Psicothema 2012. Vol. 24, nº 1, pp. 133-141
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Procedimientos para realizar meta-análisis de la precisión de instrumentos de clasificación binaria

Juan Botella y Huiling Huang
Universidad Autónoma de Madrid

La evaluación de la precisión en la clasificación binaria debe contemplar dos indicadores no independientes: verdaderos positivos y falsos positivos. Se han propuesto varios índices. Estos han sido estimados en algunos tests para la detección temprana o cribaje. Resumimos y revisamos los principales métodos propuestos para realizar meta-análisis que evalúan la precisión de este tipo de instrumentos. Se aplican a los resultados de 14 estudios que informan de estimaciones de la precisión del test AUDIT. La agregación directa no permite el uso de los procedimientos meta-analíticos; la estimación separada de la sensibilidad y la especificidad no reconoce que no son independientes; el método de la curva ROC-resumen trata la precisión y el umbral como factores fijos y tiene limitaciones para manejar el papel potencial de las covariables. Los modelos Normal Bivariado y Jerárquico de la curva ROC Resumen son estadísticamente rigurosos y son capaces de incorporar las covariables adecuadamente. Ambos métodos permitieron analizar la asociación entre el género de la muestra y el comportamiento del AUDIT.

Procedures for performing Meta-Analyses of the accuracy of tools for binary classification. The assessment of accuracy in binary classification tools must take into account two non-independent rates: true positives and false positives. A variety of indices have been proposed. They have been estimated for tests employed for early detection or screening purposes. We summarize and review the main methods proposed for performing a meta-analysis that assesses the accuracy of this type of tools. They are applied to the results from 14 studies that report estimates of the accuracy of the AUDIT. The method of direct aggregation does not allow the use of meta-analytic procedures; the separate estimation of sensitivity and specificity does not acknowledge that they are not independent; the SROC method treats accuracy and threshold as fixed effects and has limitations to deal with the potential role of covariates. The Normal Bivariate (NB) model and the Hierarchical Summary ROC (HSROC) model are statistically rigorous and can deal with the covariates properly. They allowed analyzing the association between the gender composition of the sample and the way the test AUDIT behaves in the example.

A veces se utilizan instrumentos para clasificar a los individuos en dos categorías; por ejemplo, para cribar o hacer una primera clasificación, rápida y sencilla. Las valoraciones de la eficacia de estos instrumentos se pueden integrar mediante procedimientos de meta-análisis (Sánchez-Meca y Botella, 2010). Aunque estos procedimientos son ampliamente utilizados en otras disciplinas, especialmente en Medicina, en Psicología son bastante desconocidos, lo cual motiva la revisión que presentamos en este artículo y que pretende su divulgación entre los psicólogos.

Instrumentos de clasificación binaria

Los resultados de estos estudios se resumen en tablas 2×2 (tabla 1). En las filas aparecen los resultados del instrumento y en las columnas las categorías «verdaderas», según un criterio independiente e incuestionable (*gold standard*). Por ejemplo, el AUDIT (*Alcohol Use Disorders Identification Test*; Babor, Higgins-Biddle, Saunders y Monteiro, 2001) es un test breve para detectar el abuso del alcohol. Se considera como positivo un resultado de 8 puntos. Sería un instrumento perfecto si proporcionase una clasificación «positiva» (X≥8) de todos los individuos con problemas de abuso y «negativa» (X<8) de todos los demás. Llamaremos casos *Target* (T) a los que cumplen con la condición que queremos detectar y casos *Normales* (N) a los que no la cumplen.

Tabla 1
Tabla de contingencia entre la clasificación proporcionada por el test (positivo o negativo) y el grupo de pertenencia (estatus verdadero del caso)

	Caso	
Resultado del Test	**Target**	**Normal**
Positivo (+)	*VP*	*FP*
Negativo (-)	*FN*	*VN*
	N_T	N_N

Fecha recepción: 9-3-11 • Fecha aceptación: 1-7-11
Correspondencia: Juan Botella
Facultad de Psicología
Universidad Autónoma de Madrid
28049 Madrid (Spain)
e-mail: juan.botella@uam.es

Las cuatro frecuencias involucradas son: *VP* (verdaderos positivos), *FN* (falsos negativos), *FP* (falsos positivos) y *VN* (verdaderos negativos). En un instrumento de eficacia perfecta *FP* y *FN* son nulas: todos los *T* son detectados por el test, pero ningún *N* da un resultado positivo.

Este esquema es similar al de la Teoría de la Detección de Señales (TDS; Swets, 1964; Wickens, 2001), muy empleada en Psicología (Logan, 2004; Swets, 1996; Swets, Dawes y Monahan, 2000). En el contexto de la TDS más conocido se asume una variable latente de evidencia con distribuciones normales de igual varianza (figura 1a). Un punto de corte (X_c) o *umbral* define la regla de clasificación de cada observación en *Señal* y *Ruido* (los *T* y *N* del contexto diagnóstico). Diferentes umbrales (desplazamientos de X_c por el eje de abscisas de la figura 1a) generan diferentes probabilidades de que el test proporcione una respuesta positiva, tanto en los *T* (verdaderos positivos) como en los *N* (falsos positivos). Los cocientes VP/N_T y FP/N_N son las proporciones de verdaderos positivos (*PVP*) y de falsos positivos (*PFP*) [estimaciones de las probabilidades P(+|T) y P(+|N)].

La *precisión* de un instrumento es el grado en que clasifica correctamente. Será mayor cuanto más separadas estén las distribuciones (figura 1a). En la TDS este parámetro es *d'* (allí se llama sensibilidad, pero aquí reservaremos ese término para otro concepto). Para un mismo valor de *d'* se pueden emplear diferentes umbrales. La figura 1b representa en el espacio ROC (*Receiver Operating Characteristic*) los pares de valores de *PFP* y *PVP*. En abscisas es la probabilidad de una de las formas de error (*PFP*) y en ordenadas la de una de las formas de acierto (*PVP*). El rendimiento ideal generaría un punto (0;1). La curva representa todos los pares de valores que se obtendrían para el caso de la figura 1a. A medida que el umbral se desplaza hacia la derecha se reduce *PVP* y se incrementa *PVN*. El balance o compromiso entre *PVP* y *PVN* bajo diferentes umbrales genera una curva ROC monótonamente creciente.

Cuando el umbral es implícito hay que asumir que es variable. Sin embargo, incluso en contextos diagnósticos con un umbral explícito (como el valor X≥8 para el AUDIT) éste puede estar sujeto a cambios entre estudios, debidos a diferencias en el contexto y el procedimiento en que se aplican.

En un contexto diagnóstico, la capacidad del instrumento para detectar un *T* [o P(+|T)] se llama *Sensibilidad* (*S*) y la de identificar un *N* [o P(-|N)], se llama *Especificidad* (*E*) (Franco y Vivo, 2007). Sus estimaciones son:

$$S = \frac{VP}{VP + FN} = \frac{VP}{N_T} \tag{1}$$

$$E = \frac{VN}{VN + FP} = \frac{VN}{N_N} \tag{2}$$

Sin embargo, emplear estos indicadores supone valorar simultáneamente dos cantidades, y puede ocurrir que un instrumento tenga mejor *S* que otro, pero tenga peor *E*. Por el contrario, disponemos de diversos indicadores que resumen la información en un único valor (véase Sánchez-Meca, Marín-Martínez y Chacón-Moscoso, 2003). Entre ellos destaca la *Razón de Ventajas* (*RV*; en inglés, *Odds Ratio*) para respuestas positivas. En el contexto de los instrumentos diagnósticos se llama *Razón de Ventajas Diagnóstica* (*RVD*; *Diagnostic Odds Ratio, DOR*).

La *ventaja* de una proporción es el cociente entre ésta y su complementaria. Si *S*= 0.80, la ventaja de la respuesta positiva sobre la negativa, en presencia de la condición, es 0.80/0.20= 4 (para un *T*, la probabilidad de un resultado positivo es cuatro veces mayor que la de un resultado negativo).

La *RVD* se define como el cociente entre las ventajas de *S* y (1-*E*):

$$RVD = \frac{S/(1-S)}{(1-E)/E} = \frac{S \cdot E}{(1-E)\cdot(1-S)} = \frac{VP \cdot VN}{FP \cdot FN} \tag{3}$$

RVD refleja en un único número la precisión del test; si el rendimiento de una prueba diagnóstica es mínimo (clasificación al azar) será igual a 1, mientras que será mejor cuanto más supere al valor 1. Otras formas de caracterizar el rendimiento (como el índice *d*, muy empleado en meta-análisis y con una equivalencia directa con

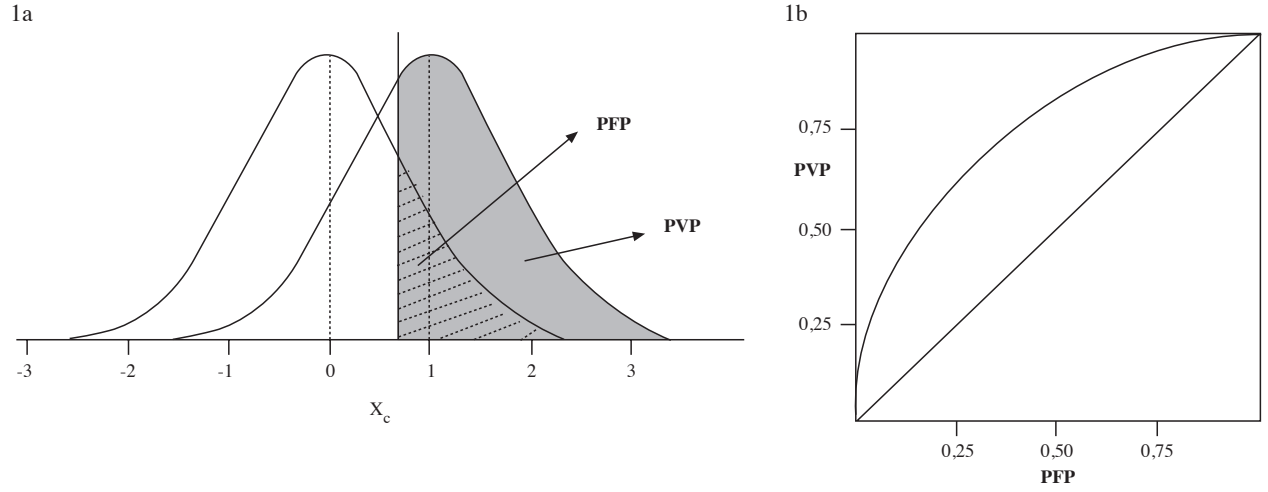

Figura 1. *(a) Distribuciones de la variable de evidencia, para casos T (curva derecha) y casos N (curva izquierda). Las áreas a la derecha del punto de corte (X_c) representan, respectivamente, la Proporción de Verdaderos Positivos (PVP) y la Proporción de Falsos Positivos (PFP). (b) Representación en el espacio ROC de la curva que se obtendría con los pares de valores de PVP y PFP al desplazar X_c en todo su rango*

RVD; Hasselblad y Hedges, 1995; Sánchez-Meca, Marín-Martínez y Chacón-Moscoso, 2003) se utilizan poco en el meta-análisis de la precisión de instrumentos de clasificación.

Es habitual no trabajar directamente con *RVD*, sino con su logaritmo. Para su ponderación en el meta-análisis se suele emplear el inverso de la varianza del logaritmo. Su varianza es, aproximadamente (Fleiss, 1994):

$$\sigma^2(LogRVD) = (1/VP) + (1/FP) + (1/FN) + (1/VN) \quad (4)$$

El intervalo de confianza se puede obtener mediante ($z_{\alpha/2}$ es el valor de la distribución normal tipificada con percentil $\alpha/2$):

$$RVD \cdot e^{\pm z_{\alpha/2} \cdot \sigma(LogRVD)} \quad (5)$$

Como ejemplo, en la figura 2 reproducimos los valores de 14 estudios que valoran la capacidad diagnóstica del AUDIT (tomados del meta-análisis de Berner, Kriston, Bentele y Härter, 2007). Incluye las frecuencias de *VP*, *FP*, *FN* y *VN*, más *S* y *E* (entre paréntesis, los intervalos de confianza). También incluye el *forest plot* (obtenido con *Review Manager*, 2008). En la figura 3 se representan los pares de valores *S* y (1-*E*) en el espacio ROC (figura realizada con *METADisc*; Zamora, Abraira, Muriel, Khan y Coomarasamy, 2006).

Study	TP	FP	FN	TN	Sensitivity	Specificity
01	34	22	36	144	0.49 [0.36, 0.61]	0.87 [0.81, 0.92]
02	170	40	163	960	0.51 [0.46, 0.57]	0.96 [0.95, 0.97]
03	9	48	1	131	0.90 [0.55, 1.00]	0.73 [0.66, 0.80]
04	16	42	36	238	0.31 [0.19, 0.45]	0.85 [0.80, 0.89]
05	573	496	181	5704	0.76 [0.73, 0.79]	0.92 [0.91, 0.93]
06	41	6	15	72	0.73 [0.60, 0.84]	0.92 [0.84, 0.97]
07	115	131	166	3139	0.41 [0.35, 0.47]	0.96 [0.95, 0.97]
08	187	193	73	1474	0.72 [0.66, 0.77]	0.88 [0.87, 0.90]
09	72	47	16	167	0.82 [0.72, 0.89]	0.78 [0.72, 0.83]
10	58	6	47	150	0.55 [0.45, 0.65]	0.96 [0.92, 0.99]
11	22	6	3	148	0.88 [0.69, 0.97]	0.96 [0.92, 0.99]
12	7	8	3	243	0.70 [0.35, 0.93]	0.97 [0.94, 0.99]
13	21	15	2	55	0.91 [0.72, 0.99]	0.79 [0.67, 0.87]
14	23	3	43	159	0.35 [0.24, 0.48]	0.98 [0.95, 1.00]

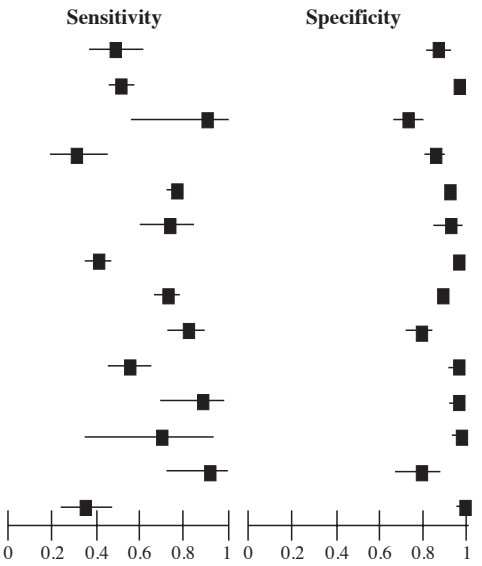

Figura 2. *Forest plot de los 14 estudios del AUDIT (figura realizada con Review Manager). Las cuatro frecuencias aparecen con sus iniciales en inglés (TP, verdaderos positivos; FP, falsos positivos; FN, falsos negativos; TN, verdaderos negativos)*

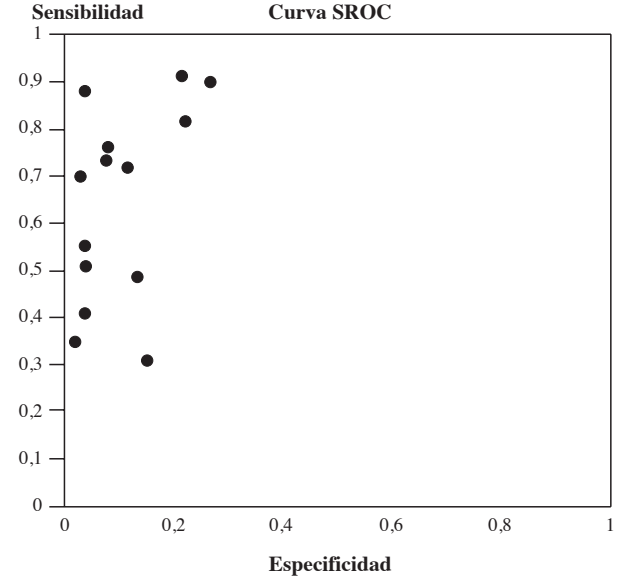

Figura 3. *Representación de los valores de S y (1-E) de los 14 estudios del AUDIT en el espacio ROC*

Procedimientos de integración meta-analítica

Como en cualquier otro ámbito empírico, se pueden aplicar procedimientos de meta-análisis para integrar los resultados de estudios que valoran la eficacia de los instrumentos diagnósticos de clasificación binaria. Con ellos se pretende responder a las tres cuestiones principales que se abordan con el meta-análisis: la estimación combinada de su eficacia, la valoración de la variabilidad observada en los indicadores y la exploración y el análisis de los factores que explican esa heterogeneidad (Sánchez-Meca y Botella, 2010). Naturalmente, los procedimientos meta-analíticos se deben adaptar a las peculiaridades de este tipo de estudio, tal y como se ha hecho con otros aspectos de la calidad de los instrumentos de medida, como por ejemplo en el meta-análisis de la fiabilidad de los tests, o Generalización de la Fiabilidad (Botella y Suero, en prensa; Botella, Suero y Gambara, 2010; Sánchez-Meca y López-Pina, 2008). Se han propuesto diversos procedimientos para dar una respuesta única y combinada a la pregunta sobre el rendimiento de los instrumentos. Algunos de ellos ya han sido reconocidos como inapropiados, mientras que otros han sido superados posteriormente. Nosotros los incluimos porque conocerlos permite interpretar los meta-análisis ya publicados en los que se han empleado.

Agregación directa

El primer (y desaconsejado) procedimiento consiste en agregar las frecuencias de todos los estudios para cada casilla. A continuación se obtienen los indicadores a partir de la tabla de agregados. En el ejemplo del AUDIT la *S* y *E* del agregado son .632 y .923, respectivamente, mientras que la *RVD* es igual a 20.652.

El método de agregación directa ha sido raramente utilizado. La principal razón para desecharlo como método meta-analítico es que en realidad niega las posibilidades que cualquier procedimiento meta-analítico debe tener para analizar la variabilidad observada en los estudios. Agregando las casillas de la tabla de contingencia tratan todos los casos como si fueran iguales, ignorando las variables moderadoras que pudieran ser factores explicativos de esa variabilidad. Por otro lado, también se ignora la (necesaria) relación entre *S* y *E*.

Integración separada de S y E (S/E)

En esta estrategia meta-analítica (tampoco recomendada) se hacen estimaciones separadas e independientes de *S* y *E*, aplicando los procedimientos habituales para proporciones (Lipsey y Wilson, 2001). Las varianzas se pueden estimar mediante:

$$\hat{\sigma}^2(S) = \frac{S \cdot (1-S)}{VP + FN} \tag{6}$$

$$\hat{\sigma}^2(E) = \frac{E \cdot (1-E)}{VN + FP} \tag{7}$$

Los valores combinados se obtienen como un promedio de las estimaciones de *S* y *E*, ponderando por el inverso de la varianza [$w_i = 1/\sigma^2(\hat{\theta}_i)$], donde $\sigma^2(\hat{\theta}_i)$ se sustituye por σ^2 (*S*) y σ^2 (*E*), respectivamente. Es decir, si disponemos de *J* estimadores de un mismo parámetro θ (en este caso, *S* y *E*, respectivamente), la estimación combinada o total se obtendría mediante (Botella y Gambara, 2002; Hedges y Olkin, 1985):

$$\hat{\theta}_T = \sum_{i=1}^{J} w_i \hat{\theta}_i \Big/ \sum_{i=1}^{J} w_i \tag{8}$$

Como el lector habrá advertido ya, se asume implícitamente un modelo de efectos fijos.

En el ejemplo del AUDIT estas estimaciones son $S_T = 0.660$ y $E_T = 0.939$. Combinando estos valores se pueden obtener indicadores globales; así, la RVD_T sería 29.881. Estas estimaciones combinadas se acompañan de contrastes de homogeneidad (Botella y Gambara, 2002). En nuestro ejemplo se rechaza la H_0 de homogeneidad tanto para *S* [Q(13)= 273.37, p<.001], como para *E* [Q(13)= 260.48, p<.001] y para los valores de *RVD* [Q(13)= 97.86, p<.001].

Respecto a las limitaciones del procedimiento *S/E*, la principal es que ignora la relación entre *S* y *E*, lo que puede dar lugar a resultados engañosos, o incluso francamente inadecuados. Veamos un ejemplo que, por simplicidad, tiene solo dos estudios primarios que, además, tienen el mismo tamaño muestral total y de los subgrupos. Los resultados del primer estudio son $S_1 = 0.80$ y $E_1 = 0.20$, mientras que los del segundo son $S_2 = 0.20$ y $E_2 = 0.80$. Mientras que una representación gráfica en el espacio ROC y unos sencillos cálculos nos mostrarán que los puntos que representan a estos estudios pertenecen a una misma curva de precisión en el espacio ROC, la estimación separada nos proporciona los valores totales $S_T = 0.50$ y $E_T = 0.50$, que generan un punto sobre la diagonal que representa la ejecución a nivel del puro azar (otra versión de la famosa *Paradoja de Simpson*).

La curva ROC-resumen de Moses, Shapiro y Littenberg (MSL)

El procedimiento de Moses, Shapiro y Littenberg (1993; *MSL* a partir de aquí) sí tiene en cuenta la relación entre *S* y *E*. Si los estudios no se diferencian ni en la precisión ni en el umbral, las estimaciones de *S* y *E* deberían ser independientes. Si tienen la misma precisión pero diferentes umbrales (algo bastante probable, incluso con umbrales explícitos) entonces es esperable una correlación negativa entre *S* y *E* (o positiva entre *S* y 1-*E*). En estudios con umbrales elevados (figura 1a) se obtendrán valores bajos de *S* y altos de *E*, mientras que en estudios con umbrales bajos ocurrirá lo contrario. Si la variabilidad observada en *S* y *E* se debe a oscilaciones en el umbral (el llamado *efecto umbral*) esos valores mostrarán cierta covariación.

Un objetivo del método *MSL* es obtener una estimación combinada de la precisión del instrumento, sabiendo que la precisión de los diferentes estudios podría variar y, además, el efecto umbral podría estar presente. Si los estudios tienen la misma precisión pero distinto umbral, entonces los puntos de los estudios deberían pertenecer a una única curva ROC, simétrica respecto a la contradiagonal [(0;1) a (1;0)]. Las desviaciones respecto a esa curva serían meras fluctuaciones aleatorias de muestreo. En el método MSL se ajusta una curva y se informa de la precisión involucrada en esa curva. En el proceso de ajuste se determina si la curva es simétrica. Si lo es, el valor de la precisión es único, pero si no lo es, hay que elegir un valor representativo, con algún criterio. El efecto umbral se valora mediante la correlación entre *S* y 1-*E*.

Para estudiar el ajuste se asume la distribución logística, algo no demasiado arriesgado, ya que es muy parecida a la distribución normal, pero con algunas propiedades que la hacen más manejable. Primero se obtienen los *logit* de las proporciones de *VP* y *FP*. El logit de una proporción (o de una probabilidad) es igual al logaritmo natural de su ventaja. Es decir, se obtendría (sus estimaciones):

$$V = \text{Logit}(S) = \log\big(S/(1-S)\big) = \log\big(VP/FN\big)$$
$$U = \text{Logit}\big(1-E\big) = \log\big((1-E)/E\big) = \log\big(FP/VN\big)$$

Asumiendo una distribución logística, es fácil demostrar que *U* y *V* están linealmente relacionados. En *MSL* se analiza la relación entre su suma y su diferencia:

$$SU = V + U \text{ y } DI = V - U$$

Tras demostrar que si la curva es simétrica la relación lineal entre *DI* y *SU* tendrá pendiente cero, proponen estudiarlo mediante la regresión de *DI* sobre *SU* (adviértase que *DI* no es más que Log*RVD*):

$$DI = A + B \cdot SU \tag{9}$$

donde *B* representa la tasa de cambio de la precisión (la interpretación de *A* se expone unas líneas más abajo). Los parámetros esti-

mados para la recta permiten trazar la curva ROC mediante (Moses et al., 1993, ecuación 1, página 1297):

$$S = \left[1 + e^{-A/(1-B)}\left(\frac{1-E}{E}\right)^{(1+B)/(1-B)}\right]^{-1} \quad (10)$$

Si la precisión es constante, *B* es nula. Esta hipótesis implica que *RVD* es constante para cualquier umbral. Si se mantiene esta hipótesis e^{A} proporciona la estimación combinada de *RVD*. En el ejemplo del AUDIT el modelo lineal es $DI' = 3.46 - 0.161 \cdot SU_i$ (ajuste mediante ponderación por el inverso de las varianzas de Log*RVD*; fórmula 4). Como la pendiente no es estadísticamente significativa (t= 1.059; *p*= .311) se mantiene que la precisión no depende del umbral, por lo que *RVD* es igual a $e^{3.46}$= 31.817 (un valor parecido al obtenido en el apartado dedicado al procedimiento *S*/*E*).

La correlación de Spearman entre los logit de *S* y (1-*E*) es .424 (*p*= .131). Adviértase que esta correlación es positiva porque se refiere a (1-*E*). Con frecuencia se informa de la correlación entre *S* y *E*, en cuyo caso sería negativa, con el mismo valor absoluto. Concluimos que no hay efecto umbral y que la precisión del instrumento (estimada como *RVD*= 31. 817) es constante a lo largo del rango de umbrales involucrados, dado que la curva ROC es simétrica.

Si la pendiente fuera significativa se concluiría que la precisión depende del umbral y la curva sintetizada sería asimétrica. En estas situaciones proponen informar del estadístico Q*, que es el punto donde la curva cruza la diagonal que va desde la esquina superior izquierda hasta la esquina inferior derecha (valores con *S*= *E*).

Modelo Normal Bivariado (NB)

Reitsma, Glas, Rutjes, Scholten, Bossuyt y Zwinderman (2005) proponen modelar los resultados de los estudios mediante una regresión Normal Bivariada (*NB*), en la que *S* y *E* se mantienen como valores separados pero se incluyen simultáneamente. Por un lado, *S* y *E* se incluyen como efectos aleatorios; por otro, se añade un parámetro que refleja la eventual correlación entre ellos. Otra fuente de variación es el muestreo aleatorio, que genera más variación cuanto menor es el tamaño muestral y se modela mediante variables binomiales.

Representaremos por $\theta_{S,i}$ y $\theta_{E,i}$ las transformaciones logit de las *S* y *E* del estudio *i*. Se asume que los logit(*S*) y logit(*E*) se distribuyen normalmente con valores medios $\theta_{S,i}$ y $\theta_{E,i}$, mientras que sus varianzas inter-estudios son σ_s^2 y σ_E^2. A ello se añade la covarianza entre los valores de $\theta_{S,i}$ y $\theta_{E,i}$, que se representa por σ_{SE}.

Por tanto, el modelo *NB* es:

$$\begin{pmatrix}\theta_{S,i}\\ \theta_{E,i}\end{pmatrix} \sim N\left(\begin{pmatrix}\theta_s\\ \theta_E\end{pmatrix}, \Sigma\right), \text{ siendo } \Sigma = \begin{pmatrix}\sigma_S^2 & \sigma_{SE}\\ \sigma_{SE} & \sigma_E^2\end{pmatrix} \quad (11)$$

Exige estimar 5 parámetros: θ_S, θ_E, σ_S^2, σ_E^2 y σ_{SE}. La variación inter-estudios se modela como un efecto aleatorio en el *nivel de test*. El modelo *NB* es jerárquico por contener otro nivel correspondiente a la variación intra-estudio, el *nivel de estudio*. En dicho nivel se asume que la *S* y la *E* del estudio *i* siguen distribuciones binomiales y se modelan en pares por tener unas mismas características propias del estudio (especialmente el umbral).

Hemos empleado el procedimiento NonLinear MIXED (Proc NLMIXED) de SAS (SAS Institute, 2008) para ajustar este modelo a nuestro ejemplo del AUDIT. Los principales resultados aparecen en la tabla 2.

Los valores mostrados entre paréntesis para $\hat{S}$ y $\hat{E}$ se han obtenido haciendo la transformación inversa del logit: $1/(1+e^{-\hat{\theta}})$; los valores de las varianzas y la correlación se refieren a los logit.

En nuestro ejemplo las estimaciones de *S* y *E* son muy parecidas a las obtenidas mediante el procedimiento *S*/*E*: 0.660 y 0.939, respectivamente. La correlación entre *S* y *E* es también parecida (-0.424; allí era positiva porque se refería a *S* y 1-*E*). El signo negativo probablemente refleja oscilaciones en el umbral, a pesar de que el valor nominal sea único (X≥8).

Con los parámetros estimados se puede construir una curva ROC-resumen, previa su transformación inversa a la métrica de *S* y *E*, como en la figura 4, obtenida mediante *Review Manager* (2008).

En la figura 4 la región de confianza cubre un área grande, lo que probablemente refleja una excesiva heterogeneidad entre los estudios. Efectivamente, recordemos (véase el apartado dedicado al procedimiento *S*/*E*) que hemos rechazado la hipótesis de homogeneidad de *S*, *E* y *RVD*. Lo indicado aquí es explorar el papel de

Tabla 2
Ajuste del modelo Normal Bivariado (NB). Resultados de la aplicación del procedimiento NLMixed de SAS al ejemplo del AUDIT

Estimaciones de los parámetros	Estadísticos de ajuste	
θ_S: 0.6416 ($\hat{S}$= 0.6551)	-2 Log Likelihood	234.2
θ_E: 2.4201 ($\hat{E}$= 0.9183)	AIC	244.2
σ_S^2: 0.7788 (del logit)	AICC	247.0
σ_E^2: 0.7139 (del logit)	BIC	247.4
σ_{SE}: -0.4002 (entre los logit)		

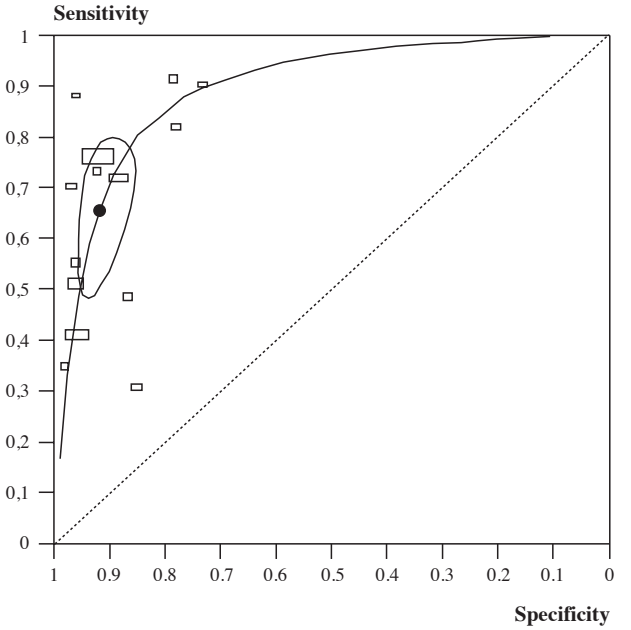

Figura 4. *Punto combinado, curva ROC-resumen y región de confianza en el espacio ROC convencional. Los estudios se representan por rectángulos con tamaño proporcional al tamaño de su muestra*

las variables moderadoras. En el ejemplo del AUDIT disponemos del valor de la variable moderadora «proporción de hombres» en 13 de los estudios.

Representando por *Z* a la covariable, el modelo *NB* se formula como:

$$\begin{pmatrix} \theta_{S,i} \\ \theta_{E,i} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_S + v_S Z_i \\ \theta_E + v_E Z_i \end{pmatrix}, \Sigma \right) \quad (12)$$

donde v_S y v_E (que se tratan como efectos fijos) son coeficientes que representan los efectos de la covariable *Z* sobre logit(*S*) y logit(*E*).

También es posible incluir la covariable en solo uno de los dos indicadores, si se considera apropiado. Se pueden crear variables *dummy* para codificar la covariable *Z* y ejecutar el procedimiento NLMIXED. En nuestro ejemplo la covariable es cuantitativa. Con el valor de [-2·log(verosimilitud)] que nos proporciona este procedimiento podemos calcular la reducción entre el modelo sin covariable y el presente modelo. En el ejemplo del AUDIT con el sexo como covariable en *S* y *E* es de 217.6 cuando no se incluye la covariable y 212.6 cuando ésta se incluye en *S* y *E*. La reducción de 217.6-212.6= 5 se distribuye χ^2 con 2 grados de libertad (*p*= .08), por lo que concluimos que el sexo tiene una asociación marginalmente significativa con la eficacia del AUDIT. En la tabla 3 presentamos los parámetros y el ajuste tanto del modelo sin covariables (algo diferentes a los de la tabla 2, pues se refieren a 13 estudios) como del modelo con la covariable sexo.

Aunque la asociación con el sexo no sea significativa, vamos a interpretar el modelo alcanzado. La estimación de *S* se reduce considerablemente y la de *E* se incrementa un poco. Hay que tener en cuenta que se refiere al caso en que la proporción de hombres es 0 (muestras solo de mujeres). Los intervalos de confianza (no incluidos en la tabla) indican que v_S es significativamente diferente de 0, mientras que v_E no lo es. El valor estimado de logit(*S*) para estudios con muestras compuestas solo por hombres es de (-0.7283+1.8354)= 1.1071, que implica *S*= .7516. Naturalmente, ambas varianzas (σ_S^2 y σ_E^2) se reducen al incluir la covariable. La covarianza es negativa y supone una correlación de -0.49; esa misma correlación era de -0.61 cuando no se incluye la covariable.

Tabla 3
Parámetros estimados y ajuste del modelo NB al ejemplo del AUDIT, sin covariables y con la covariable sexo (proporción de hombres en la muestra)

	Modelo sin covariables		**Modelo con covariable sexo**	
Estimaciones de los parámetros	θ_S: 0.7056 (Ŝ= 0.6694)		θ_S: -0.7283 (Ŝ= 0.3256)	
	θ_E: 2.4613 (Ê= 0.9214)		θ_E: 3.5152 (Ê= 0.9711)	
	σ_S^2: 0.8271		σ_S^2: 0.5302	
	σ_E^2: 0.7590		σ_E^2: 0.6095	
	σ_{SE}: -0.4797		σ_{SE}: -0.2785	
			v_S: 1.8354	
			v_E: -1.3703	
Ajuste	-2 Log Likelihood	217.6	-2 Log Likelihood	212.6
	AIC	227.6	AIC	226.2
	AICC	230.6	AICC	232.4
	BIC	230.4	BIC	230.2

Modelo jerárquico de la curva ROC resumen (HSROC)

El modelo jerárquico de la curva ROC-resumen (*HSROC*; *Hierarchical Summary Receiver Operating Characteristic*) ha sido propuesto por Gatsonis y Paliwal (2006), Rutter y Gatsonis (1995, 2001) y Macaskill (2004). También se trata de un modelo multinivel, pero en lugar de modelar directamente *S* y *E* ajusta los datos de clasificación de los estudios mediante una regresión logística de efectos aleatorios.

Incluyen dos niveles, correspondientes a la variación intra-estudio e inter-estudio, respectivamente. En el nivel 1, la frecuencia de *VP* del estudio primario *i* se denota por y_{i1}, y la frecuencia de *FP* se denota por y_{i0}. Para cada estudio *i*, se asume que la frecuencia de resultados positivos en los grupos *T* (j= 1) y *N* (j= 0) siguen la distribución binomial $y_{ij} \sim B(n_{ij}, \pi_{ij})$, *j*= 0,1, donde n_{ij} representa el tamaño de la muestra de la población *j* del estudio *i*, y π_{ij} la probabilidad de generar un resultado positivo. Este nivel se formula como:

$$\text{logit}(\pi_{ij}) = (\theta_i + \alpha_i X_{ij}) \cdot e^{-\beta X_{ij}} \quad (13)$$

donde X_{ij} codifica la pertenencia, asumiendo el valor 0.5 para la población *T* y -0.5 para la población *N*; de esta forma queda incorporada la variabilidad intra-estudio. El valor θ_i es el umbral de clasificación, en la escala logit, del estudio *i*; es conceptualmente análogo al valor *SU* en el modelo de *MSL*. El valor α_i es el log(*RVD*) del estudio *i*. Representa la eficacia de la clasificación; refleja cómo de cerca pasa la curva ROC del punto óptimo (*S*= 1, 1-*E*= 0). El parámetro de escala, β, es análogo a la pendiente del modelo *MSL*; permite que la exactitud varíe en función de los cambios de umbral. Es decir, *S* y *E* pueden cambiar a distinta velocidad. El parámetro β se incorpora como un efecto fijo. El nivel 2 modela la variación inter-estudios de θ_i y α_i como efectos aleatorios. Se asume que siguen distribuciones normales, permitiendo incorporar covariables al nivel del estudio en la media de θ y/o α. Este nivel se formula como:

$$\alpha_i \sim N(\Lambda + \lambda Z_i, \sigma_\alpha^2) \qquad \theta_i \sim N(\Theta + \gamma Z_i, \sigma_\theta^2)$$

donde Z_i es el vector de covariables. Estas se incluyen como efectos fijos, siendo λ y γ sus coeficientes.

De nuevo hemos ajustado el modelo sin covariables mediante el procedimiento NLMIXED a nuestro ejemplo. En la tabla 4 se muestran los parámetros estimados y los índices de ajuste.

A diferencia del modelo bivariado, el HSROC proporciona estimaciones directas de la precisión media (Λ= log*RVD*), del parámetro de escala (β), del umbral medio (Θ) y de las varianzas que

Tabla 4
Ajuste del modelo HSROC. Resultados de la aplicación del procedimiento NLMixed de SAS al ejemplo del AUDIT

Estimaciones de los parámetros	**Estadísticos de ajuste**	
Λ: 3.1011	-2 Log(likelihood)	234.2
Θ : -0.9227	AIC	244.2
β: -0.04346	AICC	247.0
σ_α^2: 0.6911	BIC	247.4
σ_θ^2: 0.5730		

representan los efectos aleatorios de la precisión y el umbral. La curva ROC-resumen se construye mediante la siguiente relación entre *S* y *E* (modelo sin covariables):

$$\text{logit}\left(\text{Sensibilidad}\right) = \Lambda e^{-\beta/2} - e^{-\beta}\ \text{logit}\left(\text{Especificidad}\right) \tag{14}$$

También se puede expresar como:

$$S = \frac{1}{1+\exp\left[-\left(\Lambda e^{-\beta/2} - e^{-\beta}\ \text{logit}\left(\text{Especificidad}\right)\right)\right]} \tag{15}$$

Esta relación implica variaciones del umbral, θ, mientras que la exactitud, α, se mantiene fija en su media, Λ.

Con las estimaciones de las medias se calculan los valores esperados de *S* y 1-*E*:

$$E\left(\textit{Sensibilidad}\right) = 1/\left(1+\exp\left\{-\left[\left(\Theta+0.5\Lambda\right)\right]e^{-0.5\beta}\right\}\right) \tag{16}$$

$$E\left(\textit{1 - Especificidad}\right) = 1/\left(1+\exp\left\{-\left[\left(\Theta-0.5\Lambda\right)\right]e^{0.5\beta}\right\}\right) \tag{17}$$

En nuestro caso *S* y *E* son 0.6551 y 0.9183, respectivamente. Es decir, para aplicaciones del test AUDIT en situaciones similares a las incluidas en este meta-análisis, esos son los valores medios esperados de *S* y *E*. Con ellos se representa el punto ROC resumen. *Review Manager* (2008) permite confeccionar también la representación de estos resultados. La curva, idéntica a la obtenida con *NB* (figura 5), no proporciona la región de confianza del punto resumen.

También hemos aplicado este modelo con el sexo como covariable *Z*, con objeto de explorar su efecto potencial sobre los parámetros de umbral, de exactitud y de escala (de nuevo con los 13 estudios que proporcionan esta información). El modelo más general, en el que la covariable puede estar asociada al umbral, la exactitud y la escala, sería:

$$\text{logit}\left(\pi_{ij}\right) = \left[\left(\theta_i + \gamma Z_i\right) + \left(\alpha_i + \lambda Z_i\right)\cdot X_{ij}\right]\cdot e^{-\left(\beta+\delta Z_i\right)\cdot X_{ij}} \tag{18}$$

Hemos ajustado cuatro modelos *HSROC* con distintas especificaciones; no son más que casos particulares de este modelo general:

Modelo 1. Es el modelo nulo, sin covariables, expresado en la ecuación 13. La precisión y el umbral se distribuyen, $\alpha_i \sim N\left(\Lambda,\sigma_\alpha^2\right)$ y $\theta_i \sim N\left(\Theta,\sigma_\theta^2\right)$.

Modelo 2. La covariable *Z* puede afectar al umbral; la posición del punto resumen se puede desplazar por la curva en función de *Z*. Sus distribuciones y el modelo son:

$$\alpha_i \sim N\left(\Lambda,\sigma_\alpha^2\right)\ \ \theta_i \sim N\left(\Theta+\gamma Z_i,\sigma_\theta^2\right)$$
$$\text{logit}\left(\pi_{ij}\right) = \left[\left(\theta_i+\gamma Z_i\right)+\alpha_i\cdot X_{ij}\right]\cdot e^{-\beta\cdot X_{ij}}$$

Modelo 3. La covariable *Z* pueda afectar a la exactitud; la posición de la curva puede variar con *Z*. Sus distribuciones y el modelo son:

$$\alpha_i \sim N\left(\Lambda+\lambda Z_i,\sigma_\alpha^2\right)\ \ \theta_i \sim N\left(\Theta,\sigma_\theta^2\right)$$
$$\text{logit}\left(\pi_{ij}\right) = \left[\theta_i+\left(\alpha_i+\lambda Z_i\right)\cdot X_{ij}\right]\cdot e^{-\beta\cdot X_{ij}}$$

Modelo 4. La covariable *Z* podría afectar al umbral y a la exactitud; la posición de la curva y el umbral pueden variar con *Z*. Sus distribuciones y el modelo son:

$$\alpha_i \sim N\left(\Lambda+\gamma Z_i,\sigma_\alpha^2\right)\ \ \theta_i \sim N\left(\Theta+\lambda Z_i,\sigma_\theta^2\right)$$
$$\text{logit}\left(\pi_{ij}\right) = \left[\left(\theta_i+\gamma Z_i\right)+\left(\alpha_i+\lambda Z_i\right)\cdot X_{ij}\right]\cdot e^{-\beta\cdot X_{ij}}$$

Modelo 5. La covariable *Z* puede afectar al umbral, la exactitud y la escala; la posición y la forma de la curva pueden variar con *Z*. Sus distribuciones son las siguientes, mientras que el modelo es el expresado en (18):

$$\alpha_i \sim N\left(\Lambda+\lambda Z_i,\sigma_\alpha^2\right) \qquad \theta_i \sim N\left(\Theta+\gamma Z_i,\sigma_\theta^2\right)$$

Tabla 5
Ajuste del modelo HSROC con los 13 estudios del AUDIT que informan del sexo. Incluye el modelo sin covariables y varios modelos con influencia de la covariable sexo en diferentes elementos del modelo (véase texto)

		Modelos con el sexo como covariable (efecto de la covariable en…)			
	Modelo 1 (sin covariable)	**Modelo 2 (umbral)**	**Modelo 3 (precisión)**	**Modelo 4 (umbral y precision)**	**Modelo 5 (umbral, precisión y escala)**
Parámetros	Λ= 3.2053	Λ= 3.1603	Λ= 2.7211	Λ= 2.6408	Λ= 1.9297
	Θ= -0.9121	Θ= -2.1289	Θ= -0.8613	Θ= -2.0744	Θ= -1.7578
	β= -0.04292	β= -0.00934	β= 0.02470	β= 0.06970	β= 0.5734
	σ_α^2= 0.6256	σ_α^2= 0.6158	σ_α^2= 0.5895	σ_α^2= 0.5802	σ_α^2= 0.5348
	σ_θ^2= 0.6361	σ_θ^2= 0.4245	σ_θ^2= 0.6368	σ_θ^2= 0.4235	σ_θ^2= 0.4359
		γ= 1.6070	λ= 0.5437	λ= 0.5772	λ = 1.3778
				γ = 1.6119	γ = 1.2266
					δ = -0.6003
Ajuste					
-2 Log(likel.)	217.6	212.6	217.2	212.2	211.9
AIC	227.6	224.6	229.2	226.2	227.9
AICC	230.6	229.0	233.6	232.4	236.4
BIC	230.4	228.0	232.6	230.2	232.4

Los resultados (tabla 5) indican que la bondad de ajuste del modelo 2 (que incorpora el sexo como covariable del parámetro de umbral) es significativamente mejor que la del modelo nulo. Sin embargo, el sexo no tiene efectos sobre la exactitud ni la escala. En resumen, proponemos una única curva ROC-resumen, la del modelo 2, dado que el efecto de *Z* solo se manifiesta en cambios del umbral. El umbral cambia de unos estudios a otros, pero lo hace a lo largo de esta curva, sin que cambie la forma ni la posición de la misma (figura 5).

El valor de γ (1.607) significa que hay un efecto negativo de la covariable. A mayor proporción de hombres, menor umbral implícito a pesar de que la regla es explícita (X≥8). En muestras solo de hombres el umbral implícito es 1.607 unidades de escala logit menor que en muestras solo de mujeres.

El efecto del sexo sobre el umbral podría estar reflejando que los hombres y mujeres utilizan criterios distintos al responder a preguntas relacionadas con el consumo de alcohol. Un mismo patrón de consumo es considerado más grave y patológico en las mujeres. Esta interpretación es coherente con un contexto social en el que los hombres beben más y a niveles iguales de consumo su conducta es más tolerada. Además, cuando un hombre valora su nivel de consumo de alcohol es más probable que tome como referencia el de otros hombres. Igualmente, los patrones de las mujeres son más desconocidos. Cuando contestan al AUDIT hombres y mujeres con los mismos niveles de alcoholismo, las mujeres tienden a obtener una puntuación menor. Para alcanzar la misma puntuación de clasificación (X≥8) un hombre tiene que tener un patrón de consumo menos extremo que el de las mujeres.

Discusión

El método de Agregación Directa elimina toda la capacidad del meta-análisis para interpretar los resultados. Se consideran todos los casos iguales, ignorando factores de estudio como la prevalencia, la composición de las muestras, los procedimientos empleados y la relación entre *S* y *E*.

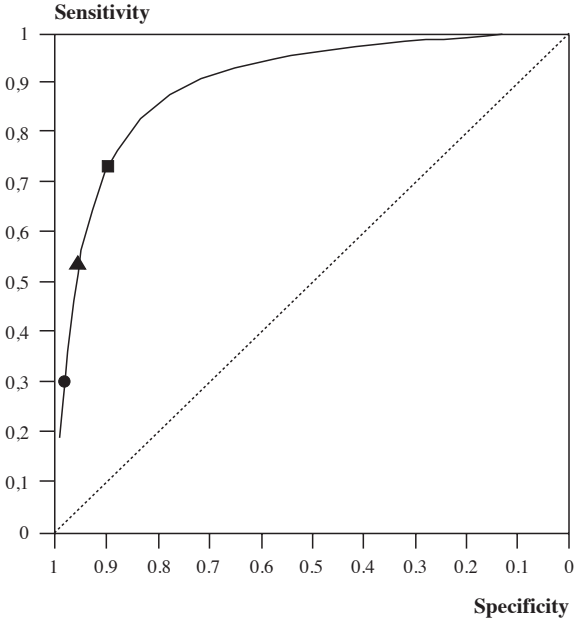

Figura 5. *La curva ROC-resumen del modelo HSROC con covariable Z que afecta al umbral. Las marcas representan los puntos en el espacio ROC para muestras con el 100% (cuadrado), 50% (triángulo) y 0% (círculo) de hombres, respectivamente. Los tres puntos forman parte del modelo ajustado, pero implican diferentes umbrales*

El método de la estimación separada (*S*/*E*) permite aplicar algunos procedimientos típicamente meta-analíticos, pero ignora la relación entre los valores de *S* y *E*, un factor imprescindible para comprender el comportamiento del test en el espacio ROC. No es esperable que *S* y *E* sean independientes. No lo serán cuando el umbral sea implícito, pero incluso cuando sea explícito y no cambie nominalmente de unos estudios a otros puede haber diferencias en las muestras, en los evaluadores o en otros factores de aplicación, que permitan oscilaciones imperceptibles. Si en los diferentes estudios se emplean distintos umbrales, los pares de valores de *S* y *E* mostrarán una correlación negativa (figura 1a). Otra razón para descartar este procedimiento es que en realidad no proporciona una curva ROC, sino la estimación del par (*S*,*E*) que mejor resume los estudios. Construir una curva solo a partir de ese punto exige añadir supuestos arbitrarios.

El método *MSL* tiene a su favor que contempla la correlación entre *S* y *E*, pero tiene varios inconvenientes. El primero es que se basa en un indicador combinado de la eficacia, la *RVD*, en el que se pierden las informaciones específicas de *S* y *E*. Un mismo valor de *RVD* puede proceder de múltiples pares (*S*,*E*). El segundo inconveniente es que no reconoce la fiabilidad imperfecta en la variable predictora de la regresión (suma de los logit). Tampoco contempla el estudio de la heterogeneidad y su eventual explicación mediante covariables. Por último, no se reconoce la diferente precisión de las estimaciones de *S* y *E*; aunque con frecuencia se pondera por el inverso de la varianza de Log*RVD*, esto supone tener en cuenta el número total de participantes en cada estudio. Sin embargo, dos estudios con el mismo tamaño total pueden tener diferentes porcentajes de las poblaciones de *T* y *N* (los modelos *NB* y *HSROC* superan esta dificultad, incluyendo el tamaño específico de cada grupo, N_T y N_N).

Aunque los procedimientos *S*/*E* y *MSL* son insuficientes o inadecuados, se emplean en las fases de exploración y para describir los resultados de los estudios.

Los métodos *NB* y *HSROC* son los más adecuados para realizar meta-análisis de estudios de clasificación binaria, especialmente para instrumentos diagnósticos. Harbord, Deeks, Egger, Whiting y Sterne (2007) han mostrado que estos modelos son equivalentes en muchas situaciones. Ambos son adecuados, ya que reconocen y diferencian entre la variación intra-estudio e inter-estudio mediante un segundo nivel de efectos aleatorios (en la línea que se va imponiendo en el meta-análisis), permitiendo además incorporar la correlación entre *S* y *E*.

Aunque parten de puntos teóricos distintos, tienen mucho en común. En primer lugar, en ambos modelos los niveles correspondientes a la variación intra-estudio son exactamente iguales, asumiendo la distribución binomial para los casos con resultados positivos en cada grupo. En segundo lugar, en ambos se asume una distribución normal de los parámetros del segundo nivel y se estiman las medias y varianzas de dichos parámetros. Además, se pueden realizar utilizando el mismo procedimiento, NLMIXED. En realidad, lo que han demostrado Harbord et al. (2007) es que bajo ciertas condiciones, los dos modelos son equivalentes; mejor dicho, son diferentes parametrizaciones del mismo modelo unificado y deben proporcionar inferencias estadísticas parecidas. Incluso proporcionan fórmulas de conversión de los parámetros entre los modelos cuando no se incluyen covariables a nivel de estudio.

Esto no significa que sean intercambiables. Afrontan problemas diferentes y con distintas flexibilidades. En el análisis de este tipo de escenarios hay dos pares de valores interdependientes: por un lado *S* y *E*, por el otro la precisión y el umbral. Si se fijan (o estiman) valores de *S* y *E*, entonces la precisión y el umbral quedan determinados. Por el contrario, si se fijan (o estiman) la precisión y el umbral, entonces *S* y *E* quedan determinados. Como el procedimiento *NB* modela *S* y *E*, permite incorporar covariables a estos estadísticos, mientras que la precisión y el umbral vienen determinados y no se pueden incorporar covariables asociadas a éstas. Este sería el caso en que las características de los estudios o de las muestras que afectan solo al grupo *T* o *N* (por ejemplo, la comorbilidad de los T puede afectar a *S*, pero no tiene nada que ver con *E*).

Por el contrario, como el procedimiento *HSROC* modela directamente la precisión y el umbral, se pueden incorporar covariables explicativas de ellos, pero *S* y *E* quedan determinados. Esto es lo adecuado cuando estudiamos si alguna covariable afecta a la precisión o el umbral (con efectos simultáneos sobre *S* y *E*). En las situaciones en las que es más aplicable el *HSROC*, el *NB* no es adecuado. En nuestro ejemplo del AUDIT, el efecto de la covariable sexo solo se puede manifestar con el modelo *HSROC*. Además, éste permite eliminar el efecto aleatorio de la exactitud, dependiendo de las necesidades. El modelo *NB* no tiene esta flexibilidad, debido a la dependencia entre *S* y *E*.

En resumen, hoy disponemos de modelos muy potentes para realizar meta-análisis de la precisión de los instrumentos de clasificación binaria, construidos como modelos jerárquicos y que se plantean como modelos de efectos aleatorios.

Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Ciencia y Tecnología, Proyecto PSI2009-12071. Agradecemos a dos revisores anónimos de *Psicothema* por sus sugerencias, que han ayudado a mejorar el manuscrito.

Referencias

Babor, T.F., Higgins-Biddle, J.C., Saunders, J.B., y Monteiro, M.G. (Eds.) (2001). *AUDIT: The alcohol use disorders identification test: Guidelines for use in primary care, 2nd edition*, WHO Document No. WHO/MSD/MSB/01.6a, Geneva, Switzerland: World Health Organization, 2001.

Berner, M.M., Kriston, L., Bentele, M., y Harter, M. (2007). The alcohol use disorders identification test for detecting at-risk drinking: A systematic review and meta-analysis. *Journal of Studies on Alcohol and Drugs*, *68*, 461-473.

Botella, J., y Gambara, H. (2002). *¿Qué es el meta-análisis?* Madrid: Biblioteca Nueva.

Botella, J., y Suero, M. (en prensa). Managing heterogeneity of variances in studies of internal consistency generalization. *Methodology*.

Botella, J., Suero, M., y Gambara, H. (2010). Psychometric inferences from a meta-analysis of reliability and internal consistency coefficients. *Psychological Methods*, *15*, 386-397.

Fleiss, J.L. (1994). Measures of effect size for categorical data. En H. Cooper y L.V. Hedges (Eds.), *The handbook of research synthesis*. Nueva York: Russell Sage Foundation.

Franco, M., y Vivo, J. (2007). *Análisis de curvas ROC. Principios básicos y aplicaciones*. Madrid: La Muralla.

Gatsonis, C., y Paliwal, P. (2006). Meta-analysis of diagnostic and screening test accuracy evaluations: Methodological primer. *American Journal of Roentgenology*, *187*, 271-281.

Harbord, R.M., Deeks, J.J., Egger, M., Whiting, P., y Sterne J.A. (2007). A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics, 8*(2), 239-251.

Hasselblad, V., y Hedges, L.V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin*, *117*, 167-178.

Hedges L.V., y Olkin, I. (1985). *Statistical Methods of meta-analysis*. Orlando: Academic Press.

Lipsey, M.W., y Wilson, D.B. (2001). *Practical meta-analysis*. Thousand oaks, CA: Sage Pub.

Logan, G.D. (2004). Cumulative progress in formal theories of attention. *Annual Review of Psychology*, *55*, 207-234.

Macaskill, P. (2004). Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *Journal of Clinical Epidemiology*, *57*, 925-932.

Moses, L.E., Shapiro, D., y Littenberg, B. (1993). Combining independent studies of a diagnostic test into a summary ROC curve: Data-analytical approaches and some additional considerations. *Statistics in Medicine*, *12*, 1293-1316.

Reitsma, J.B., Glas, A.S., Rutjes, A.W., Scholten, R.J., Bossuyt, P.M., y Zwinderman A.H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology*, *58*, 982-990.

Review Manager (RevMan) (2008) [Computer program]. *Version 5.0*. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration.

Rutter, C.M., y Gatsonis, C.A. (1995). Regression methods for meta-analysis of diagnostic test data. *Academic Radiology, 2*(Suppl 1), S48-56.

Rutter, C.M., y Gatsonis, C.A. (2001). A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine*, *20*, 2865-2884.

Sánchez-Meca, J., y Botella, J. (2010). Revisiones sistemáticas y meta-análisis: herramientas para la práctica profesional. *Papeles del Psicólogo*, *31*(1), 7-17.

Sánchez-Meca, J., y López-Pina, J.A. (2008). El enfoque meta-analítico de generalización de la fiabilidad. *Acción Psicológica*, *5*, 37-64.

Sánchez-Meca, J., Marín-Martínez, F., y Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods*, *8*, 448-467.

SAS Institute Inc. (2008). *The SAS System for Windows. Version 9.2* Cary, NC: SAS Institute Inc.

Swets, J.A. (1964). *Signal detection and recognition by human observers*. Nueva York: Wiley.

Swets, J.A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics*. Mahwah: Lawrence Erlbaum Associates.

Swets, J.A., Dawes, R.M., y Monahan, J. (2000). Psychological Science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*(1), 1-26.

Wickens, T.D. (2001). *Elementary signal detection theory*. Nueva York: Oxford University Press.

Zamora, J., Abraira, V., Muriel, A., Khan, K., y Coomarasamy, A. (2006). Meta-Disc: A software for meta-analysis of test accuracy data. *BMC Medical Research Methodology, 6*, 31.