Medicina Clínica

La evaluación de la fiabilidad en las observaciones clínicas: el coeficiente de correlación intraclase

Por Luis Prieto a, Rosa Lamarca b, Alfonso Casado b

a Unitat de Recerca en Serveis Sanitaris. Institut Municipal d'Investigació Mèdica (IMIM). Barcelona.Facultat de Psicologia i Ciències de l'Educació Blanquerna. Universitat Ramon Llull. Barcelona

b Unitat de Recerca en Serveis Sanitaris. Institut Municipal d'Investigació Mèdica (IMIM). Barcelona

Med Clin (Barc). 1998;110:142-5.

En el transcurso de la investigación clínica es frecuente, al igual que en otras disciplinas, la evaluación de la fiabilidad de las medidas realizadas, pudiéndose distinguir dos tipos de situaciones diferentes: *a)* aquellas en las que se determina el grado de estabilidad o consistencia conseguido en los resultados cuando se repite la medición con el mismo instrumento en condiciones idénticas, y *b)* aquellas en las que se determina hasta qué punto los resultados obtenidos con diferentes instrumentos de medida o con diferentes observadores concuerdan, o son equivalentes[1].

En la tabla 1 se exponen los resultados de dos mediciones consecutivas de la presión arterial sistólica efectuadas en 10 pacientes, realizadas con el mismo esfigmomanómetro (en mmHg) y por 2 médicos diferentes (datos hipotéticos).

TABLA 1

Comparación de dos mediciones consecutivas de la presión arterial sistólica en 10 pacientes, realizadas con el mismo esfigmomanómetro por dos médicos diferentes

Paciente	Médico A (mmHg)	Médico B (mmHg)	Diferencia (B-A)
1	135	140	5
2	140	145	5
3	130	135	5
4	145	150	5
5	140	145	5
6	150	160	10
7	140	145	5
8	135	140	5
9	140	145	5
10	135	145	10
Media	139	145	6
DE	5,68	6,67	2,11

La última columna de la tabla presenta la diferencia entre ambas mediciones. En este caso, la fiabilidad de la medida puede determinarse valorando hasta qué punto los resultados obtenidos por los médicos A y B concuerdan, son equivalentes.

De los diferentes métodos que existen para valorar la fiabilidad a través de la comparación del acuerdo/desacuerdo producido en diferentes mediciones, los índices de kappa y kappa ponderado han demostrado ser los estadísticos más ventajosos para variables que se expresan cualitativamente (variables nominales y ordinales respectivamente)[1,2]. Cuando los resultados del proceso de evaluación se expresan con mediciones que implican variables de carácter cuantitativo continuo, como es el caso de la tabla 1, es frecuente el uso del coeficiente de correlación producto-momento de Pearson *(r)*[2-4]. Se ha demostrado, sin embargo, que esta

estrategia es incorrecta[1,4] ya que la *r* de Pearson únicamente mide la intensidad de la asociación lineal entre dos variables y no proporciona tampoco información sobre el acuerdo observado. Podemos obtener, por ejemplo, un coeficiente de correlación igual a 1 cuando una medida ofrece resultados dos veces mayores que los de una segunda medida; la gran correlación permite una perfecta predicción de los valores obtenidos con una medida a partir de la otra, sin embargo no existe ningún acuerdo entre los resultados de una y otra. Dada la naturaleza continua de los datos, cabe señalar aquí que si la primera medida ofrece un resultado igual a 75 no esperaremos que la segunda medida ofrezca un valor idéntico, pero el acuerdo será ciertamente mayor si el resultado es 74 o 78 que si es 150. A pesar de que la *r* de Pearson no es un buen indicador de la fiabilidad de las mediciones realizadas, los problemas que involucran su uso ocupan un espacio de cierta importancia en la bibliografía médica clínica[5,6].

Para cuantificar la fiabilidad de las mediciones asociadas a las variables cuantitativas continuas, el índice estadístico que se debe utilizar es el coeficiente de correlación intraclase (CCI)[1-3,7]. Fleiss y Cohen[8] han demostrado que el CCI es matemáticamente equivalente a los índices kappa y kappa ponderado. El CCI actualmente no se puede calcular de forma automática con los paquetes estadísticos más extendidos (SAS, SPSS y BMDP). Su uso en la práctica clínica se ha visto por lo tanto limitado, siendo sustituido por otras medidas no adecuadas pero sí disponibles en dichos paquetes (p. ej., la *r* de Pearson). La aparente dificultad que supone su cálculo manual ha impedido, además, detallar en algunas ocasiones las operaciones necesarias para ello[1].

En el presente artículo ofrecemos una revisión sobre el origen, significado y cálculo del CCI, proporcionando además una formulación notablemente simple que permite su estimación en el caso quizás más habitual: dos mediciones u observaciones por sujeto estudiado.

El coeficiente de correlación de Pearson

Hemos observado que para evaluar la fiabilidad de las mediciones repetidas es común emplear la *r* de Pearson. La *r* para los datos obtenidos por el médico A y el médico B de la tabla 1 es igual a 0,95. Sin embargo, podemos apreciar que la medición del médico B es sistemáticamente mayor que la medición del médico A. Como ya habíamos indicado, el coeficiente de correlación refleja la intensidad de la asociación lineal entre dos variables, muy alta en este caso, pero no proporciona información adecuada sobre el acuerdo producido al ignorar la diferencia sistemática ocurrida. Para paliar en parte este problema, la *r* de Pearson puede combinarse con un test de la t de Student con datos apa reados para comparar las medias de las dos distribuciones y cuantificar la magnitud del sesgo general producido entre los médicos A y B. Sin embargo, el test de la t no proporciona tampoco información sobre el acuerdo obtenido entre las dos medidas para cada uno de los individuos[9,10].

El coeficiente de correlación intraclase

El CCI es una aproximación más adecuada para valorar la concordancia entre las medidas de los médicos A y B. El concepto básico subyacente al CCI fue introducido originariamente por Fisher[11,12] como una formulación especial de la *r* de Pearson bajo condiciones de igualdad de medias y variancias de las distribuciones implicadas. La obtención del CCI que permite evaluar

la concordancia general entre dos o más métodos u observaciones diferentes se basa en un modelo de análisis de la variancia (ANOVA) con medidas repetidas[7], tal como se describe en el anexo. La idea general para el ejemplo que nos ocupa es que la variabilidad total observada en la tabla 1 puede dividirse en tres componentes: *a)* la variabilidad debida a las diferencias entre los pacientes ($*^2_P$); *b)* la variabilidad debida a las diferencias entre los observadores (médicos A y B) ($*^2_O$), y *c)* una variabilidad (residual), inexplicable (aleatoria), asociada al error inherente a toda medición ($*^2_R$).

El CCI se define como la proporción de variabilidad total debida a la variabilidad de los pacientes.

$CCI = *^2_P *^2_P + *^2_O + *^2_R$

Como en el caso de cualquier proporción, los valores del CCI pueden oscilar entre 0 y 1: donde el 0 indica ausencia de concordancia y el 1, la concordancia o fiabilidad absoluta de los resultados obtenidos. Si bien el valor del CCI que define una fiabilidad satisfactoria es arbitrario y depende del uso que de ella se haga, en general, se ha indicado que valores del CCI por debajo del 0,4 representan baja fiabilidad, que valores entre 0,4 y 0,75 representan una fiabilidad entre regular y buena, y que valores por encima de 0,75 representan una fiabilidad excelente[7].

Cabe destacar que la formulación del anexo, que insistimos equivale a la formulación original del CCI, difiere de la presentada por los autores de una referencia comúnmente citada sobre los estadísticos de la concordancia[2], en la que se omite (¿por error?) la n en el denominador de la ecuación del CCI.

Cálculo simple del CCI para dos observaciones

Para el caso más habitual, como el citado ejemplo de la tabla 1, en el que sólo se comparan dos métodos u observadores, se ha desarrollado una fórmula que permite obtener el CCI de una manera más sencilla[10]. La fórmula, detallada en la tabla 2, sólo requiere conocer las desviaciones estándar de la medida A (DE_A) y de la medida B (DE_B), la desviación estándar de la diferencia entre A y B (DE_{B-A}), la media de la diferencia entre A y B (X&SHY;$_{B-A}$) y el número total de pacientes evaluados (n).

TABLA 2

Fórmula simplificada del coeficiente de correlación intraclase (CCI) y cálculo para los datos de la tabla 1

$$CCI = \frac{(DE_A)^2 + (DE_B)^2 - (DE_{B-A})^2}{(DE_A)^2 + (DE_B)^2 + (\bar{X}_{B-A})^2 - (DE_{B-A})^2/n}$$

$$CCI_{tabla\ 1} = \frac{(5{,}68)^2 + (6{,}67)^2 - (2{,}11)^2}{(5{,}68)^2 + (6{,}67)^2 + (6)^2 - (2{,}11)^2/10} = 0{,}64$$

Aplicando la fórmula a los datos del ejemplo de la tabla 1, se obtiene un CCI igual a 0,64 (ta bla 2). Los cálculos derivados a partir del ANOVA para el mismo ejemplo se presentan en el anexo. El valor del CCI obtenido es claramente inferior al valor 0,95 obtenido con la *r* de Pearson, demostrando que la existencia de una alta asociación entre dos variables no es necesariamente signo de una alta concordancia.

Modelo de efectos aleatorios o fijos

La estimación del CCI presentada para el ejemplo de la tabla 1 se ha desarrollado bajo la condición de un *modelo de efectos aleatorios.* Este modelo es apropiado cuando los observadores implicados en la medición, los médicos A y B en el ejemplo, representan una «muestra» aleatoria de la población de posibles observadores (p. ej., otros médicos del hospital) que en el futuro harán uso del instrumento evaluado (el esfigmomanómetro).

Cuando los observadores que intervienen en el estudio son los únicos que participarán en el mismo, es necesario un *modelo de efectos fijos.* En el caso, por ejemplo, en el que deseamos valorar la concordancia de las mediciones de la frecuencia cardíaca obtenidas en 10 pacientes por 2 médicos diferentes: un médico A, jefe del servicio, y un médico B, en período de formación. El objetivo es conocer la magnitud de las discrepancias entre ambos y determinar si el médico A puede ser sustituido por el médico B en esa tarea. Por lo tanto, estaremos exclusivamente interesados en la comparación de los resultados obtenidos por estos 2 médicos y no desearemos generalizar los resultados al conjunto de médicos que miden la frecuencia cardíaca en el hospital. La tabla 3 presenta datos hipotéticos para este ejemplo.

TABLA 3

Comparación de dos determinaciones de la frecuencia cardíaca en 10 pacientes, realizadas por dos médicos diferentes: un médico A, jefe del servicio, y un médico B, en período de formación (modelo de efectos fijos)

Paciente	Médico A (ppm*)	Médico B (ppm)	Diferencia (B-A)
1	75	80	5
2	74	84	10
3	76	81	5
4	79	83	4
5	82	92	10
6	83	88	5
7	85	90	5
8	87	92	5
9	87	92	5
10	88	93	5
Media	81,6	87,5	5,9
DE	5,3	5,04	2,18

*Pulsaciones por minuto.

La tabla 4 ofrece la fórmula simplificada del CCI que nosotros hemos obtenido para la comparación de dos métodos u observadores bajo la condiciones del *modelo de efectos fijos.*

TABLA 4

Fórmula simplificada del coeficiente de correlación intraclase (CCI) para un modelo de efectos fijos y cálculo para los datos de la tabla 3

$$CCI = \frac{(DE_A)^2 + (DE_B)^2 - (DE_{A-A})^2}{(DE_A)^2 + (DE_B)^2 + (\bar{X}_{A-A})^2/2 - (DE_{A-A})^2/2n}$$

$$CCI_{\text{ejemplo}} = \frac{(5,3)^2 + (5,04)^2 - (2,18)^2}{(5,3)^2 + (5,04)^2 + (5,9)^2/2 - (2,18)^2/2 \times 10} = 0,69$$

A pesar de que la *r* de Pearson indica una alta asociación (*r* = 0,96), aplicando la fórmula a los datos de la tabla 3, se obtiene un CCI igual a 0,69, lo que indica una concordancia moderada entre los resultados obtenidos por los médicos A y B. Los cálculos derivados a partir del ANOVA para este ejemplo se encuentran en el anexo.

Los lectores interesados pueden disponer también de las fórmulas necesarias para estimar el intervalo de confianza del CCI[7,11,14].

Limitaciones del CCI

La restricción más importante al uso del CCI es que se trata de una prueba paramétrica y cuando utilizamos pruebas paramétricas debemos considerar las posibles violaciones de las asunciones subyacentes, en nuestro caso normalidad de las distribuciones de las variables, igualdad de variancias e independencia entre los errores producidos por los observadores.

Otra restricción importante del CCI, al igual que cualquier otro coeficiente de correlación, es que es dependiente de la variabilidad de los valores observados. Si los pacientes varían poco en sus puntuaciones (muestra homogénea), el CCI tiende a ser bajo ya que compara la variancia entre pacientes con la variancia total observada, que incluye la variancia de los pacientes, la variancia de los métodos u observadores y el error aleatorio. Si la muestra es heterogénea, el CCI tiende a ser mayor. Una ilustración de esta limitación se puede ver en la diferente concordancia interobservador que suele haber entre la presión sistólica y la diastólica; dado que la presión sistólica permite observar un rango más amplio de valores, el CCI suele ser mayor que para la diastólica, aparentemente indicando que es más difícil de evaluar que la sistólica, si bien se trata de un artefacto estadístico.

Como en muchos otros índices estadísticos, no hay tampoco que olvidar que los resultados obtenidos mediante el CCI están expresados en términos absolutos. Diferencias sistemáticas de cinco unidades para todas las observaciones proporcionarán el mismo CCI tanto si realizamos las mediciones en metros como si las realizamos en kilómetros. Por tanto, el investigador debe tener siempre en cuenta la significación clínica que las diferencias observadas tienen para su estudio.

Alternativas al CCI

Bland y Altman propusieron un método gráfico alternativo sencillo, aunque subjetivo, para evaluar la concordancia de dos métodos de medida, de modo que el resultado no sea dependiente de la naturaleza del grupo escogido para el estudio[4]. Bland y Altman representan gráficamente la diferencia de los pares de valores observados frente a su valor medio y definen unos «límites de acuerdo» combinando la media y la desviación estándar de las diferencias (d y s, respectivamente) como $d \pm 2\,s$. Este tipo de gráfico permite una identificación de las diferencias extremas, así como una valoración de la tendencia mediante un análisis de regresión lineal; éste puede ser fácilmente explicado a personas no iniciadas en estadística y permite una valoración de la significación clínica de los resultados obtenidos. Sin embargo, la estimación del grado de acuerdo observado sigue siendo totalmente subjetiva, sólo se aplica a pares de comparaciones y no proporciona un índice objetivo como el CCI que puede ser especialmente necesario en el caso de comparar más de dos métodos.

Si bien existen otras alternativas para determinar la concordancia de diferentes métodos de medida, éstas han sido menos difundidas en el campo de la medicina. El lector interesado puede encontrar información más detallada en otras referencias[11,13].

Conclusión

Al igual que otros autores[1,2,3,10], recomendamos el uso del CCI para cuantificar la fiabilidad de las mediciones clínicas, ya sea repitiendo la medición con el mismo instrumento en las mismas condiciones[15], o bien determinando la concordancia de las valoraciones de diferentes

instrumentos u observadores en las mismas condiciones. Los resultados ofrecidos por la *r* de Pearson para estos propósitos pueden resultar equívocos, tal como demuestran los ejemplos ofrecidos en este trabajo.

El CCI proporciona, sin embargo, un único índice que facilita dicha estimación. Las ecuaciones para el cálculo del CCI se basan en el análisis de la variancia (ANOVA), pero su obtención directa no es posible con los paquetes estadísticos habituales.

Se propone una fórmula simplificada, para el caso de dos métodos u observadores, haciendo el CCI accesible para cualquier investigador. Tanto la fórmula para el modelo de efectos aleatorios, previamente enunciada[10], como la fórmula para el modelo de efectos fijos, obtenida por nosotros, derivan de la fórmula original propuesta por Bartko[9], eliminando los errores contenidos en revisiones previas del CCI[2].

Agradecimiento

Queremos agradecer los valiosos comentarios del Dr. Jordi Alonso a las versiones previas de este trabajo.

Anexo

Obtención del CCI a partir de análisis de la variancia (ANOVA)

En los estudios que evalúan la concordancia, cada sujeto es examinado por más de un observador, siendo adecuado realizar un análisis de la variancia (ANOVA) para medidas repetidas (generalización de la prueba de la t de Student para datos apareados). Este tipo de diseño permite aislar la variabilidad entre pacientes y concentrarnos en la variabilidad debida al factor de interés (fig. 1); en el análisis de la variancia simple, la variabilidad entre pacientes está incluida en la variabilidad residual.

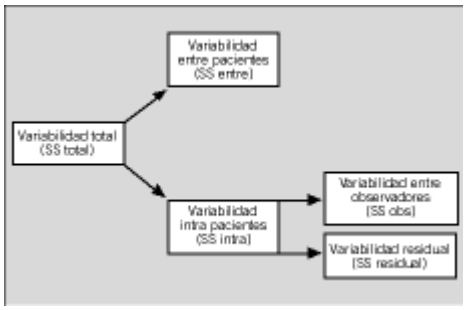

Fig. 1. Descomposición de la variabilidad total para medidas repetidas para un factor.

El CCI se interpreta como una proporción de la variabilidad total observada, expresándose en función de las variancias de cada término que pueden ser obtenidas a través de una tabla ANOVA para medidas repetidas con efectos fijos o aleatorios de los observadores (tabla A).

TABLA A
Tabla ANOVA para medidas repetidas

Fuente de variación	Grados de libertad	Suma de cuadrados	MS	E (MS) Efectos aleatorios	E (MS) Efectos fijos
Entre pacientes	n-1	$k\sum(\bar{X}_i-\bar{X}..)^2$	CM_p	$\sigma^2_e+k\sigma^2_p$	$\sigma^2_e+k\sigma^2_p$
Intra pacientes Observador	k-1	$n\sum(\bar{X}_j-\bar{X}..)^2$	CM_o	$\sigma^2_e+k\sigma^2_o$	$\sigma^2_e+n\sum\alpha^2_j$
Residual	(n-1)(k-1)	$\sum(X_{ij}-\bar{X}_i.-\bar{X}_j.-\bar{X}..)^2$	CM_e	σ^2_e	σ^2_e
Total	nk-1	$\sum(X_{ij}-\bar{X}..)^2$			

TABLA B

Tabla ANOVA para datos de presión arterial

Fuente de variación	Grados de libertad	Suma de cuadrados	MS
Entre pacientes	9	670,00	74,44
Intra pacientes			
Observador	1	180,00	180,00
Residual	9	20,00	2,22
Total	19	870,00	

TABLA C

Tabla ANOVA para mediciones de la frecuencia cardiaca

Fuente de variación	Grados de libertad	Suma de cuadrados	CM
Entre pacientes	9	459,45	51,05
Intra pacientes			
Observador	1	174,05	174,05
Residual	9	21,45	2,38
Total	19	654,95	

Las fórmulas que a continuación se detallan para estimar los CCI son la generalización de 2 a k observadores de las fórmulas presentadas con anterioridad en el artículo.
El CCI en el caso de modelo de efectos aleatorios se define como:

$$CCI = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_o + \sigma^2_R}$$

Basándose en la tabla ANOVA y realizando las transformaciones algebraicas correspondientes, el estimador del coeficiente se determina del siguiente modo:

$$CCI = \frac{n\,(CM_p - CM_R)}{nCM_p + kCM_o + (nk - n - k)\,CM_R}$$

Si los observadores tienen un efecto fijo, el coeficiente de correlación intraclase y su estimador se definen como:

$$CCI = \frac{\sigma^2_p}{\sigma^2_p + \frac{1}{k}\sum \rho^2_j + \sigma^2_R} \qquad \overline{CCI} = \frac{n\,(CM_p - CM_R)}{nCM_p + (k-1)\,CM_o + (n-1)(k-1)\,CM_R}$$

Del estudio hipotético de la presión arterial sistólica (tabla 1), en el cual K = 2 (observadores) y n = 10 (pares de datos), se obtiene la siguiente tabla ANOVA:

siendo el coeficiente de correlación intraclase estimado,

$$\overline{CCI} = \frac{10\,(74,44 - 2,22)}{10*74,44 + 2*1,80 + (10*2 - 10 - 2)*2,22} = 0,64$$

En el caso del modelo de efectos fijos, datos sobre medición de la frecuencia cardiaca (tabla 3), los resultados de aplicar un ANOVA se resumen a continuación:

Bibliografía

1. *Hernández Aguado I, Porta Serra M, Miralles M, García Benavides F, Bolúmar F* La cuantificación de la variabilidad en las observaciones clínicas. Med Clin (Barc) 1990; 95: 424-429[Medline]

2. *Kramer MS, Feinstein AR* Clinical Biostatistics LIV. The biostatistics of concordance. Clin Pharmacol Ther 1981; 29: 111-123[Medline]

3. *Candela Toha AM* Validación de aparatos y métodos de medida: concordancia sí, correlación no. Med Clin (Barc) 1992; 99: 314[Medline]

4. *Bland JM, Altman DG* Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1: 307-310[Medline]

5. *Mora Ripoll R, Ascaso Terrén C, Sentís Vilalta J* Uso actual de la estadística en investigación biomédica: una comparación entre revistas de medicina general. Med Clin (Barc) 1996; 106: 451-456[Medline]

6. *Mora Ripoll R, Ascaso Terrén C, Sentís Vilalta J* Tendencias actuales en la utilización de la

estadística en medicina. Estudio de los artículos originales publicados en Medicina Clínica (1991-1992). Med Clin (Barc) 1995; 104: 444-447[Medline]
7. *Fleiss JL* The design and analysis of clinical experiments. Nueva York: John Wiley & Sons, Inc., 1986
8. *Fleiss JL, Cohen J* The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 1973; 33: 613-619
9. *Bartko JJ* General methodology II. Measures of agreement: a single procedure. Stat Med 1994; 13: 737-745[Medline]
10. *Deyo RA, Diehr PD, Patrick DL* Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Controll Clin Trials 1991; 12 (Supl): 142-158
11. *Bravo G, Potvin L* Estimating the reliability of continuous measures with Cronbachs alpha or the intraclass correlation coefficient: towards the integration of two traditions. J Clin Epidemiol 1991; 44: 381-390[Medline]
12. *Fisher RA* On the «probable error» of a coefficient of correlation deduced from a small sample. Metron 1921; 1: 1-32
13. *Müller R, Büttner P* A critical discussion of intraclass correlation coefficients. Stat Med 1994; 13: 2.465-2.476[Medline]
14. *McGraw KO, Wong SP* Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996; 1: 30-46
15. *Alonso J, Prieto L, Antó JM* La versión española del SF-36 Health Survey (Cuestionario de Salud SF-36): un instrumento para la medida de los resultados clínicos. Med Clin (Barc) 1995; 104: 771-776[Medline]