Artículo de Revisión

R F S
REVISTA FACULTAD DE SALUD
http://journalusco.edu.co/index.php/RFS

R.F.S Revista Facultad de Salud
Enero - Junio de 2016;8(1): 42-47

Reproducibilidad de las mediciones clínicas

Reproducibility of clinical measurements

Claudia Marcela Rojas[1], Jorge Puerta[1], Julio Gomez[1], José Andrés Calvache[2*]

Resumen

En la investigación clínica existe un gran interés por garantizar la calidad de cualquier medición como base en la toma de decisiones. La calidad de una medición depende de dos propiedades: validez y fiabilidad. Mientras la validez expresa el grado en el que realmente se mide el fenómeno de interés, la fiabilidad de una medición es evaluada calculando la reproducibilidad de múltiples mediciones.

Este ensayo aborda los conceptos de validez y fiabilidad de una medición como base de la interpretación clínica y epidemiológica de la misma. Adicionalmente discute conceptos de acuerdo, concordancia y correlación estadística y presenta un breve resumen de los métodos disponibles para la evaluación de la reproducibilidad tanto numéricos como gráficos.

Palabras clave: Validez, fiabilidad, reproducibilidad, acuerdo, concordancia.

Abstract

In the clinical setting as in research there is great interest in ensuring the quality of measurings as a basis for decision making. The quality of a measurement depends of two properties: The validity and reliability. While the validity expresses the degree of the measurement of the phenomenon of interest, the accuracy of a measurement is evaluated calculating the reproducibility of multiple measurements.

This paper discusses the concepts of validity and reliability of a measurement and its clinical and epidemiological interpretation. Additionally discusses concepts of statistical agreement, concordance and correlation and presents a brief summary of the available methods for assessing the reproducibility both numerical and graphical.

Key words: Validity, reliability, reproducibility, agreement, concordance

1 Universidad del Cauca, Facultad Ciencias de la Salud, Médicos residentes del programa de Anestesiología.
2 Universidad del Cauca, Facultad Ciencias de la Salud, Profesor Departamento de Anestesiología. Erasmus University Medical Centre, Department of Anesthesiology, Rotterdam, The Netherlands.
* Correspondencia: MD. José Andrés Calvache. Correo electrónico: jacalvache@unicauca.edu.co

Recibido: 10/12/2015 - Revisado: 18/01/2016 - Aceptado: 30/03/2016

Introducción

El proceso de medición está presente de forma universal en la práctica médica diaria y siempre tiene sujeto un grado de incertidumbre. Si bien existen variables relativamente fáciles de medir (por ejemplo, la estatura de una persona en centímetros) otras representan un reto (por ejemplo, el grado de dolor, o la satisfacción con la atención en salud) desde el punto de vista objetivo.

Cabe señalar que, tanto en el ámbito de investigación como en el escenario clínico, existe un gran interés por garantizar la calidad de dichas mediciones como base en la futura toma de decisiones. La "calidad" de una medición depende de dos propiedades: i) la validez y ii) la fiabilidad. Mientras que la validez expresa el grado en el que realmente se mide el fenómeno de interés (proximidad con la verdad o carencia de sesgo), la fiabilidad indica hasta que´ punto se obtienen los mismos valores al efectuar la medición en más de una ocasión, bajo condiciones similares [1]. Algunos autores utilizan sinónimos para expresar la validez y la fiabilidad de una medición. Para clarificar, estos conceptos se presentan en la Tabla 1.

Estas dos propiedades (validez y fiabilidad) no necesariamente están relacionadas. Si el monitor utilizado para medir la presión sistólica -por un desperfecto- ofrece un resultado 10 mmHg por encima del valor real, su dato no será válido, sin embargo, si es medido en múltiples ocasiones su valor será reproducible o fiable (siempre estará 10 mmHg por encima del real pero muy cerca uno a otro). Afortunadamente, ciertas estrategias utilizadas para mejorar la validez también mejoran la fiabilidad (o precisión) de las estimaciones. Para profundizar en estas estrategias de mejoramiento, muy útiles en el proceso de planificación de una investigación, recomendamos consultar a Hulley et al [2].

Fuentes de variabilidad en una medición

Es de esperar que en la medición de una variable de interés clínico, y por tanto útil para la toma de decisiones, la variabilidad sea baja y que cada medición sea lo más próxima a la verdad. Sin embargo, existen fuentes de variabilidad en las mediciones que introducen diversos grados de error en la misma [3, 4]. Estos son:

a) La variabilidad de los observadores
b) La variabilidad propia del instrumento de medida
c) La variabilidad debida a medir en momentos diferentes en el tiempo

Existen factores conocidos o predecibles que pueden afectar la confiabilidad de una medición. Muchas fuentes de error pueden ser minimizadas mediante una adecuada

Tabla 1. Diferencias entre validez y fiabilidad de una medición*

	Validez	Fiabilidad
Sinónimos en español Sinónimos en ingles	Exactitud *Accuracy, validity*	Fiabilidad**. *Reliability***. Precisión. *Precision.* Reproducibilidad. *Reproducibility.*
Significado	El grado en el cual una variable o medición representa o mide lo que realmente pretende medir	El grado en que una variable o medición tiene aproximadamente el mismo valor cuando se mide en múltiples ocasiones en condiciones similares
Como cuantificarla	Comparación con un estándar de referencia (*gold estándar*)	Comparación entre mediciones repetidas
Afectada por	Errores sistemáticos (sesgos) a partir de: - Observador - Sujeto Instrumento	Errores aleatorio a partir de: - Observador - Sujeto Instrumento

* Tomado y modificado sin permiso de Hulley et al. [2]

** Detalles de sus diferencias en el texto.

planeación, entrenamiento y definiciones operacionales claras e inspección de los equipos. No obstante, aun tomando las precauciones necesarias, la respuesta humana variable y los ambientes impredecibles e incluso inmodificables, constituyen una parte inevitable de la medición. Muchos instrumentos (en particular los mecánicos) están sujetos a algún grado de "ruido" de fondo y fluctuación aleatoria en su desempeño, así como cambios con el paso del tiempo. Las mediciones de los evaluadores y evaluados pueden influenciarse por características personales variables, tales como motivación, cooperación, fatiga, entrenamiento y certificación de los observadores y factores ambientales tales como el ruido y la temperatura [2].

El reto más grande a la fiabilidad lo constituye la medición de una respuesta de naturaleza inherentemente inestable (por ejemplo, la presión arterial que al medirse se debe esperar una variabilidad de medición a medición). Cuando una respuesta es exageradamente inestable, no existe una medición única capaz de representarla de manera precisa. Es importante que profesionales de la salud entiendan la naturaleza teórica y práctica de las variables de respuesta para así poder interpretar de manera apropiada las fuentes de error al valorar la confiabilidad de las mediciones [4].

Para evaluar la validez de una medición, los estudios comparan su resultado con un estándar de medición (*gold* estándar) [1, 2, 4]. En la práctica epidemiológica se constituyen en un diseño de investigación denominado test diagnóstico. De acuerdo al nivel de medición de la variable en estudio sus resultados se expresan de forma diferente, por ejemplo, cuando consideramos variables dicotómicas (como la clasificación por un test diagnóstico como sano o enfermo) sus resultados se expresan en términos de sensibilidad y especificidad [2].

Por otra parte, la evaluación de la reproducibilidad se enfoca desde una perspectiva diferente. A continuación entraremos en detalles sobre cómo evaluar apropiadamente la fiabilidad (o precisión) y su utilidad en la práctica clínica e investigativa.

Evaluación de la fiabilidad

Diversos autores describen que para evaluar la fiabilidad se pueden usar dos tipos de medidas: las de fiabilidad (*reliability*) y las de acuerdo (*agreement*). Frecuentemente estos dos tipos de medidas son utilizadas indistintamente, sin embargo, hay que tener en cuenta algunas consideraciones que las diferencian. [5]

Fiabilidad es una característica que cuantifica la relación existente entre la variabilidad de los mismos sujetos (diferentes evaluadores o diferentes tiempos) y la variabilidad total de la muestra en estudio. De esta forma, esta medida cuantifica la capacidad de una medición de distinguir "entre sujetos" a pesar del error de medición.

$$\text{Fiabilidad} = \frac{\text{Variación entre participantes}}{\text{Variación entre participantes} + \text{Error de medición}}$$

Para que una medida sea fiable, el error de medición debe ser relativamente pequeño en relación a la variabilidad entre participantes, y así éstos pueden ser apropiadamente distinguidos y diferenciados. Si los resultados de una medición son muy alejados uno de otro, existiendo una gran variación entre los participantes, la habilidad de la medida para distinguir entre ellos no será afectada de forma importante por el error de medición (el cual será pequeño). Sin embargo, si los resultados de la medición son muy cercanos entre los participantes (baja variabilidad) el error de medición influirá de forma importante en la capacidad de distinguir los sujetos y la fiabilidad será baja.

En contraste a la fiabilidad, el "acuerdo" no presta mayor atención a la variación intra o entre sujetos sino que se concentra en el error de medición. Es de esperar que al repetir una medición su resultado sea lo más cercano posible al primer resultado, de tal manera que tengan un "alto acuerdo" o que sean concordantes. En términos generales, la concordancia es el grado en que dos o más observadores, métodos, técnicas u observaciones están de acuerdo sobre el mismo fenómeno observado [4]. Adicionalmente, la concordancia es una propiedad de las mediciones en las cuales sus resultados "acuerdan" o se corresponden entre sí.

En la práctica, confiamos en los resultados de un instrumento al medir la presión arterial sistólica, siempre y cuando, su resultado sea muy cercano al repetir la medición en similares condiciones. De la misma forma, cuando la medición de cierta variable requiere de métodos invasivos y buscamos evaluar un nuevo método no invasivo, buscamos que sus resultados tengan un alto acuerdo o que sean altamente concordantes y así poder reemplazar el método inicial [5].

Correlación o concordancia

Es importante resaltar la diferencia entre los conceptos de correlación y concordancia. La correlación refleja el grado de asociación entre dos grupos de datos o la consistencia de la posición dentro de dos distribuciones. El término concordancia, en cambio, hace referencia a que los valores obtenidos por dos diferentes mediciones sean los mismos o muy cercanos, no sólo proporcionales unas con otras. De esta forma, la concordancia es el grado en que dos o más observadores, métodos, técnicas u observaciones, están de acuerdo sobre el mismo fenómeno observado (acuerdan).

No necesariamente datos de mediciones que se correlacionan bien son concordantes entre sí. Por ejemplo, si buscamos comparar dos instrumentos para medir la presión sistólica, en el caso de un esfingomanometro clásico versus uno digital, se asume que el digital tiene un desperfecto en su funcionamiento y siempre arroja un resultado 15 mmHg por encima del valor real. En esta situación, la correlación (lineal) de los resultados de los dos instrumentos es perfecta y cercana a 1 ($r \tilde{} 1$). Sin embargo, su concordancia es nula puesto ninguno de los resultados de los dos instrumentos acuerdan entre sí.

R.F.S Revista Facultad de Salud
Enero - Junio de 2016;8(1): 42-47

Debido a esto, el coeficiente de correlación de Pearson (r) no es una medida eficiente para evaluar la concordancia entre dos mediciones, y por ende, para una interpretación aplicable a la práctica clínica. Los métodos estadísticos para determinar reproducibilidad deben incluir estimados de concordancia que puedan usarse de manera conjunta con la correlación [4]. Por lo tanto, la concordancia adquiere importancia cuando se desea conocer si con un método o instrumento nuevo, diferente al habitual, se pueden obtener resultados equivalentes de tal manera que eventualmente uno y otro puedan ser reemplazados o intercambiados. Esta decisión puede obedecer a que uno de ellos es más sencillo, menos costoso y, por lo tanto, más costo-efectivo, o porque uno de ellos resulta más seguro para el paciente, entre otras razones.

En síntesis, la concordancia no evalúa la validez o la certeza sobre una u otra observación con relación a un estándar de referencia dado, sino cuán acordes están entre sí las observaciones sobre el mismo fenómeno. La concordancia tiene lugar cuando se desea conocer si con un nuevo método o instrumento de medición, diferente al habitual, se obtienen resultados equivalentes.

Métodos estadísticos para la evaluación de la fiabilidad y reproducibilidad

En un estudio enfocado en la evaluación de la reproducibilidad existen ciertas características que se deben tener en cuenta (Tabla 2).

Estos métodos se pueden clasificar de acuerdo con el tipo de variable analizada, algunos se enfocan predominantemente en la fiabilidad y otros en el acuerdo o concordancia de las mediciones.

Variables cualitativas. En relación a las medidas de concordancia, cuando la unidad de medida se encuentra en la escala categórica, la fiabilidad se puede valorar con la medición de concordancia. El índice más sencillo de concordancia es el "porcentaje de concordancia", que representa la proporción de observaciones en las que existe acuerdo. Cuando se desea además tener en cuenta la proporción de concordancia esperada por el azar, se utiliza el estadístico kappa (k) [4]. Su interpretación general es descrita así: < 0.20 pobre concordancia; 0.21 – 0.40 débil; 0.41 – 0.60 moderada; 0.61 – 0.80 buena; 0.81 – 1.00 muy buena concordancia.

Tabla 2. Aspectos a evaluar en un estudio de reproducibilidad

Reproducibilidad	Indica en qué grado un instrumento proporciona resultados similares cuando se aplica a una misma persona en más de una ocasión, pero en idénticas condiciones.
Acuerdo	Indica el grado en el cual las mediciones son similares o diferentes.
Concordancia intraobservador	Evalúa el grado de acuerdo entre las mediciones realizadas por un mismo observador bajo las mismas condiciones.
Concordancia interobservador	Evalúa el grado de acuerdo entre las mediciones realizadas por observadores distintos sobre una misma medición bajo las mismas condiciones.
Concordancia entre instrumentos diferentes de medición	Cuando existen diferentes métodos de medida para un mismo fenómeno, es interesante estudiar hasta qué punto los resultados obtenidos con ambos instrumentos son equivalentes.

* Tomado y modificado sin permiso de Barton [5].

Variable cuantitativas. El coeficiente de correlación intraclase (CCI) es un índice que fluctúa entre 0,0 y 1,0 y es calculado usando estimaciones derivadas de un análisis de la varianza [6]. También ha sido llamado índice de fiabilidad [7]. Entre más cercano a uno, su resultado representa una mayor concordancia. Existe una escala categórica que sirve para interpretar sus resultados así: CCI < 0.20 pobre concordancia; 0.21 – 0.40 débil; 0.41 – 0.60 moderada; 0.61 – 0.80 buena; 0.81 – 1.00 muy buena concordancia.

El CCI tiene varias ventajas estadísticas: primero, puede ser usado para valorar concordancia entre dos o más puntajes; segundo, no requiere el mismo número de evaluadores para cada medición, facilitando la realización de estudios clínicos. Tercero, aunque el CCI está diseñado primariamente para usar con variables de intervalo/razón, puede ser usado con datos en la escala ordinal cuando se presume que los intervalos entre las mediciones son equivalentes [4, 7].

Un último método, que tiene amplia utilidad clínica, es el análisis gráfico de la concordancia descrito por Bland y Altman [8, 9]. En este, se construye una gráfica que presenta la diferencia entre los métodos en estudio contra el puntaje promedio de cada par de mediciones. En general, si las mediciones son concordantes, la diferencia entre ellas debería ser cero. La variabilidad (o dispersión) de las diferencias alrededor de cero en un rango definido, ayuda a decidir si el error observado en el acuerdo es aceptable en la práctica, y si podríamos "tolerar" una diferencia. Por lo tanto, a este rango de variabilidad de las diferencias se le denomina (límites del acuerdo) y usualmente se construye al 95% (siguiendo los principios de los intervalos de confianza) [4]. (Tabla 3, Figura 1).

Conclusión

Al margen del análisis de la validez, la evaluación de la fiabilidad es útil en la práctica clínica y de investigación puesto que todas las mediciones siempre están sujetas a diversos grados de error. Existen varios métodos para su abordaje útiles en el proceso de investigación y en la interpretación de resultados relacionados.

Tabla 3. Método de evaluación del acuerdo de Bland y Altman

Sujeto	Medición 1*	Medición 2*	Promedio de la medición 1 y 2	Diferencia de las mediciones (M1-M2)
1	9,6	9,7	9,65	-0,1
2	8,5	9,0	8,75	-0,5
3	9,9	9,3	9,60	0,6
4	12,4	11,8	12,10	0,6
5	10,1	7,6	8,85	2,5
6	11,7	10,1	10,90	1,6
7	8,8	8,3	8,55	0,5
8	7,6	8,2	7,90	-0,6
9	**6,1**	**6,4**	**6,25**	**-0,3**
10	9,5	8,5	9,00	1,0
11	7,4	7,7	7,55	-0,3
12	8,5	8,7	8,60	-0,2
13	12,9	10,9	11,90	2,0
14	9,8	9,8	9,80	0,0
15	8,8	8,9	8,85	-0,1

* Las mediciones 1 y 2 representan la medición en milímetros de diámetro de la vía aérea subglótica medida por ecografía por dos evaluadores diferentes.
*Promedio de las diferencias=0,44 Desviación estándar del promedio de las diferencias=0,95.
Límites del acuerdo = -1.41 – 2.30
Fuente: Elaboración de los autores.

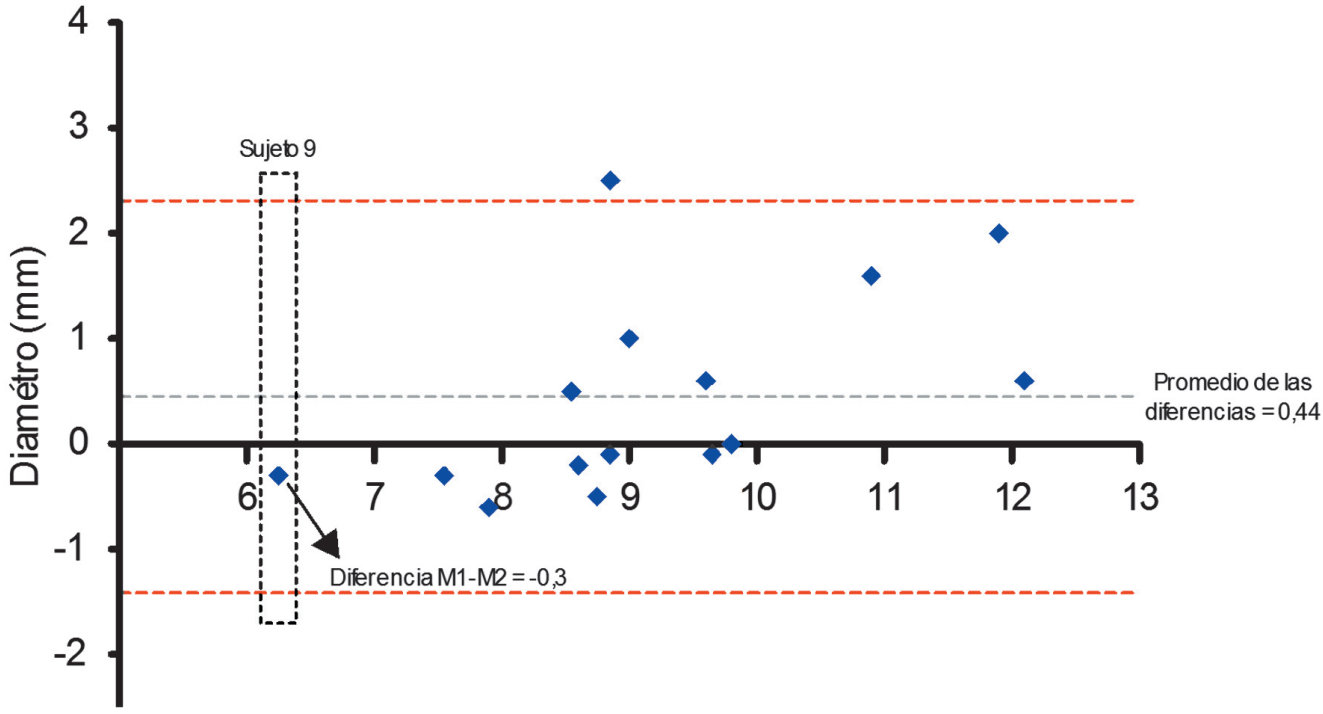

Figura 1. Método gráfico de evaluación del acuerdo de Bland y Altman.

Fuente: Elaboración de los autores.

Referencias

1. PITA S, PÉRTEGA S, RODRÍGUEZ E. La fiabilidad de las mediciones clínicas: El análisis de concordancia para variables numéricas. Cuadernos de atención primaria. 2003;10(4):290–6.
2. HULLEY SB, CUMMINGS SR, BROWNER WS, GRADY DG, NEWMAN TB. Designing Clinical Research. Lippincott Williams & Wilkins; 2013. 378 p.
3. BARTLETT JW, FROST C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound Obstet Gynecol. 2008 Apr;31(4):466–75.
4. CORTÉS É, RUBIO J, GAITÁN H. Statistical methods for evaluating diagnostic test agreement and reproducibility. Rev Colomb Obstet Ginecol. 2010;61(3):247–55.
5. BARTON B PJ. Medical Statistics: A Guide to SPSS, Data Analysis and Critical Appraisal. 2nd ed. 2014. (BJM Books).
6. SHROUT PE, FLEISS JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979 Mar;86(2):420–8.
7. SZKLO M, NIETO F. Epidemiology: Beyond the Basics. 3rd ed. Jones and Bartlett.; 2014.
8. BLAND M, ALTMAN D. Statistical Methods For Assessing Agreement Between Two Methods Of Clinical Measurement. Lancet. 1986;327(8476):307–10.
9. BLAND JM, ALTMAN DG. Measuring agreement in method comparison studies. Stat Methods Med Res. 1999 Jun;8(2):135–60.