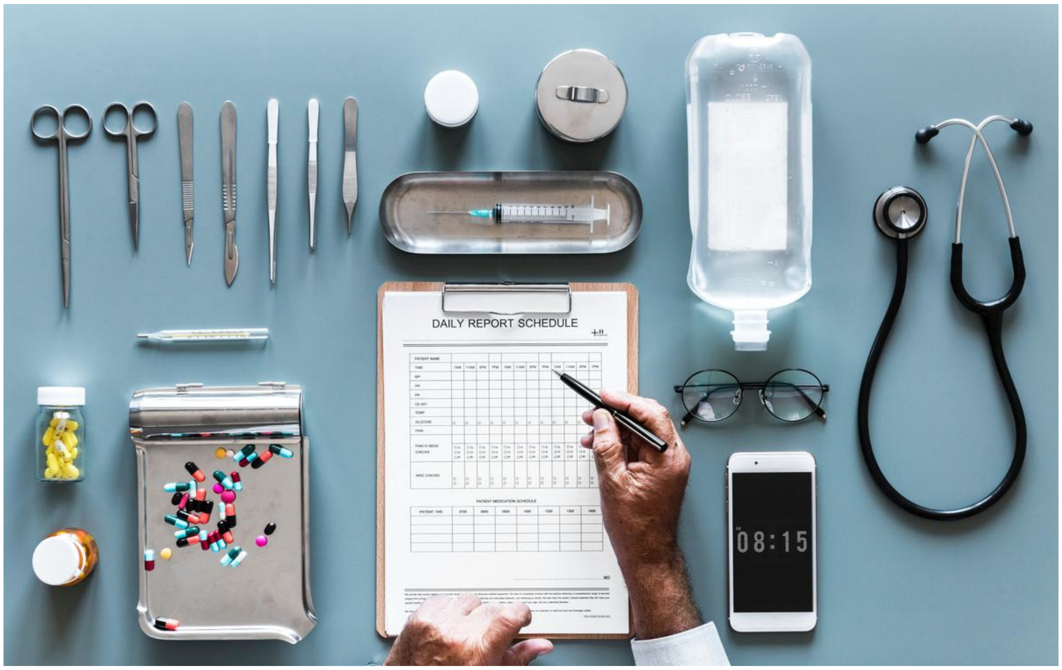
EVALUATION OF THE RELIABILITY OF CUANTITATIVE BIOMEDICAL DATA MEASUREMENTS

Rubén Prieto Paredes
Master´s Degree in Pharmacological Research



Index

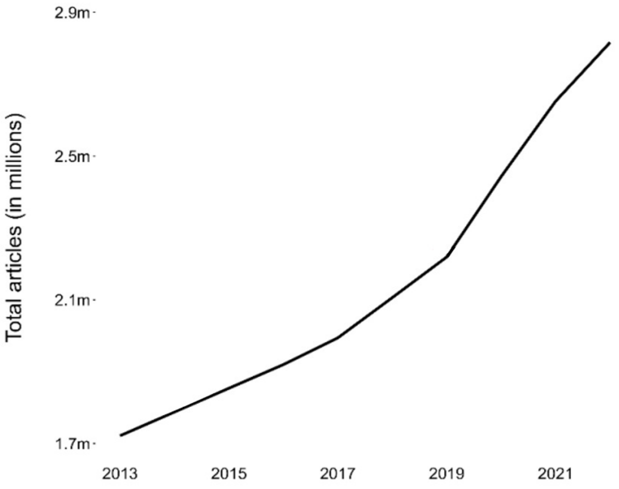
1. Introduction:

1.1. Research today and the need to do quality science without bias.

Biomedical research is a field of study that focuses on the application of scientific principles and research methods to understand, diagnose, treat and prevent diseases in. The fundamental objective of biomedical research is to improve human health and advance knowledge of the biological processes that underlie diseases. In addition, they have to comply with intrinsic scientific and ethical protocols in clinical studies carried out on human beings. This compliance ensures that the rights and well-being of those who participate are respected.

We could divide this discipline into 4 research subgroups: basic or preclinical research (related to studies in cells, tissues and animal models), clinical research (focused on the safety of medical interventions and treatments), translational research (which combines the basic research with the clinic), and epidemiological research (which studies particular phenomena by studying a certain group of people).

Each group has specific characteristics that mean that each subdiscipline has its own methodologies and analyses, but what they all have in common is the large volume of results obtained from each of them. We must keep in mind that thousands and thousands of articles indexed in Scopus and Web Science are published each year , growing at an exorbitant rate, causing the total number of articles in 2022 to be 47% higher than in 2016. This existing publication speed generated by the great growth of publications can compromise the ability of scientists to be rigorous when analyzing information. If scientific rigor declines, the term "science" is devalued, so the objective must be to combat biases in publication (Hanson *et al.* , 2023) (Picture 1).



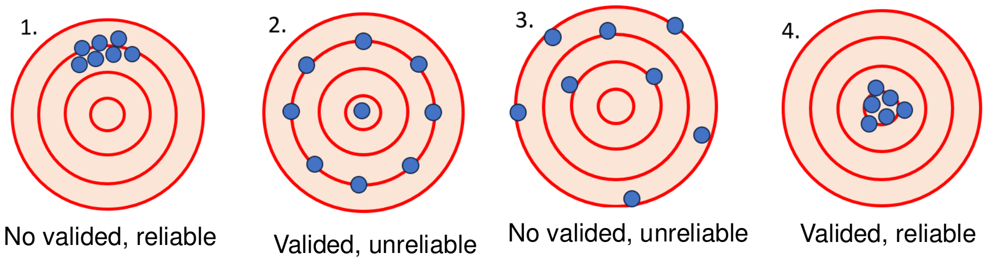
Picture 1: Gráfico en el que se representa el incremento de la publicación de artículos en revistas indexadas

Due to the enormous amount of results that are produced, we must develop statistical methodologies that allow us to compare the information and decide if the results are reliable or not. In this way, the results must be verifiable, consistent, reproducible and must be able to be subjected to reliability filters as a fundamental principle of the precision of the study.

1.2 Application of confidence and validation in biomedical studies

In any research process, given the large number of sources of potential errors, it is necessary for researchers to try to reduce those related to the measurement of the variables to provide greater confidence in the results and conclusions of their study. In other words, an instrument is reliable, precise or reproducible when the measurements made with it generate the same results at different times, scenarios and populations if applied under the same conditions.

However, it must be assumed that in everyday clinical practice, reliability is combined with another concept, which is validity, giving rise to various scenarios, from valid and reliable measurements to those that lack validity and reliability, as in the case of observations or observers who agree only due to the effect of chance, in such a way that the greater the precision of a measurement, the greater statistical power there will be in the sample under study (Picture 2) (Manterola, C. et al . , 2018) .



Picture 2. Possible validity and reliability scenarios. 1). All measurements are similar, but they are far from reality. 2). The measurements capture the entire spectrum of the phenomenon, but they are very different from each other. 3). The entire phenomenon is not captured and the measurements are very different from one another. 4). All measurements are similar and adjust to the reality of what is being measured

Before considering whether the instrument measures what we want to measure, we must ensure that the instrument measures "something" in a reproducible way: if the measuring instrument does not offer reproducible results, then the measuring instrument is not reliable, and it is pointless to ask ourselves. the problem of validity.

If we want to apply these 2 concepts to biomedical research and clinical studies, reliability refers to the ability to repeat measurements and obtain similar results. Validity, on the other hand, ensures that measurements actually measure what they are intended to measure. Both are key to making informed decisions and effectively applying the results in medical practice. These principles ensure the quality and reliability of results, which is essential to advance scientific knowledge and to translate discoveries into effective clinical applications. Reproducibility not only implies the repetition of the results, but also the consistency in obtaining them under different conditions and by different researchers.

In biomedical studies, the lack of reproducibility can undermine confidence in the results, generating erroneous conclusions and making it difficult for the finding to be applied to clinical practice, in such a way that its implementation in medical use generates ineffective or even harmful treatments or interventions. Reproducibility is essential for building a robust body of scientific knowledge and ensuring that findings can be reliably applied in clinical practice.

2. Numerical analysis of the reliability of quantitative measurements

2.2 Reliability study:

In clinical measurements, it is often necessary to compare a new measurement technique with an established one to see if they agree sufficiently, or for example to corroborate that measurements made by two measuring instruments offer results that match them for the new one to replace the old one. old, or if an observational study carried out by two researchers following the same analysis methodology offers the same results. To evaluate all this, there are different tests and methods to corroborate how reliable the different measurements are. We are going to break down some of these in this work. (Bland , J.M & Altman, D., 1986)

The concept of reliability and the various indices used to estimate it are better understood if the measurement model used is made explicit. For a random variable the simplest model is:

$$X = Y + \varepsilon$$

X= Measurement result

Y= Magnitude to measure

ε = Measurement error

The magnitude to measure (Y) of the model can be decomposed into the sum of its mean µ (the mean of the measurements is a constant) and the variable ε Y $_{that}$ contains all the variability of Y around the mean.

$$X = \mu + \varepsilon_Y + \varepsilon$$

Although both ε $_{Y}$ and ε are random variables, ε $_{Y}$ represents the variability of the variable to be measured (in our case the variability of the biomedical parameter that we are going to measure), while ε represents the measurement error.

If we try to calculate the expected value of the model, assuming independence between ε $_{and}$ and ε , we obtain:

$$E(X) = \mu + \mathrm{E}(\varepsilon)$$

The variable E(ε) is called systematic error or bias. Therefore, from this expression it follows that a measure of validity is:

$$\mathrm{E}(\varepsilon) = E(X) - \mu$$

We can also calculate the model variance Var(ε), which tells us the random error (due to the need to conserve units, the random error is defined as the standard deviation of ε .

$$Var\ (X) = Var\ (\varepsilon_Y) + Var\ (\varepsilon)$$

The measurement variance has two components: one is the variance of the variable itself and the other is that of the measuring instrument. Measurement variance is therefore not a good indicator of measurement stability. However, if the magnitude to be measured were constant, it would be a good indicator of the stability of the measurement, that is, if the variance of εY were zero, as would be the case if one were trying to estimate the precision of a balance using some standard weights.

We can summarize all this in that to evaluate the reliability or reproducibility coefficient, a series of formulas can be used, which, for their calculation, include the variability of the subject and the measurement error, being able to generate a generalized equation that would be:

$$C = \frac{Subject\ variability}{Subject\ variability + Measurement\ error} = \frac{Var(\varepsilon_Y)}{Var(\varepsilon_Y) + Var(\varepsilon)} = \frac{{\sigma^2}_{\varepsilon_Y}}{{\sigma^2}_{\varepsilon_Y} + {\sigma^2}_{\varepsilon}}$$

(Latour, J. *et al* ., 1997; Manterola, C. *et al* ., 2018)

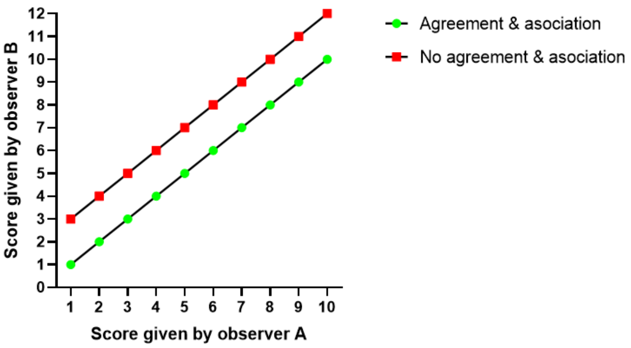
3. Methods to evaluate reliability in quantitative measurements:

3.1. Intraclass correlation coefficient (ICC)

The intraclass correlation coefficient allows measuring the general agreement between two or more measurements that involve quantitative variables, obtained with different measuring instruments or evaluators. When it is measured, there will always be a certain error that will depend randomly on variables such as the instrument, the manipulation, the subject evaluated, etc. Any measurement is the result of adding the real value of what we want to measure and a random error. A reliable instrument will be one whose random error is small enough to allow us to consider that the result obtained is not attributable to measurement errors (Picture 3).

It is based on an analysis of variance model with repeated measures. Its use is only possible if there is normality of the distributions of the variables, equality of variances and independence between the errors produced by the observers. This coefficient measures the agreement between different observers or repeated measurements. We can say that a high ICC indicates good reproducibility.

Although it is considered a type of correlation, unlike most other correlation measures, it operates on data structured in groups, rather than data structured as paired observations.



Picture 3: The dashed line represents a perfect association with little agreement between observers; and the solid line represents a perfect association and agreement between observers. Picture designed by my own by Graph Pad Prism 8.

In order for the ICC to be used, the following conditions must be met that there is normality in the distributions of the variables, homocestacy (equality of variances) and independence between the errors produced by the observers.

A limitation of the CCI is that it depends on the variability of the observed values. If patients vary little in their measurements (homogeneous sample), the ICC tends to be low, since it compares the variance between patients with the total observed variance. If the sample is heterogeneous, the ICC tends to be higher.

The equation to obtain the CIC would be:

$$CIC = \frac{(\sigma_S^2)}{(\sigma_S^2) + \left(\sigma_j^2\right) + (\sigma_e^2)}$$

- Intersubject variability due to the differences between them ($\sigma^2{}_s$).
- Intrasubject variability, due to differences in measurements from the same subject ($\sigma^2{}_e$).
- Residual variability, it is inexplicable (random), it is linked to measurement errors ($\sigma^2{}_j$).

By applying the method, values between 0 and 1 can be obtained, where 0 means lack of agreement and 1, agreement or absolute reliability. Conventionally, the following values are accepted:

- < 0.40 →Poor.
- 0.40-0.59 →Enough.
- 0.60-0.74 →Good.
- 0.75-1→ Excellent

3.2. Intraclass correlation coefficient applied to a real clinical case

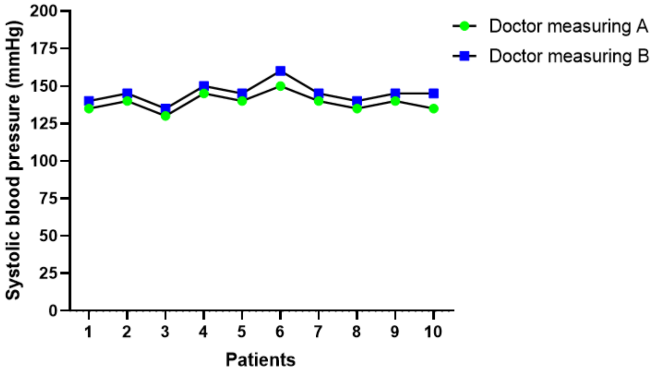
As we have already mentioned, in the clinical setting, it is sometimes necessary to assess the reliability of the measurements or observations made, so we are going to see a numerical example of two consecutive measurements of systolic blood pressure, which requires carrying out this method to study the correlation, since that the measurement data are quantitative continuous variables (Table 1).

Patient	Doctor measuring A (mmHg)	Doctor measuring B (mmHg)
1	135	140
2	140	145
3	130	135
4	145	150
5	140	145
6	150	160
7	140	145
8	135	140
9	140	145
10	135	145

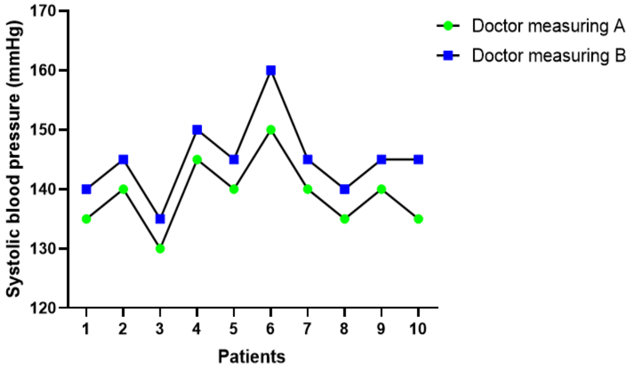
Table 1: Consecutive measurements of systolic blood pressure in 10 patients, performed by two doctors with the same sphygmomanometer. Data obtained from (Prieto, L., 1998).

If we try to graphically represent the data obtained we see that the measurements in the two measurements are very similar (Picture 4) but if we modify the Y axis and zoom in on the graphs we see that there is a considerable difference, it seems that the measurement A has tended to measure more towards smaller data and measurement B has tended to give larger results (Picture 5). But what level of agreement has this ?

Picture 4



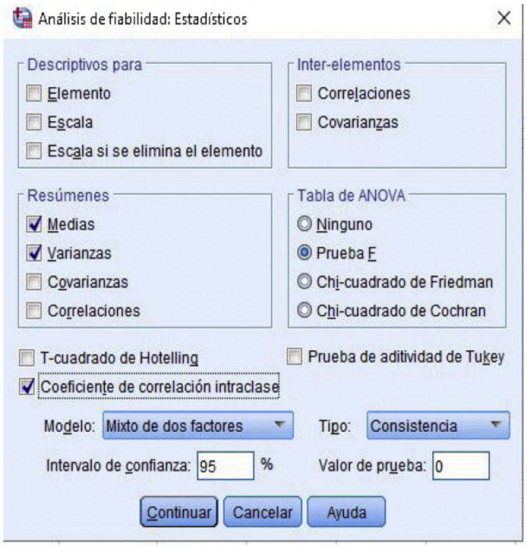
Picture 5



Picture 4 and 5: Graphs representing patients' blood pressure measurements in mmHg. In green the measurement A and in blue the measurement B. Pictures designed by my own by GraphPad Prism 8.

To calculate the CIC in IBM SPSS Statistics we follow the following command from the application's drop-down menu. The Spanish version would be the following “path” to obtain the CIC:

Analyze → Scale → Reliability analysis → (here we choose the variables that we are going to evaluate) → Statistics → Intraclass correlation coefficient → Model & Type → Continue → Accept: obtaining the values of the intraclass correlation coefficient (Picture 6).



Análisis de fiabilidad: Estadísticos

Descriptivos para
□ Elemento
□ Escala
□ Escala si se elimina el elemento

Inter-elementos
□ Correlaciones
□ Covarianzas

Resúmenes
☑ Medias
☑ Varianzas
□ Covarianzas
□ Correlaciones

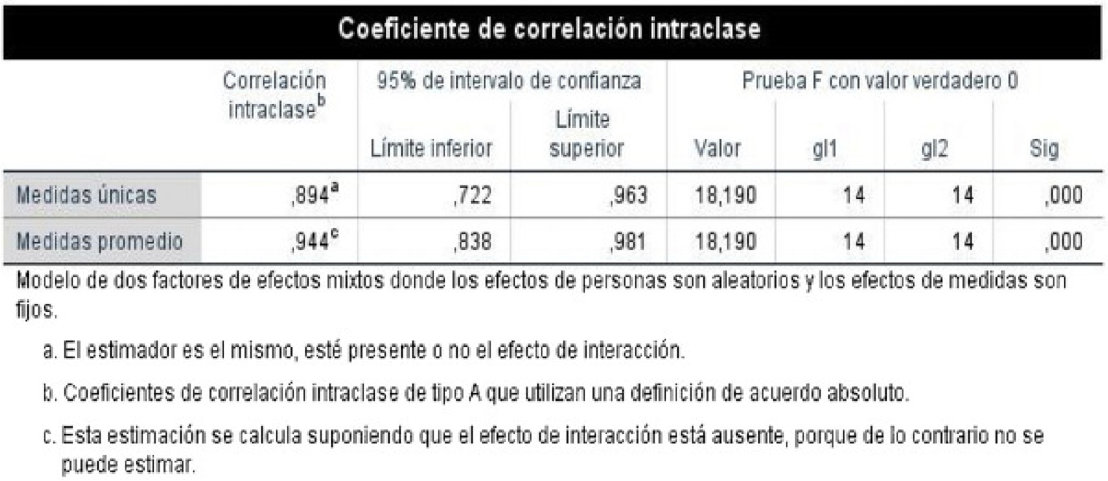
Tabla de ANOVA
○ Ninguno
● Prueba F
○ Chi-cuadrado de Friedman
○ Chi-cuadrado de Cochran

□ T-cuadrado de Hotelling
□ Prueba de aditividad de Tukey
☑ Coeficiente de correlación intraclase

Modelo: Mixto de dos factores
Tipo: Consistencia
Intervalo de confianza: 95 %
Valor de prueba: 0

Picture 6: One of the drop-down menus that appears while we try to obtain the CIC using the IBM SPSS Statistics application (Pérez, JM, & Martin, PP, 2023).

We have commented that when the CIC is in the interval between 0.75 and 1, it is considered an excellent correlation, therefore, we can say that both measurements are valid and correlate (Picture 7).



Coeficiente de correlación intraclase

	Correlación intraclase[b]	95% de intervalo de confianza		Prueba F con valor verdadero 0			
		Límite inferior	Límite superior	Valor	gl1	gl2	Sig
Medidas únicas	,894[a]	,722	,963	18,190	14	14	,000
Medidas promedio	,944[c]	,838	,981	18,190	14	14	,000

Modelo de dos factores de efectos mixtos donde los efectos de personas son aleatorios y los efectos de medidas son fijos.

a. El estimador es el mismo, esté presente o no el efecto de interacción.

b. Coeficientes de correlación intraclase de tipo A que utilizan una definición de acuerdo absoluto.

c. Esta estimación se calcula suponiendo que el efecto de interacción está ausente, porque de lo contrario no se puede estimar.

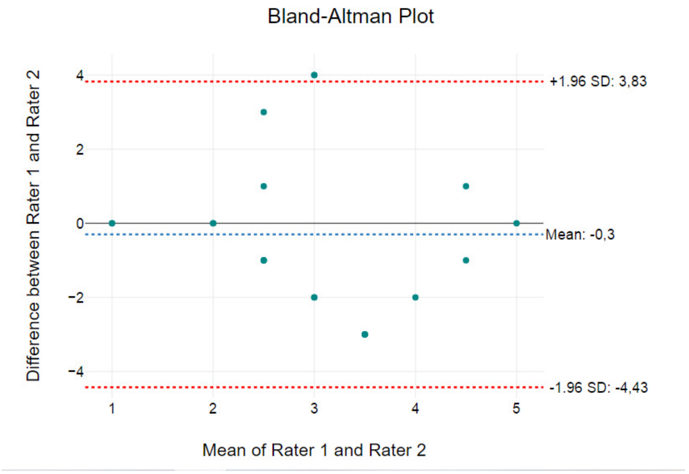
Picture 7: Results obtained in the reliability analysis in IBM SPSS Statistics (Pérez, JM, & Martin, PP, 2023).

.

3.3. Bland and Altman method

Bland and Altman method can be used, which consists of graphically representing, in a scatter diagram, the mean of the two measurements as the best estimate of the true value, against the absolute difference between the two values . The graph provides a visual representation of the difference between two measurements on the x-axis and the average of the two measurements on the x-axis.

The graph includes a horizontal line at the mean difference and two lines, called limits of agreement, at a distance of 1.96 standard deviations above and below the first. If the differences between pairs of observations follow approximately a normal distribution and the values tend to be stable over the entire measurement range, 95% of those differences are expected to fall within the limits of agreement. This allows the degree of agreement between the two methods to be assessed graphically, in a simple way (Picture 8).



Picture 8: Bland and Altman graph taken as an example of how the results obtained by two meters would be represented by these methods. Image obtained from DataTab (https://datatab.es/tutorial/bland-altman-plot).

As I have already mentioned, this method is very useful to measure the agreement between measurements regardless of the scale of the measurements. But it is also a very useful method since it allows us to know more information while at the same time obtaining the agreement. We can stand out :

- **Identify any systematic bias**: The graph can be used to identify any systematic bias or random error in the data. For example, if the mean difference between the two measurements is consistently positive or negative, it may indicate a systematic bias in one of the measurement techniques. Additionally, if the spread of the points on the graph is greater than the standard deviation, this may indicate the presence of random error in the data.

- **Finding outliers in data: Another important aspect of** Bland -Altman plots is that they can be used to identify outliers in data. Outliers can have a significant impact on the results of a study, and it is important to identify them to understand the overall agreement between the two measurement techniques. Outliers can be identified by looking for points that fall outside the lines that represent the standard deviation of the mean difference.

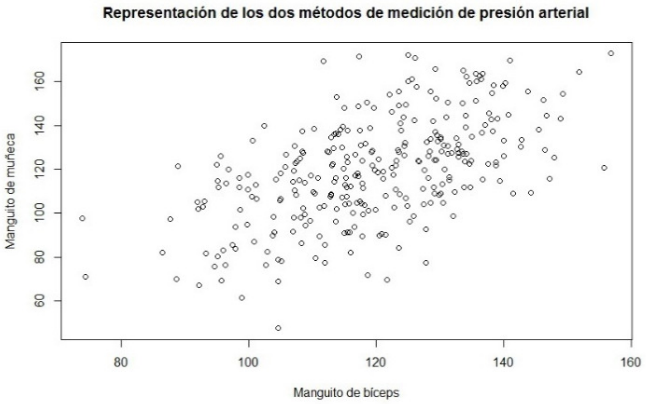
Something that is important to keep in mind when carrying out this method is that the degree of dispersion must be uniform. It may happen that the agreement is acceptable in a certain range of values, but not in another (for example, very high or low values), in which the dispersion is unacceptable. This effect can sometimes be corrected by transforming the data (for example, logarithmic transformation), although the usefulness of the measurement in that interval must always be considered.

3.4 Bland and Altman method applied to a real case

Suppose we want to assess the reliability of a new wrist blood pressure monitor to measure blood pressure. We took a sample of 300 healthy schoolchildren and measured their pressure twice. The first with a conventional arm cuff, obtaining an average systolic pressure of 120 mmHg. and a standard deviation of 15 mmHg. The second, with a new wrist blood pressure monitor, with which we obtained an average of 119.5 mmHg. and a standard deviation of 23.6 mmHg. The question we ask ourselves is the following: considering the arm cuff as a reference standard, is the determination of blood pressure with the wrist blood pressure monitor reliable?

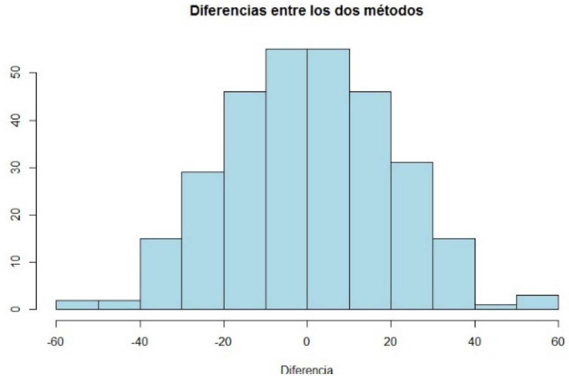
The logical thing is to think that the two methods will not always coincide, so the first thing we must ask ourselves is how much it is reasonable for them to differ to give validity to the results. This difference must be defined before comparing the two methods and establishing the sample size necessary to make the comparison. In our case we are going to consider that the difference should not be greater than one standard deviation of what was obtained with the reference method, which is 15 mmHg.

The first step we can take is to examine the data. To do this, we make a dot diagram representing the results obtained with the two methods (Picture 9). It seems that there is a certain relationship between the two variables since we observe that the cloud of points is very close to having a correct proportion that the data on the Y axis is the same or very similar to the X axis (For example X= 120, Y=120), so that the measurements increase and decrease in the same direction.



Picture 9: Point cloud representation of the biceps cuff measurements of the 300 schoolchildren. Y-axis is one of the sphygmomanometers and X-axis is another sphygmomanometer.

Another possibility is to examine what the differences are like. If there was good agreement, the differences between the two methods would still be normally distributed around zero. We can check this by making the histogram with the differences of the two measurements (measurement 1 – measurement 2). Indeed, it seems that its distribution conforms quite well to a normal one (Picture 9).



Picture 9: Histogram representing the differences in measurements between the two sphygmometers .

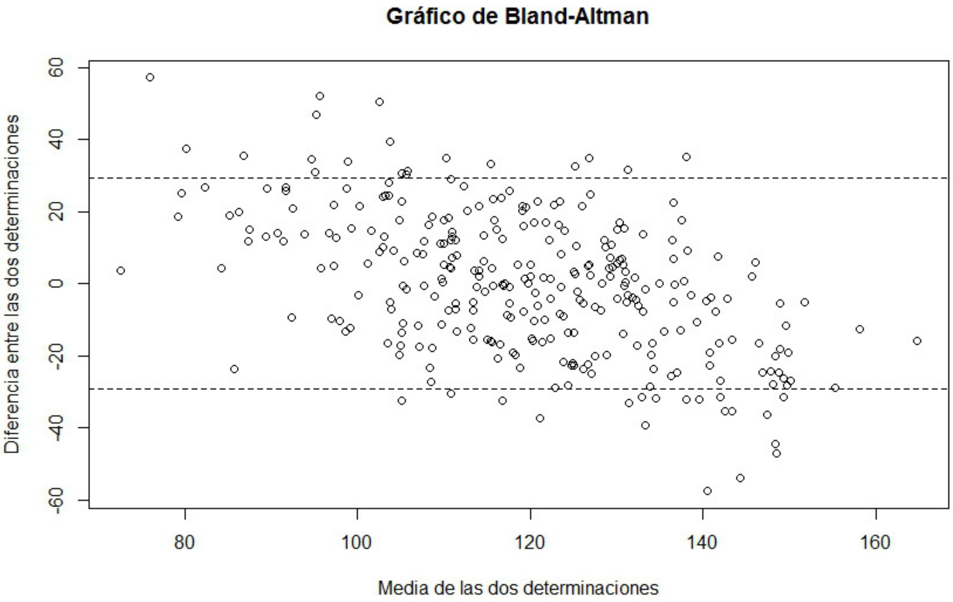
What will give us more information will be to represent the average of each pair of measurements against their difference. In this representation, the Y axis corresponds to the differences between the paired values of the measurement of the reference sphingometer and the new sphingometer , that is, the subtraction of both values, that is, the measurement of the student minus the measurement with the other sphygmometer in the same school (measurement A – measurement B). In this way when the result of the difference is 0

$$Y\ axis = \ Sphygmeter\ A - \ Sphygmeter\ B$$

While the X axis represents the respective value of the average of both (A+ B)/ 2).

$$X\ axis = \frac{Sphygmeter\ A + \ Sphygmeter\ B}{2}$$

As can be seen, the points are grouped, more or less, around a line (at zero) with a degree of dispersion that will be determined by the extent of the differences in results between the two methods. The greater this degree of dispersion, the worse the agreement between the two methods. In our case, we have drawn the lines that coincide with one standard deviation below and above the zero mean, which were the limits that we considered acceptable between the two methods to consider a good agreement. We can see that there are quite a few points that fall outside the limits, so we would have to assess whether the new method reproduces the results reliably (Picture 10) (Molina, M, 2015).



Picture 10: Bland -Altman graph representing the measurements of the schoolchildren with the two sphygmometers following the previously mentioned formulas.

4. Articles to expand information

In my opinion, the most advisable thing is to read clinical and biomedical cases and see what statistical analysis methodology researchers follow in their research to carry out the correct reliability and validation study. I am attaching in this way several papers that I think could be of interest and the reason for their choice.

To explore more applications of CIC in biomedicine. An example could be the following:

- Fernández, MG, & Escobar, JZ (2012). Reliability and correlation in the evaluation of knee mobility using goniometer and inclinometer. *Physiotherapy* , *34* (2), 73-78.

We can also study how the CIC could be applied in clinical trials.

- Thompson, D.M., Fernald, D.H., & Mold, J.W. (2012). Intraclass correlation coefficients typical of cluster-randomized studies: estimates from the Robert Wood Johnson Prescription for Health projects. *The Annals of Family Medicine* , *10* (3), 235-240.

- Brewer, BW, Van Raalte, JL, Petitpas, AJ, Sklar, JH, Pohlman, MH, Krushell , RJ, ... & Weinstock, J. (2000). Preliminary psychometric evaluation of a measure of adherence to clinic-based sport injury rehabilitation. *Physical Therapy in Sport* , *1* (3), 68-74.

To learn more about the CIC and to see another analysis, Person 's correlation , to study to what extent the CIC is better or worse than other methods, comparing it in a clinical case.

- Molina, CGE, Rodríguez, VMV, de Celis, EMR, Rodríguez, EB, Ávila, GG, & Ruiz, CEC (2006). Intraclass correlation coefficient vs. Pearson correlation of capillary glycemia by reflectometry and plasma glycemia. Internal Medicine of Mexico, 22(3), 165-171.

To expand information about the Bland and Altman method we could explore:

- Mantha, S., Roizen, M.F., Fleisher, L.A., Thisted, R., & Foss, J. (2000). Comparing methods of clinical measurement: reporting standards for Bland and Altman analysis. *Anesthesia & Analgesia* , *90* (3), 593-602.

- Taffe , P. (2021). When can the Bland & Altman limits of agreement method be used and when it should not be used. *Journal of clinical epidemiology* , *137* , 176-181.

It should be noted that there are more analyses, for example, when you want to do a validation study of qualitative measures, a study that is often done to study the correlation of qualitative measures is the kappa index . To learn more we could explore:

- by Ullibarri Galparsoro, L., & Pita Fernández, S. (1999). Agreement measures: the Kappa index. Cad Aten Primary, 6, 169-171.

And to see a practical example in the clinic about the Kappa index we could explore:

- Bes-Rastrollo , M., Pérez Valdivieso, JR, Sánchez-Villegas, A., Alonso, AMGM, & Martínez-González, MA (2005). Validation of participants' self-reported weight and body mass index from a cohort of college graduates. Rev Esp Obes , 3(6), 183-9.

5. **Bibliography**

- **Websites:**

  - Cabo, J. (last view 11/19/2023). Biomedical research. General concepts. CEF – Health Management. https://www.gestion-sanitaria.com/1-investigacion-biomedica-conceptos-generales.html

  - Unknown author (published 03/17/2023, last view 11/19/2023). The importance of biomedical research. Bind. The Internet University. https://www.unir.net/salud/revista/investigacion-biomedica/

  - Molina, M (published 03/13/2015, last view 11/27/2023). Another stone not to trip over. The Bland -Altman method for measuring agreement. https://anestesiar.org/2015/otra-piedra-con-la-que-no-tropezar-el-metodo-de-bland-altman-para-medir-atrabajo/#:~:text=El%20m%C3%A9todo%20de%20Bland%2DAltman%20para%20medir%20agreement,-by%20Manuel%20Molina&text=Dice%20el%20refr%C3%A1n%20que%20el,de%20give us%20account%20of%20it.

  - Cabo, J. (last view 11/22/2023). Datatab . Tutorials. Bland -Altman plot https://datatab.es/tutorial/bland-altman-plot .

- **Papers:**

  - Bland, J.M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* , *327* (8476), 307-310.

  - Correa-Rojas, J. (2021). Correlation coefficient _ Intraclass : Intraclass Correlation Coefficient: Applications to estimate the temporal stability of a measuring instrument. *Sciences Psychological* , *15* (2).

  - Hanson, M.A., Barreiro, P.G., Crosetto, P., & Brockington, D. (2023). The strain on scientific publishing. *arXiv preprint arXiv:2309.15884* .

  - Latour, J., Abraira , V., Cabello, JB, & Sánchez, JL (1997). Clinical measurements in cardiology: validity and measurement errors. *Spanish Journal of Cardiology* , *50* (2), 117-128.

  - Manterola, C., Grande, L., Otzen , T., García, N., Salazar, P., & Quiroz, G. (2018). Reliability, precision or reproducibility of measurements. Assessment methods, usefulness and applications in clinical practice. *Magazine chilena de infectologia* , *35* (6), 680-688.

  - Pérez, JM, & Martin, PP (2023). Intraclass correlation coefficient. *Family medicine. SEMERGEN* , *49* (3), 101907.

  - Prieto, L. (1998). The evaluation of reliability in clinical observations : the intraclass correlation coefficient . Med Clin (Barc), 110, 142-145.

  - Solano - Flores, G., & Nelson - Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching* , *38* (5), 553-573.