La concordancia entre dos *tests* clínicos para casos binarios: problemas y solución

Juan Carlos Azzimonti Renzo† [1]

El Dr. JUAN CARLOS AZZIMONTI RENZO (1949-2003) falleció repentinamente el 14 de octubre de 2003 cuando estaba en plena actividad.

Había egresado como Ingeniero industrial en la Universidad de Buenos Aires en 1971. Radicado en Posadas en 1976, al margen de sus tareas como ingeniero se orientó al estudio, investigación y práctica estadística.

Desde el año 2000 tenía a su cargo la cátedra de Bioestadística para la carrera de Farmacia, de la misma facultad.

Su preocupación por la enseñanza lo motivó a publicar en 2001 el libro *Bioestadística aplicada a Bioquímica y Farmacia,* editado por la Editorial UNAMisiones, obra que también está disponible en http://fceqyn.unam.edu.ar/bioc así como en www.biometria.freesservers.com, lo que habla de su generosidad y espíritu docente.

Quedaron en su computadora trabajos pendientes de ser enviados a publicar: *Clinical Ratios fo the Odds applied to Diagnosis Quality: a new approach* y *Clinical Ratio fo the Odds applied to Diagnosis Quality: special cases,* que seguramente aportarían al bioquímico clínico y el proyecto de un curso de enseñanza de Bioestadística a distancia.

Hemos perdido a un espíritu inquieto, y a sus generosos aportes en el área de la estadística en Bioquímica Clinica.

Dra. Ana Haedo

1. Profesor Titular. Cátedras de Bioestadística para Bioquímica y Farmacia. Facultad de Ciencias Exactas, Químicas y Naturales. Universidad Nacional de Misiones, Argentina.

Acta Bioquímica Clínica Latinoamericana

Incorporada al Chemical Abstract Service.

Código bibliográfico: ABCLDL.

ISSN 0325-2957

Resumen

Este informe puede ser considerado como un recurso en el análisis de concordancia entre *tests* clínicos, observadores, tasas, jueces o expertos. Cuando no se puedan obtener los resultados verdaderos de dos *tests* clínicos, la diferencia entre ellos puede ser reflejada estudiando la concordancia que tienen entre sí. Los métodos estadísticos diseñados para hacer este tipo de estudio no son suficientes para decidir si hay concordancia, pues presentan problemas o paradojas desde un punto de vista clínico. La solución de estos problemas se muestra usando el método de visión dual. La concordancia general es una dualidad y necesita ser estudiada en dos etapas. En la primera etapa se propone una condición para la aceptación de la concordancia: ambos métodos deben tener la misma sensibilidad y especificidad. Pero como esto no es suficiente, en la segunda etapa se propone otra condición: el nivel de concordancia encontrado debe ser mayor o igual a un valor clínico definido por los clínicos. Cuando ambas etapas muestren resultados satisfactorios, el nuevo método puede reemplazar al método viejo. Así se muestra que la concordancia estadística no es lo mismo que la concordancia clínica.

Palabras clave: **método de visión dual * concordancia clínica * concordancia estadística * tabla de concordancia * tabla diagnóstica * nivel de concordancia * sensibilidad * especificidad**

Summary

AGREEMENT BETWEEN TWO CLINICAL TESTS IN BINARY CASES: PROBLEMS AND SOLUTION

This report can be considered a resource for the analysis of agreement among clinical tests, observers, judges or experts. When the true results of two clinical tests cannot be obtained, the difference between them can be reflected by studying their agreement. Normal statistical procedures

tackling this are not enough to decide whether there is agreement, because they show problems or paradoxes from the clinical viewpoint. The solution to these problems is introduced by using the dual vision method. The overall agreement is a duality, and needs to be studied in two steps. In the first step, a condition for the acceptance of the agreement is proposed: both methods need to have the same nosologic sensitivity and specificity. But since this is not enough, another condition is proposed in the second step: the level of agreement should be greater than a critical value defined by the clinicians. When both steps show satisfactory results, the new method can replace the old one. In this way, it is showed that statistical agreement is not the same as clinical agreement.

Keywords: dual vision method * Clinical agreement * Statistical agreement * Agreement table * Diagnostic Table * Level of agreement * sensitivity * specificity

Introducción

El análisis de concordancia se usa para comparar diversos métodos clínicos entre sí cuando son aplicados al mismo grupo de individuos (muestras apareadas) (1). El caso más simple es la comparación de dos métodos binarios, cuyos resultados se pueden presentar en una Tabla de Concordancia tal como se muestra en la Tabla I. Generalmente, el objetivo es determinar si el método usual del laboratorio (método viejo) puede ser reemplazado por uno nuevo. No se trata de determinar cuál de ellos es el mejor, sino tan sólo ver si uno puede ser intercambiado por otro. Esto es más económico porque se trata de medir con el nuevo *test* al mismo individuo medido en la rutina diaria. En cambio, para determinar cuál es el mejor, se necesita usar el *test* de referencia o *gold standard*, generalmente de mayor costo y poca disponibilidad.

La manera tradicional de efectuar este estudio es mediante la aplicación de modelos estadísticos que realizan la comparación de dos proporciones, o bien la asociación matemática entre dos factores. Estos factores pueden ser *tests* clínicos, observadores, jueces, métodos de diagnóstico, etc. En este informe se analiza el caso de dos métodos o *tests* de laboratorio cuyos resultados se expresan en forma binaria (positivo / negativo). Se lo puede usar cuando a cada individuo se lo estudia con dos métodos clínicos a la vez y sus valores verdaderos son desconocidos.

Tabla I. *Tabla de concordancia.*

	Método 1		
Método 2	(+)	(-)	Total
(+)	a	b	a + b
(-)	c	d	c + d
Total	a + c	b + d	N

Donde el significado de cada frecuencia es:

a: Los sujetos que muestran ambos resultados (+) (concordancia en positivos).
b: Los sujetos que muestran (-) con el primer método y (+) con el otro (discordancia).
c: Los sujetos que muestran (+) con el primer método y (-) con el otro (discordancia).
d: Los sujetos que muestran ambos resultados (-) (concordancia en negativos).

Nivel de concordancia (tasa de concordancia): $\lambda = (a + d) / N$ (expresado como porcentaje).

El modelo más difundido es el de McNemar (2)(3) cuyo objetivo es la comparación de dos proporciones apareadas. Este modelo fue mejorado con el llamado G-*test* y optimizado con las correcciones de Williams (4). La forma no paramétrica de hacer la comparación es usando el método Q de Cochran (1)(4). Sin embargo, Conover WJ (5) mostró que Q se reduce al *test* de McNemar sin la corrección de Yates (1)(5). Estos tres métodos se enfocan en analizar la diferencia entre los dos tipos de discordancias. La hipótesis nula de trabajo es que no hay diferencia entre las discordancias (b - c = 0), y por eso ambos métodos concuerdan (Ho : b = c). El *test* estadístico rechaza o no esta hipótesis. Pero este tipo de análisis presenta dos problemas básicos (6):

a) Cuando $b \approx c$ la Ho no será rechazada y se pensará que no hay evidencia estadística suficiente como para rechazar la concordancia entre ambos métodos. Por ejemplo, si b = c = 198 y N = 400, habrá 396 discordancias en 400 casos. Y lo mismo ocurrirá cuando b = c = 2, estadísticamente hablando. Pero desde un punto de vista clínico no es lo mismo tener 396 discordancias en 400 casos, que tener tan sólo 4.
b) No se toma en cuenta el tamaño muestral N. Por ejemplo, si b = 25 y c = 10 en 400 casos, la hipótesis nula (Ho) se rechaza con los tres métodos estadísticos. Y lo mismo ocurre cuando b = 25 y c = 10 en 400 millones de casos. Pero no es lo mismo tener 35 discordancias en 400 casos que en 400 millones desde un punto de vista clínico.

Estos dos problemas muestran que las conclusiones estadísticas de estos tres modelos no son aceptables desde una perspectiva clínica. En otras palabras, la concordancia estadística difiere de la concordancia clínica expresada mediante el nivel de concordancia obtenido en el estudio.

APÈNDICE 1

Propósito: Estudiar la concordancia entre dos métodos de diagnóstico Nivel de concordancia = l

Introduzca los valores para:	**Entrar datos**	
Nº de positivos en ambos métodos	330	individuos
Nº de negativos en ambos métodos	340	individuos
Nº de (+) en Método 1 y (-) en Método 2	55	individuos
Nº de (-) en Método 1 y (+) en Método 2	45	individuos
Máximo Nº de discordancias en 100 casos	10	

Total de individuos = 770
Valor de λ_{critico} = 90%

Mc Nemar o Cochran Q-*test*	**Q =**	**1,00**
Mc Nemar-*test* con correción de Yates	**χ^2 =**	**0,81**
G-*test* con correción de Williams	**G_{adj} =**	**1,00**

DO = Odds de discordancia

Valores Observados

DO = 0,1493

Tabla de concordancia

		Método 1 (+)	**Método 1** (–)	
Método 2	(+)	330	45	375
	(–)	55	340	395
		385	385	770

Valores esperados

		Método 1 (+)	**Método 1** (–)	
Método 2	(+)	187,5	187,5	375
	(–)	197,5	197,5	395
		385	385	770

Tasa de concordancia = 0.8701 Concordancia esperada = 0,5000

Kappa =	0,740
Yules (Y) =	0,741
Phi =	0,741

45,3333

Visión estadística

$\chi^2(0,05; 1) = 3.841$ ==> Estadísticamente resulta no significativo

Cuando la concordancia sea rechazada debe hacerse un análisis más incisivo (vea debajo).

Visión Clinica

DO lím. superior (95%) = 0,176

Valor de $DO_{\text{crítico}}$ = 0,111

Valor observado de DO = 0,149 DO lím. inferior (95%) = 0,122

La concordancia clínica es aceptable cuando $DO_{\text{crítico}}$ cae en el intervalo o es mayor que el límite superior

λ lím. superior (95%) = 89.39%

Valor de $\lambda_{\text{crítico}}$ = 90,00%

Valor observado de λ = 87,01% λ lím. inferior (95%) = 84.64%

La concordancia clínica es aceptable cuando $\lambda_{\text{crítico}}$ cae en el intervalo o es menor que el límite inferior

APÈNDICE 1 (cont,)

Un análisis más incisivo - Preguntas clínicas:

¿La potencial diferencia de sensibilidad es aceptable para esta enfermedad? 2,3%

Sensibilidad (S)			
S_1 =	0,88000	Límite inferior ΔS =	-0,02514
S_2 =	0,85714	Límite superior ΔS =	0,07085
Valor absoluto de ΔS =	0,02286	Clínicamente	
SE(ΔS) =	0,02449		
p =	0,86842	z =	0,93197
SE(p) =	0,02453	Estadísticamente	

¿La potencial diferencia de especificidad es aceptable para esta enfermedad? 2,2%

Especificidad (E)			
E_1 =	0,86076	Límite inferior ΔE =	-0,02450
E_2 =	0,88312	Límite superior ΔE =	0,06921
Valor absoluto de ΔE =	0,02236	Clínicamente	
SE(ΔE) =	0,0239		
p =	0,87179	z =	0,93378
SE(p) =	0,02394	Estadísticamente	

¿La potencial diferencia del Ind, de Youden es aceptable para esta enfermedad? 0,0%

Indice de Youden (Y)			
Y_1 =	0,74076	Límite inferior ΔY =	-0,04329
Y_2 =	0,74026	Límite superior ΔY =	0,04429
Valor absoluto de ΔY =	0,00050	Clínicamente	
SE(ΔS) =	0,02234		
p =	0,74051	z =	0,02237
SE(p) =	0,02234	Estadísticamente	

La otra manera de estudiar la concordancia es analizando si ambos métodos son independientes desde un punto de vista estadístico. La hipótesis nula plantea la independencia y cuando es rechazada por el *test* estadístico se dice que ambos métodos están asociados y dicho grado de asociación se compara con una tabla de aceptabilidad de la concordancia. El modelo más difundido de este enfoque es el método de Kappa, o Cohen-Kappa (7)(8). El valor obtenido del índice Kappa se compara con los criterios siguientes (8):

a) Si es nulo se dice que la concordancia es pobre
b) Si está entre 0 y 0,2 la concordancia es muy leve
c) Si está 0,2 y 0,4 la concordancia es leve
d) Si está entre 0,4 y 0,6 es moderada
e) Si está entre 0,6 y 0,8 es substancial
f) Y entre 0,8 y 1,0 es casi perfecta

La idea para obtener este índice es remover de la concordancia observada, la parte de concordancia debida al azar. Sin embargo, en la década del 90 se presentaron dos paradojas clínicas (9) y una solución para las mismas basada en las probabilidades marginales para los casos de concordancia en resultados positivos y negativos (10)(11). Sin embargo, esta solución no fue aceptada masivamente. Y por lo tanto se buscó otro modelo para solucionar el problema: el método phi (8)(11), llamado también concordancia de independencia-azar. Este es el método recomendado por la Asociación Médica Americana en su guía de usuarios (8), porque se dice que tiene ciertas ventajas basadas en su relación con el índice Odds Ratio. El valor de phi se compara con el mismo criterio de aceptabilidad anterior. En cambio, en Epidemiología sigue vigente el método corregido de Kappa desarrollado por Feinstein AR y Cicchetti DV (11).

Sin embargo, en este enfoque también hay problemas:

1) El método llamado phi en clínica, es lo mismo que el método de Yule (índice de coligación) de ciencias sociales, que se usa desde principios del siglo pasado (12)(13).
2) En ciencias sociales el método phi existe y es diferente al anterior (13).
3) Existe una aproximación entre el método de Yule y el de Kappa ajustado (14).
4) Los tres métodos Kappa, phi y Yule no resuelven definitivamente las paradojas de Feinstein AR y Cicchetti DV (9)(10).
5) Se presentaron tres paradojas más que tampoco son resueltas por estos métodos (15).

Las fallas de los modelos Kappa, phi y Yule se pueden ilustrar con las paradojas siguientes:

1ª paradoja: Ocurre cuando con diferentes valores de concordancia casual, los valores de Kappa, phi y Yule –para idénticos valores de concordancia observada (nivel de concordancia)– pueden ser dos veces más pequeños en una instancia que en otra (9).

2ª paradoja: Ocurre cuando totales marginales desbalanceados producen mayores valores de Kappa, phi y Yule, que totales más balanceados (9).

3ª paradoja: Kappa, phi y Yule pueden ser nulos, no importa el valor obtenido del nivel de concordancia en el experimento realizado (15).

4ª paradoja: Kappa, phi y Yule pueden ser negativos, no importa el valor obtenido del nivel de concordancia en el estudio efectuado (15).

5ª paradoja: Cuando b o c son nulos, Yule tiende a uno (esto es concordancia perfecta), no importa el valor obtenido del nivel de concordancia en el estudio realizado (15).

Las cinco fallas anteriores muestran que ninguno de esos modelos se ajusta al concepto de concordancia clínica, sino al de concordancia estadística. Los valores observados en el experimento se expresan con el nivel de concordancia (λ). Esa es la realidad en la cual debe basarse un estudio clínico. Los modelos mencionados son una teoría matemática que trata de explicar esa realidad. Pero si la teoría muestra fallas como las anteriormente descriptas, entonces se debe buscar otro modelo que no las tenga tal como el modelo de visión dual. Este es el objetivo del presente informe.

Modelo de visión dual

Este procedimiento (6) fue propuesto para solucionar los problemas de los modelos estadísticos orientados al análisis de comparación de dos proporciones apareadas. Y luego fue aplicado para solucionar las cinco paradojas clínicas (15) de los modelos basados en el análisis de la independencia estadística. Consiste en efectuar dos pasos siguiendo un diagrama lógico tal como se muestra en el Apéndice 1.

El concepto clínico básico es: *Dos métodos clínicos concuerdan cuando tienen la misma sensibilidad y especificidad.* Si uno de los dos métodos (por ejemplo el Método 1) es el de referencia, entonces la Tabla de Concordancia se transforma en una Tabla Diagnóstica (o tabla de la verdad) – ver Tabla II. Y se pueden calcular la sensibilidad (S_2) y la especificidad (E_2) del otro método. Si ahora se supone que el otro (Método 2) es el de referencia, se pueden calcular ambos índices para el primer método (S_1 y E_1). La única manera de que cada par de índices coincidan es cuando b = c. Por lo tanto, el primer paso es verificar el supuesto básico, aplicando una visión estadística al problema. Esto es, verificar si el número de los dos tipos de discordancias obtenidos cumple la condición: $b \approx c$. Notar que es el mismo análisis planteado por McNemar. Para ello, el mejor método es el G-*test* porque es el más poderoso estadísticamente hablando (4). Sin embargo, no se puede aplicar cuando una de las discordancias es nula. Entonces, la segunda mejor alternativa es el Q-*test* que no tiene esa dificultad.

Tabla II. *Tabla diagnóstica y sus principales índices de calidad.*

	Enfermedad (resultados reales)		
Resultados del *test*	Sí	No	Total
Positivo (+)	vp *verdadero positivo*	fp *falso positivo*	T+
Negativo (-)	fn *falso negativo*	vn *verdadero negativo*	T-
Total	TD	TnD	N

Donde N es el número de sujetos investigados y
T+ = vp + fp : Total de sujetos diagnosticados positivos
Sensibilidad = vp / TD = S
T- = vn + fn : Total de sujetos diagnosticados negativos
Especificidad = vn / TnD = E
TD = vp + fn : Total de sujetos enfermos
Prevalencia = TD / N
TnD = fp + vn : Total de sujetos no enfermos
Eficiencia = (vp+vn) / N = A

En este primer paso hay dos resultados posibles. Si la hipótesis nula no es rechazada entonces no hay evidencia suficiente como para pensar que la sensibilidad y especificidad de ambos métodos son diferentes. Y por lo tanto se puede pasar a la etapa siguiente. En cambio, si hay evidencia estadística como para creer que la sensibilidad y especificidad son diferentes, esta primera etapa continúa efectuando un análisis más incisivo.

Para hacer el análisis más incisivo hay que tener en cuenta el tipo de enfermedad que se está analizando. Todas las enfermedades se pueden clasificar en tres tipos de acuerdo al riesgo de cometer una la tabla siguiente (1)(15):

Tabla III. *Clasificación de las enfermedades.*

Tipo	*Descripción*	*Ejemplo*	*Índice*
I	Falso negativo más peligroso que falso positivo	Infarto de miocardio	Sensibilidad
II	Falso positivo más peligroso que falso negativo	Cáncer irreversible	Especificidad
III	Restantes casos (ambos errores son peligrosos)	SIDA	Youden (16)

En toda enfermedad que sea curable si es detectada y tratada a tiempo, como por ejemplo, sífilis en primer estadío, infarto, SARS (*Severe Acute Respiratory Syndrome*) (17), etc., el principal peligro para el paciente es que ésta no sea diagnosticada cuando todavía es curable, es decir, un falso negativo. Esta clase de enfermedades se catalogan como del Tipo I y requieren del método clínico con máxima sensibilidad. Toda enfermedad que haya tomado un carácter irreversible en el paciente –sífilis en estadío IV, metástasis cerebral, etc.– tiene como principal peligro para el paciente un falso positivo, y por lo tanto se requiere máxima especificidad. Casos como el SIDA donde ambos errores son peligrosos requieren máximo índice de Youden (o bien máxima eficiencia (1)), serían catalogados como una enfermedad del Tipo III. Aquí un falso positivo es más peligroso para el paciente, pero un falso negativo es más peligroso para la sociedad. Por lo tanto, no es sencillo discernir cuál equivocación es la peor (todo depende del punto de vista del profesional actuante).

La idea básica es que no se puede aplicar un mismo criterio - maximizar *diagnostic accuracy* o eficiencia (8) (11)(18) para cualquier enfermedad sin tomar en cuenta el grado de avance de la misma en el paciente.

Entonces, en el análisis más incisivo se analiza cuál es la variación potencial de la sensibilidad, que podría tener lugar debido al intercambio de métodos en las enfermedades del Tipo I. Mientras que el cambio potencial de la especificidad se analiza para las enfermedades del Tipo II, y en el índice de Youden para los casos del Tipo III. Cuando este cambio potencial no sea aceptable desde el punto de vista clínico entonces la concordancia debería ser rechazada, y cuando sea aceptable se pasa a realizar la segunda etapa del método.

Por ejemplo, se puede obtener un nivel de concordancia del 90% pero eso no significa que las variaciones de sensibilidad y especificidad no sean importantes. Tomando los ejemplos de [6-Apéndice 2]: en el Caso A la sensibilidad puede variar un 16%, lo que significa que 16 pacientes de cada 100 no serán detectados, y esto puede ser dañino en las enfermedades del Tipo I. En el Caso B la especificidad puede variar un 26% lo cual sería peligroso para las del Tipo II. Existe el peligro de informarle a 26 pacientes de cada 100 que tienen una enfermedad incurable, cuando no es así.

En esta primera etapa cuando la concordancia sea rechazada estadísticamente por el G-*test* (o bien por el Q-*test*) se tiene una clara indicación de variación en sensibilidad y especificidad. Sólo el clínico puede juzgar si esta variación potencial puede ser peligrosa para el paciente. Por otra parte, cuando no sea rechazada, o bien el clínico considere que la variación potencial no es importante, entonces se debe pasar a la etapa siguiente. El hecho de que en esta etapa la concordancia no sea rechazada, no implica que sea aceptable porque el nivel observado puede ser bajo. Por eso se necesita analizar el problema con una visión dual y aplicar siempre ambas etapas del método – ver [6-Apéndice 2].

La segunda etapa consiste en juzgar si el nivel de concordancia alcanzado es suficiente como para efectuar el intercambio de métodos. La idea es compararlo con un nivel mínimo admisible ($\lambda_{crítico}$) definido por los clínicos. Por ejemplo, un nivel del 90% puede ser aceptable para ciertas enfermedades, pero no ser suficiente para otras más peligrosas como el SIDA, SARS, ciertos tipos de cáncer, etc., que pueden requerir niveles superiores al 95% (6).

El objetivo de esta segunda etapa del método de visión dual es verificar que el nivel de concordancia (λ) sea lo suficientemente grande como para tener una concordancia aceptable desde un punto de vista clínico. Para ello basta encontrar el intervalo de confianza del 95% para el nivel de concordancia alcanzado y ver si el valor crítico es menor o cae dentro del intervalo, en cuyo caso se considera aceptable la concordancia observada entre ambos métodos y el intercambio puede ser efectuado.

El intervalo de confianza del 95% para el nivel de concordancia se puede estimar con:

$$\text{Límite superior} = \lambda + 1.96\ SE\ (\lambda)$$
$$\text{Límite inferior} = \lambda - 1.96\ SE\ (\lambda)$$

Donde $SE\ (\lambda) = [\lambda\ (100 - \lambda)\ /\ (N)]^{1/2}$, y $E\ (\lambda) \approx \lambda$, con tal de que N sea lo suficientemente grande (N > 50). Para muestras pequeñas se puede usar el modelo *Student* en lugar del de Gauss (1).

Está disponible gratis un algoritmo para resolver todos los cálculos de este nuevo procedimiento (19) (20). Sólo se necesita ingresar las cuatro frecuencias para la tabla y el *criterio clínico* adoptado para obtener la decisión final sobre la concordancia. De esta forma

se pueden hacer pequeños cambios en el criterio clínico, para ver qué pasa con la concordancia de una manera fácil y rápida. Esto puede ser de ayuda cuando se deban investigar los valores críticos a ser adoptados para cada enfermedad.

Ejemplos de aplicación

Problema 1: En este problema se analizan 400 muestras con dos métodos, los resultados obtenidos se simulan en tres casos (Tabla IV). En el Caso 1 la concordancia es rechazada por la evidencia estadística directamente. Entonces se debe hacer el análisis más incisivo. Éste muestra que la variación potencial de sensibilidad es del 5,6%, de especificidad del 5,4% y de Youden del 0,2%. De acuerdo al criterio clínico de la etapa 2 el intervalo de confianza del 95% para el nivel de concordancia es (89,3; 94,7)%. Por lo tanto, si el valor crítico es del 95% la concordancia debe ser rechazada y el intercambio de métodos no es aceptable. Debe notarse que, si se tratase de una enfermedad del Tipo III, una variación del 0,2% del índice de Youden es muy pequeña como para ser tomada en cuenta. Y si, además, el criterio clínico establece un valor crítico del 90% entonces ahora la concordancia sería aceptable. Todo depende del tipo de enfermedad y del criterio para establecer el valor umbral. Si se aplicasen los modelos basados en la independencia estadística de los métodos resultaría: Kappa = 0,84, phi = 0,842 y Yule = 0,85. Entonces para los tres casos la concordancia sería "casi perfecta" sin que importe el tipo de enfermedad tratada y toda la discusión clínica anterior pasaría inadvertida (estadística reemplaza a clínica).

En el Caso 2 la concordancia es aceptable desde el punto de vista estadístico. Las variaciones potenciales de los tres índices clínicos son menores al 3%. En la segunda etapa resulta que el nivel de concordancia observado tiene un intervalo del 95% dado por: (79,9; 87,1)%, que no resulta suficiente para un nivel crítico del 95% o del 90% y, por lo tanto, la concordancia se rechaza. Aquí no hay dudas del rechazo clínico de la concordancia, sin embargo Kappa ≈ Yule ≈ phi = 0,67, lo que significa una concordancia "substancial". Otra vez el criterio clínico pasaría inadvertido ante la concordancia estadística.

En el Caso 3 la concordancia es aceptable desde ambos puntos de vista. No hay evidencia como para pensar que sensibilidad y especificidad varíen significativamente. Desde el punto de vista clínico se obtuvo un nivel de concordancia entre 94,4% y 98,1% con un nivel de confianza del 95% lo cual es una concordancia muy buena para casi todos los casos. Por su parte, los modelos basados en la independencia también coinciden pues: Kappa ≈ Yule ≈ phi = 0,93. Este es un caso donde la concordancia es muy buena para todos los modelos.

Problema 2: En este problema se muestran tres casos paradójicos desde el punto de vista clínico, si el análisis se hace con los modelos basados en la independencia estadística (Tabla V).

En el Caso 4 se observa a simple vista que tener tan sólo 10 concordancias en 100 casos es inaceptable desde un punto de vista clínico. No se necesita de ningún análisis estadístico para concluir que el intercambio de métodos no debe hacerse. Sin embargo, si se aplica el Modelo de Yule (llamado phi en medicina (8)(11) se obtiene un valor que tiende a 1. Lo que significaría una concordancia perfecta.

En el Caso 5 se observa a simple vista que tener 390 concordancias en 400 casos es bastante aceptable (λ = 97,5%). Sin embargo, Kappa, phi y Yule son negativos.

En el Caso 6 se observa un nivel de concordancia muy alto del 96,8% y, sin embargo, Kappa, Yule y phi son nulos, lo que indicaría una concordancia muy pobre.

Estos tres casos paradójicos se pueden resolver fácilmente aplicando el modelo de visión dual. En el Caso 4 donde a simple vista la concordancia es inaceptable, el

Tabla IV. *Tres casos para analizar la concordancia.*

Caso 1

		Método 1		
		Sí	No	
Método 2	+	180	22	202
	+	10	188	198
	−	190	210	400

Gadj = 4.54 > 3,841
Rechazo estadístico
Concordancia rechazada

$\lambda_{crítico}$ = 95%

Caso 2

		Método 1		
		Sí	No	
Método 2	+	160	36	196
	+	30	174	204
	−	190	210	400

Gadj = 0.54 < 3,841
No hay rechazo estadístico

λ 95% CI (80,87)%
$\lambda_{crítico}$ > 87%
Concordancia rechazada

Caso 3

		Método 1		
		Sí	No	
Método 2	+	185	10	195
	+	5	200	205
	−	190	210	400

Gadj = 1.64 < 3,841
No hay rechazo estadístico

λ 95% CI (94.4,98.0)%
$\lambda_{crítico}$ cae en el intervalo
Concordancia aceptada

Tabla V. *Tres casos paradójicos en el análisis de la concordancia.*

Caso 4

	Método 1		
	Sí	No	
Método 2	2	90	92
+	0	8	8
–	2	98	100

Q = 90
Rechazo estadístico
Concordancia rechazada

$\lambda_{crítico}$ = 95%

Caso 5

	Método 1		
	Sí	No	
Método 2	390	5	395
+	5	0	5
–	395	5	400

Q = 0
No hay rechazo estadístico

λ 95% CI (96,99)%
$\lambda_{crítico}$ < 96%
Concordancia aceptada

Caso 6

	Método 1		
	Sí	No	
Método 2	3.600	60	3.660
+	60	1	61
–	3.660	61	3.721

Q = 0
No hay rechazo estadístico

λ 95% CI (96,97)%
$\lambda_{crítico}$ < 96%
Concordancia aceptada

rechazo estadístico es muy fuerte Q = 90 > 3,841, y el rechazo clínico es evidente porque el nivel de concordancia observado está entre 4% y 16% lo cual es inaceptable. Si se hace el análisis más incisivo se ve que las variaciones potenciales son muy grandes y peligrosas (variación de: sensibilidad 97,8%, de especificidad 91,8%). Todo lo cual está en contradicción con un valor "casi perfecto" mostrado por el índice de Yule (notar que OR es infinito).

En cambio, en los Casos 5 y 6 la situación es diametralmente opuesta. En el Caso 5, Q = 0 y no hay rechazo estadístico, además, la concordancia observada se estima entre el 96% y 99%, lo cual es ideal. Sin embargo, esto contradice a valores negativos de Kappa, phi y Yule. En el Caso 6 no hay rechazo estadístico porque Q = 0. Lo mismo que las variabilidades potenciales de los tres índices diagnósticos principales. Y la visión clínica del problema muestra un nivel de concordancia alcanzado entre el 96 y 97%, Suficiente para casi todas las enfermedades. Sin embargo, Kappa, phi y Yule son nulos.

Conclusiones

La actual propuesta se puede resumir en dos etapas:

Etapa 1: Verificar que la sensibilidad y especificidad sean similares en ambos métodos usando el G-*test* (o el Q-*test*). Cuando la concordancia verifique esta condición, se debe realizar la etapa siguiente. Si no la verifica hay que efectuar un análisis más incisivo para ver qué pasa con las variabilidades potenciales de los principales índices diagnósticos.

Etapa 2: Verificar la condición clínica. Esto es, que el valor obtenido del nivel de concordancia λ no sea menor que $\lambda_{crítico}$ (esto es, que no sea menor que el límite inferior del intervalo)

Clínicamente la concordancia será aceptable cuando se verifiquen estas dos condiciones.

Este procedimiento sólo se puede usar si los resultados del método de diagnóstico (o un *test* clínico) se pueden transformar en un caso binario. Cuando sea posible obtener el verdadero estado del paciente conviene usar una Tabla Diagnóstica (la mejor opción para analizar la capacidad o calidad de los *tests* clínicos). Cuando no se puedan obtener los valores verdaderos, el procedimiento de visión dual puede ser usado para ver si el nuevo método puede reemplazar al viejo, pero no para decidir cuál de ellos es la mejor opción. Usualmente, una magnitud clínica puede ser transformada en un caso binario adoptando un punto de corte adecuado para separar un caso positivo de uno negativo. Por eso, este procedimiento es más general que los habituales, y se puede emplear en casi cualquier *test* clínico.

Entonces, para tener una concordancia clínica aceptable entre dos métodos apareados los principales supuestos son:

- Es posible transformar los resultados del método de diagnóstico en un caso binario.
- La sensibilidad y especificidad de ambos métodos deben ser iguales.
- La variabilidad potencial de sensibilidad, especificidad y Youden no debe ser peligrosa de acuerdo a la enfermedad estudiada.
- El nivel de concordancia observado debe ser lo suficientemente alto.

Las principales ventajas son:

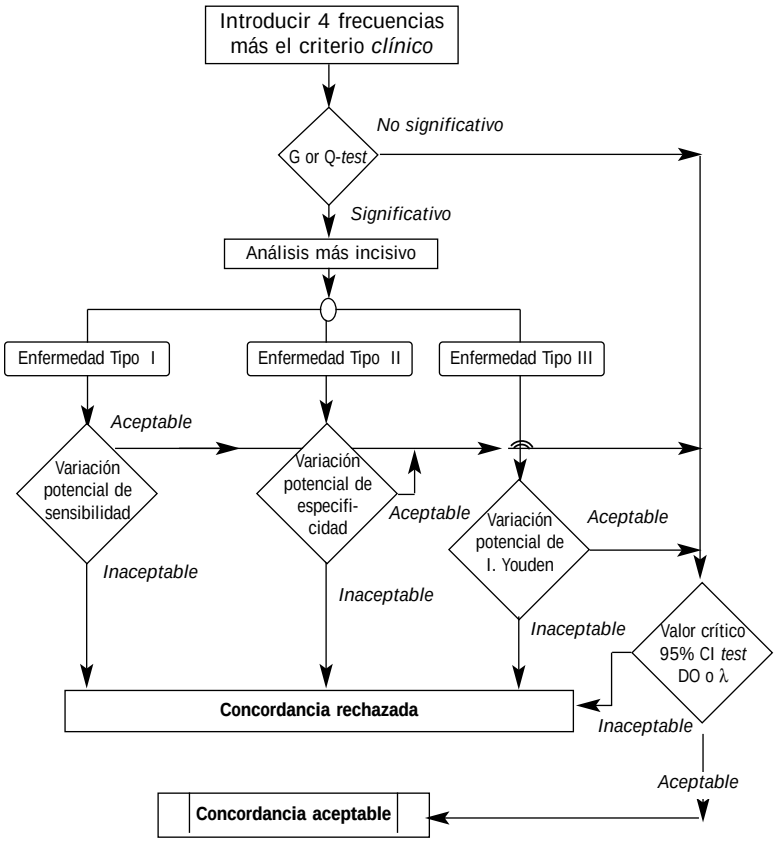
- La decisión final acerca de la concordancia se basa en un criterio clínico, en lugar de uno estadístico, y como la responsabilidad de la decisión es clínica, no se la puede delegar en estadística.
- Explica la diferencia entre los conceptos de concordancia estadística y clínica.
- Este procedimiento es más confiable que cual-

quiera de los *tests* estadísticos usuales, porque resuelve casos paradójicos desde el punto de vista clínico (como los cinco ejemplos vistos).
– Tiene una aplicación más general en el campo clínico.
– Los valores verdaderos no son necesarios por lo que el costo disminuye.
– Los problemas de sesgo debidos al espectro o a la selección (16) se evitan porque los mismos individuos se miden dos veces.
– Puede ser una solución para comparar dos métodos de referencia.

La realidad clínica se basa en los datos medidos (nivel de concordancia). Cuando los modelos teóricos (G-*test*, Q-*test*, Chi cuadrado-*test*, Kappa, phi o Yule) no basten para explicarla cabalmente, entonces se debe buscar otro modelo que solucione esos problemas, tal como el método de visión dual propuesto como solución en este informe.

APÈNDICE 2.

Diagrama de flujo para el procedimiento de visión dual.



Referencias bibliográficas*

1. Azzimonti Renzo JC, Bioestadística aplicada a Bioquímica y Farmacia. Ed. Universitaria de la UNaM, 2ª Edición, 2003. Disponible en http://www.fceqyn.unam.edu.ar/bio. Fecha de consulta 16-7-2003.
2. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947; 12: 153-7.
3. Armitage P, Berry G. Statistical Methods in Medical Research. 2ª Ed. Oxford: Blackwell Scientific Publication; 1987.
4. Sokal RR, Rohlf J. Biometry: the principles and practice of statistics in biological research. 2ª Ed. W. Freeman & Co.; 1981.
5. Conover WJ. Practical Nonparametric Statistics. 3rd Edition, New York: John Wiley & Sons; 1999.
6. Azzimonti Renzo JC. The agreement between two diagnostic methods in binary cases: a proposal. Scand J Clin Lab Invest 2002; 62: 391-8.
7. Uebersax JS. Statistical Methods for Rater Agreement; 2003. Disponible en http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm. Fecha de consulta 15-6-2003.

* Como el autor falleció durante el período de evaluación del presente trabajo, algunas citas no pudieron ser modificadas según las normas de ABCL.

8. Guyatt G, Rennie D. User's Guides to the Medical Literature, JAMA & Archives Journals, 2002, AMA Press.
9. Feinstein AR, Cicchetti DV. High agreement but low kappa, I: the problems of two paradoxes. J Clin Epidemiol 1990, 43: 543-9.
10. Cicchetti DV, Feinstein AR. High agreement but low kappa, II: resolving the paradoxes, J Clin Epidemiol 1990; 43: 551-8.
11. Feinstein AR. Principles of Medical Statistics. Chapman and Hall, 2002.
12. Liebetrau Albert M. Measures of association, Newbury Park, CA: Sage Publications. Quantitative Applications in the Social Sciences, Series No. 32, 1983.
13. Garson, DG, Social Science Computer Review, (http://www2.chass.ncsu.edu/garson/pa765/assoc2x2.htm), 2003.
14. Walter SD. Hoehler's adjusted kappa is equivalent to Yule's Y, 2001, J Clinical Epidemiol 54: 1072.
15. Azzimonti Renzo JC. Failures of Common Measures of Agreement in Medicine and the Need for a Better Tool: Feinstein's Paradoxes and the Dual Vision Method, Scand J Clin Lab Invest 2003; 63: 207-16.
16. Youden WJ. Index for rating diagnostic tests. Cancer 1950; 3: 32-5.
17. Lee C. Diagnostic Test for Killer Pneumonia Developed. Medscape from WebMD, Reuter, April 10, 2003
18. Knottnerus JA, van Weel C, Muris JWM. Evaluation of diagnostic procedures. Br Med J 2002; 324: 477-80.
19. Azzimonti Renzo JC. Disponible en www.bioestadistica.com.ar/concordancia.xls, 2003.Fecha de consulta 1-7-2003.
20. Medal, Algorithms for medicine, Release 10.1. Chapter Contributed Algorithms, 2003. Disponible en www.medal.org. Fecha de consulta 10-6-2003.

Aceptado para su publicación el 29 de julio de 2004