EVIDENCIAS EN PEDIATRÍA
Toma de decisiones clínicas basadas en las mejores pruebas científicas
www.evidenciasenpediatria.es

Basics of Evidence-Based Medicine

Evaluation of the accuracy of diagnostic tests (2). Continuous variables

Ochoa Sangrador C[1], Molina Arias M[2]
[1]Department of Paediatrics. Hospital Virgen de la Concha. Zamora. Spain.
[2]Department of Gastroenterology. Hospital Infantil Universitario La Paz. Madrid. Spain.

Correspondence: Carlos Ochoa Sangrador, cochoas2@gmail.com

English key words: accuracy, repeatability, diagnostic test, reliability of diagnostic tests, intraclass correlation coefficient, Bland and Altman plot.
Palabras clave en español: precisión, reproducibilidad, pruebas diagnósticas, fiabilidad de las pruebas diagnósticas, coeficiente de correlación intraclase, método de Bland y Altman.

Reception date: September 1, 2017 • **Acceptance date:** September 4, 2017
Publication date: September 6, 2017

Evid Pediatr. 2017;13:45.

HOW TO CITE THIS ARTICLE

Ochoa Sangrador C, Molina Arias M. Evaluación de la precisión de las pruebas diagnósticas (2). Variables continuas. Evid Pediatr. 2017;13:45.

This article is available at http://www.evidenciasenpediatria.es/EnlaceArticulo?ref=2017;13:45.
 • ISSN: 1885-7388

Evaluation of the accuracy of diagnostic tests (2). Continuous variables

Ochoa Sangrador C[1], Molina Arias M[2]

[1]*Department of Paediatrics. Hospital Virgen de la Concha. Zamora. Spain.*

[2]*Department of Gastroenterology. Hospital Infantil Universitario La Paz. Madrid. Spain.*

Correspondence: Carlos Ochoa Sangrador, cochoas2@gmail.com

INTRODUCTION

In previous articles in this series, we have addressed how to assess the validity of a diagnostic test. We have also reviewed how to assess its accuracy or reliability. To date, we have discussed the methods used to measure the accuracy of discrete data, nominal (kappa statistic) and ordinal (weighted kappa statistic). In this article, we will broach the methods that apply to continuous data: the within-subject standard deviation, the intraclass correlation coefficient and the Bland and Altman method.

CONTINUOUS VARIABLES

Within-subject standard deviation

When the result of a test is measured on a continuous scale, we can estimate the measurement error by calculating the variability that exists between repeated measurements in the same subjects. The parameter that best reflects such variability is the within-subject standard deviation (excluding the variability observed between subjects). To calculate it, we need a set of subjects to undergo at least two measurements each. Table 1 presents the results of performing two repeated transcutaneous bilirubin measurements in newborns with jaundice.[1] The within-subject standard deviation can be calculated easily using software that performs analysis of variance (ANOVA). ANOVA breaks down the variation present in the set of measurements (estimated based on the squared differences of each value and the mean of all subjects) into several components: the variation in measurements taken in different subjects (between rows in Table 1) and the variation in the residuals, which in one-way ANOVA corresponds to the variation in the measurements taken in each subject (between columns in Table 1).

Table 2 shows the ANOVA for the data in Table 1. The parameter called mean square of the residuals (MSr) is the residual or within-subject variance (which depends on the differences

TABLE 1. RResults of two repeated transcutaneous bilirubin measurements (Jaundice-Meter 101, Minolta Air Shields) in the anterior surface of the thorax in 20 newborns with jaundice. Data retrieved from a larger study.[3]

Subjects	1st measurement	2nd measurement	Difference	Mean
1	14	16	-2	15.0
2	14	14	0	14.0
3	17	17	0	17.0
4	14	15	-1	14.5
5	15	14	1	14.5
6	18	19	-1	18.5
7	16	16	0	16.0
8	12	12	0	12.0
9	19	19	0	19.0
10	9	10	-1	9.5
11	15	16	-1	15.5
12	18	18	0	18.0
13	17	18	-1	17.5
14	15	15	0	15.0
15	9	9	0	9.0
16	14	14	0	14.0
17	17	18	-1	17.5
18	18	18	0	18.0
19	20	20	0	20.0
20	10	11	-1	10.5

TABLE 2. One-way analysis of variance for the data in Table 1

Source of variation	Degrees of freedom	Sum of squares	Mean square	
Between patients	19	371.5000	19.5526	MSp
Residual	20	6.0000	0.3000	MSr
Total	39	377.5000		

MSp: mean square of the patients; **MSr:** mean square of the residuals.

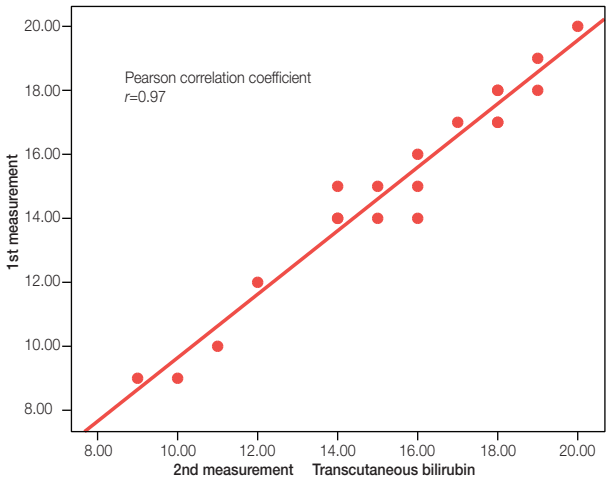
between repeated measures in each subject). If we take the square root of the MSr, we obtain the within-subject standard deviation (s_w). The s_w can also be calculated using the results of ANOVA in designs with more than two measurements per subject.

We can use the s_w to quantify the margin of error in our measurements. Thus, we can estimate that the difference between a specific measurement and the true value will not be greater than 1.96 times the s_w in 95% of observations (assuming that the data follow a normal distribution, 95% of the measurements will be contained in the interval formed by the actual value plus and minus 1.96 times the standard deviation). It also allows us to estimate that the difference between two measurements for the same subject will not exceed 2.77 times the s_w in 95% of observations.[2,3] In our example, the s_w is 0.54 (square root of 0.3), so the estimated difference from the true value would be of less than 1.05 (1.96 × 0.54) and the difference between two measurements would be of less than 1.49 (2.77 × 0.54).

Intraclass correlation coefficient

If only two measurements are taken per subject, the most intuitive way to compare them is to plot measurement pairs in a scatter diagram, assess whether there is a linear association between them, and calculate the corresponding correlation coefficient. Figure 1 shows the scatter diagram for the data in Table 1. The Pearson correlation coefficient (r) for these data is 0.97 (the closer r is to 1, the stronger the correlation).

FIGURE 1. Scatter plot and linear correlation for the data in Table 1



However, the presence of a strong linear association with a high correlation coefficient does not prove a strong agreement between the measurements, but only that the points in the plot fit a straight line well. The correlation coefficient is largely dependent on inter-subject variability and thus changes substantially based on the characteristics of the sample for which it is calculated, and is especially sensitive to the presence of extreme values. If one of the measurements is systematically greater than the other, the correlation coefficient will be very high, despite the fact that the measurements never agree. These pitfalls can be avoiding by using the intraclass correlation coefficient.

The intraclass correlation coefficient (ICC) estimates the agreement between two or more repeated measurements. The calculation of the ICC is based on a repeated measures ANOVA model, applying different formulas based on the design and objectives of the study.[4] In the simplest scenario, we would estimate the variability of the measurements without taking into account the variability contributed by different raters (one-way random effects model). Choosing this model, and using the results of ANOVA, we can calculate the ICC with the following formula:

$$CCI = \frac{CMp - CMr}{CMp + (k-1)CMr'}$$

where k stands for the number of observations per subject, MSp for the mean square between patients (which depends on the differences in measurements between subjects) and MSr for the mean square of the residuals (which depends on the differences between repeated measurements in each subject).

Using the data of the ANOVA in Table 2, the ICC will be:

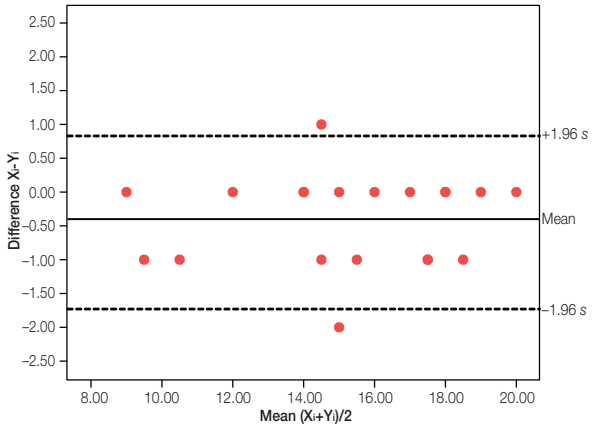
$$CCI = \frac{19.55 - 0.30}{19.55 + (2-1)0.30} = 0.96$$

In our example, there is hardly any difference between the ICC and the Pearson correlation coefficient (r). If the ICC were much smaller than r, one would assume that there is a systematic change between one measurement and the other, which may result from a learning effect. In this case, the measurements would not have been made under the same circumstances, so the conditions required for performing a reliability analysis would not be met.[5]

Bland and Altman method

An alternative approach to analysing the agreement between two repeated observations measured on a continuous scale is the graphical method described by Bland and Altman.[6] It consists of plotting the difference of each pair of measurements against the mean of the two measurements (Figure 2).

FIGURE 2. Bland and Altman method applied to the data in Table 1



The points tend to cluster around zero in the axis representing the difference between paired measurements, and the greater the dispersion around zero, the lesser the agreement between the two measurement methods. One possible way to assess agreement is to draw horizontal lines at the level of the maximum difference that would be acceptable from a clinical standpoint, and check whether the points, or most of the points, are grouped between these two horizontal lines. An alternative approach is to estimate the standard deviation of the differences and the interval in which we would expect to find 95% of them.

This method can also be used to assess the magnitude of the differences and their association with the magnitude of the measurement. When the variability in the measurements is not constant, but changes as the magnitude of the measurement increases or decreases, the calculation becomes complicated.[7] If there is a significant correlation between the differences and the means, the variability will not be constant (there may be an acceptable agreement in a specific value interval, but not in others). In this case, a logarithmic transformation of the data can be attempted, or else the variability can be analysed separately for various data intervals, although we should always hold reservations about the validity of measurements in these intervals.

REFERENCES

1. Ochoa Sangrador C, Marugán Isabel VM, Tesoro González R, García Rivera MT, Hernández Calvo MT. Evaluación de un instrumento de medición de la bilirrubina transcutánea. An Esp Pediatr. 2000;52:561-8.
2. Bland JM, Altman DG. Measurement error. BMJ. 1996;312:1654.
3. Altman DG, Bland JM. Comparing several groups using analysis of variance. BMJ. 1996;312:1472.
4. Fleiss JL. The design and analysis of clinical experiments. New York: John Wiley & Sons 1986. p. 1-32.
5. Bland JM, Altman DG. Measurement error and correlation coefficients. BMJ. 1996;313:41-2.
6. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1:307-10.
7. Bland JM, Altman DG. Measurement error proportional to the mean. BMJ. 1996;313:106.