Revista Colombiana de Psiquiatría
ISSN: 0034-7450
revista@psiquiatria.org.co
Asociación Colombiana de Psiquiatría
Colombia

Campo-Arias, Adalberto; Herazo, Edwin
Concordancia intra- e interevaluadores
Revista Colombiana de Psiquiatría, vol. 39, núm. 2, 2010, pp. 424-432
Asociación Colombiana de Psiquiatría
Bogotá, D.C., Colombia

Concordancia intra- e interevaluadores

Adalberto Campo-Arias[1]
Edwin Herazo[2]

Resumen

Introducción: En psiquiatría, los estudios de concordancia intra e interevaluador son importantes para medir la confiabilidad o reproducibilidad de las evaluaciones (entrevistas o escalas heteroaplicadas). *Objetivo:* Presentar algunos principios sobre el proceso de validación de entrevistas diagnósticas o escalas heteroaplicadas y el manejo y comprensión de las pruebas estadísticas más útiles para estos fines. *Método:* Revisión de literatura. *Resultados:* Se entiende por concordancia el grado de acuerdo o de desacuerdo entre las evaluaciones hechas a un mismo sujeto de forma sucesiva por parte de un evaluador o entre dos o más entrevistadores. Este proceso es de la validación de instrumentos, ya sea para identificar posibles casos o confirmar la presencia de un trastorno mental. En la concordancia interevaluador, dos o más psiquiatras realizan una entrevista de manera independiente y casi simultánea a una persona y así se puede estimar el grado de acuerdo, convergencia o concordancia (o lo contrario) entre las evaluaciones y los consiguientes diagnósticos. La concordancia intraevaluador es el grado de acuerdo en el diagnóstico que tiene en el tiempo un mismo evaluador. La prueba kappa de Cohen se usa para estimar la concordancia y se esperan, por lo general, valores superiores a 0,50; pero es necesario conocer la prevalencia esperada del trastorno mental, el número de evaluadores o evaluaciones y el número de categorías o casillas diagnósticas posibles.

Palabras clave: psicometría, escalas, reproducibilidad de resultados, estudios de validación, revisión.

Title: Intra- and Inter-Rater Concordance

Abstract

Introduction: Intra- and inter-rater concordance studies are important in order to measure the reliability or the reproducibility of evaluations (interviews or scales applied by a rater) in psychiatry. *Objective:* To present some principles regarding the validation process of diagnostic interviews or scales applied by a rater, and regarding the handling and comprehension of more useful statistical tests. *Method:* Review of literature. *Results:* Concordance is understood as the grade of agreement or disagreement among evaluations made to the same subject successively by an evaluator or among two or more interviewers. This process is part of the

[1] Médico psiquiatra. MSc (c). Grupo de Investigación del Comportamiento Humano, Instituto de Investigación del Comportamiento Humano. Bogotá, Colombia.
[2] Médico psiquiatra. Grupo de Investigación del Comportamiento Humano, Instituto de Investigación del Comportamiento Humano, Bogotá, Colombia.

validation of instruments, scale reliability, in order to identify possible cases or to confirm the presence of a mental disorder. Inter-rater concordance refers to the case when two or more psychiatrists realize an interview independently and almost simultaneously to a person; this can help to estimate the grade of agreement, convergence or concordance (and disagree, divergence or discordance) among the evaluations and the consequent diagnostics. Intra-rater concordance is the grade of agreement on the diagnosis made by the same rater in different times. Cohen's kappa is used to estimate concordance, and values higher than 0.50 are expected in general. To reliably estimate Cohen's kappa is necessary to know previously the expected prevalence of mental disorder, the number of evaluations or raters, and the number of possible diagnosis categories.

Key words: Psychometrics, scales, reproducibility of results, validation studies, review.

Introducción

El desarrollo de la psiquiatría en las últimas décadas guarda una estrecha relación con la implementación sistemática de criterios diagnósticos para el uso de los proveedores de servicios en salud mental, a pesar de las limitaciones de estos criterios (1,2).

La estandarización de los criterios diagnósticos se complementó con el diseño y validación de entrevistas diagnósticas, con el propósito de estandarizar la mayor parte del proceso de evaluación de las personas en la práctica clínica y en investigación epidemiológica (3). Se diseñaron entrevistas estructuradas o semiestructuradas para la aplicación por personas sin formación técnica o profesional en salud mental o por personas con entrenamiento y experiencia clínica formal (4).

No obstante, las discusiones académicas iniciales sobre el posible impacto negativo de este abordaje de los síntomas de las personas que consultan por trastornos mentales (1), la revisión cuidadosa de las manifestaciones clínicas y la presentación de un diagnóstico, provisional o uno más revisado, demanda profesionales en salud mental bien entrenados y estudiosos y, en particular, psiquiatras, que en última instancia son los profesionales llamados a dirimir las controversias diagnósticas en salud mental. El diagnóstico en casi todos los contextos de la medicina, y en especial de la psiquiatría, exige un juicio clínico cuidadoso para dar a los síntomas una connotación no adaptativa o disfuncional, es decir, importancia o relevancia clínica (5).

La *concordancia* se entiende como el grado de acuerdo, o desacuerdo, entre las evaluaciones que una persona realiza en forma sucesiva a otra persona o entre dos o más entrevistadores que hacen una evaluación a un mismo sujeto (6,7). Este proceso hace parte de la validación de instrumentos, de comprobar la confiabilidad, ya sea para identificar posibles casos o confirmar la presencia de un trastorno mental (6).

El objetivo de esta revisión es presentar algunos principios por considerar en el proceso de validación

de entrevistas diagnósticas o escalas heteroaplicadas, y el manejo y comprensión de las pruebas estadísticas más útiles para estos fines.

Principios

El objetivo central de una entrevista diagnóstica es definir qué persona reúne criterios para un trastorno mental y cuál es el trastorno mental específico, independientemente de quién lleve a cabo la entrevista (8). La determinación de la concordancia en el caso de las escalas auto- o heteroaplicadas, que habitualmente dan puntuaciones, se realiza mediante el procedimiento que se conoce como prueba-reprueba (*test-retest*, en inglés) (9). Este proceso se vale de pruebas estadísticas, como la correlación de Pearson (10,11), el coeficiente de correlación intraclase (12), el coeficiente de concordancia de Lin (13) o el coeficiente de concordancia de Altman y Bland (14).

Tanto en la validación de escalas como en la validación de entrevistas, la segunda evaluación se realiza con un periodo definido, según el trastorno que se evalúe. Se supone que durante éste los síntomas se mantienen estables, con pocas o pequeñas variaciones, y que las condiciones de medición o entrevista son similares (15).

Es importante tener presente que en psiquiatría el cambio de diagnóstico con el paso del tiempo es un fenómeno frecuente. Este hecho se puede relacionar con evaluaciones diagnósticas no estructuradas o estandarizadas, cambios en los criterios diagnósticos o la misma historia natural de los trastornos mentales que se evalúan; es decir, la inestabilidad o cambios que muestra el conjunto de síntomas en el tiempo (15,16).

Otro punto que se debe tener presente es que si dos o más profesionales realizan una entrevista a la misma persona en forma independiente o el mismo evaluador hace dos o más entrevistas en un periodo, se debe tomar uno de los evaluadores o una de las evaluaciones como criterio de referencia (*gold standard*) (17). Los criterios de referencia perfectos son excepcionales en medicina, más aún en psiquiatría (18). Sin embargo, se parte del supuesto de que este criterio que se considera punto de comparación hace una clasificación perfecta de los diagnósticos, sin errores; es decir, con 100% de sensibilidad y 100% de especificidad (17-20). Esta comparación con un criterio de referencia hace parte, igualmente, de la validación criterio (concurrente) de cualquier escala o instrumento (21,22).

Concordancia intra- e interevaluadores u observadores

Si dos o más psiquiatras realizan una entrevista de manera independiente y casi simultánea a una persona se puede estimar el grado de acuerdo, convergencia o concordancia (y de desacuerdo, divergencia o

discordancia) entre las evaluaciones y los consiguientes diagnósticos, si se toma uno de los evaluadores como criterio de referencia. Se parte del hecho de que ambos profesionales tienen el mismo entrenamiento; a esta estimación se le llama *concordancia interevaluadores o interobservadores* (6,23,24).

Por otra parte, si un psiquiatra realiza dos o más evaluaciones a una misma persona con el propósito de conocer o confirmar un diagnóstico con el uso de una técnica idéntica, se puede establecer la concordancia en el diagnóstico que tiene en el tiempo el mismo evaluador. A este cálculo se le conoce como concordancia intraevaluador o intraobservador (7,25). Sin duda, la concordancia intraevaluador tiene el sesgo que induce la memoria del evaluador que puede recordar detalles de la evaluación precedente que él mismo realizó (26).

Pruebas estadísticas para concordancia inter- o intraevaluador con resultados cualitativos

El diagnóstico en psiquiatría es, por lo general, cualitativo o categórico, o sea que se llega a la conclusión de que la persona reúne criterios o no para un trastorno mental al momento de la evaluación o en algún momento de la vida (5). El número de categorías diagnósticas posibles es *K* y el número de evaluadores es *M*. Si se compara la evaluación de un evaluador con la de otro evaluador que se toma como criterio de referencia y sólo son posibles dos diagnósticos, es decir $K=2$ y $M=2$, se puede construir una tabla de contingencia de 2×2, con un mínimo de cuatro casillas (tetracórica), para observar la concordancia entre evaluadores (24,27-29).

A continuación un ejemplo de un estudio que investigaba la concordancia interevaluador. En una investigación que se realizó en Navarra, España, Landa y colaboradores cuantificaron la concordancia en la identificación de un trastorno mental entre pediatras y los profesionales de salud mental. En la investigación se incluyeron 207 niños o adolescentes, menores de 16 años; hallaron una concordancia observada (*Po*) para la presencia de un trastorno mental del 64,3% y un valor de kappa media de Cohen de 0,58, con un intervalo de confianza del 95% entre 0,51 y 0,66 (30).

Pocos estudios se realizan para explorar la concordancia intraevaluador. Por ejemplo, Conradsson y colaboradores evaluaron en 45 adultos mayores en Umea, Suecia, las puntuaciones que dio el mismo evaluador en una escala para cuantificar equilibrio, de uno a tres días después de la primera aplicación. Este instrumento consta de 14 preguntas, con un patrón de respuesta politómico (Likert), con cinco opciones que se califican de cero a cuatro. Informaron la concordancia intraevaluador para cada pregunta mediante el

coeficiente de *K* ponderada e intervalo de confianza del 95%. Los valores de *K* (*kappa*) ponderada se encontraron entre 0,55 y 0,83 (31).

De la misma forma, es posible diseñar una tabla de contingencia más compleja, en la que se compara, por ejemplo, la concordancia en el diagnóstico específico entre un grupo de pacientes que reúne criterios para varias categorías posibles (*K*>2), por ejemplo, un trastorno depresivo (trastorno depresivo mayor, trastorno distímico, trastorno depresivo debido a una condición médica, trastorno depresivo debido al uso de sustancia o medicamento, o trastorno depresivo no especificado). Y a la vez participan dos evaluadores o se realizan evaluaciones separadas en el tiempo (*M*=2). Para esta situación, la tabla de contingencia *K*×*M* será de 5×2 (25).

A manera de ejemplo de un estudio de concordancia de más de dos categorías diagnóstica, Lin y colaboradores observaron la concordancia en 579 adultos, tras responder una escala disponible en Internet para identificar trastornos depresivos (trastorno depresivo mayor, trastorno depresivo menor, síntomas depresivos subsindrómicos y ausencia de trastorno depresivo), entre las aplicaciones que se realizaron cada dos semanas (dos, cuatro o más semanas) e informaron los valores de *K* ponderada: 0,80, 0,42 y 0,51, a la segunda semana, a la cuarta semana y más semanas, respectivamente (32).

La concordancia entre dos evaluadores o entre dos o más observaciones del mismo evaluador puede ser real o producto o resultado del azar. Por ello, además, de la concordancia observada, es necesario estimar hasta qué grado de acuerdo lo media el azar o la probabilidad (33,34). La prueba estadística que se usa para este propósito es la prueba *K* de Cohen (35). Cuando se calcula a partir de dos categorías posibles y dos evaluadores únicamente, *K*=2 y *M*=2, de una tabla de contingencia 2×2, se llama *K* media de Cohen o, simplemente, *K*. Sin embargo, en los casos en los que se calcula con *K*>2 (o con un resultado ordinal) o *M*>2 se estima un valor de *K* ponderada (24,27,36) (véase Tabla 1).

Tabla 1. Tabla para el cálculo de K cuando con hay más de dos posibilidades de diagnóstico (K>2)

Clasificación evaluador 2	Clasificación evaluador 1*			Totales
	Diagnóstico 1	Diagnóstico 2	Diagnóstico 3	
Diagnóstico 1				
Diagnóstico 2				
Diagnóstico 3				
Totales				

* Se toma como criterio de referencia.

La *K* media de Cohen se puede calcular con la Fórmula 1. No obstante, los programas estadísticos más usados en el mundo, como Epi-Info (37), PASW (anteriormente, SPSS) (38), SAS (39) o STATA (40), lo estiman más rápidamente e informan el intervalo de confianza del 95%, como una medida de precisión de la estimación (41-43). Los valores de *K* pueden encontrarse entre cero y uno, a mayor cercanía con el uno mayor concordancia en las mediciones que se realizaron por el mismo evaluador o diferentes evaluadores. La forma como se interpreta de manera racional este coeficiente se presenta en la Tabla 2 (44,45). La interpretación de la prueba estadística debe considerar la utilidad clínica de las mediciones en estudio (46).

Tabla 2. Interpretación cualitativa de los valores de K

Valores de K	Interpretación
Entre 0 y 0,20	Deficiente
Entre 0,21 y 0,40	Pobre
Entre 0,41 y 0,60	Aceptable
Entre 0,61 y 0,80	Buena
Entre 0,81 y 1,00	Excelente

Fórmula 1

$$K = Po - Pe / 1 - Pe$$

La $Po = a + d/n$ (véase Tabla 3)

La $Pe = a + b/n\ [(a + c)/n + (b + d)/n + (c + d)/n]$

Po = Frecuencia o prevalencia observada

Pe = Frecuencia o prevalencia esperada

Consideración importante

Al igual que la sensibilidad, la especificidad y los valores predictivos que se calculan con los datos de una tabla de contingencia de 2×2, la prueba K es directamente proporcional a la frecuencia o prevalencia del trastorno mental que se estudia (47,48). En consecuencia, la K puede ser baja, no obstante el alto valor para la concordancia observada (Po) (49). La K suele ser más robusta cuando la prevalencia del trastorno que se investiga es alta y debilitarse cuando la prevalencia es baja (50,51).

Tamaño de la muestra

El tamaño de la muestra se ignora con frecuencia en los estu-

Tabla 3. Tabla de contingencia de 2×2

Clasificación evaluador 2	Clasificación evaluador 1*		Totales
	Presente	Ausente	
Presente	A	B	$a + b$
Ausente	C	D	$c + d$
Totales	a + c	b + d	$a + b + c + d\ (n)$

* Se toma como criterio de referencia.

dios de validación u observación del desempeño psicométrico de los instrumentos en salud (18). El cálculo de la muestra para un estudio de concordancia y el cálculo de un valor *K* debe considerar varios puntos: el número de evaluadores o evaluaciones (52) y el número de categorías o casillas diagnósticas posibles (53). De la misma forma, se debe ponderar la prevalencia esperada o estimada del trastorno mental, como si se fuera a estimar la sensibilidad o la especificidad, y se parte de una tabla de contingencia 2×2 (54). Tener muy presente este punto permite contar con un número suficiente de participantes en cada casilla de la tabla por construir (48,50-52,55). Con esto se logra un mejor grado de precisión, con un intervalo de confianza más estrecho (42-44,56).

Conclusiones

Los estudios de concordancia inter- e intraevaluador son importantes para medir la confiabilidad o reproducibilidad de las evaluaciones (entrevistas o escalas) en psiquiatría. Para las evaluaciones con resultados categóricos (cualitativos), la concordancia más allá del azar se estima con el coeficiente de *K* de Cohen (media o ponderada). El coeficiente de *K* se puede encontrar entre cero y uno, y por lo general se espera entre 0,41 y 0,60 o más. La prevalencia del trastorno o trastornos que se investigan puede afectar la estimación del coeficiente. Es necesario un tamaño de muestra razonable para contar un valor de *K* lo suficientemente preciso.

Referencias

1. Acton SG, Zodda JJ. Classification of psychopathology. Goals and methods in an empirical approach. Theory Psychol. 2005;15(3):373-99.
2. Rogler LH. Making sense of historical changes in the diagnostic and statistical manual of mental disorders: five propositions. J Health Soc Behav. 1997;38(1):9-20.
3. Páez F, Nicolini H. Las entrevistas para el diagnóstico clínico en psiquiatría. Salud Mental. 1996;19(Supl 2):19-25.
4. Calinou I, McClellan J. Diagnostic interviews. Cur Psychiatry Rep. 2004;6(2):88-95.
5. Eaton WW, Hall AL, MacDonald R, McKibben J. Case identification in psychiatric epidemiology: a review. Int Rev Psychiatry. 2007;19(5):497-507.
6. Carrasco JL, Jover L. Métodos estadísticos para evaluar la concordancia. Med Clin (Barc). 2004;122(Supl 1):28-34.
7. Alarcón AM, Muñoz S. Medición en salud: Algunas consideraciones metodológicas. Rev Med Chile. 2008;136(1):125-30.
8. Othmer E, Othmer SC. DSM-IV-TR. La entrevista clínica. Fundamentos. Tomo I. Madrid: Masson; 2001.
9. Sánchez R, Echeverry J. Validación de escalas de medición en salud. Rev Salud Pública. 2004;6(3):302-18.
10. Pearson K. Determination of the coefficient of correlation. Science. 1909;30(757):23-5.
11. Spearman C. Correlation calculated from faulty data. Br J Psychol. 1910;3:271-95.
12. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420-8.
13. Lin L. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989;45(1):255-68.
14. Bland JM, Altman DG. Statistical methods for assessing agreement between

two methods of clinical measurement. Lancet. 1986;1(8476):307-10.
15. Lecrubier Y. Refinement of diagnosis and disease classification in psychiatry. Eur Arch Psychiatry Clin Neurosci. 2008;258 Suppl 1:6-11.
16. Neighbors HW, Trieweiler SJ, Ford BC, Muroff JR. Racial differences in DSM diagnosis using a semi-structured instrument: The importance of clinical judgment in the diagnosis of African Americans. J Health Soc Behav. 2003;44(3):237-56.
17. Rielgelman RK, Hirsch RP. Definición de enfermedad: la prueba de oro. Bol Of Sanit Panam. 1991;111(6):534-38.
18. Knottnerus JA, van Weel C, Muris JWM. Evaluation of diagnostic procedures. BMJ. 2002;324(7335):477-80.
19. López-Jiménez F, Rohde LEF, Luna-Jiménez MA. Problemas y soluciones en la interpretación de pruebas diagnósticas. Rev Invest Clin. 1998;50(1):65-72.
20. Castro-Jiménez MA, Cabrera-Rodríguez D, Castro-Jiménez MI. Evaluación de tecnologías diagnósticas: conceptos básicos en un estudio con muestreo transversal. Rev Colomb Obstet Ginecol. 2007;58(1):45-52.
21. Morgan GA, Gliner JA, Harmon RJ. Measurement validity. J Am Acad Child Adolesc Psychiatry. 2001;40(6):729-31.
22. Bland JM, Altman DG. Validating scales and indexes. BMJ. 2002;324(7337):606-7.
23. Ludbrook J. Statistical techniques for comparing measurers and methods of measurements: a critical review. Clin Exp Pharmacol Physiol. 2002;29(7):527-36.
24. Watkins MW, Pacheco M. Interobserver agreement in behavioral research: importance and calculation. J Behav Educ. 2000;10(4):205-12.
25. Kramer HC, Periyakoil VS, Noda A. Kappa coefficients in medical research. Stat Med. 2002;21(14):2109-29.
26. Ibáñez C, Maganto C. El proceso de evaluación clínica: cogniciones del evaluador. Summa Psicol UST. 2009;6(1):81-99.
27. McGinn T, Wyer PC, Newmann TB, Keitz S, Leipzig R, For GG, et al. Tips for learners for evidence-based medicine: 3. Measures of observer variability (kappa statistic). CMAJ. 2004;171(11):1369-73.
28. Álvarez-Martínez HE, Pérez-Campos E. Utilidad clínica de la tabla 2x2. Rev Eviden Invest Clin. 2009;2(1):22-7.
29. Colimon K-M. Programa de estudio y programa de control. En: Colimon KM. Fundamentos de epidemiología. 3ª edición. Medellín: Corporación para Investigaciones Biológicas; 2010. p. 123-124.
30. Landa N, Goñi A, García de Jalón E, López-Goñi JJ. Concordancia en el diagnóstico entre pediatra y salud mental. An Sist Sanit Navar. 2009;32(2):161-8.
31. Conradsson M, Lundin-Olsson L, Lindelöf N, Littbrand H, Malmqvist L, Gustafson Y, et al. Berg Balance Scale: Intrarater test-retest reliability among older people dependent in activities of daily living and living in residential care facilities. Phys Ther. 2007;87(9):1155-63.
32. Lin CC, Bai YM, Liu CY, Hsiao MC, Chen JY, Tsai SJ, et al. Web-based tools can be used reliably to detect patients with major depressive disorder and subsyndromal depressive symptoms. BMC Psychiatry. 2007;7:12.
33. Schuster C. Kappa as a parameter of a symmetry model for rater agreement. J Educ Behav Stat. 2001;26(3):331-42.
34. Barnhart HX, Song J, Haber MJ. Assessing intra, inter and total agreement with replicated readings. Stat Med. 2005;24(9):1371-84.
35. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(3):37-46.
36. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968;70(4):213-20.
37. Epi-Info 3.5.1. Atlanta: Centers for Disease Control and Prevention (CDC); 2008.
38. PAWS 18.0. Chicago: SPSS. Inc; 2009.
39. SAS 9. SAS Institute Inc.; 2009.
40. STATA 11 for windows. College Station: StataCorp LP; 2009.
41. Herrera AN, Quintero C, Sanchez R. Algunas estadísticas de uso frecuente

en investigación en salud (1ª parte). Rev Colomb Anest. 1998;26:225-32.
42. Montori VM, Kleinbart J, Newman TB, Keitz S, Wyer PC, Moyer V, et al. Measures of precision (confidence intervals). CMAJ. 2004;171(6):611-5.
43. Cepeda-Cuervo E, Aguilar W, Cervantes V, Corrales M, Díaz I, Rodríguez D. Intervalos de confianza e intervalos de credibilidad para una proporción. Rev Colomb Estat. 2008;31(2):211-28.
44. Abraira V. El índice kappa. Semergen. 2000;27(5):247-9.
45. McGinn T, Wyer PC, Newmann TB, Keitz S, Leipzig R, Guyatt G, et al. Understanding and calculating kappa. CMAJ [Internet]. 2004 [citado 2010 Ene 26];171(11):1-9. Disponible en: www.cmaj.ca/cgi/data/171/11/1369/DC1/1.
46. Cepeda MS, Pérez A. Estudios de concordancia. En: Ruiz A, Gómez C, Londoño D. Investigación clínica: epidemiología clínica aplicada: Bogotá: Centro Editorial Javeriano, CEJA; 2001. p. 287-301.
47. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990;43(6):543-9.
48. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol. 1993;46(5):422-9.
49. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol. 1990;43(6):551-8.
50. Streiner DL. Learning how to differ: agreement and reliability statistics in psychiatry. J Can Psychiatry. 1995;40(2):60-6.
51. Guggenmoos-Holzmann I. The meaning of kappa: Probabilistic concepts of reliability and validity revisited. J Clin Epidemiol. 1996;49(7):775-82.
52. Cantor AB. Sample-size calculations for Cohen's Kappa. Psychol Methods. 1996;1(2):150-3.
53. Streiner DL. Diagnosing tests: Using and misusing diagnostic and screening tests. J Pers Assess. 2003;81(3):209-19.
54. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. J Clin Epidemiol. 2005;58(8):859-62.
55. Kramer HC, Bloch DA. A note on case-control sampling to estimate kappa coefficients. Biometrics. 1990;46(1):49-59.
56. Scotto MG, Garcés AT. Interpretando correctamente en salud pública estimaciones puntuales, intervalos de confianza y contrates de hipótesis. Salud Publica Mex. 2003;45(6): 505-11.

Conflicto de interés: los autores manifiestan que no tienen ningún conflicto de interés en este artículo.

Recibido para evaluación: 28 de enero del 2010
Aceptado para publicación: 27 de abril del 2010

Correspondencia
Adalberto Campo-Arias
Grupo de Investigación del Comportamiento Humano
Instituto de Investigación del Comportamiento Humano
Calle 58 No. 5-24, oficina 202
Bogotá, Colombia
campoarias@comportamientohumano.org