to point out the limitations of the work itself but also to try to find arguments capable of refuting the hypotheses initially postulated to be true. In other words, the investigator personally should try to be the strictest judge of what he or she is proposing as scientific truth. In the words of the Nobel Prize winner Richard Feynman,[10] ''the idea is to try to give all the information to allow others to judge the value of your contribution: not only the information that may lead judgment in one concrete direction or other''. Integrity is a key principle of scientific thinking.

References

1. García Soler P, Camacho Alonso JM, González Gómez JM, Milano Manso G. Monitorización no invasiva transcutánea de la concentración de hemoglobina en pacientes críticos pediátricos con riesgo de sangrado. Med Intensiva. 2017:41.
2. Good IJ. Studies in the history of probability and statistics XXXVII. A M Turing's statistical work in World War II. Biometrika. 1979;66:393–6.
3. Kramer MS, Feinstein AR. Clinical biostatistics. LIV. The biostatistics of concordance. Clin Pharmacol Ther. 1981;29:111–23.
4. Lin L, Hedayat AS, Sinha B, Yang M. Statistical methods in assessing agreement: models, issues, and tools. J Am Stat Assoc. 2002;97:257–70.
5. Pita Fernández S, Pértega Díaz S. La fiabilidad de las mediciones clínicas: el análisis de concordancia para variables numéricas. Atención Primaria en la red. Fisterra.com. Available from: http://www.fisterra.com/mbe/investiga/conc_numerica/conc_numerica.asp [accessed 12.6.04; consulted October 2016].
6. Phillips MR, Khoury AL, Bortsov AV, Marzinsky A, Short KA, Cairns BA, et al. A noninvasive hemoglobin monitor in the pediatric intensive care unit. J Surg Res. 2015;195:257–62.
7. Dewhirst E, Naguib A, Winch P, Rice J, Galantowicz M, McConnell P, et al. Accuracy of noninvasive and continuous hemoglobin

measurement by pulse co-oximetry during preoperative phlebotomy. J Intensive Care Med. 2014;29:238–42.
8. Amano I, Murakami A. Use of non-invasive total hemoglobin measurement as a screening tool for anemia in children. Pediatr Int. 2013;55:803–5.
9. Jung YH, Lee J, Kim HS, Shin SH, Sohn JA, Kim EK, et al. The efficacy of noninvasive hemoglobin measurement by pulse co-oximetry in neonates. Pediatr Crit Care Med. 2013;14:70–3.
10. Feynman RP. *La ciencia del culto al cargamento*. Discurso inaugural 1974 en Caltech.