Artículo de revisión
Vol. 64
Núm. 1

Equivalencia y acuerdo en validación de instrumentos: una revisión metodológica práctica

Equivalence and agreement in validation studies: A practical methodological review

Silvina Dell'Era[1a], Vanina Pagotto[1b]

Resumen

El análisis estadístico adecuado es esencial en estudios de validación de instrumentos que cuantifican variables continuas frente a un estándar de referencia. Este artículo describe enfoques estadísticos para evaluar la equivalencia entre instrumentos de medición que combinan métodos gráficos y pruebas estadísticas. Se ejemplifica su aplicación en un estudio que evaluó la exactitud de una pulsera de medición de actividad física (*Xiaomi Mi Band 4*) para contar pasos caminados en diferentes actividades en pacientes con enfermedades respiratorias crónicas, y se comparó con un método de referencia basado en videofilmación. Se emplearon intervalos de confianza frente a zonas de equivalencia predefinidas, pruebas TOST (*two one-sided tests*) y se calcularon indicadores de acuerdo grupal e individual como el error medio (ME), el error porcentual medio (MPE), el error porcentual absoluto medio (MAPE) y la raíz del error cuadrático medio (RMSE). Asimismo, se discutieron algunos errores frecuentes como el uso inapropiado de gráficos de dispersión o correlaciones para evaluar la exactitud. Se concluyó que la elección de métodos estadísticos apropiados es un aspecto clave para asegurar la validez clínica y metodológica en estudios de equivalencia entre instrumentos de medición que cuantifican variables continuas y un método de referencia.

Abstract

Accurate statistical analysis is essential in validation studies of instruments that quantify continuous variables against a reference standard. This article describes statistical approaches to evaluate equivalence between measurement instruments combining graphical methods and statistical tests. Its application is exemplified through a study that assessed the accuracy of a physical activity tracking wristband (Xiaomi Mi Band 4) for counting steps walked during different activities in patients with chronic respiratory diseases, and it was compared with a video-based reference method. Confidence intervals were used alongside predefined equivalence zones, TOST (two one-sided tests) procedures were applied, and both group-level and individual-level indicators of agreement were calculated, such as the mean error (ME), mean percentage error (MPE), mean absolute percentage error (MAPE), and root mean squared error (RMSE). In addition, some common errors were also discussed, such as the inappropriate use of scatter plots or correlations to assess accuracy. The article concludes that selecting appropriate statistical methods is a key aspect to ensure clinical and methodological validity in equivalence studies between measurement instruments that quantify continuous variables and a reference method.

[1]Universidad Hospital Italiano, Secretaría de Investigación. Ciudad Autónoma de Buenos Aires, Buenos Aires, Argentina

ORCID: 0000-0001-9186-6229[a], 0000-0003-0309-2660[b]

Palabras clave
Métodos
Estudio de Validación
Exactitud de los Datos
Estudio de Evaluación
Análisis de Datos

Keywords
Methods
Validation Study
Data Accuracy
Evaluation Study
Data Analysis

Fecha de recibido: 29/07/2025

Fecha de aceptado: 28/08/2025

Comunicación con:
Vanina Pagotto
vanina.pagotto@hospitalitaliano.org.ar
(54) 11 4959 0200, interno 9353

Cómo citar este artículo: Dell'Era S, Pagotto V. Equivalencia y acuerdo en validación de instrumentos: una revisión metodológica práctica. Rev Med Inst Mex Seguro Soc. 2026;64(1):e6772. doi: 10.5281/zenodo.17477647

Licencia CC 4.0

Copyright:© 2026 Instituto Mexicano del Seguro Social

Introducción

En el contexto de la investigación clínica y de la práctica médica, es habitual que se desarrollen instrumentos alternativos para cuantificar variables continuas como el número de pasos, la distancia recorrida, el peso o la composición corporal. Estos nuevos instrumentos suelen tener ventajas en términos de accesibilidad, costo, portabilidad y facilidad de uso,[1] pero para que puedan ser adoptados de manera confiable, es necesario establecer su validez frente a un método de referencia o estándar de oro.[2,3,4] La pregunta que motivó esta investigación fue ¿cómo evaluar de forma adecuada si un nuevo instrumento de medición genera resultados equivalentes a los de un método de referencia? Este interrogante adquiere particular relevancia frente a la creciente disponibilidad de dispositivos portátiles o tecnologías utilizadas en el seguimiento de pacientes, especialmente aquellos con enfermedades crónicas.[4,5,6] La correlación entre mediciones de diferentes instrumentos no permite asegurar que los 2 métodos sean intercambiables; se requiere un análisis que considere si las diferencias observadas son lo suficientemente pequeñas como para ser clínicamente irrelevantes.[2,7] Este artículo de revisión tiene como propósito describir los métodos estadísticos apropiados para evaluar la equivalencia entre instrumentos de medición continua y cuantificar el acuerdo entre ellos, con lo que se evitan errores metodológicos frecuentes.[3,8,9]

Metodología

Se presenta una revisión aplicada con base en un estudio que comparó el conteo de pasos con un reloj pulsera de medición de actividad física de bajo costo (*Xiaomi Mi Band 4* [XMB4]), el nuevo método de medición, frente a un método de referencia basado en la videofilmación, en 33 pacientes con enfermedades respiratorias crónicas durante diferentes actividades.[4] El XMB4 demostró ser exacto para medir pasos durante caminatas de 10 y 30 metros, pero no así en las de 5 metros. El estudio completo donde se realizó la validación del XMB4 en esta población y de donde se obtuvieron los datos, se encuentra publicado.[4]

A continuación, se describe el análisis realizado para evaluar la equivalencia de ambos métodos con un enfoque adecuado y asimismo se señala qué estrategias analíticas no resultan válidas y se justifica por qué.

Resultados

Se presentan a continuación los métodos adecuados para evaluar la equivalencia estadística. Los siguientes enfoques pueden ser utilizados complementariamente: el método del intervalo de confianza, el método *two one-side test* (TOST) y la cuantificación de los errores de medición con los indicadores de acuerdo.

Método del intervalo de confianza

El método del intervalo de confianza (o test de equivalencia del 95%) permite testear gráficamente la equivalencia estadística. La zona de equivalencia debe ser seleccionada considerando un porcentaje a ambos lados de la media de la medición del criterio o método de referencia. El método evalúa si el intervalo de confianza del 90% de las mediciones del nuevo método (instrumento a validar) cae completamente dentro de la zona de equivalencia definida, con lo que concluye que los métodos de medición son estadísticamente equivalentes.

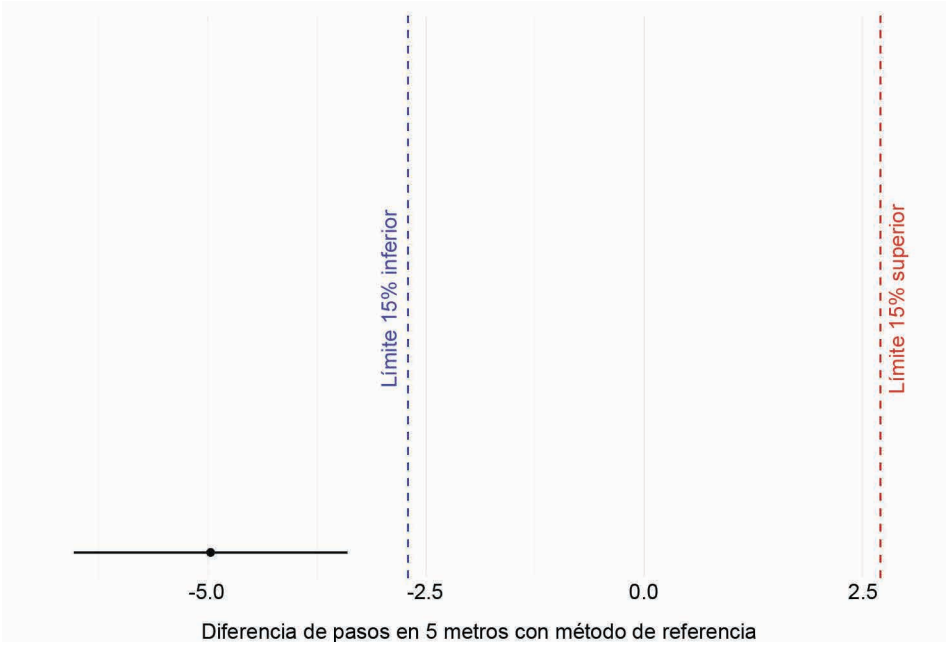
Un paso clave durante el diseño del estudio de validación del instrumento de medición es determinar cuál será la zona o región de equivalencia que se considera aceptable en la población de interés. Para definirla, se contempla la bibliografía publicada, la experiencia clínica y la utilidad que se le dará al nuevo instrumento de medición; debe determinarse, asimismo, el delta o la diferencia aceptable entre ambos métodos antes de comenzar con la recolección de los datos.

La visualización gráfica de este método puede hacerse de 2 formas: considerando la diferencia entre los métodos y evaluando si esa diferencia se encuentra dentro de la zona de equivalencia definida, o considerando los valores absolutos y evaluando si la media y el intervalo de confianza del 90% del nuevo método quedan contenidos dentro de la zona de equivalencia (lo cual se calcula como la media del método de referencia ± el porcentaje de error definido por los investigadores como aceptable o equivalente).

En el ejemplo, se presenta la equivalencia estadística utilizando el método del intervalo de confianza —graficando la diferencia entre los métodos— entre el XMB4 y el método de referencia (videofilmación) para cuantificar los pasos realizados por los pacientes durante 2 caminatas de diferentes distancias: una de 5 (figura 1) y otra de 10 metros (figura 2). La zona de equivalencia seleccionada en el estudio fue de ±15% con base en bibliografía previa.[8] En el estudio original[4] se presentan los resultados del método de intervalo de confianza con los valores absolutos.

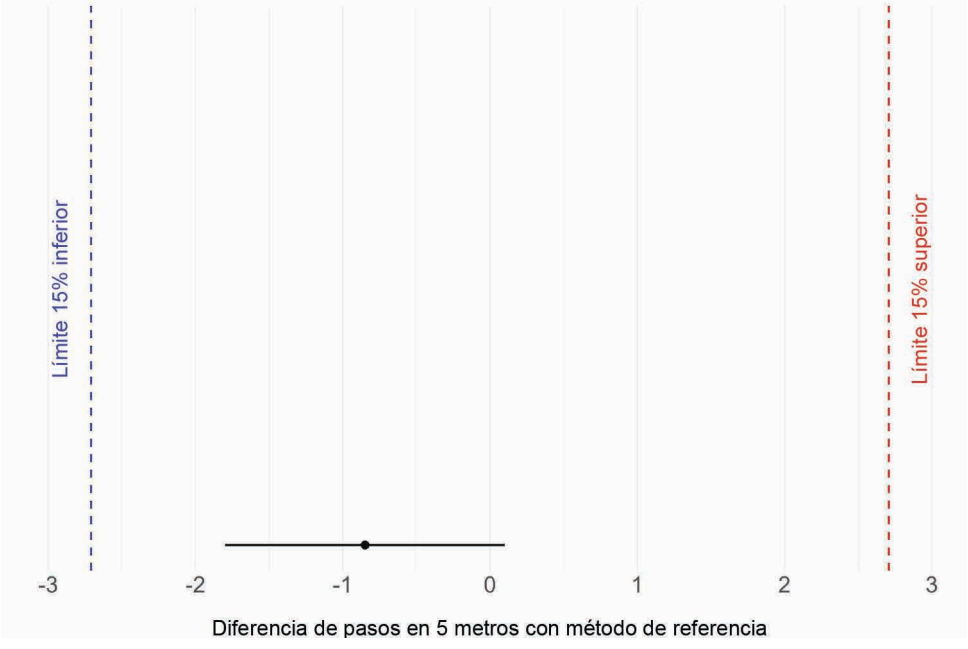
En el ejemplo, se evidencia que los métodos no son equivalentes para cuantificar los pasos durante las caminatas de 5 metros y se observa en la figura 1 que la media de la diferencia entre los métodos y su intervalo de confianza cae por fuera de los límites aceptables (en este caso por debajo

Figura 1 Comparación de métodos para la medición de pasos en 5 metros mediante el método del intervalo de confianza entre el XMB4 y la videofilmación



En la figura se representa la equivalencia mediante la diferencia de pasos cuantificados por el nuevo método (XMB4) y el método de referencia (videofilmación) para pasos caminados en 5 metros con márgenes de equivalencia seleccionados por los autores como aceptables de ±15% (zona de equivalencia) representados como líneas punteadas, y en azul el límite inferior y rojo el superior; el pequeño círculo representa la media de la diferencia entre XMB4 y el método de referencia y las barras los intervalos de confianza del 90% (IC 90%)

Figura 2 Comparación de métodos para la medición de pasos en 10 metros mediante el método del intervalo de confianza entre el XMB4 y la videofilmación



En la figura se representa la equivalencia mediante la diferencia de pasos cuantificados por el nuevo método (XMB4) y el método de referencia (videofilmación), para pasos caminados en 10 metros con márgenes de equivalencia seleccionados por los autores como aceptables de ±15% (zona de equivalencia) representados como líneas punteadas, y en azul el límite inferior y rojo el superior; el pequeño círculo representa la media de la diferencia entre XMB4 y el método de referencia y las barras los intervalos de confianza del 90% (IC 90%)

del límite inferior, con lo que se subestiman por lo tanto con el nuevo método los pasos caminados respecto al método de referencia). Por el contrario, se observa en la figura 2 que el intervalo de confianza de la diferencia entre los métodos cae dentro de los límites de equivalencia establecidos, por lo que se concluye que ambos métodos son equivalentes.

Método TOST

El método *two one-side test*, o también llamado TOST por sus iniciales en inglés, es una prueba estadística que se utiliza para evaluar la equivalencia entre 2 mediciones con el que se determina si la diferencia entre ellas se encuentra dentro de un rango de equivalencia definido previamente. En el método TOST se realizan simultáneamente 2 pruebas unilaterales o a una cola: una para evaluar si la diferencia observada entre los métodos es menor o igual que el límite inferior establecido y otra para verificar si es mayor o igual al límite superior.

El método TOST implica hacer 2 pruebas unilaterales:

Prueba 1: H_{01}: μ_D ≤ -Δ frente a H_A_1: μ_D> -Δ

Prueba 2: H_{02}: μ_D ≥ Δ frente a H_A_2: μ_D < Δ

donde H_{01} es la primera hipótesis nula, H_A_1 es la primera hipótesis alternativa, μ_D es la diferencia media entre los dos métodos (por ejemplo, XMB4 - videofilmación), y Δ es el margen de equivalencia definido como clínicamente aceptable.

De esta manera se establecen las hipótesis nula y alternativa:

H_0: μ_D ≤ -Δ o μ_D ≥ Δ (no equivalencia)

Ha: -Δ < μ_D < Δ (equivalencia)

donde la hipótesis nula es de no equivalencia y, por lo tanto, si las 2 pruebas unilaterales son significativas, se rechaza la hipótesis nula de no equivalencia y se concluye equivalencia si ambas pruebas unilaterales son estadísticamente significativas (por ejemplo, si ambos valores de *p* son < 0.05), o si el intervalo de confianza del 90% se encuentra completamente contenido dentro de [-Δ, +Δ].

En la figura 3 se muestra esquemáticamente el TOST. Si ambas pruebas resultan significativas, se rechaza la hipótesis nula de no equivalencia y se concluye que las mediciones son equivalentes, ya que las diferencias observadas entre los métodos caen tanto por encima del límite inferior como por debajo del superior, es decir dentro de los límites establecidos como equivalencia.

Utilizando los mismos datos y los límites de equivalencia seleccionados de ±15%, se empleó el método TOST para evaluar la equivalencia entre ambos métodos en la medición (XMB4 y videofilmación) de los pasos caminados en distancias de 5 y de 10 metros.

Para los pasos en 5 metros, la prueba de equivalencia fue significativa para el límite superior (estadístico t -0.6702; $p < 0.001$); sin embargo, no fue significativa para el límite inferior (estadístico t -3.749; $p = 0.999$). Una observación importante es que para que el TOST determine equivalencia global, ambos límites (inferior y superior) deben ser significativos; por lo tanto, no se demostró equivalencia entre el XMB4 y la videofilmación para contabilizar pasos caminados en una distancia de 5 metros, lo que concuerda con lo observado en el método gráfico del intervalo de confianza.

Para los pasos registrados en caminatas de 10 metros, la prueba de equivalencia fue significativa tanto para el límite superior (estadístico t 3.23; $p = 0.001$), como para el límite inferior (estadístico t 6.17; $p < 0.001$), y por lo tanto se concluye que ambos métodos de medición son estadísticamente equivalentes. En la figura 4 se presentan los resultados del TOST para 5 y 10 metros.

Medidas de acuerdo global e individual

Luego de evaluar la equivalencia, es útil cuantificar el error de medición del método nuevo en relación con el de referencia. Para ello se calculan 2 indicadores de validación que reflejan el acuerdo a nivel grupal (la media de error y la media de error porcentual) y 2 indicadores de acuerdo individual (el error porcentual absoluto medio y el error cuadrático medio).

La media de error (ME) se calcula al promediar la diferencia entre el método de referencia y el nuevo método. En el ejemplo, la diferencia entre la videofilmación y el XMB4:

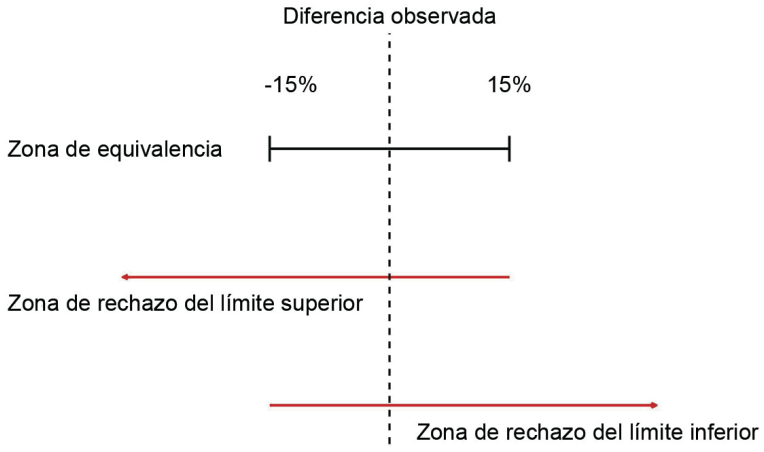
$$ME = (1/n) \times \Sigma (\text{videofilmación}_i - \text{XMB4}_i)$$

La media de error porcentual (MPE) estandariza esta diferencia como un porcentaje respecto al valor del criterio de referencia. Se estima al promediar los errores porcentuales individuales dividido el valor del criterio de referencia [videofilmación-XMB4]/videofilmación). Este indicador permite capturar el grado de sobreestimación o subestimación general para el dispositivo XMB4, comparado con la videofilmación:

$$MPE = (1/n) \times \Sigma [(\text{videofilmación}_i - \text{XMB4}_i) / \text{video filmación}_i] \times 100$$

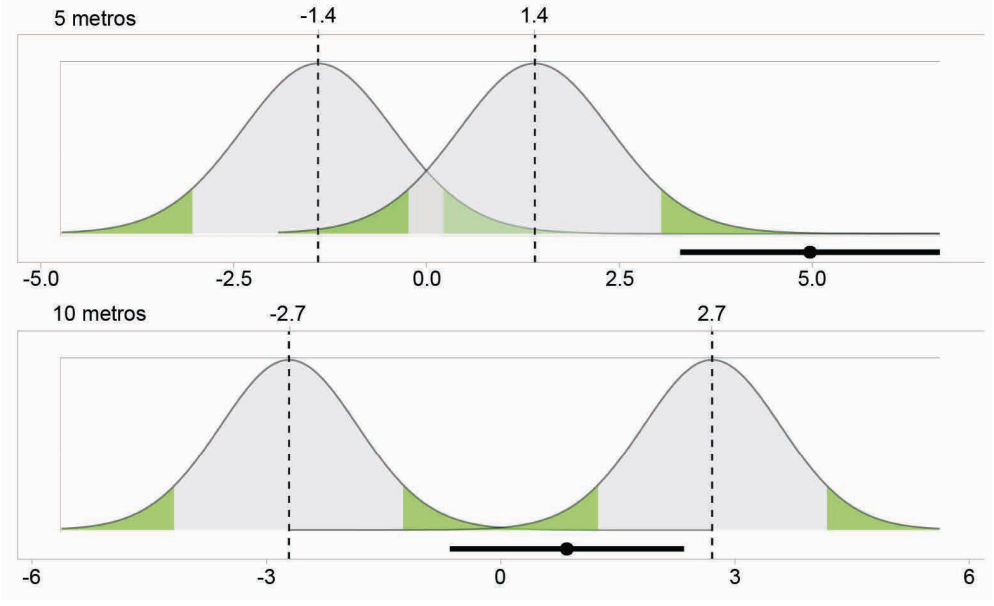
El error porcentual absoluto medio (MAPE) se calcula

Figura 3 Zona de equivalencia y de rechazo de las hipótesis nulas del TOST



En la figura 3 se muestran los límites de la zona de equivalencia (-15% a 15%) definidos para el ejemplo, indicados por las líneas negras verticales de la barra, dentro de las cuales las diferencias observadas se consideran equivalentes. La línea punteada vertical representa la diferencia observada y su relación con estos límites de equivalencia. Se representan las zonas donde se rechaza cada una de las hipótesis de no equivalencia, la del límite superior e inferior. En el caso del límite superior, la hipótesis nula testeada es que la diferencia observada entre ambos métodos de medición es mayor o igual que dicho límite, y la zona de rechazo de la hipótesis nula de este límite (indicada por línea roja hacia la izquierda) incluye valores menores que el límite superior de equivalencia, lo cual indica en caso de rechazo que la diferencia entre los métodos se encuentra por debajo del límite superior. Para el límite inferior, la hipótesis nula testeada es que la diferencia es menor o igual que dicho límite. La zona de rechazo de esta hipótesis (línea roja hacia la derecha) incluye valores por arriba del límite inferior de equivalencia que indican en el caso de rechazo que la diferencia entre los métodos se encuentra por encima del límite inferior

Figura 4 Prueba de equivalencia (TOST) entre la pulsera *Xiaomi Mi Band 4* y la videofilmación en las pruebas de 5 y 10 metros (arriba y abajo, respectivamente)



Las curvas representan la distribución bajo la hipótesis nula de no equivalencia y las líneas verticales discontinuas señalan los márgenes de equivalencia preestablecidos. El punto indica la diferencia media observada y la barra negra su intervalo de confianza del 90%. Las áreas verdes corresponden a las regiones de rechazo de la hipótesis nula de no equivalencia; el color verde que se aprecia en la zona de intersección de las curvas proviene únicamente de la superposición gráfica y no tiene un significado estadístico. En la prueba de 5 metros, el intervalo de confianza excede los márgenes de equivalencia y no puede concluirse equivalencia; en la de 10 metros, el intervalo se encuentra completamente contenido dentro de los márgenes, lo que permite concluir equivalencia entre los métodos

promediando los errores porcentuales absolutos individuales dividido el valor del criterio de referencia (|[videofilmación-XMB4]/videofilmación|). Esta medida permite obtener los errores que representan tanto la sobreestimación como la subestimación a nivel de cada participante:

$$MAPE = (1/n) \times \Sigma \left|(videofilmación_i - XMB4_i) / videofilmación_i\right| \times 100$$

Valores similares de MPE y MAPE indican que el instrumento tiende a subestimar o sobrestimar consistentemente a nivel grupal e individual. Por el contrario, si los errores se cancelan (algunas subestimaciones y algunas sobreestimaciones), el MPE será bajo, pero el MAPE reflejará mejor el error real. El MAPE también facilita la comparación con otros estudios, y se consideran como buenos los valores entre 10 y 20% y como aceptables entre 20 y 30%,[10] o como excelentes de 0 a 5%, buenos de 5 a 10%, aceptables entre 10 y 15% y pobres > 15%,[11] según la literatura consultada.

El error cuadrático medio (RMSE) se calcula como la raíz cuadrada del promedio de los errores elevados al cuadrado, y proporciona una medida de la dispersión del error.[11] Todos estos diferentes indicadores se estiman para proporcionar una descripción completa del error de medición.

En el cuadro I se comparan los errores de medición de los pasos caminados censados por el método nuevo (XMB4) y el método de referencia (videofilmación) para las distancias de 5 y 10 metros. Se observa que el dispositivo presenta menor error de medición para las distancias de 10 metros.

Aproximaciones estadísticas inadecuadas

Por último, cabe mencionar que hay otras aproximaciones estadísticas que si bien son de amplio uso en los estudios de validación, deberían evitarse. Se describen a continuación estas aproximaciones junto con la justificación de sus limitaciones.

Una de ellas son los gráficos de dispersión, los cuales no evalúan la concordancia ni las diferencias entre métodos, sino que solo muestran las variaciones entre mediciones en un rango determinado y no dan información sobre discrepancias individuales. En el mismo sentido, los coeficientes de correlación indican el grado de asociación lineal entre 2 variables, pero no reflejan la concordancia ni identifican sesgos sistemáticos o errores de medición.

El coeficiente de correlación intraclase (CCI), si bien es una medida de acuerdo entre métodos con datos continuos, no permite la comparación entre un método a validar con uno de referencia, sino que su aplicación es útil para evaluar la estabilidad temporal de instrumentos de medida, es decir, la confiabilidad en 2 momentos diferentes en el tiempo entre las observaciones.

Por otro lado, los conocidos gráficos de Bland y Altman si bien permiten visualizar la distribución de la diferencia entre los métodos o instrumentos de medición con los límites de concordancia establecidos con una media de las diferencias de ±1.96 desvío estándar, tampoco son adecuados cuando está disponible el método de referencia, ya que su objetivo es estimar la concordancia entre 2 métodos de medición independientes, y el supuesto es que ambos métodos tienen el mismo nivel de validez y error, y buscan identificar si existen diferencias sistemáticas o aleatorias entre ellos (figura 5 y figura 6).

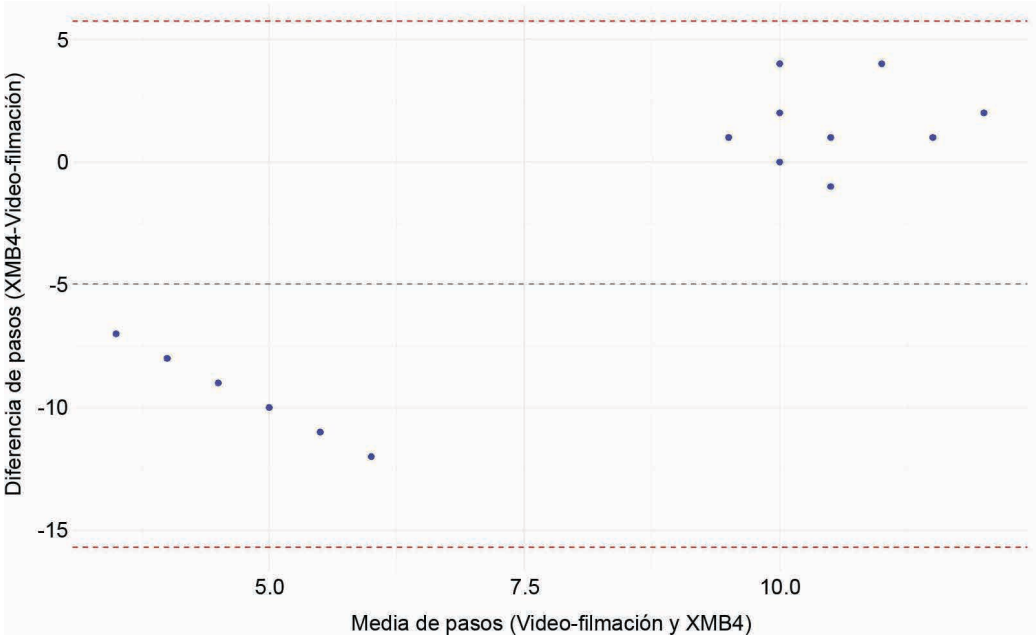
Tampoco es apropiado usar pruebas estadísticas como t-test o ANOVA para evaluar las diferencias entre un método nuevo y uno de referencia, ya que evalúan diferencias en medias y no el grado de acuerdo entre 2 métodos. Otro aspecto es que estas pruebas no evalúan si las diferencias son relevantes o si están dentro de un rango aceptable para que el método nuevo sea considerado equivalente al de

Cuadro I Comparación de la cantidad de pasos caminados registrados por el *Xiaomi Mi Band 4* y observados en la videofilmación en caminatas de 5 y 10 metros (*n* = 33)

Pasos en la caminata	ME (DE)	MPE (DE)	MAPE (DE)	RMSE
5 metros	4.97 (5.5)	54% (7.1%)	68.3% (41.3%)	7.44
10 metros	0.85 (3.31)	4.8% (0.6%)	12.4% (18%)	3.42

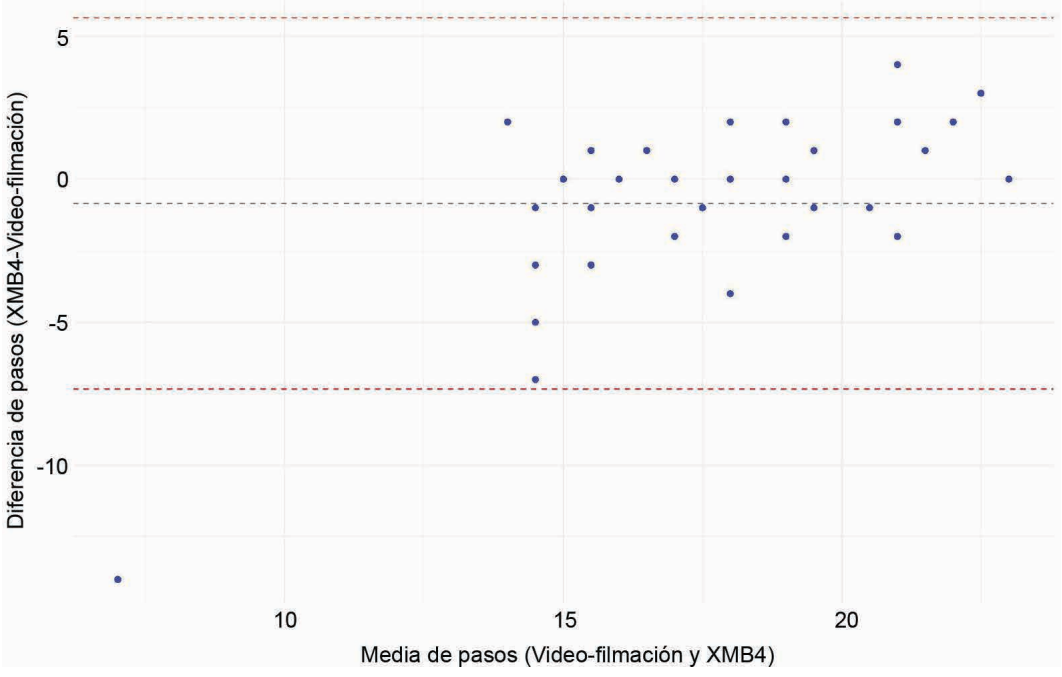
Se comparan los errores de medición entre el método nuevo (Xiaomi Mi Band 4) y el método de referencia (videofilmación) para distancias de 5 y 10 metros. Para las caminatas de 5 metros, el error promedio (ME) fue de 4.97 pasos y el error porcentual (MPE) de 54%, y son consistentes con subestimación. Para las de 10 metros, un error promedio (ME) de 0.85 pasos y un error porcentual medio (MPE) de 4.8% indican un bajo nivel de subestimación. El error porcentual absoluto medio (MAPE), que mide la magnitud total del error sin considerar la dirección, fue 68.3% para la distancia de 5 metros y de 12.4% para la de 10 metros, en este último caso considerado como bueno. El error cuadrático medio (RMSE) refleja mayor variabilidad en distancias de 5 metros (7.44) frente a 10 metros

Figura 5 Gráficos de Bland y Altman para pasos caminados en una distancia de 5 metros



En el eje X se representa el promedio de los valores entre ambos métodos y en el eje Y las diferencias absolutas (videofilmación *Xiaomi Mi Band 4*). La línea negra punteada corresponde a la media de la diferencia entre los métodos y las líneas discontinuas rojas a los límites de concordancia (±1.96 desvío estándar)

Figura 6 Gráficos de Bland y Altman para pasos caminados en una distancia de 10 metros



En el eje X se representa el promedio de los valores entre ambos métodos y en el eje Y las diferencias absolutas (videofilmación *Xiaomi Mi Band 4*). La línea negra punteada corresponde a la media de la diferencia entre los métodos y las líneas discontinuas rojas a los límites de concordancia (±1.96 desvío estándar)

referencia, sino que simplemente demuestran si existe una diferencia estadísticamente significativa.

Discusión

La comparación entre 2 métodos de medición continua exige un enfoque metodológico específico que va más allá del análisis de diferencias estadísticas o correlaciones. En particular, cuando uno de los métodos se considera un estándar de referencia (estándar de oro),[2,3,4] el objetivo debe ser establecer si el nuevo instrumento es equivalente, es decir, si puede reemplazar al estándar sin que la diferencia entre ambos afecte las decisiones clínicas o de investigación.

Este principio fue aplicado rigurosamente en el presente estudio, donde se utilizó una zona de equivalencia predefinida (±15%) y se implementó un análisis por intervalo de confianza del 90% en el que se siguieron los marcos conceptuales recomendados.[7,8,9] La combinación de este enfoque con el uso de métricas complementarias como el MAPE y el RMSE permitió una caracterización robusta del desempeño del dispositivo evaluado (*Xiaomi Mi Band 4*). Este tipo de abordaje, aún poco utilizado en estudios de validación, permite determinar no solo si 2 métodos están correlacionados, sino si sus mediciones son suficientemente similares como para ser intercambiables en la práctica.[7,8,9]

La revisión de artículos publicados sobre validación de métodos de medición revela que gran parte de la literatura actual no sigue este marco metodológico. En numerosos casos, como en los estudios que evaluaron podómetros en niños y adolescentes,[10,11,12] se observa la utilización del coeficiente de correlación de Pearson o Spearman como criterio de equivalencia. Esta estrategia es inadecuada, ya que la correlación evalúa asociación lineal, pero no acuerdo. De hecho, pueden existir correlaciones altas incluso en presencia de sesgo sistemático.[13,14,15]

Otros trabajos, como los que compararon la medición de movimiento como saltos, actividad física con dispositivos móviles o plataformas de fuerza,[11,16,17] o los que contrastaron medición de composición corporal,[18,19] o que compararon variables respiratorias en pacientes ventilados,[20] si bien incorporan métricas como el CCI,[21] el error estándar o gráficos de Bland y Altman, tampoco definen de manera explícita una zona de equivalencia ni aplican pruebas estadísticas como el TOST. En algunos de estos casos, se aplicó una zona de equivalencia solo en una de las variables mientras que en las restantes se recurrió nuevamente a correlaciones u otras métricas, lo cual generó inconsistencias analíticas dentro del mismo estudio.

En el grupo de estudios que compararon tomografía computada con otros métodos diagnósticos en mediciones óseas[22] o dispositivos para medición de movimiento, cinética articular o actividad física diaria,[17,23] se aplicaron t-tests y análisis de correlación, sin prueba de equivalencia ni margen predefinido. En estos casos, aunque se cuente con un método de referencia válido, los métodos utilizados no permiten afirmar que ambos sean intercambiables. Esto ilustra una problemática generalizada: la confusión entre ausencia de diferencia estadística y equivalencia clínica, un error conceptual ampliamente documentado.[7,8,24]

Varios estudios utilizaron este método y describieron los márgenes de equivalencia con base en la literatura, tanto en el método de intervalos de confianza[4,5,10] como en el TOST.[5,25,26] Por el contrario, se observan estudios donde se aplicaron métodos adecuados, pero no hay justificación basada en estudio previos del margen de equivalencia,[27] o no se incluyeron errores como MAPE que permita comparar con otros estudios o RMSE.[18,28]

Como se mencionó, la falta de definición del margen de equivalencia antes del análisis fue un problema frecuente. Algunos estudios lo omitieron completamente.[20,29] Este margen debe establecerse a priori, en función de la variabilidad esperada, estándares previos o consensos clínicos.[2]

A estas debilidades se suma una importante heterogeneidad en la forma en que se reportan los resultados. Algunos estudios no indican claramente el método estadístico utilizado, no reportan intervalos de confianza o no explicitan cómo se definió la equivalencia.[30] Esta falta de transparencia metodológica limita la reproducibilidad y la capacidad de comparación entre estudios.[7,8,24]

En la misma línea, se menciona que el uso de gráficos de Bland y Altman no es apropiado cuando uno de los métodos tiene error de medición despreciable, ya que sus supuestos no se cumplen.[7]

Con respecto al tamaño de la muestra, en estudios de equivalencia, el cálculo del mismo es un aspecto crítico, dado que la potencia estadística depende no solo de la magnitud del efecto real, sino también de los límites del intervalo de equivalencia (Δ) que se definan a priori.[30] A diferencia de los contrastes tradicionales de hipótesis, donde tamaños pequeños pueden llevar a resultados no significativos que a veces se interpretan erróneamente como ausencia de efecto, en pruebas de equivalencia un margen más estrecho exige muestras considerablemente mayores para alcanzar potencia adecuada. Por ejemplo, simulaciones recientes han mostrado que reducir el margen de equivalencia en un 25% puede implicar casi duplicar el número de participantes necesarios para mantener el 80% de potencia.[30]

Este artículo busca justamente evitar estos errores comunes. Se definió un margen de equivalencia (±15%) antes del análisis, se aplicaron TOST e intervalos de confianza del 90%, y se incorporaron métricas complementarias de error grupal e individual. Además, se evitó el uso de correlaciones, t-tests o gráficos de Bland-Altman, que, aunque populares, no responden adecuadamente a la pregunta de equivalencia.

Mediante este ejemplo se evidencia que validar un instrumento no se reduce a aplicar técnicas estadísticas convencionales, sino que exige comprender los fundamentos de los estudios de equivalencia, definir márgenes clínicamente relevantes y reportar con transparencia. Promover la adopción de guías metodológicas claras es esencial para mejorar la calidad, la reproducibilidad y el impacto de la evidencia producida en este campo.[7,8,24]

Conclusiones

Este artículo destaca la importancia de utilizar enfoques estadísticos apropiados en estudios de validación de instrumentos que cuantifican variables continuas frente a un método de referencia. Por medio de un ejemplo aplicado, se demuestra que el uso de pruebas de equivalencia, como el método TOST y el análisis por intervalos de confianza dentro de márgenes predefinidos, permiten establecer si un nuevo instrumento puede considerarse clínicamente intercambiable con el estándar de referencia.

Además, la incorporación de métricas complementarias como el error medio, el MAPE y la RMSE proporciona una visión más completa del desempeño del instrumento, tanto a nivel grupal como individual. Este enfoque supera las limitaciones de los métodos comúnmente utilizados, como los coeficientes de correlación o los gráficos de Bland y Altman, que, si bien pueden describir asociación o sesgo, no permiten evaluar equivalencia clínica.

Es importante que los investigadores definan a priori los márgenes de equivalencia, seleccionen métodos estadísticos adecuados y reporten sus resultados con transparencia. La adopción de este enfoque metodológico riguroso no solo mejora la calidad científica de las publicaciones, sino que también favorece decisiones clínicas más seguras y basadas en evidencia.

Declaración de conflicto de interés: las autoras han completado y enviado la forma traducida al español de la declaración de conflictos potenciales de interés del Comité Internacional de Editores de Revistas Médicas, y no fue reportado alguno relacionado con este artículo.

Referencias

1. Shei RJ, Holder IG, Oumsang AS, et al. Wearable activity-trackers-advancedtechnologyoradvanced marketing?Eur J Appl Physiol. 2022;122(9):1975-90.doi: 10.1007/s00421-022-04951-1
2. Dixon PM, Saint-Maurice PF, Kim Y, et al. A Primer on the Use of Equivalence Testing for Evaluating Measurement Agreement. Med Sci Sports Exerc. 2018;50(4):837-45. doi: 10.1249/MSS.0000000000001481
3. Giurgiu M, von Haaren-Mack B, Fiedler J, et al. The wearable landscape: Issues pertaining to the validation of the measurement of 24-h physical activity, sedentary, and sleep behavior assessment. J Sport Health Sci. 2024;14:101006. doi: 10.1016/j.jshs.2024.101006
4. Dell'Era S, Gimeno-Santos E, Chain NAF, et al. Exactitud del Xiaomi Mi Band 4 para contabilizar pasos en adultos con enfermedades respiratorias crónicas. Estudio de concordancia. Respirar. 2024;16(2):101-12. doi: 10.55720/respirar.16.2.1
5. Kim J, Kenyon J, Billingsley H, et al. Validity of the Actigraph-GT9X accelerometer for measuring steps and energy expenditures in heart failure patients. PLoS One. 2024;19(12): e0315575. doi: 10.1371/journal.pone.0315575
6. Hibbing PR, Pilla M, Birmingham L, et al. Evaluation of the Garmin Vivofit 4 for assessing sleep in youth experiencing sleep disturbances. Digit Health. 2024. doi: 10.1177/20552076241277150
7. Taffé P, Zuppinger C, Burger GM, et al. The Bland-Altman method should not be used when one of the two measurement methods has negligible measurement errors. PLoS One. 2022;17(12):e0278915. doi: 10.1371/journal.pone.0278915
8. Welk GJ, Bai Y, Lee JM, et al. Standardizing Analytic Methods and Reporting in Activity Monitor Validation Studies. Med Sci Sports Exerc. 2019;51(8):1767-80. doi: 10.1249/MSS.0000000000001966
9. Ialongo C. The logic of equivalence testing and its use in laboratory medicine. Biochem Med (Zagreb). 2017;27(1):5-13. doi: 10.11613/BM.2017.001
10. Mayorga-Vega D, Casado-Robles C, Guijarro-Romero S, et al. Criterion-Related Validity of Consumer-Wearable Activity Trackers for Estimating Steps in Primary School children under Controlled Conditions: Fit-PersonStudy. J Sports Sci Med. 2024;23(1):79-96. doi: 10.52082/jssm.2024.79
11. Casado-Robles C, Mayorga-Vega D, Guijarro-Romero S, et al. Validity of the Xiaomi Mi Band 2, 3, 4 and 5 Wristbands for Assessing Physical Activity in 12-to-18-Year-Old Adolescents under Unstructured Free-Living Conditions. Fit-Person Study. J Sports Sci Med. 2023;22(2):196-211. doi: 10.52082/jssm.2023.196
12. Hao Y, Ma XK, Zhu Z, et al. Validity of Wrist-Wearable Activity Devices for Estimating Physical Activity in Adolescents: Comparative Study. JMIR Mhealth Uhealth. 2021;9(1):e18320. doi: 10.2196/18320
13. Ummels D, Bijnens W, Aarts J, et al. The Validation of a Pocket Worn Activity Tracker for Step Count and Physical Behavior in Older Adults during Simulated Activities of Daily Living. Geron-

tol Geriatr Med. 2020;6:2333721420951732. doi: 10.1177/2333721420951732
14. Kwon S, Wan N, Burns RD, et al. The Validity of Motion Sense HRV in Estimating Sedentary Behavior and Physical Activity under Free-Living and Simulated Activity Settings. Sensors (Basel). 2021;21(4). doi: 10.3390/s21041411
15. Viciana J, Casado-Robles C, Guijarro-Romero S, et al. Are Wrist-Worn Activity Trackers and Mobile Applications Valid for Assessing Physical Activity in High School Students? Wearfit Study. J Sports Sci Med. 2022;21(3):356-75. doi: 10.3390/s21041411
16. Silva JC, Silva KF, Torres VB, et al. Reliability and validity of My Jump 2 app to measure the vertical jump in visually impaired five-a-side soccer athletes. Peer J. 2024;12:e18170. doi: 10.7717/peerj.18170
17. Matlary RED, Holme PA, Glosli H, et al. Comparisonof free-living physical activity measurements between ActiGraph GT3X-BT and Fitbit Charge 3 in young people with haemophilia. Haemophilia. 2022;28(6):e172-80. doi: 10.1111/hae.14624
18. Sullivan K, Metoyer CJ, Hornikel B, et al. Agreement Between A 2-Dimensional Digital Image-Based 3-Compartment Body Composition Model and Dual Energy X-Ray Absorptiometry for The Estimation of Relative Adiposity. J Clin Densitom. 2022;25(2):244-51. doi: 10.1016/j.jocd.2021.08.004
19. Majmudar MD, Chandra S, Yakkala K, et al. Smartphone camera based assessment of adiposity: a validation study. NPJ Digit Med. 2022;5(1):79. doi: 10.1038/s41746-022-00628-3
20. Shinozaki K, Yu PJ, Zhou Q, et al. An Automation System Equivalent to the Douglas Bag Technique Enables Continuous and Repeat Metabolic Measurements in Patients Undergoing Mechanical Ventilation. Clin Ther. 2022;44(11):1471-9. doi: 10.1016/j.clinthera.2022.09.004
21. Correa-Rojas J. Coeficiente de correlación intraclase: aplicaciones para estimar la estabilidad temporal de un instrumento de medida. Cienc Psicol. 2021;15(2):e1220. doi: 10.22235/cp.v15i2.2318
22. Nazaroff J, Mark B, Learned J, et al. Measurement of acetabular wall indices: comparison between CT and plain radiography. J Hip Preserv Surg. 2021;8(1):51-7. doi: 10.1093/jhps/hnab008
23. Villa G, Cerfoglio S, Bonfiglio A, et al. Validation of a Commercially Available IMU-Based System Against an Optoelectronic System for Full-Body Motor Tasks. Sensors (Basel). 2025;25(12):3736. doi: 10.3390/s25123736
24. Johnston W, Judice PB, Molina García P, et al. Recommendations for determining the validity of consumer wearable and smartphone step count: expert statement and checklist of the INTERLIVE network. Br J Sports Med. 2021;55(14):780-93. doi: 10.1136/bjsports-2020-103147
25. Courtney JB, Nuss K, Lyden K, et al. Comparing the activPAL software's Primary Time in Bed Algorithm against Self-Report and van derBerg's Algorithm. Meas Phys Educ Exerc Sci. 2021;25(3):212-26. doi: 10.1080/1091367x.2020.1867146
26. Tinsley GM, Park KS, Saenz C, et al. Deuterium oxide validation of bioimpedance total body water estimates in Hispanic adults. Front Nutr. 2023;10:1221774. doi: 10.3389/fnut.2023.1221774
27. McCarthy C, Tinsley GM, Yang S, et al. Smartphone prediction of skeletal muscle mass: model development and validation in adults. Am J Clin Nutr. 2023;117(4):794-801. doi: 10.1016/j.ajcnut.2023.02.003
28. Katz MJ, Wang C, Nester CO, et al. T-MoCA: A valid phone screen for cognitive impairment in diverse community samples. Alzheimers Dement (Amst). 2021;13(1):e12144. doi: 10.1002/dad2.12144
29. Cheng X, Liu J, Wang Y, et al. Comparison of Students' Physical Activity at Different Times and Establishment of a Regression Model for Smart Fitness Trackers. Sensors (Basel). 2025;25(6). doi: 10.3390/s25061726
30. Gutierrez NM, Cribbie R. Effect Sizes for Equivalence Testing: Incorporating the Equivalence Interval. Methods in Psychology. 2022;9:100127. doi: 10.31234/osf.io/5buz9