MHSalud
Revista en Ciencias del Movimiento Humano y Salud
URL: http://www.revistas.una.ac.cr/mhsalud

DOCUMENTO DE ACTUALIZACIÓN AL ARTÍCULO TITULADO:

COMPARACIÓN DE DOS MÉTODOS PARA MEDIR LA COMPOSICIÓN CORPORAL DE FUTBOLISTAS PROFESIONALES COSTARRICENSES

Fecha de recepción de la actualización: 02/06/2020

Fecha de aceptación de la actualización: 20/08/2020

Acerca de la actualización presentada:

En el mes de junio de 2020 las siguientes personas autoras: José Moncada, Braulio Sánchez-Ureña, Felipe Araya-Ramírez, Luis Blanco-Romero, Carmen Crespo-Coco solicitan la incorporación de un documento de actualización referido al artículo titulado "COMPARACIÓN DE DOS MÉTODOS PARA MEDIR LA COMPOSICIÓN CORPORAL DE FUTBOLISTAS PROFESIONALES COSTARRICENSES" publicado por la Revista MHSalud en el volumen 12, número 2 del año 2016.

Razón de la actualización: Quienes presentan la actualización plantean que, a partir de un nuevo análisis de los datos, se identificaron aspectos de mejora con respecto a los resultados reportados en el artículo de 2016. Para esto han preparado un documento explicativo de esta valoración así como recomendaciones para su adecuada interpretación. A continuación se presenta el documento con la actualización realizada.

¿CÓMO EVALUAR ESTADÍSTICAMENTE LOS PUNTAJES DE MÉTODOS DISTINTOS QUE MIDEN EL MISMO CONSTRUCTO?

José Moncada-Jiménez[1] 0000-0001-9807-5163

Braulio Sánchez-Ureña[2] 0000-0001-8791-6836

Felipe Araya-Ramírez[3] 0000-0002-4226-7817

Luis Blanco-Romero[4] 0000-0002-2810-1941

Carmen Crespo-Coco[5] 0000-0002-7634-2002

1 Universidad de Costa Rica, Escuela de Educación Física y Deportes, San José, Costa Rica, jose.moncada@ucr.ac.cr

2 Universidad Nacional, Escuela de Ciencias del Movimiento Humano y Calidad de Vida, Heredia, Costa Rica, braulio.sanchez.urena@una.cr

3 Universidad Nacional, Escuela de Ciencias del Movimiento Humano y Calidad de Vida, Heredia, Costa Rica, felipe.araya.ramirez@una.cr

4 Universidad Nacional, Escuela de Ciencias del Movimiento Humano y Calidad de Vida, Heredia, Costa Rica, luis.blanco.romero@una.cr

5 Universidad de Extremadura, Facultad de Ciencias del Deporte,Laboratorio de Fisiología del Ejercicio, Mérida, España, carmencrespococo@gmail.com

Resumen

El propósito de esta actualización es reconocer la existencia de técnicas estadísticas modernas utilizadas para evaluar la concordancia entre dos métodos que supuestamente miden el mismo constructo. Se utilizaron datos primarios provenientes del estudio publicado en la Revista MHSalud (Sánchez-Ureña, Araya-Ramírez, Blanco-Romero, & Crespo-Coco, 2016), en el que se compararon dos métodos para estimar la composición corporal de futbolistas costarricenses. Los investigadores deseaban comprender la asociación entre los puntajes del método considerado como "estándar de oro", absorciometría con rayos X de doble energía (DXA), y el método indirecto de pliegues cutáneos. Con los análisis originales, los autores concluyeron que ambos métodos eran intercambiables, pero en esta actualización, se utilizaron técnicas de análisis de concordancia de Lin y gráficos de Bland-Altman y se determinó que existe una pobre concordancia entre ambos métodos ($\rho c = 0.70$, $IC_{95\%} = 0.60, 0.77$). Se concluye que los dos métodos no son intercambiables entre sí cuando se desea medir la cantidad de grasa corporal de futbolistas costarricenses.

Palabras claves: medición, correlación, métodos estadísticos

HOW TO STATISTICALLY EVALUATE THE SCORES OF DIFFERENT METHODS THAT MEASURE THE SAME CONSTRUCT?

Abstract

The purpose of this update is to acknowledge the existence of modern statistical techniques used to evaluate the agreement between two methods that supposedly measure the same construct. Primary data from the study published in the Revista MHSalud (Sánchez-Ureña, Araya-Ramírez, Blanco-Romero, & Crespo-Coco, 2016) were used, in which two methods were compared to estimate the body composition of Costa Rican soccer players. The researchers aimed at understanding the association between the scores of the "gold standard" method, dual-energy X-ray absorptiometry (DXA), and the indirect skinfold method. With the original analyses, the authors concluded that both methods were interchangeable; however, in this update, Lin's concordance coefficient and Bland-Altman plots were used to find a poor concordance between both methods ($\rho c = 0.70$, 95% CI = 0.60, 0.77). In conclusion, the two methods are not interchangeable when they are used to measure the body fat mass of Costa Rican soccer players.

Key words: measurement, correlation, statistical methods

Introducción

La literatura estadística "tradicional" generalmente le presenta al investigador varias técnicas de análisis estadístico para resolver el problema de si un método novedoso proporciona información similar o lo más cercana posible a la de un método establecido como criterio o estándar de oro (en inglés, "gold standard"). Este se conoce como un problema de acuerdo o concordancia (en inglés, "agreement", "concordance") entre dos metodologías. Este no es un tema novedoso, pues por ejemplo se ha discutido previamente en el campo de la medicina desde inicios de los años 80's (Altman & Bland, 1983).

Debido a que en los cursos de estadística universitaria en carreras afines a las Ciencias del Movimiento Humano (CMH), incluyendo a la Educación Física, se enseñan métodos para comparar promedios o estudiar asociaciones o correlaciones entre dos variables (McLaughlin, 2013; Moncada-Jiménez, 2004), esos métodos se utilizan para resolver todos los problemas, cuando en realidad se debería buscar la técnica o el conjunto de técnicas más apropiadas para la situación en particular. La teoría subyacente al problema de la concordancia es que ambos métodos podrían ser equivalentes (i.e., proporcionar los mismos resultados con cierto grado de exactitud) al ser utilizados para medir a un individuo, y que la posterior selección de uno u otro método se realizará tomando en consideración criterios no sólo estadísticos sino también prácticos, como por ejemplo, el costo del instrumento o la dificultad de la técnica en sí misma, entre otros (Baumgartner & Jackson, 1998; Miller, 1989). De esta forma, los investigadores tienen a su disposición un variado conjunto de herramientas estadísticas que utilizan indiscriminadamente. Algunas de éstas proporcionan información parcial acerca del grado de equivalencia entre los métodos, mientras que otras, proveen mayor certeza de que la equivalencia está garantizada (McLaughlin, 2013).

El objetivo del presente trabajo es reconocer la existencia de técnicas estadísticas comúnmente utilizadas en las CMH, cuando se desea evaluar la concordancia entre dos métodos que supuestamente miden el mismo constructo. Para cumplir con el objetivo pedagógico de este trabajo, se utilizará la información proporcionada por los autores de un estudio publicado en el año 2016 (Sánchez-Ureña, Araya-Ramírez, Blanco-Romero, & Crespo-Coco, 2016), en el que se comparan dos métodos con los que se estima la composición corporal en futbolistas costarricenses. Para este documento solamente se utilizó la información de la cantidad relativa de grasa corporal total (% grasa), determinados por absorciometría con rayos X de doble energía (DXA, por sus siglas en inglés) y con la técnica de los panículos adiposos o pliegues cutáneos (Heymsfield, Lohman, Wang, & Going, 2005). Como se observa, el problema incluye la utilización de dos variables medidas de forma continua, por lo que se debe utilizar estadística paramétrica adecuada después de haber evaluado los supuestos de cada técnica (e.g., normalidad, sesgo, curtosis) (Lin, Hedayat, & Wu, 2012). Si se tuviera un ejemplo en el que las variables son dicotómicas, como cuando dos personas distintas codifican estudios para un meta análisis (Kelley & Kelley, 2019), entonces se debería recurrir a estadísticos no paramétricos, como por ejemplo, las pruebas de McNemar y κ de Cohen (Cohen, 1960, 1968; McNemar, 1947).

Para contextualizar el ejercicio, en este trabajo primero se explican brevemente algunos de los métodos comúnmente utilizados para resolver el problema del acuerdo o concordancia. Sin embargo, para una revisión más profunda de las técnicas comunes enunciadas desde el siglo pasado y las nuevas tendencias, se pueden consultar los trabajos de Altman y Bland (1983), Bradley y Blackwood (1989) y de Watson y Petrie (2010).

a. Método de las diferencias

El primer método lógico de análisis en el que podemos pensar se basa en la creencia de que si no existen diferencias estadísticas (i.e., $p \leq 0.05$) en las medias aritméticas (i.e., promedios) y

las varianzas de los puntajes de un método y los del otro método, entonces podemos deducir que ambos métodos son equivalentes porque los puntajes promedio provienen de la misma distribución y comparten la misma varianza (Bradley & Blackwood, 1989). Las técnicas de análisis estadístico que permiten analizar esta creencia son las pruebas t-student de medidas repetidas o de análisis de varianza (ANOVA) de medidas repetidas (recordando que $\sqrt{F} = t$, o que $t^2 = F$). Estas técnicas no son correctas para analizar la concordancia entre métodos, ya que prueban la hipótesis de que las medias provienen de una misma distribución de puntajes, y no que cada pareja de puntajes de una persona son idénticos aunque fueran medidos con diferentes métodos. Por lo tanto, las pruebas de diferencias entre medias no son apropiadas para determinar concordancia. A pesar de esto, se utilizan constantemente, como por ejemplo, en el estudio reciente de Klepin, Wing, Higgins, Nichols, y Godino (2019), quienes usaron el test de Bradley-Blackwood para determinar la concordancia entre los promedios y las varianzas de medidas repetidas de puntajes de consumo máximo de oxígeno ($VO_{2máx}$) y las de un acelerómetro comercial.

b. Método de las asociaciones

El segundo método lógico de análisis en el que podemos pensar se basa en la creencia de que, si todos los puntajes obtenidos con un método se asocian o correlacionan con los puntajes obtenidos con el otro método, entonces ambos métodos comparten el mismo constructo, y, por lo tanto, son equivalentes o concordantes. Las técnicas de análisis estadístico que permiten analizar esta creencia son las pruebas basadas en correlación o asociación, una técnica inventada por Sir. Francis Galton (Galton, 1888) y popularizada y conocida como el coeficiente de correlación producto-momento de Pearson (r) (Stigler, 1989). Sin embargo, se ha encontrado que la r de Pearson únicamente determina la magnitud o intensidad de la asociación lineal entre los puntajes de los dos métodos, sin proporcionar información acerca del grado de acuerdo o concordancia observado entre ambos (Bland & Altman, 1986).

El uso incorrecto de la r de Pearson posiblemente proviene de una comprensión superficial de los conceptos de confiabilidad y concordancia. Para Kottner y Streiner (2011), la concordancia se refiere a que dos puntajes provenientes de dos métodos distintos son idénticos o muy similares, o en qué grado difieren. En ese contexto, el grado absoluto de error de la medición es lo que interesa analizar; es decir, lo que interesa es saber si el error aleatorio o el error sistemático afectan la concordancia entre los métodos (Watson & Petrie, 2010). Por lo tanto "no importa la variabilidad entre sujetos o la distribución de los puntajes en la población" (Kottner & Streiner, 2011), p. 701). Por su parte, la confiabilidad se refiere a la relación entre la variabilidad de los puntajes de un mismo sujeto con respecto a la variabilidad total de todos los puntajes de la muestra. En ese contexto, lo que interesa analizar es la habilidad de los puntajes para discriminar entre los sujetos (Kottner & Streiner, 2011). Así, por ejemplo, si un grupo de evaluadores calificara a los estudiantes de un curso con nota 10, la concordancia sería perfecta, pero la confiabilidad de la escala será de cero porque no existiría varianza entre los puntajes de los sujetos. Por lo tanto, tratar de establecer la concordancia entre dos métodos utilizando la r de Pearson sería erróneo.

c. Método de la confiabilidad

El tercer método lógico de análisis en el que podemos pensar se basa en la creencia de que, si todos los puntajes obtenidos con un método son consistentemente similares a los puntajes obtenidos con el otro método, entonces ambos métodos son equivalentes o concordantes. La técnica de análisis estadístico que permite analizar esta creencia es el coeficiente de correlación intraclase (CCI) (Fisher, 1925). El CCI es la razón entre la varianza entre las muestras ("between") y la varianza total ("between" y "within") (Lin, Hedayat, Sinha, & Yang, 2002). De acuerdo con Koo y Li (2016), los valores menores a 0.5 son considerados como pobres, entre 0.5 y 0.75 son moderados, entre 0.75 y 0.90 son buenos, y mayores a 0.90 son excelentes. El CCI puede interpretarse erróneamente cuando no se consideran las varianzas entre y dentro ("within") de los participantes, lo cual puede ocultar el error de medición de las variables (Looney, 2000). A pesar de esto, el CCI se utiliza comúnmente; por ejemplo, cuando se trata de validar un dispositivo inercial de captura de movimientos y se le compara con un goniómetro (García-Rubio, Pino, Olivares, & Ibáñez, 2019).

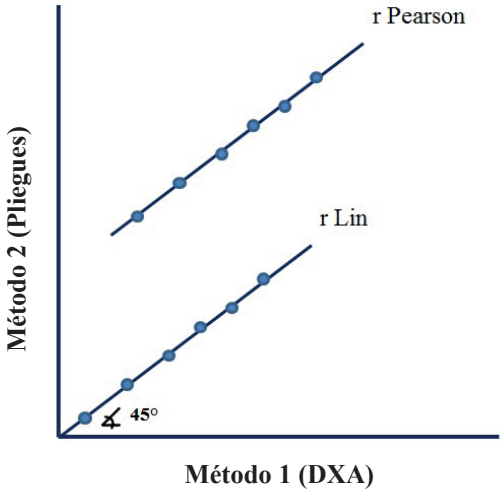
Por lo tanto, como se mencionó anteriormente, se ha discutido que los conceptos de equivalencia, concordancia o acuerdo y correlación no son lo mismo, pues una alta correlación (indicada por una magnitud alta) no necesariamente implica que exista una concordancia entre dos métodos que se utilicen para medir un constructo determinado. Incluso, se acostumbra graficar las diferencias entre parejas de puntajes para conocer el patrón de la asociación y el potencial acuerdo entre los métodos. Esto se realiza con el gráfico de Bland-Altman (Bland & Altman, 1986).

d. Método de la concordancia

Un cuarto método recomendado actualmente, es el coeficiente de concordancia (ρc) propuesto por Lin (1989). Este coeficiente es una modificación de la r de Pearson, en donde no sólo se representa conceptualmente como la forma en que las parejas de puntajes provenientes de dos métodos diferentes se ubican en la línea de mejor ajuste de un gráfico de dispersión, sino también qué tan lejos se encuentra la línea de mejor ajuste del origen a 45°, que representaría una línea de ajuste perfecto. Se ha indicado que este coeficiente tiene la ventaja que es robusto con al menos 10 pares de puntajes (Lin, 1989, 2000). El coeficiente de Lin se interpreta de la siguiente manera con respecto a la fuerza de la concordancia: a) casi perfecto ($\rho c > 0.99$), b) sustancial ($\rho c > 0.95$-0.99), c) moderado ($\rho c = 0.90$-0.95), y d) pobre ($\rho c < 0.90$) (McBride, 2005). En el ejemplo de la concordancia entre el estándar de oro (DXA) y el método de pliegues, se comprendería que aunque la correlación de Pearson fuera perfecta (i.e., $r = 1.0$) ya que los puntos están sobre la línea de mejor ajuste, no necesariamente indicaría que existe una concordancia entre ambos métodos (Figura 1) (Watson & Petrie, 2010).

Figura 1.

Representación conceptual de un coeficiente de correlación de Pearson y un coeficiente de concordancia de Lin. El coeficiente de correlación de Pearson muestra la magnitud de la asociación entre los métodos, mientras que el coeficiente de Lin muestra la magnitud y la concordancia entre los métodos.



Análisis del artículo de Sánchez-Ureña et al. (2016)

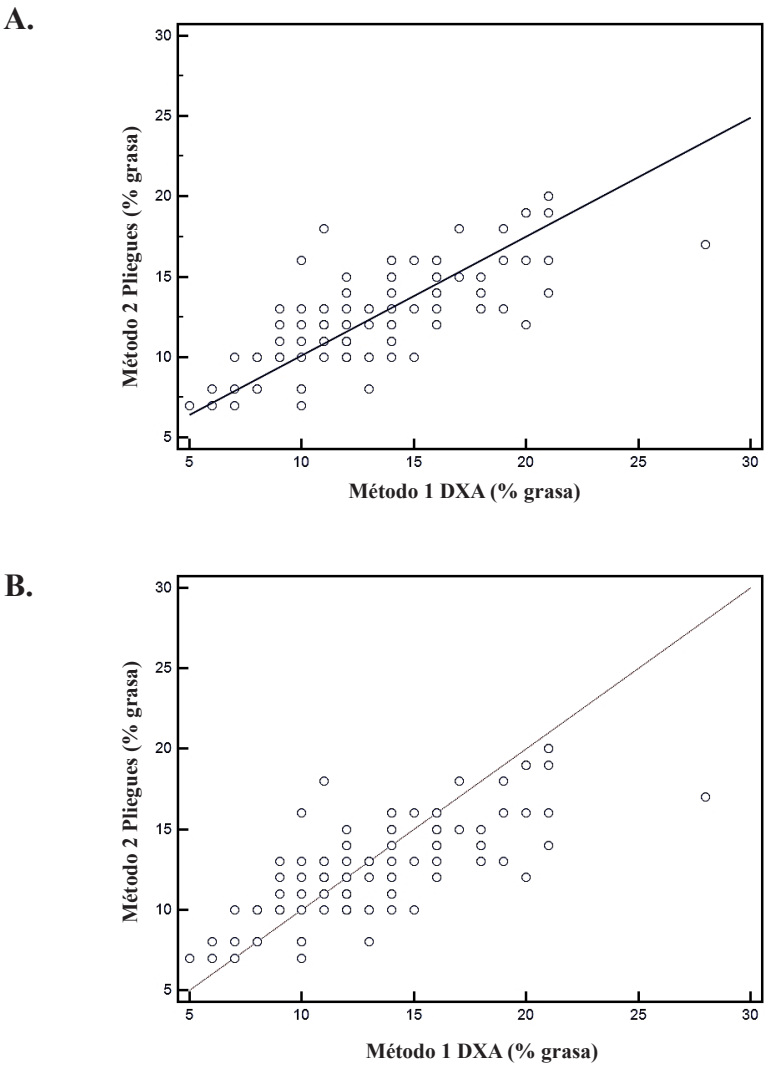
En el artículo de Sánchez-Ureña et al. (2016), se determinó que al medir a los jugadores de fútbol con la técnica de DXA, la diferencia promedio en la cantidad relativa de grasa corporal total era de 0.76% ($t = 2.89$, $p = 0.005$), en donde los valores promedio eran mayores cuando se utilizaba la técnica del DXA que cuando se utilizaba la estimación por medio de los pliegues cutáneos. En síntesis, esto quiere decir las distribuciones de los puntajes son diferentes; en otras palabras, no se obtienen los mismos valores promedio cuando se mide el porcentaje de grasa corporal con DXA y cuando se miden por medio de los pliegues cutáneos.

Cuando los investigadores correlacionaron los porcentajes de grasa corporal obtenidos con ambos métodos obtuvieron una asociación estadística ($r = 0.75$, $p = 0.001$), indicando un coeficiente de determinación $r^2 = 0.56$ (Sánchez-Ureña et al., 2016). Así, la asociación fue estadísticamente diferente de cero y la magnitud moderada, y además, la proporción de variación de los puntajes obtenidos con DXA se explican por los puntajes obtenidos por medio de los pliegues cutáneos; es decir, que entre mayor sea r^2 (i.e., más cercano a 1.0) significa que se podrían predecir mejor los puntajes de DXA con base en las mediciones de pliegues por medio de un análisis de regresión lineal simple (Moncada-Jiménez, 2005). Sin embargo, al integrar ambos hallazgos, los investigadores concluyen erróneamente (como aclararemos más adelante) que ambas técnicas pueden ser utilizadas indistintamente para estimar la grasa corporal debido a la "alta asociación" (Sánchez-Ureña et al., 2016, pág. 1) entre ambas técnicas. Esta afirmación es incorrecta porque no solamente se encontraron

diferencias en las medias entre ambas técnicas, sino también porque la asociación no eliminaba esas diferencias (aunque matemáticamente el valor de r de Pearson fue positivo y estadísticamente significativo). Esa es la razón primordial por la cual los estudios de concordancia no deben utilizar técnicas estadísticas basadas en diferencia de promedios o en la r de Pearson.

Figura 2.

*Diagramas de dispersión que muestran la línea de mejor ajuste de la correlación de Pearson (*r = 0.75, p = 0.001, Panel **A***) y el coeficiente de concordancia de Lin (*ρc = 0.70, $IC_{95\%}$ = 0.60, 0.77, Panel **B***) entre los valores obtenidos mediante DXA y pliegues cutáneos.*

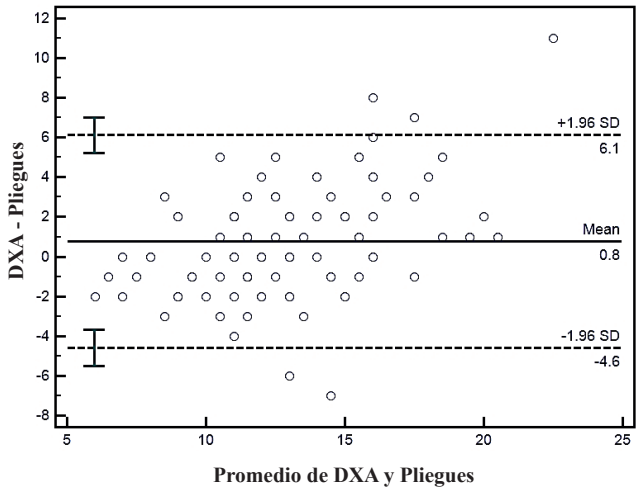


Dado lo anterior, los autores del estudio someten de nuevo los datos crudos a análisis, por lo cual se pudo calcular la prueba t-student, la r de Pearson, el CCI y el coeficiente de concordancia de Lin. Los análisis se realizaron con la versión 20.0 del SPSS, con el software del National Institute of Water and Atmospheric Research (NIWA) de Nueva Zelanda (https://niwa.co.nz/our-services/online-services/statistical-calculators) y con el MedCalc Statistical Software versión 18.6 (MedCalc Software bvba, Ostend, Belgium). Así, se obtuvieron valores del CCI = 0.70 ($IC_{95\%}$ = 0.58, 0.79) y del coeficiente de concordancia de Lin, ρc = 0.70 ($IC_{95\%}$ = 0.60, 0.77), y aunque ambos coeficientes son iguales, los $IC_{95\%}$ son diferentes, siendo el intervalo de ρc más conservador y el del CCI más liberal. De hecho, el CCI se considera moderado (Koo & Li, 2016), pero el ρc se considera pobre (McBride, 2005). Por lo tanto, en el estudio, el grado de concordancia entre los métodos sería considerado como pobre, por lo que no se podría afirmar que ambas metodologías para estimar la proporción de grasa corporal total son intercambiables, concordantes o que brindan la misma información (Figura 2).

Finalmente, se presenta el gráfico de concordancia de Bland-Altman en el que se muestra un sesgo sistemático entre los métodos (Figura 3). Este es un error de medición, en el que los valores tienden a ser consistentemente altos o bajos debido a algún factor extraño (conocido o desconocido), que afecta las mediciones de la misma forma. De acuerdo con Watson y Petrie (2010), este error ocurre cuando el instrumento de medición no se ha calibrado correctamente o cuando el administrador de la prueba consistentemente sobre estima los valores, lo que finalmente produce un sesgo. Para reducir o eliminar ese sesgo, lo que se recomienda es entrenar apropiadamente al personal que recolecta datos, estandarizar los procedimientos de medición, y calibrar apropiadamente los equipos de medición (Watson & Petrie, 2010).

Figura 3.

Límites de concordancia de Bland-Altman entre el porcentaje de grasa corporal medido con DXA y pliegues cutáneos.



Conclusión

El propósito de la presente actualización fue reconocer la existencia de técnicas estadísticas modernas utilizadas para evaluar la concordancia entre dos métodos que supuestamente miden el mismo constructo. Gracias a la calidad profesional de los autores originales y la disposición del cuerpo editorial de la Revista MHSalud, se reanalizaron los datos del estudio publicado previamente para obtener dos nuevas conclusiones: a) los métodos DXA y pliegues cutáneos no proporcionan resultados concordantes en el porcentaje de grasa corporal total de futbolistas costarricenses, y b) se recomienda utilizar el coeficiente de concordancia de Lin y el método gráfico de Bland-Altman para estudios donde se requiera conocer la concordancia entre dos métodos distintos utilizados para medir un mismo constructo. Así, se concluye que el grupo de datos analizados no son concordantes entre sí. La ciencia debe ser objetiva, observable, medible, cuantificable y reproducible; pero, además, debe ser educativa en su divulgación. Esta actualización proporciona herramientas de análisis que pueden aplicarse en los procesos de enseñanza-aprendizaje para mejorar la divulgación de la ciencia.

Referencias

Altman, D. G., & Bland, J. M. (1983). Measurement in Medicine: The Analysis of Method Comparison Studies. *Journal of the Royal Statistical Society. Series D (The Statistician), 32*(3), 307-317. https://doi.org/10.2307/2987937

Baumgartner, T. A., & Jackson, A. S. (1998). *Measurement for evaluation in physical education and exercise science* (6th ed.). WCB/McGraw-Hill.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 327*(8476), 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8

Bradley, E. L., & Blackwood, L. G. (1989). Comparing Paired Data: A Simultaneous Test for Means and Variances. *The American Statistician, 43*(4), 234-235. https://doi.org/10.1080/00031305.1989.10475665

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213. https://doi.org/10.1037/h0026256

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd.

Galton, F. (1888). Co-relations and their measurement. *Proceedings of the Royal Society London, 45*, 135-145. https://doi.org/10.1098/rspl.1888.0082

García-Rubio, J., Pino, J., Olivares, P. R., & Ibáñez, S. J. (2019). Validity and Reliability of the WIMUTM Inertial Device for the Assessment of Joint Angulations. *International Journal of Environmental Research and Public Health, 17*(1), 1-9. https://doi.org/10.3390/ijerph17010193

Heymsfield, S., Lohman, T. G., Wang, Z. M., & Going, S. (2005). *Human Body Composition* (2nd ed.). Champaign, IL: Human Kinetics.

Kelley, G. A., & Kelley, K. S. (2019). Systematic reviews and meta-analysis in nutrition research. *British Journal of Nutrition, 122*(11), 1279-1294. https://doi.org/10.1017/s0007114519002241

Klepin, K., Wing, D., Higgins, M., Nichols, J., & Godino, J. G. (2019). Validity of Cardiorespiratory Fitness Measured with Fitbit Compared to VO2max. *Medicine & Science in Sports & Exercise, 51*(11), 2251-2256. https://doi.org/10.1249/mss.0000000000002041

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

Kottner, J., & Streiner, D. L. (2011). The difference between reliability and agreement. *Journal of Clinical Epidemiology, 64*(6), 701-702. https://doi.org/10.1016/j.jclinepi.2010.12.001

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. (2000). A note on the concordance correlation coefficient. *Biometrics, 56*, 324–325. https://doi.org/10.1177/1536867X0200200206

Lin, L. I., Hedayat, A. S., Sinha, B., & Yang, M. (2002). Statistical methods in assessing agreement: Models, issues, and tools. *Journal of the American Statistical Association, 97*(457), 257-270. https://doi.org/10.1198/016214502753479392

Lin, L. I., Hedayat, A. S., & Wu, W. (2012). *Statistical Tools for Measuring Agreement*. New York: Springer.

Looney, M. A. (2000). When is the intraclass correlation coefficient misleading? *Measurement in Physical Education and Exercise Science, 4*(2), 73-78. https://doi.org/10.1207/s15327841mpee0402_3

McBride, G. B. (2005). *Using Statistical Methods for Water Quality Management: Issues, Problems and Solutions*. New York, NY: Wiley.

McLaughlin, P. (2013). Testing agreement between a new method and the gold standard—How do we test? *Journal of Biomechanics, 46*(16), 2757-2760. https://doi.org/10.1016/j.jbiomech.2013.08.015

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153-157. https://doi.org/10.1007/bf02295996

Miller, D. K. (1989). *Measurement by the Physical Educator: Why and How*. New York, NY: Benchmark Press, Inc.

Moncada-Jiménez, J. (2004). Métodos estadísticos utilizados en las ciencias del movimiento humano. *Revista Educación, 28*(2), 279-287. https://doi.org/ 10.15517/revedu.v28i2.2265

Moncada-Jiménez, J. (2005). *Estadística: para las ciencias del movimiento humano*. Editorial de la Universidad de Costa Rica.

Sánchez-Ureña, B., Araya-Ramírez, F., Blanco-Romero, L., & Crespo-Coco, C. (2016). Comparación de dos métodos para medir la composición corporal de futbolistas profesionales costarricenses. *MHSalud: Revista en Ciencias del Movimiento Humano y Salud, 12*(2), 1-13. https://doi.org/10.15359/mhs.12-2.1

Stigler, S. M. (1989). Francis Galton's account of the invention of correlation. *Statistical Science*, 73-79. https://www.jstor.org/stable/2245329

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