Validez, Integridad y Monitoreo para la Enfermedad

Nigel Paneth M.D., MPH

College of Human Medicine Michigan State Univ.

paneth@msu.edu

Nicolás Padilla, M.D.

Universidad de Guanajuato, México

padilla@celaya.podernet.com.mx

Inicialmente en el sitio: www.pitt.edu/~super1/

Medición de Integridad

Terminología

Integridad es análogo a precisión

Validez es análogo a seguridad

Integridad es cuan bien un observador clasifica al mismo individuo bajo diferentes circunstancias.

Validez es cuan bien una prueba refleja su resultado bajo otra prueba de mayor seguridad conocida.

Integridad y Validez

Integridad incluye:

- **evaluación por el *mismo observador en diferente tiempo* - Integridad intraobservador**
- **evaluación de *diferentes observadores al mismo tiempo* - Integridad interobservador**

Integridad asume que todas las pruebas u observadores son iguales; validez asume que hay un estándar de "oro" con el cual la prueba o el observador será comparado.

Evaluando Integridad

¿Cómo evaluamos la integridad?

Una forma es observar el porcentaje de concordancia.

- Porcentaje de concordancia es la proporción de todos los diagnósticos clasificados en la misma manera por los dos observadores.

Ejemplo

A dos médicos se les dan 100 radiografías para verlas independientemente y son preguntados acerca de si neumonía está presente o ausente. Cuando ambos grupos de diagnósticos son comparados, se encuentra que 95% de los diagnósticos son los mismos.

- **¿Es el diagnóstico *íntegro*? ¿El porcentaje de concordancias es sufiente para indicar integridad?**
- **¿95% de concordancia entre los dos médicos y la ausencia o presencia de enfermedad en una muestra de 100 pacientes siempre indica buena concordancia?**
- **¿Te senírías tranquilo de que tu hospital estaba haciendo un trabajo consistente en las lecturas de rayos X si muestran un 95% de integridad?**

Compare las dos tablas:

Tabla 1

		MD#1	
		Si	No
MD#2	Si	1	3
	No	2	94

Tabla 2

		MD#1	
		Si	No
MD#2	Si	43	3
	No	2	52

En ambos ejemplos, los médicos concuerdan en el 95%. ¿Son los dos médicos igualmente íntegros en las dos tablas?

- **¿Cuál es la diferencia esencial entre las dos tablas?**

- **El problema surge de la facilidad de concordancia en eventos comúnes (v.gr. No teniendo neumonía en la tabla 1).**
- **Una medida de concordancia deberá tomar en cuenta la “facilidad” de concordancia debido sólo al azar.**

Uso de Kappa para Evaluar Integridad

Kappa es una prueba ampliamente usada de concordancia inter o intra observadores (o integridad) el cual corrige por concordancia al azar.

Kappa varía de + 1 a - 1

+ 1 significa que los dos observadores concuerdan perfectamente. Ellos clasifican a todos en la misma forma.

0 significa que no hay relación entre las clasificaciones de los dos observadores.

- 1 significa que los dos observadores clasifican exactamente lo opuesto. Si un observador dice si, el otro siempre dice no.

Guía del Uso de Kappa en Epidemiología y Medicina:

Kappa > .80 es considerado excelente

Kappa .60 - .80 es considerado bueno

Kappa .40 - .60 es considerado regular

Kappa < .40 es considerado pobre

Primera Forma de Calcular Kappa:

1. Calcule la concordancia *observada* (celdas en las cuales el observador concuerdan/total de celdas). En tablas 1 y 2 es de 95%

2. Calcule la concordancia *esperada* (concordancia al azar) basado en los totales marginales

En la Tabla 1 los Totales Marginales son:

OBSERVADOS		MD#1		
		Si	No	
MD#2	Si	1	3	4
	No	2	94	96
		3	97	100

OBSERVADOS		MD #1		
		Si	No	
MD#2	Si	1	3	4
	No	2	94	96
		3	97	100

ESPERADOS		MD #1		
		Si	No	
MD#2	Si			4
	No			96
		3	97	100

- ¿Cómo calculamos el N esperado por azar en cada celda?
- Asumimos que cada celda refleja las distribuciones marginales, v.gr. La proporción de respuestas si y no deberá ser la misma **dentro** de la tabla de cuatro celdas como en los totales **marginales**.

Para hacer esto, encontramos la proporción de respuestas en cada columna (3% y 97%, si y no respectivamente para MD #1) o renglón (4% y 96% si y no respectivamente para MD #2) totales marginales y aplica una de las dos proporciones al otro total marginal. Por ejemplo, 96% de los totales de renglones en la categoría "No". Por lo tanto, por azar 96% of MD #1 es "No" debería estar en la columna "No". 96% de 97 es 93.12.

ESPERADOS		MD#1		
		Si	No	
MD#2	Si			4
	No		93.12	96
		3	97	100

Por restas, llene todas las otras celdas y cada distribución de si/no refleja la distribución marginal. Cualquier celda podría ser usada para hacer el cálculo, debido a que una vez que la celda es especificada en una tabla 2 x 2 con distribuciones marginales fijadas, todas las otras celdas son también especificadas.

ESPERADOS		MD #1		
		Si	No	
MD#2	Si	0.12	3.88	4
	No	2.88	93.12	96
		3	97	100

Ahora puede ver que sólo por el azar, 93.24 de las 100 observaciones deberían concordar por los dos observadores (93.12 + 0.12)

ESPERADO		MD #1		
		Si	No	
MD#2	Si	0.12	3.88	4
	No	2.88	93.12	96
		3	97	100

Ahora comparemos la concordancia actual con la esperada

- **Concordancia esperada es 6.76% de la perfecta concordancia del 100% (100 – 93.24)**
- **Concordancia actual fue 5.0% de la perfecta concordancia del 100% (100 – 95)**
- **Así, nuestros dos observadores fueron 1.76% mejor que el azar, pero si ellos hubieran concordado perfectamente, deberían haber estado 6.76% mejor que el azar. Ellos están sólo cerca de ¼ mejor que el azar (1.76/6.76)**

Fórmula para el Cálculo de Kappa de la Concordancia Esperada

$$\frac{\text{C. observada - C. esperada}}{\text{1 - C. esperada}}$$

$$\frac{95\% - 93.24\%}{1 - 93.24\%} = \frac{1.76\%}{6.76\%} = .26$$

C=concordancia

¿Es buena la Kappa de 0.26?

Kappa > .80 es considerada excelente

Kappa .60 - .80 es considerada buena

Kappa .40 - .60 es considerada regular

Kappa < .40 es considerada pobre

En el segundo ejemplo, la concordancia observada fue también de 95%, pero los totales marginales fueron muy diferentes

ACTUAL		MD #1		
		Si	No	
MD#2	Si			46
	No			54
		45	55	100

Usando el mismo procedimiento como antes, calculamos el esperado de N, basado en los totales marginales. Por ejemplo, la celda abajo a la derecha es 54% de 55, el cual es 29.7

ACTUAL		MD #1		
		Si	No	
MD#2	Si			46
	No		29.7	54
		45	55	100

Y, por restar las otras celdas están como abajo. Las celdas que indican concordancia son resaltadas en amarillo y suman 50.4%

ACTUAL		MD #1		
		Si	No	
MD#2	Si	**20.7**	25.3	46
	No	24.3	**29.7**	54
		45	55	100

Entre las dos concordancias (C) en la fórmula:

$$\frac{\text{C. observada - C. esperada}}{\text{1 - Concordancia esperada}}$$

$$\frac{95\% - 50.4\%}{1 - 50.4\%} = \frac{44.6\%}{49.6\%} = .90$$

En este ejemplo, los observadores tienen la misma % de concordancia, pero ahora ellos están muy diferentes del azar.

Kappa de 0.90 es considerado excelente

Otra Forma de Calcular Kappa

$$\frac{2(AD - BC)}{N_1N_4 + N_2N_3}$$

donde las Ns son los totales marginales etiquetados así:

		MD#1		
		Si	No	
MD#2	Si	**A**	**B**	N_1
	No	**C**	**D**	N_2
		N_3	N_4	**total**

Mire la tabla de la diapositiva 7.

Para Tabla 1:

$$\frac{2(94 \times 1 - 2 \times 3)}{4 \times 97 + 3 \times 96} = \frac{176}{676} = .26$$

Para Tabla 2:

$$\frac{2(52 \times 43 - 3 \times 2)}{46 \times 55 + 45 \times 54} = \frac{4460}{4960} = .90$$

Note paralelismos entre:

LA RAZON DE MOMIOS (ODDS RATIO)

LA CHI-CUADRADA ESTADISTICA

LA KAPPA ESTADISTICA

Note que los productos cruzados de la tabla de cuatro celdas y su relación a los totales marginales , son centrales en las tres expresiones

Validez y Monitoreo

Tres mediciones claves de la validez

1. SENSIBILIDAD
2. ESPECIFICIDAD
3. VALORES PREDICTIVOS

Tabla de Cuatro Celdas para Evaluar la Relacion Prueba-Enfermedad

		Estado de enfermedad		
		+	-	
Resultado de la prueba	+	Enfermo, prueba positiva	No enfermo, prueba positiva	PRUEBA POSITIVA
	-	Enfermo, prueba negativa	No enfermo, prueba negativa	PRUEBA NEGATIVA
		ENFERMO	NO ENFERMO	

Sensibilidad

Nos indica que tan bien una prueba positiva detecta la enfermedad.

Es definida como la *fracción de los enfermos que dan positivos en la prueba.*

Su complemento es la tasa de falsos negativos, definida como la fracción de los enfermos que dan negativo con la prueba.

Sensibilidad y tasa de falsos negativos suman uno.

Especificidad

Nos indica que tan bien una prueba negativa no detecta la enfermedad.

Es definida como la *fracción de los no enfermos cuya prueba fue negativa*.

Su complemento es la tasa de falsos positivos, definida como la fracción de los no enfermos cuya prueba fue positiva.

Especificidad más tasa de falsos positivos dan uno.

Valores Predictivos

Valor predictivo *positivo* es la proporción de toda la gentes con pruebas positivas quienes tienen la enfermedad.

Valor predictivo *negativo* es la proporción de toda las personas con pruebas negativas quienes no tienen la enfermedad.

En general, el valor predictivo positivo es el más usado. Valor predictivo positivo y sensibilidad son, quizá, los dos más importantes parámetros en el entendimiento de la utilidad de una prueba bajo condiciones de campo.

Puntos Clave a Recordar

Sensibilidad, especificidad, falsos positivos y falsos negativos son denominadores a personas enfermas o no enfermas. (Usándolos, en totales de columnas).

Por lo contrario, valores predictivos son denominadores para el status de la prueba, positivo o negativo (usandolos en totales de renglones).

Sensibilidad y especificidad no varían de acuerdo a la prevalencia de la enfermedad en la población. Valores predictivos de una prueba, sin embargo, son ALTAMENTE DEPENDIENTES de la prevalencia ode la enfermedad en la población.

Calculando Sensibilidad, Especificidad y Valores Predictivos

Una prueba es usada en 50 personas con la enfermedad y 50 personas sin ella. Estos son los resultados:

		Enfermedad		
		+	-	
Prueba	+	48	3	51
	-	2	47	49
		50	50	100

		Enfermedad		
		+	-	
Prueba	+	48	3	51
	-	2	47	49
		50	50	100

Sensibilidad = 48/50 = 96%

Especificidad = 47/50 = 94%

Valor predictivo positivo = 48/51 = 94%

Valor predictivo negativo = 47/49 = 96%

Ahora, apliquemos esta prueba en una población donde 2% de las personas tienen la enfermedad, no el 50% como en el ejemplo anterior. Asuma que hay 10,000 personas, y la misma sensibilidad y especificidad de antes, 96% y 94% respectivamente.

		Enfermedad		
		+	-	
Prueba	+	192	588	780
	-	8	9,212	9,220
		200	9,800	10,000

¿Cuál es el valor predictivo positivo ahora?

192/780 = 24.6%

Cuando la prevalencia de la enfermedad es 50%, 94% de las pruebas positivas indican enfermedad. Pero cuando la prevalencia es sólo 2%, menos de 1 en cuatro pruebas señalan una persona con la enfermedad y 2% actualmente deberían representar una muy común enfermedad. Falsos positivos tienden a esconderse en verdaderos positivos en poblaciones, debido a que muchas enfermedades que probamos son raras.

Cambiando el Límite para una Prueba

Cuando enfermedad es definida por un límite en una prueba contínua, las características de la prueba pueden ser alteradas cambiando el límite o punto de corte.

Disminuyendo el límite mejora la sensibilidad, pero el precio es la disminución de la especificidad (v.gr. Más falsos positivos).

Aumentando el límite mejora la especificidad, per el precio es la disminución de la sensibilidad (v.gr. Más falsos negativo. Esto es especialmente importante cuando la distribución de una característica es *unimodal*, como la tensión arterial, colesterol, peso, etc. (debido a que la zona gris es grande)

Problemas con Monitoreo

1. ¿Tenemos el correcto umbral?
2. ¿Hay un tratamiento verdaderamente efectivo disponible para la enfermedad?
3. ¿Es el tratamiento más efectivo en casos monitoreados que en los no monitoreados?
4. ¿Cuáles son los eventos adversos del proceso de monitoreo?
5. ¿Cuán eficiente es el monitoreo?v.gr. ¿cuánta gente tiene que ser monitoreada para encontrar un caso?

Ejemplo

Un ensayo aleatorizado para evaluar un programa de monitoreo para cáncer de cólon es implementado. El grupo de la intervención tiene regular monitoreo, el grupo control es dejado a merced de sus recursos.

Después de cinco años, se encontró que:

1. Más casos son descubiertos en el grupo monitoreado que en el grupo control.
2. Los casos son descubiertos a un más temprano estadío del cáncer en el grupo monitoreado.
3. La sobrevida a cinco años es mayor en las personas con cáncer monitoreadas.

¿Podemos concluir que este programa de monitoreo es necesariamente efectivo?

No, el programa no necesariamente será efectivo.

Los aparantes beneficios sólo demuestran los efectos del SESGO DEPENDIENTE DEL TIEMPO.

Si es posible diagnosticar una condición en forma temprana, pero no mejorar la sobrevida después del diagnóstico, el programa de monitoreo tendrá una sobre representación de casos diagnosticados más temprano, cuya sobrevida será incrementada por exactamente el tiempo en que su diagnóstico fue hecho en forma más temprana por el programa de monitoreo. Así, ellos no se han beneficiado, pero la cantidad de tiempo que ellos saben que tienen cáncer ha aumentado.

Considere cuánto cambia el tiempo de diagnóstico en el monitoreo en el escenario de abajo:

grupo sin motorear:

	Dx				Muerte
Edad50	51	52	53	54	55

Grupo monitoreado:

	Dx				Muerte
Edad 50	51	52	53	54	55

Otros Sesgo en Monitoreo: Sesgo de Longitud de Tiempo

- Muchas enfermedades crónicas, especialmente cánceres, no progresan con la misma rapidez en todos los pacientes.
- Cualquier grupo de enfermos incluirá algunos en quienes la enfermedad se desarrolla más lento y algunos en quienes se desarrolla más rápido.
- Monitoreo preferencialmente incluirán enfermedades de desarrollo lento (mayor oportunidad de ser monitoreados) y que usualmente tienen mejor pronóstico.

En el escenario previo, la incidencia de la enfermedad es inicialmente más alta, diagnóstico es hecho más temprano, estadio de diagnóstico es más temprano y la duración de sobrevida desde el diagnóstico es más larga. Todos ellos dan la impresión de beneficiarse del monitoreo.

Sin embargo, no se benefician, como la muerte no es pospuesta.

La sóla evidencia de efectividad de un programa de monitoreo es una reducción de la morbilidad o mortalidad específica por edad total, idealmente demostrada en un ensayo aleatorizado.