Confiabilidad, precisión o reproducibilidad de las mediciones. Métodos de valoración, utilidad y aplicaciones en la práctica clínica

Carlos Manterola[1,2], Luis Grande[3,4,5], Tamara Otzen[1,2,6], Nayely García[1], Paulina Salazar[1] y Guissela Quiroz[1]

[1]Centro de Estudios Morfológicos y Quirúrgicos (CEMyQ), Universidad de La Frontera. Temuco, Chile.
[2]Programa de Doctorado en Ciencias Médicas, Universidad de La Frontera. Temuco, Chile.
[3]Departamento de Cirugía, Universidad Autònoma de Barcelona, España.
[4]Servicio de Cirugía, Hospital del Mar, Barcelona, España.
[5]Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), Barcelona, España.
[6]Facultad de Ciencias de la Salud, Universidad de Tarapacá, Chile.

Financiamiento parcial por Proyecto MEC 80170022, CONICYT.
Conflictos de interés: inexistentes.

Recibido: 23 de julio de 2018
Aceptado: 20 de noviembre de 2018

Correspondencia a:
Carlos Manterola Delgado
carlos.manterola@ufrontera.cl

Reliability, precision or reproducibility of the measurements. Methods of assessment, utility and applications in clinical practice

Reliability (accuracy, consistency and reproducibility) is a psychometric property, which is related to the absence of measurement error, or, to the degree of consistency and stability of the scores obtained through successive measurement processes with the same instrument. Thus a greater variability of results will lower the accuracy or reliability of instrument used, fact that is transverse from the laboratory to the clinical practice. It is determined by applying the reliability coefficient, which is the correlation between the scores obtained by the subjects in two parallel forms of a test. Assuming that the two forms of the test are parallel (measure the same), the scores of the subjects under study should be the same in both applications. In this way, when the correlation is 1, the reliability or precision is maximum. On the other hand, reliability could be influenced by the observer (the one that measures), the measuring instrument (by that with which it is measured), and by the observed (by what is measured). Therefore, the variability of each of these components must be taken into account when planning the measurement of the variable under study, in such a way to reduce measurement biases as much as possible. The most common ways to determine reliability are the models of parallel forms, test-retest and two halves. This manuscript focuses on the concepts of measurement and the various statistical techniques used for this, as a step prior to application in the clinic. Therefore, the aim of this manuscript is to generate a consultation document related to the reliability or reproducibility of the measurement process.

Keywords: Reproducibility of results [Mesh]; reliability; data-analysis; kappa coefficient; statistics; research.
Palabras clave: Confiabilidad; reproducibilidad; análisis de datos; coeficiente kappa.

Introducción

La confiabilidad, reproducibilidad o consistencia de las mediciones es un principio fundamental de la precisión de un estudio. En cualquier proceso de investigación, ante la gran cantidad de fuentes de potenciales errores, es necesario que los investigadores intenten reducir aquellos relacionados con la medición de las variables para proporcionar una mayor confianza en los resultados y las conclusiones de su estudio[1,2].

Dicho de otra forma, un instrumento es confiable, preciso o reproducible, cuando las mediciones realizadas con él, generan los mismos resultados en diferentes momentos, escenarios y poblaciones si se aplica en las mismas condiciones; sin embargo, se ha de asumir que en la práctica clínica cotidiana, la confiabilidad se conjuga con otro concepto que es la validez, dando origen a diversos escenarios, desde mediciones válidas y confiables hasta aquellas que carecen de validez y de confiabilidad, como en el caso de observaciones u observadores que están de acuerdo sólo por el efecto del azar (Figura 1). Por ende, a mayor precisión de una medición, mayor poder estadístico existirá en la muestra en estudio[3].

La precisión en las mediciones está influenciada por el que mide (observador), por aquello con lo que se mide (instrumento de medición); y por lo que es medido, o lo que es observado, ya sea sujeto o variable/s en estudio. Para evitar errores aleatorios y reducir al máximo los potenciales sesgos de medición, el investigador deberá tener en cuenta la variabilidad de todos y cada uno de estos componentes en el momento de planificar la medición o mediciones de la/s variable/s de un estudio y conocer la confiabilidad del propio observador (intra observador o *intra rater*) o entre observadores (inter observador), si fueran varios –de ahí la necesidad de un adecuada formación y entrenamiento de los observadores– o del instrumento de medición (intra instrumento), o entre instrumentos (inter instrumento), si se utilizasen varios –suponiendo la correcta calibración de los mismos–[2,4,5].

El objetivo de este manuscrito es generar un documento de estudio y consulta relacionado con la confiabilidad, reproducibilidad o precisión del proceso de medición.

Estadísticas utilizadas para medir confiabilidad

Para evaluar la confiabilidad se pueden utilizar una serie de fórmulas[4], las que para su cálculo, incluyen la variabilidad del sujeto y el error de medición (Figura 2). Entre ellas cabe destacar:

- el porcentaje de acuerdo, grado de concordancia o porcentaje de concordancia[6,7];
- el índice κ (kappa) de Cohen, para dos observadores[7];
- el método de Bland y Altman, para evaluar el grado de acuerdo entre dos instrumentos que tienen la misma escala de medición[8];
- el índice *k* de Fleiss cuando existen tres o más observadores[9];
- el coeficiente de correlación o r de Pearson;
- el coeficiente de correlación interclase, para instrumentos que evalúan variables continuas[10];
- el coeficiente de correlación o ρ (rho) de Spearman, para dos variables aleatorias continuas[11]; y
- el α (alfa) de Krippendorff's, para múltiples observadores y múltiples observaciones posibles[12].

Porcentaje de acuerdo

El concepto de "acuerdo entre observadores" es simple, y durante muchos años la confiabilidad entre evaluadores se midió como el porcentaje de acuerdo entre ellos. Para ello, se crea una matriz en la que las columnas representan a los observadores y las filas a las variables que se están midiendo. Las celdas entonces contienen las puntuaciones de cada uno de los observadores para cada variable. Se calcula el porcentaje de acuerdo para cada fila y se promedian las filas. El escenario es obvio cuando las respuestas son idénticas, especialmente en respuestas dicotómicas y cuando los observadores son dos[6,13], pero complicada, cuando hay más observadores o las opciones de respuesta no están limitadas a sólo dos valores. Una ventaja adicional de esta técnica es que permite identificar aquellas variables que pudieran ser problemáticas.

Índice κ de Cohen

Computo muy sólido para determinar la confiabilidad inter e intra observador. Se trata de una forma de coeficiente de correlación y, como todos ellos, puede variar de -1 a +1, donde 1 representa la concordancia perfecta entre los observadores y 0 el acuerdo que se puede esperar de forma aleatoria. Desde el punto de vista matemático, se pueden obtener valores kappa inferiores a 0 pero, en la práctica, es muy poco probable. Un valor de $k = 0$ refleja que la concordancia observada es precisamente la que se espera a causa del azar[14]. Se calcula según la siguiente formula:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

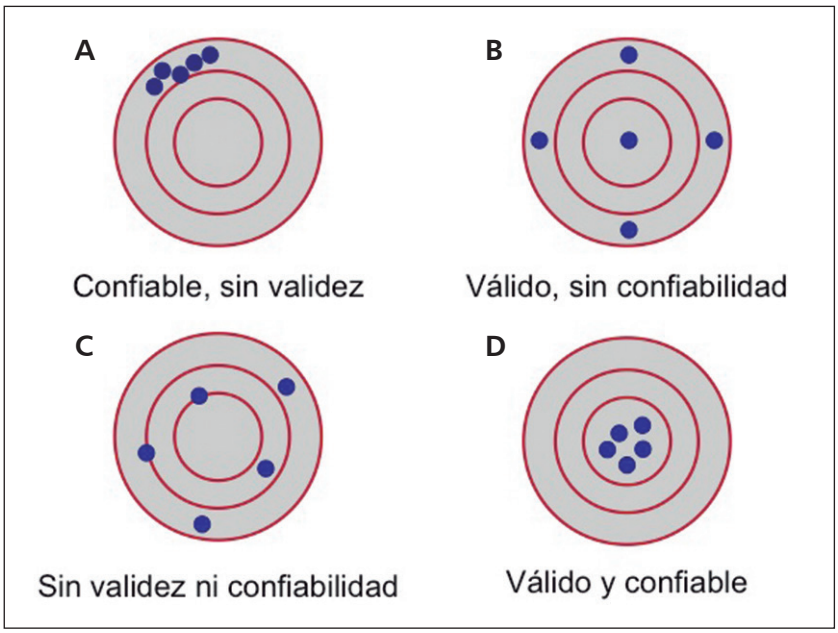

Figura 1. Posibles escenarios de validez y confiabilidad. **A)** Todas las mediciones son parecidas, pero se alejan de la realidad. **B)** Las mediciones captan todo el espectro del fenómeno, pero son muy diferentes entre ellas. **C)** No se capta todo el fenómeno y las mediciones son muy disímiles entre ellas. **D)** Todas las mediciones son parecidas y se ajustan a la realidad de lo que se está midiendo.

$$C = \frac{\text{Variabilidad del sujeto}}{\text{Variabilidad del sujeto + Error de medición}} \qquad C = \frac{\sigma^2}{\sigma^2_s + \sigma^2_e}$$

Figura 2. Fórmula de estimación de confiabilidad.

en la que Pr (a) es la proporción de concordancia observada entre observadores y Pr (e) la proporción esperada por azar. Existe, además, una categorización de los valores de kappa, que se utilizan para ordenar y definir el grado de acuerdo[1,15] (Tabla 1), aunque es muy utilizada, no está universalmente aceptada, porque es obvio que determinados valores que pudieran ser aceptables en

Tabla 1. Interpretación para los valores de confiabilidad. Categorización de los valores de kappa, que se utilizan para ordenar y definir el grado de acuerdo entre dos observadores

Valores	Interpretación
< 0,01	No acuerdo
0,01 - 0,20	Ninguna a escaso
0,21 - 0,40	Regular o razonable
0,41 - 0,60	Moderado
0,61 - 0,80	Substancial
0,81 - 1,00	Casi perfecto

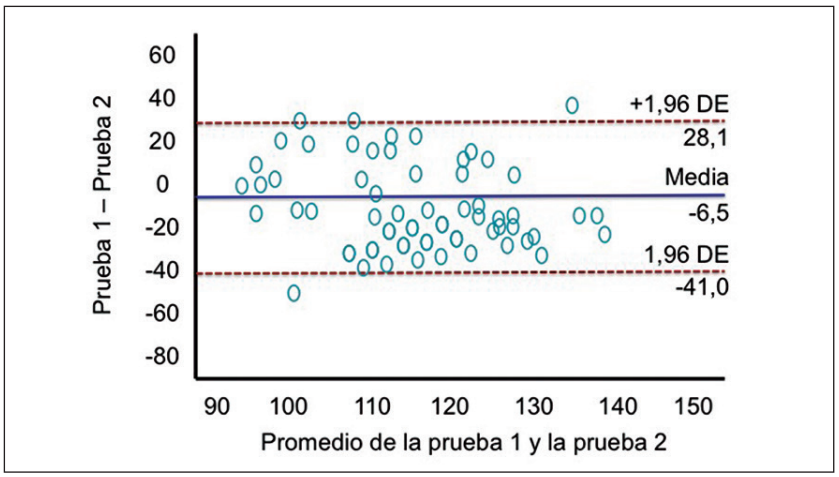

Figura 3. Gráfico de Bland y Altman. En este ejemplo, se puede apreciar que, si bien existen diferencias entre las puntuaciones obtenidas con dos instrumentos de medición, éstas se distribuyen en relación a la media más menos 2 desviaciones estándar, lo que resulta aceptable para este tipo de situaciones.

determinados ámbitos, por ejemplo, un índice de 0,45 –acuerdo moderado–, representarían un grave problema de salud en otros, por ejemplo, en los controles de calidad de un laboratorio. Muchos textos sugieren un valor mínimo de 80% para un acuerdo inter observadores aceptable.

Para evaluar el significado del índice kappa es muy interesante disponer de sus intervalos de confianza (IC). No es raro que el intervalo de confianza (IC) sea tan amplio que incluye cualquier valor, desde un acuerdo bueno a uno pobre[1]. Las fórmulas para su cálculo serían:

$$\kappa - 1{,}96 \times SE\kappa \qquad \text{y} \qquad \kappa + 1{,}96 \times SE\kappa$$

en las que 1,96 sería el factor necesario para lograr el IC al 95% y SEκ el error estándar del índice kappa.

Método de Bland y Altman

Tal como se ha dicho anteriormente este método permite evaluar el grado de acuerdo entre dos instrumentos que tienen la misma escala de medición. Para ello, se representa de forma gráfica las diferencias entre dos mediciones frente a su media[8], de forma similar a los gráficos media-diferencia de Tukey. También se puede usar para comparar una nueva técnica o método de medición con un estándar de referencia (Figura 3).

Índice k de Fleiss

A diferencia de otras medidas estadísticas de confiabilidad inter observador, como la π de Scott, la κ de Cohen y la J de Youden, pensadas para dos observadores o para un observador versus sí mismo, el índice κ de Fleiss es el más apropiado cuando se valoran la confiabilidad del acuerdo entre tres o más observadores que utilicen variables categóricas.

Su cálculo es similar al utilizado para la κ de Cohen, pero obviamente añade el cálculo del sesgo del observador (precisión-error) y el cálculo de la concordancia (calibración). La fórmula seria:

$$K = 1 - \frac{nm^2 - \sum_{i=1}^{n}\sum_{j=1}^{r} x^2ij}{nm\,(m-1)\sum_{j=1}^{r} pj\;qj}$$

En la que n: se corresponde con el número total de conductas o códigos a registrar; m: identifica el número de codificaciones; xij: define el número de registros de la conducta i en la categoría j; r: indica el número de categorías de que se compone el sistema nominal; p: es la proporción de acuerdos positivos entre codificadores; q: es la proporción de acuerdos negativos (no acuerdos) en codificadores (1-p).

El valor resultante también variará entre -1 y 1, donde 1 representa la concordancia perfecta entre los observadores y 0 el grado de acuerdo que se puede esperar del azar.

Una ventaja adicional es que el índice κ de Fleiss se puede aplicar a núcleos de observadores diferentes mientras su número sea idéntico; por ejemplo, si la variable 1 fue valorada por los observadores A, B y C, y la variable 2, por los observadores D, E y F. Una desventaja es que no existe una versión para variables categóricas ordenadas y sólo se puede utilizar con clasificaciones binarias o de escala nominal.

Coeficiente de correlación de Pearson

Es una estadística inferencial que refleja la intensidad de la asociación lineal entre dos variables cuantitativas: Debe hacerse hincapié en el término "asociación lineal", porque pueden existir variables fuertemente relacionadas, pero no de forma lineal -por ejemplo, exponencial-, en las que no procede aplicar la correlación de Pearson, extremo que parece olvidarse con cierta frecuencia.

Su fórmula es:

$$p = \frac{\sum_{i=1}^{n} (xi - x)\,(yi - y)}{(n-1)\,s_x\,s_y}$$

en la que *x* y *s*, son la media y la desviación estándar de la primera variable, la *y* y la *s,* la media y la desviación estándar de la segunda variable y n la longitud de la columna. Un inconveniente de la correlación de Pearson es que no proporciona información adecuada sobre el

Tabla 2. Comparación de las características principales de concordancia estadísticas más utilizadas para determinar confiabilidad

Prueba estadística	Aplicación	Tipo de variables	Permite:
Porcentaje de acuerdo	Confiabilidad entre 2 observadores	Dicotómicas	Calcular porcentaje de acuerdo para cada fila de observaciones
Índice Kappa de Cohen	Confiabilidad inter e intra observador	Dicotómicas	Determinar concordancia entre observadores
Método de Bland y Altman	Confiabilidad entre dos instrumentos que tienen la misma escala de medición	Continuas	Representar de forma gráfica las diferencias entre dos mediciones frente a su media
Índice κ de Fleiss	Confiabilidad inter observador (tres o más observadores)	Dicotómicas y nominales	Determinar concordancia entre observadores
CC de Pearson	Intensidad de asociación lineal entre dos variables	Continuas	Evaluar los rangos entre valores de dos variables
CIC	Confiabilidad de una variable en sujetos con un grado fijo de relación (hermanos)	Continuas	Evaluar el comportamiento de variables en grupos de estudio
CC de rango (ρ de Spearman)	Comparación de rangos de cada grupo de sujetos	Continuas	Evaluar los rangos entre valores de dos variables, sin precisar que la relación sea lineal
Alpha de Krippendorff	Confiabilidad inter observador	Dicotómicas, nominales y ordinales	Determinar concordancia entre cualquier número de observadores, y en casos en que hay datos incompletos

CC: Coeficiente de correlación. CIC: Coeficiente de correlación intraclase.

acuerdo producido al ignorar la diferencia. Para paliar en parte este problema, el coeficiente de Pearson se puede combinar aplicando una t de Student con datos pareados para comparar los promedios de las dos distribuciones y cuantificar la magnitud del sesgo general producido entre las observaciones[16]. No obstante, se ha de considerar que el t-test tampoco proporciona información sobre el grado de acuerdo obtenido entre las dos medidas para cada uno de los individuos[17].

Coeficiente de correlación intraclase (CIC)

Se trata de una estadística inferencial que evalúa el comportamiento de variables continuas en grupos de estudio, es decir, describe cómo las unidades en el mismo grupo se parecen fuertemente entre sí. Aunque se considera un tipo de correlación, a diferencia de la mayoría de las otras medidas de correlación, opera con datos estructurados en grupos, en lugar de datos estructurados como observaciones emparejadas[10].

El concepto básico subyacente al CIC fue introducido por Fisher como una formulación especial de la r de Pearson bajo condiciones de igualdad de promedios y variancias de las distribuciones involucradas. Se basa en un modelo de análisis de la variancia (ANOVA) con medidas repetidas; razón por la que se puede definir como la proporción de variabilidad total debida a la variabilidad de los sujetos en estudio[10].

Si bien se utiliza habitualmente para cuantificar el grado de confiabilidad de una variable en sujetos con un grado fijo de relación (por ejemplo, hermanos), también se puede aplicar para evaluar la reproducibilidad de las mediciones cuantitativas realizadas por diferentes observadores que miden la misma cantidad e incluso en mediciones emparejadas (Figura 4).

Coeficiente de correlación de rango o ρ de Spearman

Al igual que la correlación de Pearson, evalúa los rangos entre valores de dos variables, pero en este caso no precisa que la relación sea lineal. La correlación de Spearman evalúa cualquier relación monótona (incluso las lineales). Si no hay valores de datos repetidos, se produce una correlación Spearman perfecta de +1 ó -1 cuando cada

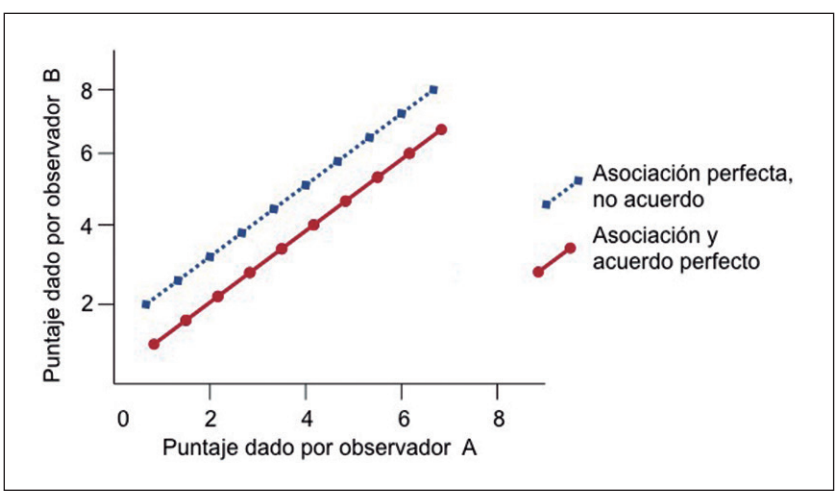

Figura 4. Gráfica explicativa del coeficiente de correlación de Pearson o CIC. La línea discontinua representa una asociación perfecta con poco acuerdo entre los observadores; y la línea continua, representa una asociación y acuerdo perfecto entre observadores.

una de las variables es una función monótona perfecta de la otra[11,18]. Su formulación es:

$$r_s = 1 - \frac{6 \sum d^2}{n\,(n^2 - 1)}$$

en la que d es la diferencia entre los rangos (x - y) y n el número de datos. Es apropiado para variables ordinales, continuas o discretas y tanto la ρ de Spearman como el τ (tau) de Kendall pueden utilizarse como un coeficiente de correlación más general[18].

Alpha de Krippendorff

Es una alternativa al κ de Cohen, para determinar confiabilidad inter observador[19], con la ventaja de que puede aplicarse a cualquier número de observadores, con datos incompletos (no son necesarias todas las parejas de datos), a cualquier número de valores disponibles para codificar una variable, sean esta binaria, nominal, ordinal, interválica, proporcional, polar o métrica circular (niveles de medición) e incluso se ajusta a pequeños tamaños de muestra[12].

Muy utilizada en el ámbito de las encuestas, entrevistas y pruebas psicológicas donde se deben comparar alternativas de los mismos fenómenos, o en estudios observacionales donde se registran sucesos no estructurados[12].

Su formulación es:

$$\alpha = 1 - \frac{D_o}{D_e}$$

donde D_0 es el desacuerdo observado y D_e el desacuerdo esperado por el azar. Como en otras fórmulas, los resultados posibles van de 0 a 1, donde 0 es un desacuerdo total y 1 el acuerdo perfecto. Según Krippendorff, un resultado de 0,667 sería el límite inferior admisible y un resultado ≥ 0,800 aceptable[19].

Estrategias para mejorar la confiabilidad

Ya hemos insistido en la necesidad de evitar las fuentes potenciales de error en cualquier proceso de investigación y se ha hecho especial énfasis en la confiabilidad, reproducibilidad o consistencia de las mediciones como principio fundamental de la precisión de un estudio. En este momento se hará un repaso a las estrategias que permiten mejorar la confiabilidad. Entre ellas, destacan la estandarización de los métodos de medición, el entrenamiento del o de los observadores, el mantenimiento en buen estado de los instrumentos de medición, la automatización de las mediciones y la repetición de las mediciones (promedios)[2].

Estandarizar los métodos de medición

Es una medida de la precisión de datos obtenidos por un solo operador trabajando siempre en las mismas condiciones (equipos, materiales y reactivos). También representa la precisión de datos obtenidos entre dos o más observadores o laboratorios que utilizan el mismo método y similares condiciones.

Entrenar al o a los observadores

La formación del o los observadores corresponde a un proceso a través del cual se adquiere madurez conceptual, empírica y tecnológica para realizar una observación en un estudio observacional. Este proceso debe ser de progresión lógica y dificultad ascendente; e implica una fase de entrenamiento, otra de adiestramiento y finalmente una evaluación del proceso[20].

Mantener los instrumentos de medición en buen estado

Esto implica mantener los instrumentos de medición calibrados (según especificaciones y calendario establecidos en el plan de mantenimiento metrológico), verificados (comparación con un patrón certificado) y en buen estado físico e higiénico.

Automatizar instrumentos y proceso de medición

Se refiere a los procesos analíticos que son realizados por equipos, con la menor participación posible de un operador. Esta estrategia permite, además, un mejor control de cada una de las operaciones implicadas, un ahorro de tiempo y una disminución de gasto de reactivos. Un ejemplo típico de esta estrategia es la que se desarrolla en los laboratorios clínicos.

Repetir las mediciones y calcular promedios de éstas

Para reducir el impacto que pudieran tener los diferentes intervalos en los que se obtienen los resultados de mediciones sucesivas de una misma magnitud o las condiciones de medición (observador, instrumento de medición, procedimiento de medición, lugar y repetición dentro de un período de tiempo corto), es una buena estrategia trabajar con promedios, desviaciones estándar y eventualmente calcular rangos. Además, esta práctica, permite validar los métodos de calibración, realizar análisis de comparación de mediciones por diferentes laboratorios (inter laboratorios) y estudiar la variabilidad de las mediciones e instrumentos.

Tipos de confiabilidad y ejemplos

Del observador

Intra observador

La confiabilidad intra observador es aquella que ocurre cuando un observador individual efectúa mediciones repetidas en un grupo de sujetos.

Ejemplos:

- Se determinó confiabilidad intra observador de un método posturográfico en 34 pacientes con neuritis vestibular. Los coeficientes de CIC para todos los parámetros y posiciones de prueba (ALL (media)) variaron entre 0,71 (IC 95%: 0,41-0,85) y 0,92 (IC 95%: 0,84-0,96). La confiabilidad absoluta (coeficiente de variación) osciló entre 3,1% (IC del 95%: 2,60-8,67) y 42,3% (IC del 95%: 40,7-74,5). Concluyéndose que el sistema posturográfico estudiado tiene buena confiabilidad intra observador absoluta relativa y satisfactoria para ALL (media)[21].
- Se evaluó la confiabilidad intra e inter-evaluador de la cuantificación de las tasas de sangre y LCR mediante RM de contraste de fase. Para ello, se utilizaron escáneres 3T de diferentes fabricantes en dos centros, estudiando las imágenes con contraste de fase de cine codificado por velocidad de los flujos de sangre y LCR en la región cervical superior. Se analizaron los datos de seis sujetos escaneados en el centro A y de cinco sujetos en el centro B por seis observadores de dos niveles diferentes de entrenamiento. Cada conjunto de datos se analizó tres veces en un orden aleatorio para un total de 33 conjuntos de datos. Se calcularon CIC, el que varió de 0,80 a 0,96 para el área de luz y de 0,97 a 0,99 para la tasa de flujo volumétrico. El ICC para medidas secundarias derivadas varió de 0,85 a 0,99. Concluyéndose que se puede lograr una alta confiabilidad intra e inter-observador de las mediciones de la tasa de flujo volumétrico entre fabricantes y niveles de entrenamiento de los observadores[22].
- Se analizó la confiabilidad intra observador e inter observador en radiografías simples de 81 pacientes, con un intervalo de cuatro años entre las observaciones; para estudiar el grado de calcificación de la aorta abdominal aplicando la escala AAC-24, para lo cual, se aplicó CIC. Se observó una CIC intra observador de 0,93 (IC 95%: 0,6-0,9); y una CIC inter observador de 0,91 (IC 95%: 0,8-0,9). Estos valores permiten concluir que el AAC-24 es un método confiable para determinar calcificación de la aorta abdominal[23]; es decir, que varios clínicos al examinar las mismas radiografías (o un clínico la misma radiografía en más de una oportunidad), tendrán un alto nivel de acuerdo en la existencia de un hallazgo particular.
- Se evaluó la confiabilidad inter observador e intra observador de los sistemas de clasificación Letournel, Sanders y Zwipp en la tomografía computarizada y la medición del ángulo de Bohler en radiología simple. Para ello, tres observadores (dos radiólogos y un cirujano ortopédico) evaluaron dos veces la tomografía y las radiografías de 51 fracturas intra articulares de calcáneo en un intervalo de 5 meses. La confiabilidad inter observador se midió utilizando el kappa de Fleiss mientras que la intra observador se midió utilizando el kappa de Cohen. Se verificaron valores medios de kappa para confiabilidad inter observador e intra observador de la clasificación de Sanders de 0,25 y 0,39 respectivamente; de la clasificación de Zwipp de 0,24 y 0,16 respectivamente; y de la de Letournel de 0,50 y 0,42, respectivamente. Concluyéndose entonces, que debe existir consciencia de la limitada confiabilidad de los sistemas de clasificación estudiados[24].

Inter-observador

La confiabilidad ínter observador es aquella que ocurre cuando dos o más observadores evalúan independientemente a un mismo sujeto de estudio.

Ejemplos:

- Se determinó validez y confiabilidad de la escala de evaluación de esfuerzo percibido pediátrico EPInfant en niños chilenos; para lo cual, se seleccionaron 31 niños de 8-12 años de edad; a los que se midió frecuencia cardíaca, esfuerzo percibido y carga de trabajo durante dos pruebas de Chester consecutivas, realizadas con un intervalo de una semana. Se estimó coeficiente r de Pearson y de CIC. Se verificó correlación entre esfuerzo percibido y criterios de referencia; y alta CIC, con una adecuada confiabilidad interobservador[25].
- Se evaluó la reproducibilidad inter observador de tres sistemas de puntuación validados en la evaluación de las úlceras del pie diabético (PEDIS [Perfusión, Extensión, Profundidad, Infección, Sensación], SINBAD [Sitio, Isquemia, Neuropatía, Infección bacteriana, Profundidad] y el de la Universidad de Texas [UT]); para lo cual, se estudiaron 37 pacientes, que fueron evaluados por un grupo de 12 observadores, midiéndose CIC. La confiabilidad para sistemas únicos fue baja para todos los sistemas (UT 0,53; SINBAD 0,44; PEDIS 0,23-0,42), pero la confiabilidad múltiple fue casi perfecta (UT 0,94; SINBAD 0,91; PEDIS 0,80-0,90). El mejor acuerdo para múltiples observadores infección: 0,83; isquemia: 0,80-0,82; ambos: 0,8. Esto significa que estos sistemas pueden ser utilizados de manera adecuada por múltiples observadores al realizar investigaciones y auditorías; no así pasa en el entorno clínico, al documentar características de una úlcera o al comunicarse entre colegas[26].
- Se evaluó la confiabilidad entre centros para en la interpretación de frotis vaginales teñidos con Gram de mujeres embarazadas. La confiabilidad de los morfotipos individuales identificados en el frotis vaginal se evaluó comparándolos con los obtenidos en un centro estándar. Se propuso un nuevo sistema de puntuación que utiliza los morfotipos más habituales del frotis vaginal; el que se comparó con los criterios de Spiegel de vaginitis bacteriana. El sistema de

puntuación (0 a 10), es una combinación ponderada de los morfotipos: lactobacilos, *Gardnerella vaginalis* o bacteroides, y varillas curvas Gram variables. Los valores de correlación de Spearman para variabilidad del intercentro fue: 0,23 para cocos Gram positivos; 0,65 para lactobacilos, 0,69 para *G. vaginalis* y 0,57 para bacteroides (concordancia moderada). Por otra parte, varillas pequeñas y varillas curvas Gram variables, obtuvieron una correlación de 0,74 y 0,85 respectivamente. Sin embargo, al no considerar las cocáceas grampositivas y combinando los morfotipos de *G. vaginalis* y bacteroides; la confiabilidad se incrementó. Así, la confiabilidad entre centros (r = 0,82), fue superior a la aplicación de los criterios de Spiegel (r = 0,61); lo que El facilita la investigación sobre vaginitis bacteriana porque proporciona graduaciones de la alteración de la microbiota vaginal que puede estar asociada con diferentes niveles de riesgo de complicaciones del embarazo[27].

Del instrumento

Intra-instrumento

La confiabilidad intra instrumento es aquella que ocurre cuando se efectúan mediciones repetidas en un grupo de sujetos con un mismo instrumento de medición.

Ejemplos:

- Se utilizó un modelo animal (tibias de cinco ovejas), para estudiar la reproducibilidad y precisión de ciertos biomarcadores de calidad de la imagen ósea con resonancia magnética en comparación con el estándar de referencia de la microtomografía computada. Se verificó que la cuantificación morfométrica y mecánica del hueso trabecular por RM fue muy reproducible, con variación inferior a 9% para todos los parámetros; confirmando que la medición de los biomarcadores estudiados con ambas técnicas son reproducibles[28].
- Se estudió la reproducibilidad de las mediciones del grosor de la capa de fibras nerviosas retinianas peripapilares, mediante la tomografía de coherencia óptica del dominio espectral en 45 sujetos normales y 33 con glaucoma. Las mediciones se repitieron tres veces durante la misma visita utilizando la desviación estándar dentro de la materia, el coeficiente de variación y el coeficiente de CIC. Éstos oscilaron entre 0,977 y 0,990 en ojos normales; y entre 0,983 y 0,997 en ojos con glaucoma; demostrando una excelente reproducibilidad para este tipo de mediciones en ojos sanos y con glaucoma con el instrumento utilizado[29].

Inter-instrumento

La confiabilidad ínter instrumento es aquella que ocurre cuando se utilizan diferentes instrumentos para realizar mediciones repetidas en un grupo de sujetos.

Ejemplos:

- Se analizó la confiabilidad del análisis inmunohistoquímico del microambiente tumoral del linfoma folicular en siete equipos de patólogos pertenecientes al Consorcio de Biomarcadores de Linfomas de Lunenburg, comparando citometría de flujo versus la puntuación manual o automática de microscopía de tinciones inmuno-histoquímicas de densidades de infiltración de células T CD3, CD4 y CD8. El grado de acuerdo entre puntuación manual y citometría fue moderado para CD3, bajo para CD4 y moderado a alto para CD8. Por otra parte, el acuerdo en la cuantificación manual en los siete laboratorios fue bajo a moderado para las frecuencias CD3, CD4 y CD8. Finalmente, la puntuación manual de la distribución arquitectónica resultó en acuerdo moderado para CD3, CD4 y CD8. Datos que sugieren realizar mediciones más objetivas en pacientes con linfoma folicular, como puntuación asistida por computadora[30].
- Se estimó la reproducibilidad de la tomografía confocal, la polarimetría láser de barrido y la tomografía de coherencia óptica para determinar el grosor de la capa de fibras ganglionares. Para ello, se estudiaron dos veces, un total de 75 ojos normales, aplicando coeficiente de Pearson para analizar la correlación entre las variables. Se constató que existe una amplia variación tanto en la medición inter observador como inter instrumentos, para la determinación del grosor de las capas de las fibras nerviosas[31].

Test re-test

Consiste en administrar una escala o instrumento de medición (test) dos veces a los mismos sujetos. La administración puede ser inmediata o con un intervalo de tiempo entre el test y el re-test. Posterior a ello, se calcula la correlación de Pearson entre las puntuaciones obtenidas en ambas aplicaciones, el resultado de lo cual, será el coeficiente de confiabilidad, que representa la estabilidad en el tiempo del instrumento. Cuando existen diferencias, se atribuyen a la consistencia interna o de muestreo de los ítems que componen la escala o instrumento de medición. No se recomienda dejar pasar demasiado tiempo entre una aplicación y la otra, como tampoco que la segunda aplicación sea muy cercana a la primera; pues en ambos casos, se puede alterar la validez interna.

Variaciones del test re-test, son el método de formas alternativas o paralelas, en el que se administran dos o más versiones equivalentes al instrumento de medición, utilizándose el coeficiente de correlación producto-momento de Pearson; y, el método de mitades partidas, en el que se requiere de una sola aplicación del instrumento de medición, pero los ítems que lo componen se dividen en dos partes y luego se comparan los resultados, para lo cual se utilizan las pruebas de Pearson y Spearman-Brown (Figura 5).

Ejemplos:

- Se analizó el comportamiento de un cuestionario de cinco ítems, auto administrado sobre vulnerabilidad a infecciones, en sujetos pertenecientes a la cohorte nacional alemana, obtenidos de un centro de Hamburgo y otro de Hannover (tiempo entre aplicaciones fue de ocho días). Se calcularon consistencia interna, porcentaje de acuerdo y kappa. La consistencia interna resultó en un α de Cronbach de 0,78. El kappa de los ítems sobre frecuencia de enfermedades infecciosas en los últimos 12 meses fue de 0,65 a 0,87; y de los ítems sobre síntomas en los últimos 12 meses, entre 0,77 y 0,90. Esto, lo hace ser prometedor y moderadamente confiable, pero requiere evaluaciones adicionales[32].

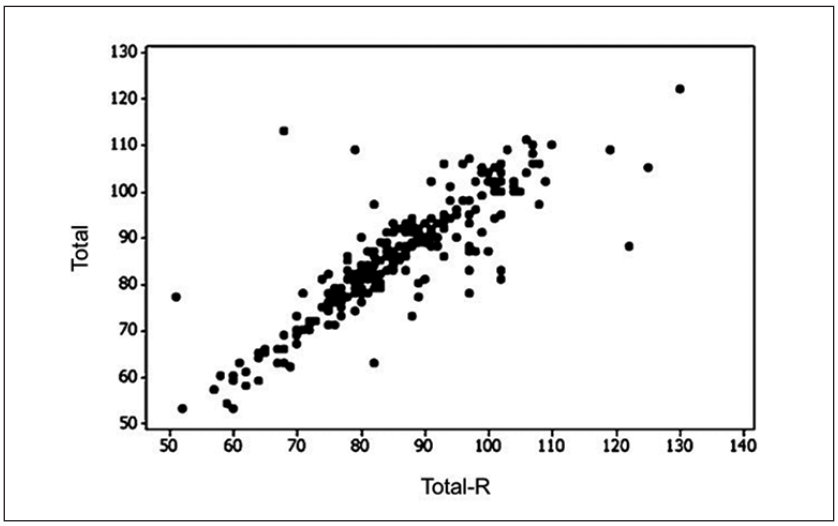

Figura 5. Diagrama de dispersión para un conjunto de datos en los que se aplica test y retest. Se aprecia una relación lineal positiva entre las dos mediciones realizadas (con el mismo instrumento de medición) al mismo grupo de sujetos.

Finalmente, y a modo de conclusión, podemos señalar que en este manuscrito hemos discutido el concepto de confiabilidad, también denominado precisión, consistencia o reproducibilidad. Hemos destacado el hecho de que la confiabilidad no sólo depende de los errores de medición, sino también de la heterogeneidad de los valores de la variables en estudio en la población en la que se realizan las mediciones; así como en los procesos relacionados con la medición, los que incluyen entre otras, las condiciones en que se encuentra el (los) instrumento (s) de medición. Sugerimos que, como las mediciones se realizan en distintos escenarios (estudios clínicos, pruebas diagnósticas, de tamizaje, etc.); y en diferentes poblaciones; se deberían informar las desviaciones estándar entre sujetos y dentro de los sujetos. También describimos los métodos utilizados para estudiar mediciones realizadas por diferentes observadores o evaluadores, y diferentes instrumentos de medición. Para finalizar con ejemplos obtenidos de la literatura científica, sobre la aplicación de algunas de estas metodologías para resolver los problemas de la investigación en diversos escenarios.

Resumen

En investigación, la confiabilidad (precisión, consistencia y reproducibilidad), corresponde a una propiedad psicométrica que dice relación con la ausencia de error de la medición; o del grado de consistencia y estabilidad de las puntuaciones obtenidas a lo largo de sucesivos procesos de medición con un mismo instrumento. Por ello, es esperable que, a mayor variabilidad de resultados, menor sea la precisión del instrumento de medición utilizado, concepto que aplica desde el laboratorio a la práctica clínica. La reproducibilidad se determina aplicando el coeficiente de confiabilidad, que es la correlación entre las puntuaciones obtenidas por los sujetos, en dos formas paralelas de una prueba (porque se supone que miden lo mismo). Por ello, asumiendo que midan lo mismo, las puntuaciones de los sujetos en estudio deberían ser iguales en ambas aplicaciones. De este modo, cuando la correlación es 1, la confiabilidad, precisión o reproducibilidad es máxima; y mientras más cercana a 0 es peor. La precisión en las mediciones está influenciada por el que mide (observador), por aquello con lo que se mide (instrumento de medición), y por lo que es medido (lo observado). Por ende, se ha de tomar en cuenta la variabilidad de cada uno de estos componentes al momento de planificar la medición de la variable en estudio; de tal modo de reducir al máximo los sesgos de medición. Así, las formas más comunes de determinar confiabilidad son: modelos de formas paralelas, test-retest y de dos mitades. Este manuscrito se centra en los conceptos de medición y las diversas técnicas estadísticas utilizadas para ello, como paso previo a la aplicación en la clínica. Por ello, el objetivo de este manuscrito es generar un documento de estudio y consulta relacionado con la confiabilidad, reproducibilidad o precisión del proceso de medición.

Referencias bibliográficas

1.- McHugh M L. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012; 22(3): 276-82. PMID: 23092060.

2.- Atmanspacher H, Bezzola Lambert L, Folkers G, Schubiger P A. Relevance relations for the concept of reproducibility. J R Soc Interface. 2014; 11(94): 20131030. doi: [10.1098/rsif.2013.1030].

3.- Goodman S N, Fanelli D, Ioannidis J P. What does research reproducibility mean? Sci Transl Med 2016; 8 (341): 341ps12. DOI: 10.1126/scitranslmed.aaf5027.

4.- Stemler S E. A Comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. practical research, assessment and evaluation. 2004. Disponible en t http://pareonline.net/getvn.asp?v=9&n=4 (acceso el 29 de marzo de 2018).

5.- Ubersax J. Kappa Coefficients. Statistical Methods for Diagnostic Agreement 2010 update. Disponible en http://john-uebersax.com/stat/kappa2.htm (acceso el 11 de abril de 2018).

6.- Landis J R, Koch G G. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159-74. DOI: 10.2307/2529310.

7.- Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960; 20: 37-46. https://doi.org/10.1177%2F001316446002000104.

8.- Bland J M, Altman D G. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1 (8476): 307-10. PMID: 2868172.

9.- Fleiss J L, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational Psychological Measurement 1973; 33: 613-9. DOI: 10.1177/001316.

10.- Hazra A, Gogtay N. Biostatistics Series Module 6: Correlation and linear regression. Indian J Dermatol 2016; 61 (6): 593-601. DOI: 10.4103/0019-5154.193662.

11.- Mondragón M A. Uso de la correlación de Spearman en un estudio de intervención en fisioterapia. Mov Cient 2014; 8 (1): 98-104. https://revistas.iberoamericana.edu.co/index.php/Rmcientifico/article/viewFile/739/645

12.- Gwet K L. On The Krippendorff's Alpha Coefficient. Disponible en http://www.agreestat.com/research_papers/onkrippendorffalpha.pdf (acceso el 12 de mayo de 2018).

13.- Bangdiwala S I, Shankar V. The agreement chart. BMC Med Res Methodol 2013; 13: 97. https://doi.org/10.1186/1471-2288-13-97.

14.- Marusteri M, Bacarea V. Comparing groups for statistical differences: how to choose the right statistical test? Biochem Med 2010; 20: 15-32. https://hrcak.srce.hr/47847.

15.- Cerda J, Villarroel L. Evaluación de la concordancia inter-observador en investigación pediátrica: Coeficiente de Kappa. Rev Chil Pediatr 2008; 79 (1): 54-8. https://scielo.conicyt.cl/pdf/rcp/v79n1/art08.pdf.

16.- Manterola C, Otzen T. Los sesgos en investigación clínica. Int J Morphol. 2015; 33 (3): 1156-64. https://scielo.conicyt.cl/pdf/ijmorphol/v33n3/art56.pdf.

17.- Bartko J J. General methodology II. Measures of agreement: a single procedure. Stat Med 1994; 13: 737-45. PMID: 8023046.

18.- de Vet H C W, Mokkink L B, Mosmuller D G, Terwee C B. Spearman-Brown prophecy formula and Cronbach's alpha: different faces of reliability and opportunities for new applications. J Clin Epidemiol 2017; 85: 45-9. doi: 10.1016/j.jclinepi.2017.01.013.

19.- Krippendorff K. Computing Krippendorff's alpha-reliability." Philadelphia: Annenberg School for Communication Departmental Papers, 2011. Disponible en: http://repository.upenn.edu/cgi/viewcontent.cgi?article=1043&context=asc_papers (acceso el 11 de mayo de 2018).

20.- Anguera M T. La observación. En C. Moreno Rosset (ed.), Evaluación psicológica. Concepto, proceso y aplicación en las áreas del desarrollo y de la inteligencia (pp. 271-308). Madrid: Sanz y Torres. 2003.

21.- Schwesig R, Fischer D, Becker S, Lauenroth A. Intraobserver reliability of posturography in patients with vestibular neuritis. Somatosens Mot Res 2014; 31 (1): 28-34. https://doi.org/10.3109/08990220.2013.822364.

22.- Koerte I, Haberl C, Schmidt M, Pomschar A, Lee S, Rapp P, et al. Inter- and intra-rater reliability of blood and cerebrospinal fluid flow quantification by phase-contrast MRI. J Magn Reson Imaging 2013; 38 (3): 655-62. doi: [10.1002/jmri.24013].

23.- Pariente-Rodrigo E, Sgaramella G A, García-Velasco P, Hernández-Hernández J L, Landeras-Alvaro R, Olmos-Martínez J M. Reliability of radiologic evaluation of abdominal aortic calcification using the 24-point scale. Radiologia 2016; 58 (1): 46-54. http://dx.doi.org/10.1016/j.rx.2015.03.002.

24.- Sayed-Noor A S, Agren P H, Wretenberg P. Interobserver reliability and intraobserver reproducibility of three radiological classification systems for intra-articular calcaneal fractures. Foot Ankle Int 2011; 32 (9): 861-6. DOI: 10.3113/FAI.2011.0861.

25.- Rodríguez-Núñez I, Manterola C. Concurrent validity and interobserver reliability of the EPInfant pediatric perceived exertion rating scale among healthy Chilean children. Arch Argent Pediatr 2016; 114 (4). DOI: 10.5546/aap.2016.eng.343.

26.- Forsythe R O, Ozdemir B A, Chemla E S, Jones K G, Hinchliffe R J. Interobserver reliability of three validated scoring systems in the assessment of diabetic foot ulcers. Int J Low Extrem Wounds 2016; 15 (3): 213-9. DOI: 10.1177/1534734616654567.

27.- Nugent R P, Krohn M A, Hillier S L. Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. J Clin Microbiol 1991; 29 (2): 297-301. https://jcm.asm.org/content/jcm/29/2/297.full.pdf.

28.- Alberich-Bayarri A, Martí-Bonmatí L, Sanz-Requena R, Sánchez-González J, Hervás Briz V, García-Martí G, et al. Reproducibility and accuracy in the morphometric and mechanical quantification of trabecular bone from 3 Tesla magnetic resonance images. Radiologia 2014; 56 (1): 27-34. http://dx.doi.org/10.1016/j.rx.2013.06.001.

29.- Wu H, de Boer J F, Chen T C. Reproducibility of retinal nerve fiber layer thickness measurements using spectral domain optical coherence tomography. J Glaucoma. 2011; 20 (8): 470-6. doi: 10.1097/IJG.0b013e3181f3eb64.

30.- Sander B, de Jong D, Rosenwald A, Xie W, Balagué O, Calaminici M, et al. The reliability of immunohistochemical analysis of the tumor microenvironment in follicular lymphoma: a validation study from the Lunenburg Lymphoma Biomarker Consortium. Haematologica 2014; 99 (4): 715-25. doi: [10.3324/haematol.2013.095257].

31.- Sánchez-García M, Rodríguez de la Vega R, González-Hernández M, González de la Rosa M. Variability and reproducibility of 3 methods for measuring the thickness of the nerve fiber layer. Arch Soc Esp Oftalmol 2013; 88 (10): 393-7. doi: 10.1016/j.oftal.2013.01.012.

32.- Castell S, Akmatov M K, Obi N, Flesh-Janys D, Nieters A, Kemmling Y, et al. Test-retest reliability of an infectious disease questionnaire and evaluation of self-assessed vulnerability to infections: findings of Pretest 2 of the German National Cohort. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2014; 57 (11): 1300-7. doi: 10.1007/s00103-014-2045-x.