Ciencia, Tecnología y Salud, Vol. 10, Núm. 1, 2023
ISSN: 2410-6356 (impreso) / 2409-3459 (electrónico)
https://doi.org/10.36829/63CTS.v10i1.1490

Artículo de revisión / Review Article

Técnicas estadísticas utilizadas en la comparación de métodos cuantitativos de medición

Statistical techniques used in method comparison for quantitative measurements

Jorge Nave [1], Federico Nave [2]*

[1]Facultad de Ciencias Químicas y Biológicas, Universidad Mariano Gálvez de Guatemala
[2]Instituto de Investigaciones Químicas y Biológicas, Facultad de Ciencias Químicas y Farmacia, Universidad de San Carlos de Guatemalaa

***Autor al que se dirige la correspondencia:** fedenave@profesor.usac.edu.gt

Recibido: 07 de junio 2022 / Revisión: 02 de septiembre 2022 / Aceptado: 06 de enero 2023

Resumen

La comparación de métodos cuantitativos de medición es una práctica importante en las áreas de salud y científico-tecnológica, así como en todo caso en el que se realizan procedimientos de medición, permite analizar muestras por el método del cual se desea conocer su desempeño analítico y por el método de referencia u otro probadamente efectivo, para determinar si pueden obtenerse los mismos resultados. Se debe considerar un procedimiento de muestreo adecuado, utilizando un diseño pareado, en el cual cada muestra debe ser analizada por ambos métodos y establecer que la variable medida cumpla con los requisitos necesarios para poder aplicar estadísticos adecuados a variables cuantitativas. Para el análisis de estas comparaciones se debe evaluar la reproducibilidad de los métodos en términos de confiabilidad y concordancia, así como un análisis de regresión que brindará información sobre los errores proporcional, constante y aleatorio, mediante el análisis de la pendiente, la intersección y la desviación estándar de los residuos, respectivamente. Las técnicas más apropiadas para la comparación de métodos cuantitativos deben considerar una combinación que incluya el procedimiento de Bland y Altman para concordancia, el coeficiente de correlación intraclase o coeficiente de correlación de concordancia para evaluar confiabilidad y al menos una técnica de regresión (regresiones lineal ordinaria, Deming o Passing-Bablok), debiéndose reportar todos los elementos necesarios para su interpretación y así tener la mejor información estadística para tomar decisiones sobre la reproducibilidad e intercambiabilidad de los métodos; en ninguna circunstancia debe usarse el coeficiente de correlación de Pearson.

Palabras clave: Reproducibilidad, confiabilidad, concordancia, acuerdo, Bland-Altman

Abstract

The methods comparison for quantitative measurements is an important practice both in the health and scientific-technological areas, as well as in any case in which measurement procedures are carried out, it allows the analysis of samples by the method whose analytical performance is desired, and by the method of reference or another that has proven effectiveness, to determine if the results can be alike. An adequate sampling procedure should be considered, using a paired design, in which each sample must be analyzed by both methods and establishing that the variable measured meets the requirements in order to apply adequate statistics for quantitative variables. For the analysis of these comparisons, the reproducibility of the methods should be evaluated in terms of reliability and agreement, as well as a regression analysis that will provide useful information on the proportional, constant and random errors, through the analysis of the slope, the intercept and the standard deviation of the residuals, respectively. The most appropriate techniques for comparing quantitative methods should consider a combination that includes the Bland and Altman procedure for concordance, the intraclass correlation coefficient or concordance correlation coefficient to assess reliability, and at least one regression technique (ordinary linear, Deming or Passing-Bablok regressions), all the elements necessary for its interpretation must be reported and thus have the best statistical information to take decisions on the reproducibility and interchangeability of the methods; under no circumstances the Pearson's correlation coefficient should be used.

Keywords: Reproducibility, reliability, concordance, agreement, Bland-Altman

Introducción

La comparación de métodos cuantitativos de medición se refiere a analizar muestras por el método del cual, se desea conocer su desempeño analítico y por el método de referencia u otro probadamente efectivo o preexistente en el laboratorio, lo cual es una práctica importante en medicina, laboratorios e investigaciones en general. Las razones de estas comparaciones pueden ser muy diversas, cambio de marca o de aparatos, aspectos económicos, rapidez o mayor accesibilidad y lo que se quiere responder es, si los dos métodos concuerdan lo suficientemente bien en sus mediciones para justificar el cambio y si pueden usarse indistintamente (Bartko, 1994; Bland & Altman, 1995b; Chen & Kao, 2021; Stevens et al., 2017). Otra razón para realizar estas comparaciones es cuando se desarrolla un nuevo método de medición, ya que se hace necesario comparar su desempeño con un método preexistente (McDemid, 2021; Morgan & Aban, 2016; Rojas et al., 2016). Debe haber evidencia que un nuevo método de medición funciona bien antes de que pueda ser usado en la práctica, sin embargo, no existe una definición clara de lo que se considera satisfactorio o lo que es un buen rendimiento (Altman, 2009). Siempre que el rendimiento del ensayo o método analítico sea estable, las diferencias sistemáticas observadas en la comparación de métodos se reflejan en los resultados cuando se utilizan muestras clínicas o de campo, lo que permite su aplicación al mundo real (Kalaria et al., 2022).

Lo que se quiere determinar es si las medidas de ambos métodos están lo suficientemente cerca de tal manera que uno pueda reemplazar al otro con suficiente precisión (Altman & Bland, 1983; Bartlett & Frost, 2008). La forma usual de efectuar estos estudios es realizar simultáneamente observaciones por ambos métodos en la misma muestra, comparar los valores obtenidos e investigar las fuentes de error o sesgo analítico (Bland & Altman, 1995b; Jensen & Kjelgaard-Hansen, 2006); específicamente, se debe obtener información sobre la naturaleza proporcional y constante del error sistemático y cuantificar el error aleatorio entre los métodos (Westgard, 1998). Por lo tanto, estos estudios permiten la investigación de las diferencias pareadas y estimar su reproducibilidad en términos de confiabilidad y concordancia (Bartlett & Frost, 2008).

Se debe analizar la confiabilidad, investigando el error analítico total, que es la suma del error aleatorio y el error sistemático, ya que ningún método proporciona una medición inequívocamente correcta; por otra parte, también se debe evaluar la concordancia o grado de acuerdo entre ambos métodos, es decir el estudio estadístico del comportamiento de las diferencias entre las mediciones, lo que bajo condiciones ideales se reflejaría en que las diferencias fueran iguales a 0, es decir que las medidas obtenidas por un método u otro, darán exactamente el mismo resultado (Bartlett & Frost, 2008; Giavarina, 2015; Jensen & Kjelgaard-Hansen, 2006).

A través del tiempo, se han desarrollado y utilizado diversas técnicas estadísticas para probar la confiabilidad y concordancia de métodos o procedimientos que generan resultados cuantitativos, pero responder qué método es el mejor, aún es tema de debate ya que casi todas las técnicas que se han propuesto han sido criticadas (Zaki et al., 2012). El análisis inapropiado de los estudios de comparación de métodos conlleva a conclusiones incorrectas acerca del rendimiento del método o instrumento que se está evaluando (Bland & Altman, 1995b; Zaki et al., 2012). El conocimiento de la metodología estadística correcta para realizar la comparación de métodos es un campo que debe considerarse de importancia y que va más allá de lo académico (McDemid, 2021), ya que se han publicado muchos artículos con aplicación incorrecta de técnicas que son reproducidas por otros investigadores (Altman & Bland, 1983). Se ha demostrado que la revisión por pares no ha impedido la publicación de artículos con técnicas estadísticas inapropiadas, ya que muchas veces estas son replicadas simplemente porque se presentan en la literatura (Altman, 2009).

Lo anterior ha sido demostrado a través de estudios de síntesis del conocimiento relacionados con revisiones de literatura científica sobre el uso de técnicas estadísticas en la comparación de métodos de medición, en los que con frecuencia se observan deficiencias y errores en la ejecución e interpretación de las técnicas estadísticas, así como aplicaciones inadecuadas (Abu-Arafeh et al., 2016; Berthelsen & Nilsson, 2006; Chhapola et al., 2015; Dewitte et al., 2002; Gerke, 2020a; Mantha et al., 2000; Zaki et al., 2012; Zaki, Bulgiba, Nordin et al., 2013).

El objetivo de esta revisión es presentar un resumen de las principales técnicas estadísticas aplicables para evaluar la confiabilidad y concordancia de métodos cuantitativos de medición que puedan ser de utilidad en el campo de trabajo del área médica y científico tecnológica cuando se miden variables cuantitativas y se tenga que evaluar alguna nueva metodología, equipos o juegos de reactivos. Cada técnica se ha desa-

rrollado explicando sus fundamentos, características, limitaciones y proporcionando algunos lineamientos clave para su correcta aplicación e interpretación; en lo que corresponde a confiabilidad se presentan los coeficientes de correlación intraclase (CCI) y coeficiente de correlación de concordancia de Lin (CCC); para evaluar la concordancia, el procedimiento gráfico de Bland y Altman y sus límites de concordancia, así como la prueba de sesgo (*t* pareada) y el análisis de regresión como una técnica complementaria. Se presenta una sección dedicada a aclarar las razones por las cuales en la comparación de métodos cuantitativos no debe realizarse la correlación de Pearson.

Contenido

Comparación estadística de métodos cuantitativos

Reproducibilidad (confiabilidad y concordancia)

En el campo de la comparación de métodos, así como en validación y en control de calidad, en los que se miden variables cuantitativas, existen muchos términos que se han usado indistintamente sin una clara diferencia entre ellos; tal es el caso de confiabilidad, fiabilidad, repetibilidad, reproducibilidad y concordancia (Bartlett & Frost, 2008; de Vet et al., 2006; Hernaez, 2015), así como exactitud, precisión, validez y generalización (*generalizability*) (Barnhart et al., 2007). Varios autores coinciden en que la forma estadística más apropiada para comparar métodos es realizando el análisis de la reproducibilidad (Bartlett & Frost, 2008; Hernaez, 2015; Lin, 1989; Martelli Filho et al., 2005; Rojas et al., 2016; Watson & Petrie, 2010); esta se refiere al análisis de la variación en las mediciones realizadas sobre un material de prueba idéntico, pero en diferentes condiciones (Bartlett & Frost, 2008; Jensen & Kjelgaard-Hansen, 2006), las cuales pueden deberse a diferentes operadores, instrumentos, laboratorios, intervalos de tiempo o métodos de medición (Hernaez, 2015; Jensen & Kjelgaard-Hansen, 2006), pero específicamente en la comparación de métodos, lo que se evalúa es si estos producen esencialmente el mismo resultado (de Vet et al., 2006).

Cuando se realizan dos mediciones que supuestamente deben ser iguales, pero las condiciones son diferentes, la diferencia entre estos valores se conoce como error de medición, cuyos componentes generalmente se desconocen, por lo que solo es posible estimar la cantidad de la medición que sea atribuible al error y la cantidad que representa una lectura precisa, lo cual constituye una medida de confiabilidad (Bruton et al., 2000). Por lo tanto, la confiabilidad mide la consistencia de dos lecturas obtenidas por dos instrumentos o métodos diferentes, pero en condiciones distintas, refiriéndose entonces a la reproducibilidad de los datos de evaluación o de los resultados (Barnhart et al., 2007; Cortés-Reyes et al., 2010; Downing, 2004; Jensen & Kjelgaard-Hansen, 2006), la cual depende de la variabilidad que se presente en la muestra (Hernaez, 2015).

La concordancia describe hasta qué punto las variables miden la misma entidad, atributo o resultado en el mismo nivel de medición, mide qué tan cerca están las puntuaciones para las mediciones repetidas y está relacionada con el error de medición, definido como sistemático y aleatorio (Hernaez, 2015; Jensen & Kjelgaard-Hansen, 2006; Watson & Petrie, 2010). En un estudio de comparación de métodos, habrá diferencias debido a la variabilidad inherente en cada uno de los métodos de medición, así como un posible sesgo entre las mediciones de los métodos (Bartlett & Frost, 2008; Jensen & Kjelgaard-Hansen, 2006). La concordancia entre las mediciones es una característica de los métodos de medición involucrados y de su rango analítico, que no depende de la población en la que se realizan las mediciones (Bartlett & Frost, 2008). Si las mediciones de los dos métodos se realizan en la misma métrica o escala, es posible cuantificar su concordancia (Bartlett & Frost, 2008; Hernaez, 2015; Vetter & Schober, 2018), la que se puede definir como el grado en que dos o más observadores, métodos, técnicas u observaciones están de acuerdo sobre la misma variable medida, evaluándose qué tan similares o cercanas están las lecturas, lo que es útil para evaluar la reproducibilidad de un ensayo, instrumento o método (Barnhart et al., 2007; Bartlett & Frost, 2008; Cortés-Reyes et al., 2010; Lin, 1989, 1992; Rojas et al., 2016).

Se han desarrollado técnicas estadísticas que miden la confiabilidad (*reliability*) y las que miden concordancia (*agreement*). Dentro de los parámetros de confiabilidad para variables cuantitativas continuas se pueden utilizar los llamados índices escalados, como los CCI y el CCC. En tanto que, entre los parámetros de concordancia para variables continuas están el gráfico de Bland y Altman con el cálculo de límites de concordancia y la prueba de *t* pareada o prueba de sesgo. Los parámetros de confiabilidad tienen un

valor adimensional en una escala entre 0 y 1 (de allí el término de escalados), mientras que los parámetros de concordancia se expresan en la escala real de medición y se les denomina no escalados (Bruton et al., 2000; de Vet et al., 2006; Hernaez, 2015; Rojas et al., 2016).

La confiabilidad y la concordancia no son propiedades fijas de los instrumentos o métodos de medición, sino que son el producto de las interacciones entre estos, los sujetos u objetos que se miden y el contexto de la evaluación. Las estimaciones de confiabilidad y concordancia se ven afectadas por diversas fuentes de variabilidad en el entorno de medición y el enfoque estadístico (Kottner et al., 2011).

Técnicas estadísticas para evaluar la confiabilidad

CCI: Aunque originalmente la correlación intraclase fue definida por Galton en 1889 (Barnhart et al., 2007), el concepto moderno del CCI fue introducido por el estadístico Ronald A. Fisher en 1950, consistente en una modificación del coeficiente de correlación de Pearson, definido como la correlación que se obtiene de mediciones divididas en clases o grupos que están relacionados, en las que se puede determinar la media aritmética (en adelante simplemente media o promedio) y desviación estándar, que son distintas para las dos clases (Koo & Li, 2016; Liljequist et al., 2019).

Varios autores coinciden en que es la técnica más apropiada para determinar la confiabilidad cuando se comparan dos métodos de medición (Barnhart et al., 2007; Bartko, 1994; Cortés-Reyes et al., 2010; Hernaez, 2015; Koo & Li, 2016; Liljequist et al., 2019; Vetter & Schober, 2018; Watson & Petrie, 2010).

A pesar que cuando se menciona el CCI se hace en singular, la realidad es que se han desarrollado diversas formas de cálculo tanto por métodos paramétricos como no paramétricos, que han resultado en al menos 10 índices, los cuales implican distintos supuestos en sus cálculos que pueden dar diferentes resultados cuando se aplican al mismo conjunto de datos, también las formas de informar los CCI pueden variar entre investigadores y darán lugar a diferentes interpretaciones (Koo & Li, 2016; Kusunoki et al., 2009; Müller & Büttner, 1994; Vetter & Schober, 2018).

Varias versiones de CCI han evolucionado a partir del análisis de varianza (ANOVA por sus siglas en inglés) de dos vías, donde las muestras de sujetos son medidas por los mismos evaluadores múltiples, utilizando el término de "evaluadores" en un sentido amplio, que abarca no solamente a evaluadores humanos, sino también a dispositivos o métodos de medición. Se tiene entonces que el CCI mide la reproducibilidad intra e interevaluadores; es decir, la reproducibilidad intra evaluadores será la medida de la variabilidad observada dentro de las muestras y la reproducibilidad interevaluadores, será la medida de variabilidad entre los evaluadores o más ampliamente, los métodos de medición (Koo & Li, 2016; Müller & Büttner, 1994). El CCI evalúa el tamaño de los componentes de la varianza entre clases y dentro de estas, describiendo la proporción de la variación total, la cual es explicada por las diferencias entre las clases (evaluadores, observadores, instrumentos o métodos de medición), representando la varianza entre los pares de valores medidos por cada evaluador o método, expresada como una proporción de la varianza total de las observaciones (Watson & Petrie, 2010).

El CCI que mejor se ajusta para la comparación de métodos cuantitativos, es aquel que considera que las muestras a analizar representan una muestra aleatoria de la población que será estudiada, en tanto que los métodos e instrumentos que se emplearán en la investigación, constituyen elementos fijos en el diseño, dado que no cambian. Esto representa un planteamiento que, en términos de diseños experimentales, se denomina modelo de efectos mixtos, lo que determina el tipo de ICC que se debe usar, ya que su definición matemática deriva del modelo de ANOVA de dos vías para efectos mixtos sin interacción, propuesto originalmente por Shrout y Fleiss (Koo & Li, 2016; Liljequist et al., 2019).

Este modelo de efectos mixtos de dos factores para el cálculo del CCI, es el que debe utilizarse en el estudio de confiabilidad para la evaluación de métodos de medición, en los cuales la misma muestra se evalúa por dos métodos, constituyendo lo que en diseños experimentales se denomina medidas repetidas (en este caso, emparejadas o en bloques), dichos métodos constituyen el efecto fijo del modelo, en tanto que, las muestras obtenidas de la población constituyen el elemento aleatorio (Barnhart et al., 2007; Koo & Li, 2016; Liljequist et al., 2019; Manterola et al., 2018).

La forma matemática para el cálculo del CCI a partir de un modelo de efectos mixtos, se desarrolla como ya se mencionó, del resultado del ANOVA sin interacción, que considera dos componentes: la variabilidad debida a las diferencias que se dan dentro de los sujetos o muestras y la debida a las diferencias entre los evaluadores, instrumentos o métodos de medición que se están comparando. A estas fuentes de variación se le suman los residuos que representan la

variación no explicada. De la tabla del ANOVA, se utilizan entonces los valores denominados cuadrado medio (*CM*), que son los estimadores de las varianzas de estos componentes. Varios autores han compilado las diversas fórmulas para el cálculo de los coeficientes de correlación intraclase, entre las que se incluye la que corresponde al modelo de efectos mixtos o de medidas repetidas para la comparación de métodos (Koo & Li, 2016; Kusunoki et al., 2009; Liljequist et al., 2019), la cual se presenta a continuación utilizando una nomenclatura aplicada a la comparación de métodos:

Elementos generales:

- Número de métodos a comparar (*k*)
- Número de muestras a analizar (*n*)

Fuentes de variación:

- Muestras (provenientes de pacientes o sujetos de estudio)
- Métodos (métodos de medición que se están comparando)
- Residuo (error, variación no explicada o debida al azar)

Parámetros del modelo (varianzas):

- $\sigma^2_{\text{muestras}}$
- $\sigma^2_{\text{métodos}}$
- $\sigma^2_{\text{residuo}}$
- $\sigma^2_{\text{total}} = \sigma^2_{\text{muestras}} + \sigma^2_{\text{métodos}} + \sigma^2_{\text{residuo}}$

Estimadores de las varianzas (según el resultado del ANOVA):

- Cuadrado medio de las muestras (CM_{muestras})
- Cuadrado medio de los métodos ($CM_{\text{métodos}}$)
- Cuadrado medio del residuo (CM_{residuo})

Formula de definición:

$$\rho = \frac{\sigma^2_{muestras}}{\sigma^2_{muestras} + \sigma^2_{métodos} + \sigma^2_{residuo}}$$

Fórmula operativa a partir del resultado del ANOVA:

$$CCI = \frac{CM_{muestras} - CM_{residuo}}{CM_{muestras} + (k-1)CM_{residuo} + \left(\frac{k}{n}\right)(CM_{métodos} - CM_{residuo})}$$

Los valores del CCI pueden oscilar entre 0 y 1, donde el 0 indica ausencia de confiabilidad (reproducibilidad entre métodos) y el 1, la confiabilidad absoluta de los resultados obtenidos (Koo & Li, 2016); aunque el CCI puede ser negativo, su interpretación sigue sin estar clara (Costa-Santos et al., 2011), se ha considerado que un valor negativo no tiene legitimidad teórica (Giraudeau, 1996), es inverosímil (Gulliford et al., 2005) o se trata de una mala y desafortunada estimación (Liljequist et al., 2019). Para su interpretación se han utilizado diversas escalas sobre lo que representan los valores del coeficiente, algunas de ellas como analogía a la interpretación del coeficiente *kappa* para evaluar concordancia entre métodos de respuesta categórica, como por ejemplo la escala de Fleiss y colaboradores (2003) con tres clasificaciones y la de Landis y Koch con seis clasificaciones (Mandeville, 2005); sin embargo, la escala propuesta por Koo y Li de cuatro clasificaciones, es la más apropiada, ya que fue desarrollada específicamente para interpretar el CCI en términos de confiabilidad aplicada a métodos cuantitativos, estableciendo que valores inferiores a .50 indican una confiabilidad pobre, entre .50 y .75 confiabilidad moderada, entre .75 y .90 buena confiabilidad y superiores a .90 excelente confiabilidad, entendiéndose como confiabilidad la reproducibilidad entre métodos (Koo & Li, 2016; Liljequist et al., 2019).

No debe perderse de vista que la estimación del CCI obtenida de un estudio de confiabilidad es solo un valor esperado del verdadero CCI, por lo que es necesario realizar el cálculo de un intervalo de confianza y probar si el valor del CCI obtenido excede significativamente los valores sugeridos mencionados anteriormente usando inferencia estadística, es decir, pruebas de hipótesis (Koo & Li, 2016).

Limitaciones: La restricción o limitación más importante al uso del CCI es que se trata de una prueba paramétrica derivada del análisis de varianza y, por lo tanto, se debe considerar que es aplicable únicamente bajo condiciones de normalidad de las distribuciones de las variables, igualdad de varianzas (homocedasticidad), independencia entre los errores producidos por los evaluadores (Barnhart et al., 2007; Müller &

Büttner, 1994) y por tratarse de un diseño de medidas repetidas, en el caso que se realicen varias repeticiones de las medidas al menos con uno de los métodos, se debe cumplir con el principio de esfericidad, es decir que las variaciones de las diferencias entre todos los pares no sean diferentes (Mishra et al., 2019; Oleson et al., 2019). Los resultados obtenidos mediante el CCI están expresados en términos absolutos, no en la escala de medición; sin embargo, la escala de medición sí influye grandemente en el resultado del coeficiente, ya que cuanto más amplio sea el rango de medición, mejor será el resultado; así también, el CCI se ve afectado por la presencia de valores atípicos (Müller & Büttner, 1994).

CCC o coeficiente de Lin: Fue desarrollado como un nuevo índice para evaluar la confiabilidad de un ensayo, instrumento o método, por medio de la reproducibilidad; en su nombre se indica el término "concordancia", debido a que para su cálculo es necesario determinar las diferencias entre mediciones, incluyendo una estimación de precisión y exactitud, asegurándose que este coeficiente es superior al CCI (Akoglu, 2018; Cortés-Reyes et al., 2010; Lin, 1989, 1992).

El CCC evalúa el grado de acuerdo entre dos lecturas de la misma muestra, modificando el coeficiente de correlación de Pearson al evaluar no solo qué tan cerca están los datos de la línea de mejor ajuste, sino también qué tan lejos se encuentran de la línea de 45° a través del origen, esta línea de 45° representa el acuerdo perfecto cuando los resultados de un método se gra-fican contra el otro; es decir, que mide el grado en que los pares caen en la línea de 45°, siendo un producto de la cantidad de acuerdo entre los evaluadores (métodos de medición) y el coeficiente de correlación de Pearson (Cortés-Reyes et al., 2010; Lin, 1989; Vetter & Schober, 2018; Watson & Petrie, 2010). El CCC se ve afectado entonces por el alejamiento de los datos de la línea de acuerdo perfecto y también por las diferencias en las lecturas entre ambos métodos (Figura 1).

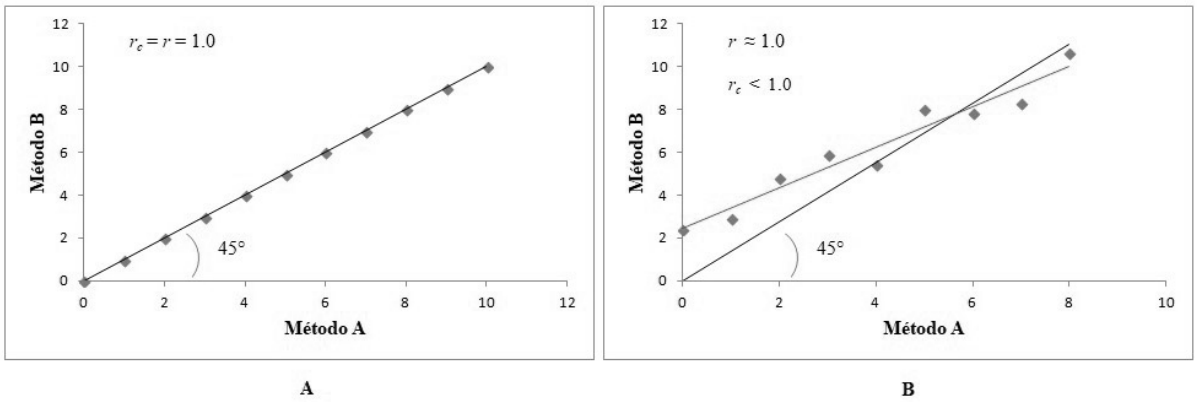
Para su cálculo, Lin determinó que si cada par de mediciones en n muestras independientes, y_1 y y_2, están en perfecto acuerdo, la suma de las diferencias al cuadrado entre ambas sería 0, considerando además las variaciones de los métodos individuales y de sus diferencias, desarrollando las correspondientes fórmulas de definición y operativa (Carrasco & Jover, 2004; Lin, 1989; Nickerson, 1997; Watson & Petrie, 2010).

Parámetros del modelo:

- μ_1 (Promedio poblacional de las mediciones del método 1)
- μ_2 (Promedio poblacional de las mediciones del método 2)
- σ^2_1 (Varianza poblacional de las mediciones del método 1)
- σ^2_2 (Varianza poblacional de las mediciones del método 2)
- σ_{12} (Desviación estándar poblacional común para ambos métodos)

Figura 1

A. Acuerdo perfecto, los datos observados no presentan diferencia entre los métodos y están en la línea de 45°. B. Alejamiento de los datos de la línea de acuerdo y diferencias en las mediciones por ambos métodos.



Estimadores del modelo:

- $\bar{y}_1$ (Promedio de las mediciones observadas por el método 1)
- $\bar{y}_2$ (Promedio de las mediciones observadas por el método 2)
- s^2_1 (Varianza de las mediciones observadas por el método 1)
- s^2_2 (Varianza de las mediciones observadas por el método 2)
- s^2_d (Varianza de las diferencias entre las observaciones del método 1 menos las observaciones del método 2)

Fórmula de definición:

$$\rho_c = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}$$

Fórmula operativa:

$$r_c = \frac{s_1^2 + s_2^2 - s_d^2}{s_1^2 + s_2^2 + (\bar{y}_1 - \bar{y}_2)^2}$$

El resultado del CCC puede variar de 0 a ± 1, similar al coeficiente de correlación de Pearson. Un valor del CCC de +1 corresponde a un acuerdo perfecto, mientras que, por el contrario, un valor CCC de -1 corresponde a un acuerdo negativo perfecto (discordancia), y un valor de 0 corresponde a falta de acuerdo (Camacho-Sandoval, 2008; Liu et al., 2016; Vetter & Schober, 2018); en términos prácticos, la discordancia (valores negativos del CCC) resulta inverosímil en un problema real, puesto que los procedimientos que se evalúan pretenden medir la misma característica (Carrasco & Jover, 2004). Lo anterior significa que cuando todos los datos obtenidos por ambos métodos caen sobre la línea de acuerdo, habrá reproducibilidad perfecta, por lo que el coeficiente determina el grado de dicha reproducibilidad, como lo refiere Lin (Cortés-Reyes et al., 2010).

En cuanto a la interpretación de los valores que puede tener el CCC, no se han propuesto muchas alternativas respecto a qué representan en términos de confiabilidad. El primer intento lo desarrolló Altman, sugiriendo que debería interpretarse como otros coeficientes de correlación como el de Pearson, con criterios de < .20 como pobre y > .80 como excelente (Akoglu, 2018). Sin embargo, fue McBride (2005) quien propuso una clasificación basada en criterios muy similares a los desarrollados por Landis y Koch, aunque de una forma más exigente, ya que el acuerdo será considerado casi perfecto para valores mayores a .99; sustancial, de .95 a .99; moderada, de .90 a .95 y pobre cuando está por debajo de .90.

Al igual que en el caso de otros coeficientes de correlación, se puede (y es deseable) realizar algún tipo de inferencia a partir del valor puntual del CCC, como el cálculo de intervalos de confianza o contrastar hipótesis, hay que tener en cuenta que los procedimientos derivados para este fin dan por supuesto que las dos variables medidas deben tener una distribución normal y ser homocedásticas (Carrasco & Jover, 2004; Lin, 1992).

El CCC presenta la ventaja que su cálculo es sencillo, fácil de usar e interpretar, lo que le confiere las propiedades deseables para la evaluación de la confiabilidad (reproducibilidad), es un estadístico robusto incluso con tamaños de muestra pequeños (Camacho-Sandoval, 2008; Lin, 1989). Además, el CCC no asume una media común para las mediciones de los métodos al principio, ni tampoco el modelo de efectos mixtos en el diseño, ya que depende directamente de los promedios y varianzas de las mediciones de cada evaluador o método y sus correspondientes diferencias (Hernaez, 2015; Liu et al., 2016).

Limitaciones: El CCC considera solamente los efectos fijos de los evaluadores o métodos de medición (Nickerson, 1997). Varios estudios comparativos del CCC con distintos diseños y modelos del CCI, han concluido que el CCC propuesto originalmente es similar al CCI, contrario a lo que creía Lin que era superior (Barnhart et al., 2007; Hernaez, 2015; Kusunoki et al., 2009); incluso, los estudios de simulación y las consideraciones teóricas, muestran que el CCC se comporta de manera bastante similar a un CCI y, por lo tanto, ambos coeficientes pueden ser utilizados indistintamente para efectos de comparación de métodos (Müller & Büttner, 1994). Otra limitación del CCC en su cálculo original, es que solo se puede aplicar para comparar dos evaluadores o métodos a la vez y que no considera la posibilidad de realizar réplicas de las mediciones de las muestras por uno o ambos métodos (Vetter & Schober, 2018). No está de más recordar que no siempre ni para todos los casos hay un consenso acerca de qué valores deberían considerarse como criterios de interpretación para este coeficiente (Cortés-Reyes et al., 2010).

Métodos estadísticos para evaluar la concordancia

Análisis de las diferencias individuales y límites de concordancia (método de Bland y Altman): En 1983, Bland y Altman describieron una técnica para analizar estudios de comparación de métodos, la cual se basó en los trabajos de Oldham (1962), quien casi dos décadas antes, había propuesto una forma de análisis, basándose en las diferencias entre dos mediciones que evita la correlación que se da por estar midiendo dos veces a los mismos individuos (o muestras), proponiendo que la mejor solución era utilizar las funciones de las medidas repetidas dadas por polinomios ortogonales, que en el caso de dos mediciones, estas corresponden a la media de las dos observaciones y su diferencia (Ludbrook, 2010; Oldham, 1962).

Bland y Altman han escrito gran cantidad de artículos sobre su método, presentando ejemplos y modificaciones para resolver algunos inconvenientes que fueron surgiendo a medida que tanto ellos, como otros autores señalaran; al mismo tiempo que ahondaron en argumentación sobre el uso incorrecto de la correlación de Pearson, análisis de regresión y coeficiente de correlación intraclase para la comparación de métodos de medición (Bland & Altman, 1990, 1995b). En su trabajo original publicado como Altman y Bland (1983), propusieron graficar las diferencias entre los valores resultantes de los dos métodos de medición frente a los promedios de esos valores; sin embargo, no fue sino hasta en 1986 que publicaron un artículo específicamente orientado a evaluar la concordancia entre dos métodos de medición clínica, en el que introdujeron el concepto de límites de concordancia como complemento a la gráfica, el cálculo de un intervalo de confianza del 95% de dichos límites para estimar su precisión y la transformación logarítmica cuando no se logra la independencia entre los dos índices que se grafican (Bland & Altman, 1986). El trabajo de Bland y Altman que fuera publicado en la revista *Lancet*, ha sido uno de los más citados en la literatura médica de los últimos tiempos, incluso por su impacto, ha sido reimpreso por otras revistas (Bartlett & Frost, 2008; Bland et al., 2012; Doğan, 2018; Ludbrook, 2010; Mansournia et al., 2021; Stevens et al., 2017).

En publicaciones posteriores, Bland y Altman realizaron una propuesta más extensa y más argumentada para el desarrollo de su método, con una explicación detallada de los aspectos estadísticos y cálculos de los límites de concordancia, su intervalo de confianza, alternativas para evitar el problema de la correlación entre los índices (falta de independencia), así como el análisis para casos cuando se tienen réplicas para uno o ambos métodos y un enfoque no paramétrico (Bland & Altman, 1999, 2003, 2007); en este último caso, se han propuesto metodologías alternativas para el cálculo de los límites de concordancia, basándose en cuantiles (Chen & Kao, 2021; Gerke, 2020b).

Además de la descripción realizada por los propios autores, dada la popularidad que ha alcanzado este método, se han publicado gran cantidad de artículos que se refieren al mismo, resumiendo y explicando la forma de realizarlo (Bruton et al., 2000; Bunce, 2009; Cortés-Reyes et al., 2010; Doğan, 2018; Ludbrook, 2010). Consiste en una gráfica para comparar dos medidas de la misma variable, en la que el eje *x* corresponde a la media de las dos medidas, en tanto que el eje *y* a la diferencia entre ambas; se grafican los datos y, si hubiera concordancia perfecta, los puntos deberían encontrarse en la línea de igualdad o línea de concordancia. En la gráfica, además de los puntos correspondientes a los valores de las diferencias y sus promedios, se trazan tres líneas horizontales principales, correspondientes al promedio de todas las diferencias, los límites inferior y superior de concordancia, y adicionalmente, se deben calcular y presentar los intervalos de confianza del 95% para estos límites de concordancia, que establecen su precisión (Carkeet & Teng Goh, 2018), lo que Ludbrook (2010) denominó límites de tolerancia con un 95% de confianza; adicionalmente, se puede calcular el intervalo de confianza del 95% para el promedio de las diferencias; para todos los intervalos de confianza se utiliza el valor de la distribución de *t* de Student para *n* - 1 grados de libertad que deja el 95% de confianza ($t_{1\text{-}\alpha/2,\ n\text{-}1\ g.l.}$) (Altman & Bland, 1983; Bland & Altman, 1986, 1999).

Las formas de cálculo de todos estos elementos y un esquema de la gráfica (Figura 2), se presentan a continuación:

Parámetros del método:

- μ_d (promedio de las diferencias)
- σ_d^2 (varianza de las diferencias)

Estimadores del método:

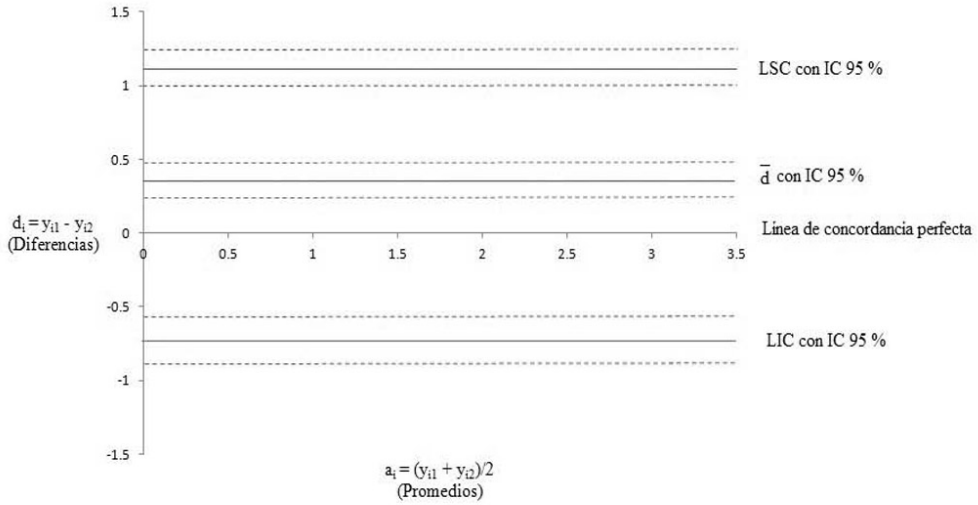
- $d_i = y_{i1} - y_{i2}$ para i = 1, ..., *n* (diferencias observadas para cada par de observaciones)
- $a_i = (y_{i1} + y_{i2})/2$ para i = 1, ..., *n* (promedio de cada par de observaciones)
- $\bar{d}$ (promedio de las diferencias de todas las observaciones)
- s_d (desviación estándar de las diferencias de todas las observaciones)
- $ES_{\bar{d}} = s_d/\sqrt{n}$ (error estándar de las diferencias)
- $\bar{d} \pm t\,(ES_d)$ (IC 95% para $\bar{d}$)
- $\bar{d} \pm 1.96\,(s_d)$ (límites de concordancia)
- $\pm t\,[1.71\,(s_d/\sqrt{n})]$ (amplitud del intervalo de confianza del 95% de los límites de concordancia)

Este procedimiento se basa en el supuesto que ambos métodos que se comparan presentan errores de medición, es decir que ninguno de los dos es el método de referencia, por lo que la media de las dos medidas es la mejor estimación que se tiene para realizar el gráfico (Bland & Altman, 1995a; Carrasco & Jover, 2004; Giavarina, 2015; Mansournia et al., 2021); sin embargo, aunque se tenga un método de referencia, siempre puede caber la duda sobre si las mediciones en realidad se realizan sin ningún error, por lo que los límites de concordancia del 95% son también una medida válida sobre la posible diferencia entre el nuevo método y el de referencia (Bland & Altman, 1995b).

Para realizar el análisis de Bland y Altman se debe de cumplir en primer lugar con la independencia de las diferencias entre los métodos con la magnitud de la medición, es decir que las diferencias no deben estar correlacionadas con los promedios de cada medición, lo que se puede determinar después de realizar el gráfico como se indica más adelante, cuando se evalúa la existencia de un sesgo proporcional (Altman & Bland, 1983; Bland & Altman, 1999, 2003; Mansournia et al., 2021). Luego se debe confirmar el supuesto de distribución normal de las diferencias entre las dos mediciones, para lo cual los datos pueden contrastarse con la distribución normal utilizando métodos clásicos como la prueba de Shapiro-Wilk o la de Kolmogorov-Smirnov, dependiendo el tamaño de la muestra (Doğan, 2018; Giavarina, 2015); aunque en principio Bland y Altman indicaron que podría ser suficiente la evaluación visual del histograma o del diagrama de cuantiles normal de las diferencias (Bland & Altman, 1986), considerando que el supuesto de una distribución normal es mucho menos importante que el supuesto de independencia de la diferencia y la magnitud (Bland & Altman, 1995b).

Figura 2

Esquema del gráfico de Bland y Altman con sus partes



$\bar{d}$ = Promedio de las diferencias, LSC = Límite superior de concordancia, LIC = Límite inferior de concordancia, IC 95% = Intervalo de confianza del 95 %

El gráfico de Bland y Altman debe interpretarse según la distribución de los puntos a lo largo de la línea de concordancia, considerando los datos que superan los límites de concordancia y sus respectivos intervalos de confianza, con base en los siguientes criterios:

- Idealmente, los dos métodos que se comparan deberían producir resultados idénticos, es decir, la diferencia entre los métodos debería ser en promedio 0, lo que equivale a que todas las diferencias estarían en la línea de concordancia. Si no hubiera concordancia perfecta, se esperaría que los puntos se encontraran en torno a la línea de concordancia distribuidos de manera aleatoria, por encima y por debajo de 0, lo que se puede comprobar por medio de una prueba de significación como la *t* pareada para evaluar el sesgo (Altman & Bland, 1983; Bunce, 2009; Jensen & Kjelgaard-Hansen, 2006; Ludbrook, 2010), ya que el método de Bland y Altman en sí no contempla la significancia estadística para determinar si un sesgo aparente es mayor que lo que se pueda atribuir al azar (Smith et al., 2010). Como alternativa a la prueba de significación, se puede determinar que el intervalo de confianza del 95% de las diferencias incluya el 0, lo que significa la inexistencia de un sesgo sistemático (Ludbrook, 2010).
- Si los puntos tienden a encontrarse por encima o por debajo de la línea de concordancia, reflejan sobreestimación o subestimación de alguno de los sistemas de medición. Esto constituye un sesgo sistemático fijo o constante, lo que indica que un conjunto de mediciones da valores que son consistentemente más altos o más bajos que el otro, en todo el rango de medición (Bland & Altman, 1986; Doğan, 2018; Kalra, 2017; Ludbrook, 2010).
- Si la diferencia de valores entre los dos métodos aumenta o disminuye en proporción al promedio de cada par de mediciones, se está frente a lo que se llama un sesgo proporcional, que hace que las medidas aumenten o disminuyan con relación a su promedio, que se interpreta como la existencia de correlación entre las diferencias y sus promedios, lo que repercute en que la desviación estándar de las diferencias sea grande y afecte la amplitud de los límites de concordancia (Doğan, 2018; Ludbrook, 2010; Mansournia et al., 2021; Zaki, Bulgiba, & Ismail, 2013). Ungerer y Pretorius (2018) indican que en presencia de un error significativo constante o proporcional, el procedimiento de comparación no es útil. Lo anterior se puede demostrar en el gráfico por una distribución de los datos en forma de embudo (Watson & Petrie, 2010) o por la existencia de una pendiente significativa al realizar un análisis de regresión entre las diferencias y sus promedios (Bartlett & Frost, 2008; Bland & Altman, 1986, 1995a, Gerke, 2020a; Ludbrook, 2010; Taffé, 2021), pudiéndose solventar este problema por medio de una transformación logarítmica de los datos y aplicar el método sobre las diferencias y promedios de los datos transformados (Altman & Bland, 1983; Bland & Altman, 1999; Bartlett & Frost, 2008), o alternativamente, utilizar en el eje y del gráfico las diferencias como porcentaje de los respectivos promedios (Bartlett & Frost, 2008; Bland & Altman, 1999, 2003; Bunce, 2009). En caso de que se compruebe un sesgo proporcional, no se puede aplicar una prueba de significación como la t pareada (Jensen & Kjelgaard-Hansen, 2006; Westgard & Hunt, 1973).
- Los límites de concordancia representan el rango dentro del cual se ubicarán aproximadamente el 95% de las diferencias entre las mediciones de los dos métodos, basándose originalmente en la regla empírica de la distribución normal (±2sd) y no constituyen estrictamente un intervalo de confianza (Bland & Altman, 1986, 1990), por lo que es mejor calcularlos a partir de lo que se denomina el coeficiente de repetibilidad (±1.96sd) (Bland & Altman, 1986, 1999, 2003), usando la distribución normal estándar para definir los límites que incluirán la diferencia entre mediciones individuales sobre el mismo sujeto por los dos métodos, con una probabilidad del 95% (Bland & Altman, 1995b). Son utilizados para decidir si la concordancia entre pares de lecturas es aceptable o no y por ende, si los métodos pueden usarse indistintamente o son intercambiables (Bland & Altman 1986, 1990); la magnitud de la diferencia que puede considerarse aceptable no es una decisión estadística (Bland & Altman, 1990, 1999), por lo que los valores de aceptación deben definirse previamente con base en consideraciones clínicas, biológi-

cas, analíticas o técnicas, de acuerdo a lo que se esté midiendo (Cortés-Reyes et al., 2010; Giavarina, 2015; Kalra, 2017; Mansournia et al., 2021). En el área médica, por ejemplo, se deben comparar con lo que se denomina la diferencia clínicamente aceptable de la medición que se trate y corresponde al investigador valorar si estas diferencias son lo suficientemente pequeñas como para considerar que los métodos son intercambiables (Giavarina, 2015; Kalra, 2017; Stevens et al., 2017; Vetter & Schober, 2018). Complementariamente, el gráfico de Bland y Altman puede contener una banda de referencia basada en el CCC en los casos en los que no se dispone de una diferencia clínicamente aceptable, que brinda pautas para una evaluación gráfica descriptiva del acuerdo, así como información útil para el reconocimiento de patrones o la identificación de valores atípicos en los datos (Kim & Lee, 2022); también se ha propuesto la elaboración de bandas de confianza puntuales y simultáneas alrededor de los límites de concordancia con un enfoque bayesiano para que el investigador pueda decidir si el desacuerdo (no el acuerdo) no es demasiado alto para que los dos métodos se consideren intercambiables (Taffé, 2023). Los intervalos de confianza del 95% de los límites de concordancia se sitúan por arriba y por abajo de los valores o líneas superior e inferior, representando la precisión o incertidumbre de estos (Bartlett & Frost, 2008; Bland & Altman, 1986). Indican el rango dentro del cual debe estar una nueva observación si se extrae de la misma población que la muestra que se ha estudiado, es decir que, como todo intervalo de confianza, tiene propósitos de estimación o predicción (Ludbrook, 2010), permitiendo estimar el tamaño del posible error de muestreo (Giavarina, 2015; Jan & Shieh, 2018). Cuanto mayor sea el número de muestras utilizadas para la evaluación de la diferencia entre los métodos, más estrechos serán los intervalos de confianza, tanto para la diferencia de medias como para los límites de concordancia (Bartlett & Frost, 2008; Bland & Altman, 2003).

- Gerke y Möller (2021) han reducido el problema de la interpretación de los límites de concordancia y sus intervalos de confianza, proponiendo el uso de un enfoque bayesiano, para la construcción de lo que denominan intervalos de credibilidad, que ofrecen una interpretación probabilística directa en términos de la credibilidad de los posibles valores de los parámetros que no tienen los intervalos de confianza, además de ser útiles cuando el supuesto de normalidad de las diferencias no se cumple.
- El gráfico de Bland y Altman también se puede utilizar para detectar valores atípicos, que son lecturas extremas ocasionales que se apartan del cuerpo principal de los datos, posiblemente causadas por errores de medición (Watson & Petrie, 2010).

Limitaciones: Es importante hacer ver que el método de Bland y Altman, como cualquier otro, también presenta problemas de aplicación o uso inapropiado y de interpretación; uno de ellos es que la gráfica no puede desentrañar los sesgos de confusión, no indica qué sistema es más preciso y sin la información adicional obtenida por la aplicación de réplicas, la gráfica puede ser engañosa (Stevens et al., 2017); aunque Bland y Altman (2007) desarrollaron una modificación de su método para cuando se tienen réplicas considerando un modelo fijo, el análisis se puede realizar aplicando un modelo de efectos aleatorios para un diseño de medidas repetidas (Doğan, 2018: Ludbrook, 2010; Myles & Cui, 2007; Stevens et al., 2017).

Otra circunstancia que puede influir en el uso inadecuado de este método, es que no tiene un enfoque basado en fórmulas que clasifique automáticamente el acuerdo en bueno o malo, o que proporcione una guía sobre qué método utilizar cuando el desacuerdo es considerable, de allí que es recomendable que el método de Bland y Altman deba usarse con precaución y complementarse con otras técnicas estadísticas, lo que da como ventaja que se logran compensar las limitaciones de las técnicas individuales (Bunce, 2009; Ludbrook, 2010; Zaki, Bulgiba, & Ismail, 2013).

Se ha determinado que el método de Bland y Altman puede fallar en la estimación de los sesgos en los métodos de medición, indicando su presencia cuando no los hay o, por el contrario, indicando que no hay sesgo cuando sí lo hay (Taffé et al., 2020), lo que se manifiesta en la línea de regresión obtenida de la gráfica de las diferencias y sus promedios mostrando erróneamente una tendencia ascendente o descendente cuando no hay sesgo o una pendiente cero cuando sí lo hay (Taffé, 2023). Para solventar este problema se

recomienda que se aplique una metodología complementaria al método de Bland y Altman, que incluye un gráfico para estimar los sesgos y otro gráfico para evaluar la precisión de los métodos que se están comparando (Taffé, 2018, 2019; Taffé et al., 2020).

Prueba de sesgo (*t* de Student pareada): La distribución de probabilidad denominada *t* de Student fue desarrollada por William S. Gosset, quien trabajaba para una empresa que no aprobaba que sus empleados publicaran sus investigaciones, por lo que, para hacerlo, utilizó el nombre de Student como pseudónimo (Wilkerson, 2008). En su trabajo original, Gosset no solo derivó y tabuló la distribución *t*, sino que presentó los primeros modelos prácticos de su aplicación que corresponden a lo que ahora se conoce como muestras pareadas, emparejadas, dependientes o relacionadas. Específicamente, analizó el efecto de dos isómeros ópticos de hiosciamina en su capacidad para inducir el sueño y en un segundo ejemplo, los rendimientos al sembrar 11 parcelas diferentes de tierra con semillas provenientes de la misma cosecha, pero sometidas a dos tratamientos distintos (Senn & Richardson, 1994; Student, 1908).

Para la comparación de métodos, se emplean observaciones emparejadas o pareadas, ya que muchas veces es difícil o imposible comparar los resultados con un estándar conocido (Bland & Altman, 1995b; Zimmerman, 1997); por lo que, en un estudio de concordancia, no interesa la media de las lecturas de cada instrumento o método, sino que los valores individuales y cada lectura del método estándar debe ser repetida por el nuevo método (Zaki et al., 2012). Esto conduce a la prueba *t* pareada que se utiliza para comparar diferencias de medias cuando las observaciones se han obtenido en pares y, por lo tanto, son dependientes, por lo que se deben determinar las diferencias individuales y realizar el análisis sobre dichas diferencias. Los datos en una prueba *t* pareada se dice que son dependientes (o relacionados), porque cada valor en la primera medición está emparejado con un valor en la segunda medición, básicamente en el caso particular que se está analizando, las mediciones corresponden a cada uno de los métodos y las muestras provienen de un mismo individuo (Hsu & Lachenbruch, 2005; Wilkerson, 2008).

El fundamento de esta prueba en la comparación de métodos es que, si estos producen resultados idénticos, la diferencia entre ambos debería ser en promedio 0 (Jensen & Kjelgaard-Hansen, 2006). Para ello, se tienen *n* pares de observaciones y cada par es independiente de los otros pares, el estadístico *t* de Student se puede calcular de la siguiente forma (Hsu & Lachenbruch, 2005):

Parámetros del modelo:

- μ_{y1}, μ_{y2} (promedios de las observaciones de cada población)
- $\mu_d = \mu_{y1} - \mu_{y2}$ (promedio de las diferencias de las poblaciones)
- σ_d^2 (varianza de las diferencias)

Estimadores del modelo:

- y_1, y_2 (observaciones emparejadas)
- $d_i = y_{1i} - y_{2i}$, para i = 1, ..., *n* (diferencias de cada par de observaciones)
- $\bar{d}$ (promedio de las diferencias)
- s_d (desviación estándar de las diferencias)

Hipótesis nula (Ho): $\mu_d = 0$

Cálculo del estadístico:

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

Luego, este estadístico se compara con la distribución *t* de Student con *n* - 1 grados de libertad y se determina el valor *p*, necesario para probar si la diferencia de medias difiere significativamente de 0 (rechazo de la Ho) (Hsu & Lachenbruch, 2005). La distribución *t* de Student es similar a la distribución normal ya que tiene forma de campana y es simétrica, pero su forma depende de los grados de libertad, que es el tamaño de la muestra menos uno (*n* - 1) (Wilkerson, 2008).

Para esta prueba se deben tener en cuenta varios elementos que son clave en su aplicación, las variables deben ser aleatorias, continuas y con distribución normal (Hsu & Lachenbruch, 2005; Martelli Filho et al., 2005). Dado que el diseño es pareado, se verifica la normalidad y aleatoriedad de los datos y no es necesario probar la igualdad de varianzas (Martínez Curbelo et al., 2016) ya que el interés en la compara-

ción de métodos es la diferencia entre las observaciones y lo más importante es que la diferencia entre las mediciones (d_i), tenga una distribución normal (Hsu & Lachenbruch, 2005). Las propiedades de robustez de la prueba *t* pareada corresponden a las de la prueba *t* de una muestra, en la que la falta de homocedasticidad entre las dos variables no afecta formalmente el análisis, pero los valores atípicos afectarán negativamente, ya que pueden considerarse como un alejamiento de la normalidad, incidiendo en la simetría de las distribuciones y la varianza de las observaciones (Hsu & Lachenbruch, 2005; Martelli Filho et al., 2005).

En la comparación de métodos, la variabilidad de las diferencias entre estos indica qué tan bien concuerdan en los resultados. La media de las diferencias pareadas revela si, en promedio, un método tendió a subestimar o sobrestimar las mediciones en relación con las mediciones del segundo método, lo que se denomina sesgo entre los métodos. Con la prueba *t* pareada se prueba la hipótesis que la media de las diferencias es 0, lo que corresponde a la ausencia de sesgo entre los métodos (Bartlett & Frost, 2008). Se debe tener claro que rechazar la hipótesis nula de no diferencia, lo que permite concluir es que los métodos no concuerdan ya que hay un sesgo significativo (Martínez Curbelo et al., 2016; Morgan & Aban, 2016; Westgard & Hunt, 1973), pero no rechazar esta hipótesis no constituye una prueba que los métodos concuerden (Zaki et al., 2012), por lo que la conclusión correcta es que no se tuvo evidencia suficiente para establecer que el sesgo sea significativo.

Lo mejor es interpretar la prueba *t* pareada desde el punto de vista de lo que representan las diferencias en la detección de errores sistemáticos (Bruton et al., 2000; Houston, 1983; Watson & Petrie, 2010), por ello se refieren a ella como una prueba de sesgo entre las mediciones (Bartlett & Frost, 2008). Una alternativa para visualizar mejor la estimación de la diferencia entre los métodos o sesgo, es reportar la misma con un intervalo de confianza, usualmente del 95% (Cardemil, 2017), de esta manera, el análisis del sesgo por medio de la prueba *t* pareada se complementa con otras técnicas (Bruton et al., 2000; Manterola et al., 2018).

Limitaciones: No hay que olvidar que, si existen errores aleatorios, estos pueden exagerar o encubrir las diferencias entre las mediciones y llevar a conclusiones erróneas (Houston, 1983; Westgard & Hunt, 1973). Por otra parte, si la media de las diferencias es 0, se puede concluir que no existe un error sistemático entre los pares de resultados; en la prueba *t* pareada, un resultado significativo sugiere que hay un error sistemático, pero un resultado no significativo indica que no hay evidencia de un error sistemático (Morgan & Aban, 2016; Watson & Petrie, 2010). En este último caso, la prueba no es concluyente, porque esta evalúa si existe un error sistemático constante (Westgard & Hunt, 1973), pero puede existir un error sistemático proporcional que cambia dentro del rango analítico (Jensen & Kjelgaard-Hansen, 2006), situación que invalida la prueba ya que incide en las dispersiones de los dos grupos que se comparan, lo que puede pasar desapercibido ya que generalmente no se comprueba la homocedasticidad (Bland & Altman, 1995b; Pandis, 2021). También debe considerarse que, en el caso de muestras pareadas, relacionadas o no independientes, disminuye la probabilidad de error tipo I y el poder de la prueba para detectar diferencias, por lo que el no rechazo de la hipótesis nula debe considerarse con prudencia (Linnet, 1999; Zimmerman, 1997).

Análisis de regresión, una técnica complementaria

Según Pearson (1930), fue Francis Galton en 1877 quien desarrolló el concepto de regresión aplicado a datos relacionados con la herencia en el hombre. Años más tarde, Galton (1886a) retomó el tema y lo presentó formalmente en un discurso ante la Sección de Antropología de la Asociación Británica para el Desarrollo de la Ciencia y publicó su primer artículo sobre este tema, planteando que la regresión puede ser establecida por una ecuación, que corresponde a la ecuación de una línea recta (Galton, 1886b).

Es un hecho que el método de regresión se originó a partir del método de mínimos cuadrados (Dhakal, 2018), el cual ha sido comprobado que es efectivo para el cálculo de las ecuaciones de regresión (Han et al., 2015). La primera referencia publicada sobre el método de mínimos cuadrados corresponde al 1805, por el matemático francés Adrien-Marie Legendre, quien estableció la regla que la suma de los cuadrados de los errores debe hacerse al mínimo para obtener los valores ajustados de las cantidades observadas (Merriman, 1877).

El modelo de regresión lineal clásico, denominado también como regresión lineal ordinaria (OLR, por sus siglas en inglés), se calcula minimizando la suma de los residuos al cuadrado en la dirección de la variable *y*, dichos residuos son las distancias o diferencias entre los valores observados con el valor sobre la recta

de forma perpendicular, lo que se denomina método de los mínimos cuadrados, mediante el ajuste a un modelo de regresión simple:

$$y = \alpha + \beta x + \varepsilon$$

Donde α es el parámetro que corresponde a la intersección, β, a la pendiente y ε, el error aleatorio (Carrasco & Jover, 2004; Stöckl et al., 1998).

Varios autores se han referido al análisis de regresión lineal como una técnica comúnmente utilizada para analizar datos en la comparación de métodos, ya que si se puede definir una relación lineal entre el método a prueba y el método de referencia, entonces la pendiente y la intersección de esta línea pueden proporcionar estimaciones de los errores sistemático proporcional y constante entre los dos métodos (Cornbleet & Gochman, 1979; Linnet, 1993; Stevens et al., 2017; Westgard & Hunt, 1973), esperando que la regresión libre de errores entre todos los pares del conjunto de datos tuviera una pendiente de 1 y una intersección de 0 (Payne, 1997), lo que constituye un elemento básico del análisis llamado prueba de identidad que es útil para determinar que la pendiente no difiere significativamente de 1 y que el intercepto no difiere significativamente de 0 (Liao et al., 2006; National Committee for Clinical Laboratory Standards [NCCLS], 2002).

El elemento del que depende que exista un acuerdo perfecto es la varianza residual (Bland & Altman, 1995b; Cornbleet & Gochman, 1979), ya que se puede rechazar un acuerdo razonablemente bueno cuando los errores residuales son pequeños, pero aceptar un acuerdo deficiente cuando los errores residuales son grandes, considerando el hecho que ambos métodos realmente se miden con error (Liao et al., 2006); la OLR es aplicable entonces cuando la comparación se hace frente a un método de referencia, que presupone mediciones libres de error, lo que supone que la desviación estándar analítica del método de referencia es 0 y que la desviación estándar analítica para el otro método es constante en todo el rango de medición (Cornbleet & Gochman, 1979; Linnet, 1993), aunque en términos estrictos no debe descartarse la probabilidad que la línea estimada pueda ser incorrecta porque nada asegura que el método de referencia presente, si bien mínimo, un error aleatorio (Carrasco & Jover, 2004).

Aun suponiendo que el método de referencia pueda brindar medidas libres de error, estrictamente no se cumple con el principio de regresión que implique una relación causa-efecto entre una variable explicativa independiente (*x*) y otra dependiente (*y*), teniéndose que ambas variables son aleatorias e independientes (Carrasco & Jover, 2004; Cornbleet & Gochman, 1979). Sin embargo, algunos autores han insistido en que la regresión lineal por el método de mínimos cuadrados, cuando se aplica a los datos de comparación de métodos, proporciona información útil sobre los errores proporcional, constante y aleatorio, mediante el análisis de la pendiente, la intersección y la desviación estándar de los residuos, respectivamente, concluyendo que los datos de regresión lineal se pueden utilizar para evaluar un nuevo método frente al método de referencia (Cornbleet & Gochman, 1979; Jensen & Kjelgaard-Hansen, 2006; Stöckl, et al., 1998). Considerando que en muchos casos la comparación no se hace frente a un método de referencia y que los supuestos necesarios para la OLR rara vez se cumplen, es razonable aplicar otras alternativas (Cornbleet & Gochman, 1979; Linnet, 1993), siendo las más recomendadas las siguientes:

- El método de regresión ortogonal ponderada de Deming que minimiza la suma de los cuadrados de las desviaciones de la línea en ambas direcciones x y y, ponderada por la relación de las varianzas analíticas de los dos métodos, que se supone constante en el rango de observaciones. Esta derivación da como resultado la mejor línea para minimizar la suma de los cuadrados de las distancias perpendiculares desde los puntos de datos a la línea (Bland & Altman, 1995b; Cornbleet & Gochman, 1979; Linnet, 1993; Jensen & Kjelgaard-Hansen, 2006; Payne, 1997; Stöckl, et al., 1998).
- El método no paramétrico de Passing y Bablok (1983), que esencialmente, usa todas las líneas rectas entre dos puntos para calcular una línea de regresión y la pendiente de la recta se calcula como la mediana de todas las pendientes posibles; esta estimación no paramétrica de los coeficientes de regresión α y β puede ser más robusta que la OLR (Baumdicker & Hölker, 2020). Este principio de estimación hace que el método sea robusto frente a valores atípicos, que es su principal ventaja (Linnet, 1993; Jensen & Kjelgaard-Hansen, 2006; Payne, 1997; Stöckl, et al., 1998). Como criterio adicional, debe evaluarse la posible desviación de la linealidad por medio de una prueba llamada suma acumulativa (CUSUM) (Bilić-Zulle, 2011;

Passing & Bablok, 1983). Este método es muy utilizado para la comparación de métodos en bioquímica clínica, farmacología y medicina de laboratorio, porque se describe en las directrices EP9-A2 para comparación de métodos del Instituto de Estándares Clínicos y de Laboratorio (Baumdicker & Hölker, 2020, NCCLS, 2002).

- El análisis de regresión estructural bivariante desarrollado por Feldmann y Shneider, por medio de estimadores de máxima verosimilitud, así como estimadores robustos para la pendiente y la intersección, lo que permite la estimación insesgada de los errores estándar de la pendiente y la intersección de la línea de calibración bivariada (Müller & Büttner, 1994).

Como elemento adicional, Linnet (1999) propuso la regresión no solo como complemento al análisis, sino como alternativa para el cálculo de muestra en todo estudio de comparación de métodos cuantitativos, indicando que los tamaños de 40 a 100 muestras que se utilizan convencionalmente en los estudios de comparación de métodos deben reconsiderarse, ya que lo que más afecta es la relación del rango analítico (valor máximo dividido por el valor mínimo), lo que para regresión implica diferencias sustanciales dependiendo de la variable que se mida o mensurando; por lo que el procedimiento de muestreo debe resultar en la inclusión de sujetos de estudio que contribuyan con mediciones en todo el rango de medición de interés y relevancia clínica o analítica (Gerke et al., 2022). Esta propuesta viene a resolver el problema sobre la forma de cálculo de muestra, ya que se evitan las contradicciones sobre si el cálculo debe hacerse para los índices de confiabilidad (CCI o CCC) (Barnhart et al., 2007; Temel & Erdogan, 2017), para el método de Bland y Altman o *t* pareada (Bartlett & Frost, 2008), considerando además que las motivaciones formales para calcular el tamaño de la muestra para estudios de concordancia han sido escasas (Gerke et al., 2022).

Limitaciones: El principal problema de la OLR es que solamente tiene utilidad para la evaluación de los sesgos o errores con base en los parámetros del modelo (Bland & Altman, 2003; Cornbleet & Gochman 1979; Jensen & Kjelgaard-Hansen, 2006) y la ecuación en sí no tiene ninguna aplicación predictiva ya que ese no es el objetivo de los estudios de comparación de métodos (Bland & Altman 1995b). No cumple con el principio de tener una variable regresora independiente y una variable dependiente (Dhakal 2018), sino que se trata de dos métodos o técnicas de medición que son variables aleatorias independientes (Altman & Bland, 1983; Stevens et al., 2017). Además, cuando se tiene un método de referencia sobre el que se va a realizar la comparación, se debe suponer que mide sin error (Müller & Büttner, 1994; Stöckl et al., 1998), lo cual no es razonable (Carrasco & Jover, 2004).

También debe considerarse que algunos métodos de medición exhiben un cambio estructural en la relación sobre el rango de medición, es decir que muestran rangos analíticos con comportamiento lineal distinto, lo cual impide la aplicación de cualquier método de regresión lineal (OLR, Deming o Passing Bablok), para ello se han desarrollado los llamados modelos segmentados, multifásicos o por partes (Kotinkaduwa & Choudhary, 2020).

El caso de la correlación de Pearson

Francis Galton (1889) fue el primero en utilizar el término co-relación o correlación aplicado a estudios antropométricos, mencionando el índice de correlación (*r*) que mide la cercanía de la relación. Sin embargo, fue Karl Pearson (1896) quien, planteó la forma de cómo determinar de la mejor manera posible la correlación, desarrollando una fórmula más directa, operativa y sencilla para el índice *r*, que él llamó coeficiente de correlación, tal y como actualmente se conoce.

La correlación es una medida inferencial que refleja la intensidad de la asociación lineal entre dos variables cuantitativas aleatorias continuas (Bruton et al., 2000; Cortés-Reyes et al., 2010; Manterola et al., 2018; Vetter & Schober, 2018), en la que el cambio en la magnitud de una variable está asociado con un cambio en la magnitud de otra variable (Vetter & Schober, 2018); los datos se inspeccionan en busca de linealidad entre dos variables continuas y se puede expresar por medio del coeficiente de correlación de Pearson (*r*) (Vetter & Schober, 2018).

La correlación se ha empleado principalmente en la literatura médica como técnica estadística para comparación de métodos, un enfoque inadecuado que aún se sigue aplicando (Altman, 2009; Bland & Altman, 1995b; Müller & Büttner, 1994), a pesar que desde 1973, Westgard y Hunt lo señalaron por primera vez, seguidos por gran cantidad de autores que han insistido que la correlación de Pearson solo mide asociación y no acuerdo o concordancia (Altman & Bland, 1983; Bartko, 1994; Jensen & Kjelgaard-Hansen, 2006; Morgan & Aban, 2016; Rojas et al., 2016; Watson &

Petrie, 2010), por lo que no puede ser interpretado como índice de confiabilidad (Watson & Petrie, 2010; Westgard & Hunt, 1973) o que justifique la intercambiabilidad de los métodos (Altman & Bland, 1983; Bland & Altman, 2003); además que su cálculo y reporte pueden resultar engañosos, porque si bien mediciones que son concordantes pueden tener un alto coeficiente de correlación, no necesariamente mediciones que se correlacionan bien, deben ser concordantes (Bland et al., 2012; Jensen & Kjelgaard-Hansen, 2006; Morgan & Aban, 2016; Müller & Büttner, 1994; Rojas et al., 2016; Schober et al., 2018).

Las bases o fundamentos que descalifican el uso de la correlación de Pearson como técnica estadística para la comparación de métodos se presentan a continuación:

- El coeficiente de correlación depende tanto de la variación entre individuos (entre los valores reales) como de la variación dentro de los individuos (error de medición), por lo que puede ser elevado si aumenta la variabilidad entre sujetos, ignorando la diferencia entre las dos mediciones (Altman & Bland, 1983).
- El coeficiente de correlación mide una relación lineal, pero no detecta ninguna desviación de la línea de 45°, solamente es sensible al error aleatorio, en tanto que ignora o no detecta cualquier sesgo sistemático entre las dos variables (Bland & Altman, 2003; Bruton et al., 2000; Lin, 1989; Westgard & Hunt, 1973).
- La correlación se utiliza para asociar dos variables que no miden la misma entidad, atributo o resultado, mientras que las estadísticas de concordancia tienen como objetivo describir el grado en que las variables miden la misma entidad, atributo o resultado y en la misma escala; por lo que un cambio en la magnitud o en la escala de medición no afecta la correlación, pero sí afecta el acuerdo (Bland & Altman 1995b; Cortés-Reyes et al., 2010; Vetter & Schober, 2018).
- La correlación solo debe aplicarse cuando los pares de datos se observan de forma independiente entre sí; no deben aplicarse a datos de medidas repetidas donde las variables no son independientes (Lin, 1989; Vetter & Schober, 2018).

Conclusiones

Cuando se comparan métodos cuantitativos de medición, debe ponerse especial atención al diseño de la investigación, estableciendo el número apropiado de muestra, siendo la técnica de regresión la que se debe tomar como criterio, utilizando el método propuesto por Linnet (1999), en el que se deben establecer el error o diferencia máxima aceptable, rango analítico en términos de "relación de rango" (límite superior/límite inferior), la variación esperada del mensurando (en desviación estándar o coeficiente de variación) y si se comparará con un método de referencia o no. El diseño de la investigación es pareado, es decir que cada muestra deberá ser medida por ambos métodos. Posteriormente se deberán comprobar los supuestos de normalidad de los datos de cada método y de las diferencias, lo cual es necesario para poder aplicar los procedimientos estadísticos más apropiados.

Lo más recomendable para el análisis es la aplicación conjunta del método de Bland y Altman para evaluar concordancia, un índice de confiabilidad (CCI o CCC) y una técnica de regresión, cuyos parámetros (pendiente e intercepto) brindarán información sobre los sesgos sistemáticos constante y proporcional que puedan estar presentes.

Al aplicar la técnica de Bland y Altman, se debe realizar el gráfico de las diferencias y sus promedios, con la línea de concordancia y límites de concordancia con sus respectivos intervalos de confianza del 95%, se debe definir previamente los límites que se considerarán aceptables con base en un criterio clínico o técnico. Realizar un análisis de tendencia de las diferencias con sus respectivos promedios, por medio de regresión lineal simple y coeficiente de correlación de Pearson, no debiendo haber linealidad o relación entre estas medidas. Indicar el valor del sesgo por medio del promedio de las diferencias y su intervalo de confianza del 95%, el cual deberá evaluarse en función de su dispersión, para establecer que no se aleje del error máximo permitido (definido para el cálculo de muestra), como la diferencia máxima esperada entre ambos métodos. También es necesario que el sesgo se evalúe considerando que su intervalo de confianza debe incluir el valor 0, o alternativamente, realizar una prueba de significancia por medio de la técnica de *t* de Student pareada.

Para evaluar la confiabilidad puede utilizarse el CCI o el CCC, reportando su respectivo intervalo de confianza del 95% y significancia estadística, interpretándose de acuerdo con la clasificación de Koo y Li (2016) o la de McBride (2005), respectivamente.

Si la comparación de métodos se hace frente al método de referencia o algún método validado en el que la medición se efectúa sin error, se aplica la técnica de análisis de regresión lineal ordinaria por el método de los mínimos cuadrados, siempre y cuando los datos tengan una distribución normal. Si no se cuenta con un método de referencia o validado, aplicar la regresión ortogonal de Deming. En todo caso, se debe reportar la ecuación, su significancia, intervalos de confianza del 95% para intercepto y pendiente, así como la prueba de identidad, para determinar que la pendiente no difiere de 1 ($\beta = 1$) y el intercepto no difiere de 0 ($\alpha = 0$). La regresión Passing-Bablok es una alternativa no paramétrica que constituye una buena opción para cualquier caso, ya que puede usarse sin restricciones referentes a tener un método de referencia o el cumplimiento de la normalidad de los datos.

Se debe hacer una interpretación integral de estas técnicas estadísticas para poder decidir sobre la equivalencia o intercambiabilidad de los métodos cuantitativos de medición y en ninguna circunstancia es justificable aplicar la correlación para comparación de métodos.

Agradecimientos

Al Licenciado Armando Cáceres por la revisión y valiosas sugerencias hechas al manuscrito.

Contribución de los autores

Coordinación, elaboración y revisión del Documento: FN
Revisión y búsqueda de literatura: JN, FN
Lectura y resumen de documentos: JN, FN
Participación en la estructura y escritura del documento: JN, FN

Materiales suplementarios

Este artículo no tiene archivos complementarios.

Referencias

Abu-Arafeh, A., Jordan, H., & Drummond, G. (2016). Reporting of methods comparison studies: A review of advice, an assessment of current practice, and specific suggestions for future reports. *British Journal of Anaesthesia, 117*(5), 569-575. https://doi.org/10.1093/bja/aew320

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine, 18*(3), 91-93. https://doi.org/10.1016/j.tjem.2018.08.001

Altman, D. G. (2009). Assessing new methods of clinical measurement. *British Journal of General Practice, 59*(563), 399-400. https://doi.org/10.3399/bjgp09X420905

Altman, D. G., & Bland, J. M. (1983). Measurement in medicine: The analysis of method comparison studies. *The Statistician, 32*(3), 307-317. https://doi.org/10.2307/2987937

Barnhart, H. X., Haber, M. J., & Lin, L. I. (2007). An overview on assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics, 17*(4), 529-569. https://doi.org/10.1080/10543400701376480

Bartko, J. J. (1994). General methodology II. Measures of agreement: A single procedure. *Statistics in Medicine, 13*(5-7), 737-745. https://doi.org/10.1002/sim.4780130534

Bartlett, J. W., & Frost, C. (2008). Reliability, repeatability and reproducibility: Analysis of measurement errors in continuous variables. *Ultrasound in Obstetrics & Gynecology, 31*(4), 466-475. https://doi.org/10.1002/uog.5256

Baumdicker, F., & Hölker, U. (2020). Method comparison with repeated measurements – Passing-Bablok regression for grouped data with errors in both variables. *Statistics and Probability Letters, 164,* Artículo 108801. https://doi.org/10.1016/j.spl.2020.108801

Berthelsen, P. G., & Nilsson, L. B. (2006). Researcher bias and generalization of results in bias and limits of agreement analyses: A commentary based on the review of 50 Acta Anesthesiologica Scandinavica papers using the Altman-Bland

approach. *Acta Anesthesiologica Scandinavica, 50*(9), 1111-1113. https://doi.org/10.1111/j.1399-6576.2006.01109.x

Bilić-Zulle, L. (2011). Comparison of methods: Passing and Bablok regression. *Biochemia Medica, 21*(1), 49-52. https://doi.org/10.11613/bm.2011.010

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, 1*(8476), 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8

Bland, J. M., & Altman, D. G. (1990). A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Computers in Biology and Medicine, 20*(5), 337-340. https://doi.org/10.1016/0010-4825(90)90013-f

Bland, J. M., & Altman, D. G. (1995a). Comparing methods of measurement: Why plotting difference against standard method is misleading. *Lancet, 346*(8982), 1085-1087. https://doi.org/10.1016/s0140-6736(95)91748-9

Bland, J. M., & Altman, D. G. (1995b). Comparing two methods of clinical measurement: A personal history. *International Journal of Epidemiology, 24*(Suppl. 1), S7-S14. https://doi.org/10.1093/ije/24.supplement_1.s7

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research, 8*(2), 135-160. https://doi.org/10.1177/096228029900800204

Bland, J. M., & Altman, D. G. (2003). Applying the right statistics: Analyses of measurement studies. *Ultrasound in Obstetrics & Gynecology, 22*(1), 85-93. https://doi.org/10.1002/uog.122

Bland, J. M., & Altman, D. G. (2007). Agreement between methods of measurement with multiple observations per individual. *Journal of Biopharmaceutical Statistics, 17*(4), 571-582. https://doi.org/10.1080/10543400701329422

Bland, J. M., Altman, D. G., & Warner, D. S. (2012). Agreed statistics: Measurement method comparison. *Anesthesiology, 116*(1), 182-185. https://doi.org/10.1097/ALN.0b013e31823d7784

Bruton, A., Conway, J. H., & Holgate, S. T. (2000). Reliability: What is it, and how is it measured? *Physiotherapy, 86*(2), 94-99. https://doi.org/10.1016/S0031-9406(05)61211-4

Bunce, C. (2009). Correlation, agreement, and Bland-Altman analysis: Statistical analysis of method comparison studies. *American Journal of Ophthalmology, 148*(1), 4-6. https://doi.org/10.1016/j.ajo.2008.09.032

Camacho-Sandoval, J. (2008). Coeficiente de concordancia para variables continuas. *Acta Médica Costarricense, 50*(4), 211-212.

Cardemil, F. (2017). Análisis de comparación y aplicaciones del método de Bland-Altman: ¿Concordancia o correlación? *Medwave, 16*(1), Artículo e6852. https://doi.org/10.5867/medwave.2017.01.6852

Carkeet, A., & Teng Goh, Y. (2018). Confidence and coverage for Bland-Altman limits of agreement and their approximate confidence intervals. *Statistical Methods in Medical Research, 27*(5), 1559-1574. https://doi.org/10.1177/0962280216665419

Carrasco, J. L., & Jover, L. (2004). Métodos estadísticos para evaluar la concordancia. *Medicina Clínica (Barc), 122*(Supl 1), 28-34. https://www.elsevier.es/es-revista-medicina-clinica-2-articulo-metodos-estadisticos-evaluar-concordancia-13057543

Chen, L. A., & Kao, C. L. (2021). Parametric and nonparametric improvements in Bland and Altman's assessment of agreement method. *Statistics in Medicine, 40*(9), 2155-2176. https://doi.org/10.1002/sim.8895

Chhapola, V., Kanwal, S. K., & Brar, R. (2015). Reporting standards for Bland-Altman agreement analysis in laboratory research: A cross-sectional survey of current practice. *Annals of Clinical Biochemistry, 52*(3), 382-386. https://doi.org/10.1177/0004563214553438

Cornbleet, P. J., & Gochman, N. (1979). Incorrect least-squares regression coefficients in method-comparison analysis. *Clinical Chemistry, 25*(3), 432-438. https://doi.org/10.1093/clinchem/25.3.432

Cortés-Reyes, E., Rubio-Romero, J. A., & Gaitán-Duarte, H. (2010). Métodos estadísticos de evaluación de la concordancia y la reproducibilidad

de pruebas diagnósticas. *Revista Colombiana de Obstetricia y Ginecología, 61*(3), 247-255. https://doi.org/10.1093/clinchem/25.3.432

Costa-Santos, C., Bernardes, J., Ayres-de-Campos, D., Costa, A., & Costa, C. (2011). The limits of agreement and the intraclass correlation coefficient may be inconsistent in the interpretation of agreement. *Journal of Clinical Epidemiology, 64*(3), 264-269. https://doi.org/10.1016/j.jclinepi.2009.11.010

Dhakal, C. P. (2018). Regression invented as statistics. *International Journal of Interdisciplinary Research and Innovations, 6*(2), 1-5.

de Vet, H. C. W., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology, 59*(10), 1033-1039. https://doi.org/10.1016/j.jclinepi.2005.10.015

Dewitte, K., Fierens, C., Stöckl, D., & Thienpont, L. M. (2002). Application of the Bland-Altman plot for interpretation of method-comparison studies: A critical investigation of its practice. *Clinical Chemistry, 48*(5), 799-801. https://doi.org/10.1093/clinchem/48.5.799

Doğan, N. Ö. (2018). Bland-Altman analysis: A paradigm to understand correlation and agreement. *Turkish Journal of Emergency Medicine, 18*(4), 139-141. https://doi.org/10.1016/j.tjem.2018.09.001

Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education, 38*(9), 1006-1012. https://doi.org/10.1111/j.1365-2929.2004.01932.x

Fleiss, J. L., Levin, B., & Cho Paik, M. (2003). *Statistical methods for rates and proportions* (3rd ed.). John Wiley & Sons.

Galton, F. (1886a). *Adress by Francis Galton, M.A., F.R.S., President of the Anthropological Institute, President of the Section*. Report of the fifty-fifth meeting of the British Association for the Advancement of Science held at Aberdeen in September 1885 (pp. 1206-1214). John Murray.

Galton, F. (1886b). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland, 15*, 246-263. https://doi.org/10.2307/2841583

Galton, F. (1889). I. Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London, 45*(273-279), 135-145. https://doi.org/10.1098/rspl.1888.0082

Gerke, O. (2020a). Reporting standards for a Bland-Altman agreement analysis: A review of methodological reviews. *Diagnostics, 10*(5), 334. https://doi.org/10.3390/diagnostics10050334

Gerke, O. (2020b). Nonparametric limits of agreement in method comparison studies: A simulation study on extreme quantile estimation. *International Journal of Environmental Research and Public Health, 17*(22), 8330. https://doi.org/10.3390/ijerph17228330

Gerke, O., & Möller, S. (2021). Bland-Altman limits of agreement from a Bayesian and frequentist perspective. *Stats, 4*(4), 1080-1090. https://doi.org/10.3390/stats4040062

Gerke, O., Pedersen, A. K., Debrabant, B., Halekoh, U., & Möller, S. (2022). Sample size determination in method comparison and observer variability studies. *Journal of Clinical Monitoring and Computing, 36*, 1241-1243. https://doi.org/10.1007/s10877-022-00853-x

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica, 25*(2), 141-151. https://doi.org/10.11613/BM.2015.015

Giraudeau, B. (1996). Negative values of the intraclass correlation coefficient are not theoretically possible. *Journal of Clinical Epidemiology, 49*(10), 1205-1206. https://doi.org/10.1016/0895-4356(96)00053-4

Gulliford, M. C., Adams, G., Ukoumunne, O. C., Latinovic, R., Chinn, S., & Campbell, M. J. (2005). Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. *Journal of Clinical Epidemiology, 58*(3), 246-251. https_//doi.org/10.1016/j.jclinepi.2004.08.012

Han, H., Ma, Y., & Zhu, W. (2015). *Galton's family heights data revisited*. arXiv:1508.02942v1 [stat.AP]. https://doi.org/10.48550/arXiv.1508.02942

Hernaez, R. (2015). Reliability and agreement studies: A guide for clinical investigators. *Gut, 64*(7), 1018-1027. https://doi.org/10.1136/gutjnl-2014-308619

Houston, W. J. (1983). The analysis of errors in orthodontic measurements. *American Journal of Orthodontics, 83*(5), 382-390. https://doi.org/10.1016/0002-9416(83)90322-6

Hsu, H., & Lachenbruch, P. A. (2005). *Paired t test*. Encyclopedia of Biostatistics, Online. John Wiley & Sons. https://doi.org/10.1002/0470011815.b2a15112

Jan, S. L., & Shieh, G. (2018). The Bland-Altman range of agreement: Exact interval procedure and sample size determination. *Computers in Biology and Medicine, 100*(1), 247-252. https://doi.org/10.1016/j.compbiomed.2018.06.020

Jensen, A. L., & Kjelgaard-Hansen, M. (2006). Method comparison in the clinical laboratory. *Veterinary Clinical Pathology, 35*(3), 276-286. https://doi.org/10.1111/j.1939-165x.2006.tb00131.x

Kalaria, T., Fenn, J., Sanders, A., Ford, C., & Gama, R. (2022). Clinical concordance assessment should be an integral component for laboratory method comparison studies: A regression transference of routine clinical data approach. *Clinical Biochemistry, 103*, 25-28. https://doi.org/10.1016/j.clinbiochem.2022.02.008

Kalra, A. (2017). Decoding the Bland-Altman plot: Basic review. *Journal of the Practice of Cardiovascular Sciences, 3*(1), 36-38. https://doi.org/10.4103/jpcs.jpcs_11_17

Kim, J., & Lee, J. H. (2022). A novel graphical evaluation of agreement. *BMC Medical Research Methodology, 22*(1), Artículo 51. https://doi.org/10.1186/s12874-022-01532-w

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Kotinkaduwa, L. N., & Choudhary, P. K. (2020). A segmented measurement error model for modeling analysis of method comparison data. *Statistics in Medicine, 39*(25), 3491-3502. https://doi.org/10.1002/sim.8677

Kottner, J., Audigé, L., Brorson, S., Donner A., Gajewski, B. J., Hróbjartsson, A., Roberts, C., Shoukri, M., & Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Journal of Clinical Epidemiology, 64*(1), 96-106. https://doi.org/10.1016/j.jclinepi.2010.03.002

Kusunoki, T., Matsuoka, J., Ohtsu, H., Kagimura, T., & Nakamura, H. (2009). Relationship between intraclass and concordance correlation coefficients: Similarities and differences. *Japanese Journal of Biometrics, 30*(1), 35-53. https://doi.org/10.5691/jjb.30.35

Liao, J. J., Capen, R. C., & Schofield, T. L. (2006). Assessing the reproducibility of an analyitical method. *Journal of Chromatographic Science, 44*(3), 119-122. https://doi.org/10.1093/chromsci/44.3.119

Liljequist, D., Elfving, B., & Skavberg Roaldsen, K. (2019). Intraclass correlation - A discussion and demonstration of basic features. *PLoS ONE, 14*(7), Artículo e0219854. https://doi.org/10.1371/journal.pone.0219854

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45*(1), 255-268. https://doi.org/10.2307/2532051

Lin, L. I. (1992). Assay validation using the concordance correlation coefficient. *Biometrics, 48*(2), 599-604. https://doi.org/10.2307/2532314

Linnet, K. (1993). Evaluation of regression procedures for methods comparison studies. *Clinical Chemistry, 39*(3), 424-432. https://doi.org/10.1093/clinchem/39.3.424

Linnet, K. (1999). Necessary sample size for method comparison studies based on regression analysis. *Clinical Chemistry, 45*(6), 882-894. https://doi.org/10.1093/clinchem/45.6.882

Liu, J., Tang, W., Chen, G., Lu, Y., Feng, C., & Tu, X. M. (2016). Correlation and agreement: Overview and clarification of competing concepts and measures. *Shangai Archives of Psychiatry, 28*(2), 115-120. https://doi.org/10.11919/j.issn.1002-0829.216045

Ludbrook, J. (2010). Confidence in Altman-Bland plots: A critical review of the method of differences. *Clinical and Experimental Pharmacology and Physiology, 37*(2), 143-149. https://doi.org/10.1111/j.1440-1681.2009.05288.x

Mandeville, P. B. (2005). Tema 9: El coeficiente de correlación intraclase (ICC). *Ciencia UANL, 8*(3), 414-416.

Mansournia, M. A., Waters, R., Nazemipour, M., Bland, M., & Altman, D. G. (2021). Bland-Altman methods for comparing methods of measurement and response to criticisms. *Global Epidemiology, 3*, Artículo 100045. https://doi.org/10.1016/j.gloepi.2020.100045

Manterola, C., Grande, L., Otzen, T., García, N., Salazar, P., & Quiroz, G. (2018). Confiabilidad, precisión o reproducibilidad de las mediciones. Métodos de valoración, utilidad y aplicaciones en la práctica clínica. *Revista Chilena de Infectología, 35*(6), 680-688. http://doi.org/10.4067/S0716-10182018000600680

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. http://doi.org/10.1097/00000539-200003000-00018

Martelli Filho, J. A., Ávila Maltagliati, L., Trevisan, F., & Lopes de Alcântara Gil, C. T. (2005). Novo método estadístico para análise da reprodutibilidade. *Revista Dental Press de Ortodontia e Ortopedia Facial, 10*(5), 122-129. https://doi.org/10.1590/S1415-54192 005000 500012

Martínez Curbelo, G., Cortés Cortés, M. E., & Pérez Fernández, A. d. (2016). Metodología para el análisis de correlación y concordancia en equipos de mediciones similares. *Universidad y Sociedad, 8*(4), 65-70.

McBride, G. B. (2005). *A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient*. National Institute of Water & Atmospheric Research.

McDemid, R. (2021). Statistics in medicine. *Anaesthesia and Intensive Care Medicine, 22*(7), 454-462. https://doi.org/10.1016/j.mpaic.2021.05.009

Merriman, M. (1877). On the history of the method of least squares. *The Analyst, 4*(2), 33-36. https://doi.org/10.2307/2635472

Mishra, P., Singh, U., Pandey, C. M., Mishra, P., & Pandey, G. (2019). Application of Student's t-test, analysis of variance, and covariance. *Annals of Cardiac Anaesthesia, 22*(4), 407-411. https://doi.org/10.4103/aca.ACA_94_19

Morgan, C. J., & Aban, I. (2016). Methods for evaluating the agreement between diagnostic tests. *Journal of Nuclear Cardiology, 23*(3), 511-513. https://doi.org/10.1007/s12350-015-0175-7

Müller, R., & Büttner, P. (1994). A critical discussion of intraclass correlation coefficients. *Statistics in Medicine, 13*(23-24), 2465-2476. https://doi.org/10.1002/sim.4780132310

Myles, P. S., & Cui, J. (2007). Using the Bland-Altman method to measure agreement with repeated measures. *British Journal of Anaesthesia, 99*(3), 309-311. https://doi.org/10.1093/bja/aem214

National Committee for Clinical Laboratory Standards. (2002). *Method comparison and bias estimation using patient samples, approved Guideline* (EP9-A2, 2nded., Vol. 22, No. 19). Clinical and Laboratory Standards Institute.

Nickerson, C. A. (1997). A note on "A concordance correlation coefficient to evaluate reproducibility". *Biometrics, 53*(4), 1503-1507. https://doi.org/10.2307/2533516

Oldham, P. D. (1962). A note on the analysis of repeated measurements of the same subjects. *Journal of Chronic Diseases, 15*(10), 969-977. https://doi.org/10.1016/0021-9681(62)90116-9

Oleson, J. J., Brown, G. D., & McCreery, R. (2019). The evolution of statistical methods in speech, language, and hearing sciences. *Journal of Speech, Language, and Hearing Research, 62*, 498-506. https://doi.org/10.1044/2018_JSLHR-H-ASTM-18-0378

Pandis, N. (2021). Why using a paired t test to assess agreement is problematic? *American Journal of Orthodontics and Dentofacial Orthopedics, 160*(5), 767-768. https://doi.org/10.1016/j.ajodo.2021.07.001

Passing, H., & Bablok, W. (1983). A new biometrical procedure for testing the equality of measurements from two different analytical methods. *Journal of Clinical Chemistry and Clinical Biochemistry, 21*, 709-720. https://doi.org/10.1515/cclm.1983.21.11.709

Payne, R. B. (1997). Method comparison: Evaluation of least squares, Deming and Passing/Bablok regression procedures using computer simulation. *Annals of Clinical Biochemistry, 34*(3), 319-320. https://doi.org/10.1177/000456329703400317

Pearson, K. (1896). Mathematical contributions to the theory of evolution. -III. Regression, hereditym,

and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 187*, 253-318. https://doi.org/10.1098/rsta.1896.0007

Pearson, K. (1930). *The life, letters and labours of Francis Galton* (Vol. IIIA Correlation, personal identification and eugenics). Cambridge at the University Press.

Rojas, C. M., Puerta, J., Gomez, J., & Calvache, J. A. (2016). Reproducibilidad de las mediciones clínicas. *Revista Facultad de Salud, 8*(1), 42-47. https://doi.org/10.25054/rfs.v8i1.1335

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia, 126*(5), 1763-1768. https://doi.org/10.1213/ANE.0000000000002864

Senn, S., & Richardson, W. (1994). The first t-test. *Statistics in Medicine, 13*(8), 785-803. https://doi.org/10.1002/sim.4780130802

Smith, M. W., Ma, J., & Stafford, R. S. (2010). Bar charts enhance Bland-Altman plots when value ranges are limited. *Journal of Clinical Epidemiology, 63*(2), 180-184. https://doi.org/10.1016/j.jclinepi.2009.06.001

Stevens, N. T., Steiner, S. H., & MacKay, R. J. (2017). Assessing agreement between two measurement systems: An alternative to the limits of agreement approach. *Statistical Methods in Medical Research, 26*(6), 2487-2504. https://doi.org/10.1177/0962280215601133

Stöckl, D., Dewitte, K., & Thienpont, L. M. (1998). Validity of linear regression in method comparison studies: Is it limited by the statistical model or the quality of the analytical input data? *Clinical Chimestry, 44*(11), 2340-2346. https://doi.org/10.1093/clinchem/44.11.2340

Student. (1908). The probable error of a mean. *Biometrika, 6*(1), 1-25. https://doi.org/10.2307/2331554

Taffé, P. (2018). Effective plots to assess bias and precision in method comparison studies. *Statistical Methods in Medical Research, 27*(6), 1650-1660. https://doi.org/ 10.1177/096228021666666

Taffé, P. (2019). Assessing bias, precision, and agreement in method comparison studies. *Statistical Methods in Medical Research 29*(3), 778-796. https://doi.org/10.1177/0962280219844535

Taffé, 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. https://doi.org/10.1016/j.jclinepi.2021.04.004

Taffé, P. (2023). Use of clinical tolerance limits for assessing agreement. *Statistical Methods in Medical Research, 32*(1), 195-206. https://doi.org/10.1177/09622802221137743

Taffé., P., Halfon, P., & Halfon, M. (2020). A new statistical methodology overcame the defects of the Bland & Altman method. *Journal of Clinical Epidemiology, 124*, 1-7. https://doi.org/10.1016/j.jclinepi.2020.03.018

Temel, G., & Erdogan, S. (2017). Determining the sample size in agreement studies. *Marmara Medical Journal, 30*, 101-112. https://doi.org/10.5472/marumj.344822

Ungerer, J. P. J., & Pretorius, C. J. (2018). Method comparison - a practical approach based on error identification. *Clinical Chemistry and Laboratory Medicine, 56*(1), 1-4. https://doi.org/10.1515/cclm-2017-0842

Vetter, T. R., & Schober, P. (2018). Agreement analysis: What he said, she said versus you said. *Anesthesia & Analgesia, 126*(6), 2123-2128. https://doi.org/10.1213/ANE.0000000000002924

Watson, P. F., & Petrie, A. (2010). Method agreement analysis: A review of correct methodology. *Theriogenology, 73*(9), 1167-1179. https://doi.org/10.1016/j.theriogenology.2010.01.003

Westgard, J. O. (1998). Points of care in using statistics in method comparison studies. *Clinical Chemistry, 44*(11), 2240-2242. https://doi.org/10.1093/clinchem/44.11.2240

Westgard, J. O., & Hunt, M. R. (1973). Use and interpretation of common statistical tests in method-comparison studies. *Clinical Chemistry, 19*(1), 49-57. https://doi.org/10.1093/clinchem/19.1.49

Wilkerson, S. D. (2008). Application of the paired t-test. *Xavier University of Louisiana's Undergraduate Research Journal, 5*(1), Article 7.

Zaki, R., Bulgiba, A., & Ismail, N. A. (2013). Testing the agreement of medical instruments: Overestimation of bias in the Bland-Altman analysis. *Preventive Medicine, 57*(Suppl), S80-S82. https://doi.org/10.1016/j.ypmed.2013.01.003

Zaki, R., Bulgiba, A., Ismail, R., & Ismail, N. A. (2012). Statistical method used to test for agreement of medical instruments measuring continuous variables in method comparison studies: A systematic review. *PLoS ONE, 7*(5), Artículo e37908. https://doi.org/10.1371/journal.pone.0037908

Zaki, R., Bulgiba, A., Nordin, N., & Ismail, N. A. (2013). A systematic review of statistical methods used to test for reliability of medical instruments measuring continuous variables. *Iranian Journal of Basic Medical Sciences, 16*(6), 803-807.

Zimmerman, D. W. (1997). Teacher's corner: A note on interpretation of the paired-samples t test. *Journal of Educational and Behavioral Statistics, 22*(3), 349-360. https://doi.org/10.3102/10769986022003349