



# Lectura crítica en pequeñas dosis

## El metaanálisis de pruebas diagnósticas

M. Molina Arias

Publicado en Internet:  
4-septiembre-2015

Manuel Molina Arias:  
mma1961@gmail.com

Servicio de Gastroenterología. Hospital Infantil Universitario La Paz. Grupo de Trabajo de Pediatría Basada en la Evidencia AEP/AEPap. Madrid. España. Editor de [www.cienciasinseso.com](http://www.cienciasinseso.com)

### Resumen

#### Palabras clave:

- Metaanálisis
- Estudios de pruebas diagnósticas
- Efecto umbral
- Curva ROC

La metodología a seguir a la hora de realizar una revisión sistemática con metaanálisis sobre estudios de tratamiento está bastante estandarizada y es, en general, conocida por los consumidores de literatura médica. Por otra parte, cada vez más nos encontramos con una clase de metaanálisis diferente, el de pruebas diagnósticas. Aunque las recomendaciones generales para su realización son similares a los metaanálisis de tratamiento, tienen aspectos específicos que conviene conocer bien para poder interpretarlos adecuadamente. Entre estos, destacamos el estudio del efecto umbral, los métodos de selección de las medidas de resumen y la forma de presenta el resultado global en forma de curva ROC resumen.

### Meta-analysis of diagnostic test studies

### Abstract

#### Key words:

- Meta-analysis
- Diagnostic test studies
- Threshold effect
- ROC curve

The methodology to be followed to do a systematic review with meta-analysis on intervention studies is quite standardized and is, in general, well known by medical literature consumers. On the other hand it is more and more frequent finding a different kind of meta-analysis: the one on diagnostic tests. Although the general recommendations to perform them are similar to those of meta-analysis on interventions, they have specific aspects important to be well known in order to be able to understand them properly. We highlight for instance the study of the threshold effect, the methods of selecting summary measures and the way to present the global finding as a ROC curve summary.

## INTRODUCCIÓN

El metaanálisis (MA) es el método de síntesis cuantitativa que se utiliza en las revisiones sistemáticas para integrar los resultados de los estudios primarios en una medida resumen de resultado. El MA comienza a utilizarse a partir de los años 1980 con

el paso de realización de revisiones narrativas a revisiones sistemáticas.

Inicialmente los MA se centran en estudios sobre tratamiento, por lo que en la actualidad está bastante bien definida la metodología que hay que seguir para hacer una revisión sistemática y un MA sobre tratamiento de forma correcta.

Cómo citar este artículo: Molina Arias M. El metaanálisis de pruebas diagnósticas. Rev Pediatr Aten Primaria. 2015;17:281-5.

Por otra parte, durante las últimas dos décadas hemos asistido a una revisión sobre el estudio de la validez de las pruebas diagnósticas, identificándose los marcadores que miden su desempeño como sensibilidad (S), especificidad (E), valores predictivos, cocientes de probabilidades, etc. Esto lleva aparejado el desarrollo del MA para evaluar la validez de las pruebas diagnósticas (MAD), que tiene muchas características comunes con el MA sobre pruebas de tratamiento, pero también algunas peculiaridades que es preciso conocer.

El proceso de revisión sistemática de diagnóstico sigue las mismas fases que las de una de tratamiento en cuanto al planteamiento de los objetivos de la revisión, la realización de una búsqueda bibliográfica completa, la selección de artículos, la evaluación de su calidad metodológica y el posterior análisis estadístico o MA propiamente dicho. Sin embargo, existen diferencias en las bases de datos que puede interesar utilizar, el tipo de descriptores a utilizar en la búsqueda, las listas de verificación útiles para valorar los trabajos primarios (QUADAS para revisiones sistemáticas de diagnóstico y STARD para los trabajos primarios de la revisión), etc.<sup>1-3</sup>.

No describiremos en este artículo todos los puntos de desarrollo de una revisión sistemática y MA sobre pruebas diagnósticas, sino que nos centraremos en tratar de explicar de forma sencilla los aspectos específicos del MAD para favorecer la lectura crítica y comprensión de estos trabajos que cada vez nos encontraremos con más frecuencia en la literatura médica. Explicaremos brevemente qué medidas de resumen pueden utilizarse para pasar a detallar algunos aspectos del estudio de heterogeneidad y de la presentación del resultado de un MAD.

## ELECCIÓN DE LA MEDIDA RESUMEN

Estamos acostumbrados a ver como en los MA sobre estudios de tratamiento se selecciona una medida de resultado resumen, ya sea un riesgo relativo, una diferencia de medias, etc. Sin embargo, en

los estudios sobre pruebas diagnósticas se estima el resultado con una pareja que define la validez de la prueba.

Es frecuente combinar S y E como valores resumen ponderados, aunque no es la medida preferida por la mayor parte de los autores que desarrollan la metodología estadística del MAD. El problema con estos indicadores es que los puntos de corte para considerar la prueba como positiva o negativa han podido ser diferentes entre los distintos estudios primarios. Otra posible causa de variación se produce cuando la valoración de la prueba depende de la subjetividad del evaluador, como ocurre con las pruebas de imagen como las radiografías. Esto dar lugar a un sesgo característico de los MAD conocido como efecto umbral, que debe valorarse siempre que se empleen estos indicadores.

Otra pareja de indicadores que tampoco es aconsejable utilizar son los valores predictivos positivo y negativo. La razón es que estos índices varían con la prevalencia del efecto, y esta generalmente variará de un estudio a otro, por lo que introducen mucha heterogeneidad<sup>3</sup>.

Quizás la pareja más fiable para combinar en un resultado resumen sea la formada por los cocientes de probabilidades positivo y negativo, que, al igual que el valor resumen de S y E, pueden calcularse mediante un modelo de efecto fijo (prueba de Mantel-Haenszel) o de efectos aleatorios (método de DerSimonian-Laird), en cuya descripción no vamos a entrar.

La ventaja de los cocientes de probabilidades es doble. Por una parte, son mucho menos susceptibles al efecto umbral. Por otra, permiten calcular la *odds* posprueba (cociente de probabilidad  $\times$  *odds* preprueba) y, a partir de ella, la probabilidad posprueba ( $odds/1 + odds$ ).

Por último, nos referiremos a la *odds ratio* diagnóstica, que puede calcularse, entre otras formas, como la razón de los cocientes de probabilidades<sup>4</sup>. Es también una medida muy robusta frente al efecto umbral y resulta de gran utilidad a la hora de calcular la curva ROC resumen que comentaremos más adelante.

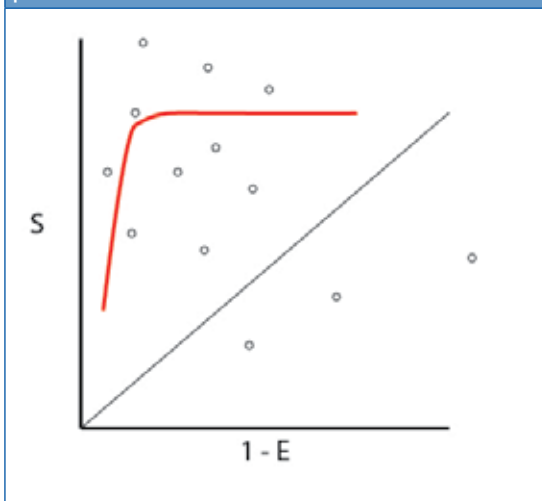
## EFECTO UMBRAL

Puede sospecharse observando la heterogeneidad de la metodología de los estudios primarios de la revisión. La forma matemática más sencilla es calculando el coeficiente de correlación de Spearman entre S y E. Si existe efecto umbral existirá una correlación inversa, tanto más fuerte cuanto mayor sea el efecto<sup>5</sup>.

Una forma gráfica de observarlo consiste en representar los puntos S-E sobre el espacio de la curva ROC hipotética del estudio, tal como vemos en la Fig. 1. Este diagrama de dispersión nos da una idea también de la heterogeneidad de los estudios primarios, que será mayor cuanto más dispersos se encuentren los puntos<sup>3</sup>.

Si existe dispersión pero al unir los puntos podemos formar una línea semejante a la curva ROC, es probable que exista un efecto umbral, aunque leves diferencias pueden estar provocadas simplemente por azar. Otros sesgos (selección, verificación, etc.) suelen aumentar más la dispersión de los puntos. Así, esta representación gráfica nos sirve para sospechar la existencia de un efecto umbral y para determinar la posible heterogeneidad de los estudios.

Figura 1. Representación del diagrama de dispersión de los pares S-E en el espacio ROC. La gran dispersión de los puntos indica heterogeneidad en los estudios primarios de la revisión.



## CURVA ROC RESUMEN (ROCr)

Es la estimación de una curva ROC común ajustada según los resultados combinados de todos los estudios primarios de la revisión y constituye la forma más idónea de expresar el resultado final del MAD. Existen diversas formas de calcularla, que no vamos a detallar aquí, basados en métodos bivariantes, jerárquicos, etc.<sup>5,6</sup>. En general, la curva ROCr se construye con un modelo de regresión empleando la *odds ratio* diagnóstica como estimador por su robustez frente a la heterogeneidad y al posible efecto umbral.

La lectura de la curva ROCr es similar a la de cualquier curva de valoración de una prueba diagnóstica individual. Los dos parámetros más empleados para su descripción son el área bajo la curva (ABC) y el índice Q (que no debe confundirse con la Q de Cochrane para valorar la heterogeneidad de los estudios primarios)<sup>7</sup>.

El ABC de la curva perfecta se aproximará a uno. Valores entre 0,5 y uno indican la capacidad discriminativa de la prueba diagnóstica, que será mayor cuanto más próximo a uno. Un valor de ABC de 0,5 nos dice que la capacidad de la prueba es similar a la de lanzar una moneda al aire. Valores inferiores a 0,5 indican que la prueba no contribuye al diagnóstico que pretende detectar.

El punto o índice Q corresponde a aquel en el que S y E se igualan. Al igual que con el ABC, su valor numérico entre 0,5 y uno indicará la eficacia global de la prueba diagnóstica, que será mayor cuanto más próximo al uno esté.

Tanto el ABC como el índice Q pueden calcularse con sus intervalos de confianza del 95%. En estos casos, siempre que el límite inferior del intervalo sea inferior a 0,5 concluiremos que la prueba no desempeña ningún papel en el diagnóstico de la enfermedad.

## EJEMPLOS PRÁCTICOS

Para terminar con el tema, vamos a ver cómo se aplica lo que hemos comentado hasta ahora fijándonos en dos trabajos reales de la literatura médica.

El primero es una revisión sistemática con MAD para valorar la precisión de la ecografía pulmonar para el diagnóstico de neumonía en el niño, utilizando la radiografía de tórax como patrón de referencia<sup>8</sup>. Los resultados globales obtenidos son una S de 96% y una E de 93%, así como unos cocientes de probabilidades positivo de 15,3 y negativo de 0,06.

Vemos que los autores seleccionan S y E, cuyo valor ponderado resumen calculan utilizando un modelo de efectos fijos (a pesar de la heterogeneidad existente, como comentaremos más adelante), mientras que emplean uno de efecto aleatorio para calcular los cocientes de probabilidades. La heterogeneidad se estudió mediante los habituales índices Q e I<sup>2</sup>, y resultó moderada (>45%) en todos los casos. Si reflexionamos, la prueba de referencia (radiografía) depende de la subjetividad del observador, por lo que está sujeto a riesgo de existencia de efecto umbral, que los autores no valoran en absoluto, ni mediante la correlación de Spearman ni mediante la representación de los puntos de los valores resumen en el espacio ROC. Quizás este dato, junto con otros datos de heterogeneidad no estadística en el emplazamiento de los estudios, edades de los pacientes, experiencia de los ecografistas, etc., podría desaconsejar la realización de un MA para calcular una medida resumen.

A pesar de todo, los autores lo llevan a cabo y calculan la curva ROCr, obteniendo un ABC del 98%, lo que les permite concluir el alto rendimiento de la prueba aunque, como hemos dicho, los resultados del MA pueden ponerse en duda por la alta heterogeneidad de los estudios primarios.

El segundo es un MAD para valorar la utilidad de los péptidos natriuréticos (BNP y NT-proBNP) para el diagnóstico del *ductus* arterioso persistente hemodinámicamente significativo en prematuros, utilizando la ecocardiografía como patrón de referencia<sup>9</sup>. Se seleccionan un total de diez trabajos con BNP y 11 con NT-proBNP y se construyen las tablas de contingencia de todos ellos para el cálculo de S y E,

obteniéndose unos valores resumen del 88% y 95% para BNP y del 90% y 84% para NT-proBNP.

En este caso los autores detectan heterogeneidad importante en la realización del estándar de referencia, así como entre las características de los pacientes de los diferentes estudios. Esta heterogeneidad se pone de manifiesto por la gran dispersión de los valores individuales en *los forest plot* y por la dispersión de las parejas de S y E en el espacio ROC.

A pesar de que llegan a realizar el MA completo, incluyendo la realización de la curva ROCr, los autores concluyen, a diferencia del estudio anterior, que no es seguro obtener una medida resumen a partir de los datos de los estudios primarios de la revisión, por lo que recomiendan realizar estudios que valoren de manera local la prueba diagnóstica en grupos más homogéneos.

Hemos vistos, pues, que el MAD tiene algunas peculiaridades en su realización diferentes al MA de pruebas de tratamiento, que tenemos que tener en cuenta a la hora de hacer lectura crítica de uno de estos trabajos. Además de las directrices generales para realizar cualquier revisión sistemática con MA, debemos pedir a los autores que estudien la posible existencia de efecto umbral como causa importante de heterogeneidad, preferir la utilización de los cocientes de probabilidad frente a la pareja S-E y, por último, buscar como resultado resumen principal el que proporciona la curva ROCr del estudio.

## CONFLICTO DE INTERESES

Los autores declaran no presentar conflictos de intereses en relación con la preparación y publicación de este artículo.

## ABREVIATURAS

**ABC:** área bajo curva • **E:** especificidad • **MA:** metaanálisis • **MAD:** metaanálisis de pruebas diagnósticas • **ROCr:** curva ROC resumen • **S:** sensibilidad.

## BIBLIOGRAFÍA

---

1. González Rodríguez MP, Velarde Mayol C. Listas guía de comprobación de estudios sobre pruebas diagnósticas incluidos en las revisiones sistemáticas: declaración QUADAS. *Evid Pediatr.* 2012;8:20.
2. González Rodríguez MP, Velarde Mayol C. Lista de comprobación de estudios sobre precisión de pruebas diagnósticas: declaración STARD. *Evid Pediatr.* 2012;8:43.
3. De Sousa MR, Ribeiro ALP. Revisión sistemática y metaanálisis de estudios de diagnóstico y pronóstico: una guía. *Arq Bras Cardiol.* 2009;92:235-45.
4. Glas AS, Lijmer JG, Prins MH, Bossel GJ, Bossuyt PMM. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol.* 2003;56:1129-35.
5. Zamora Romero J, Plana MN, Abaira Santos V. Estudios de evaluación de la validez de una prueba diagnóstica: revisión sistemática y metaanálisis. *Nefrología.* 2009;29:15-20.
6. Moses LE, Shapiro D, Littenberg B. Combining in dependent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med.* 1993;12:1293-316.
7. Identifying and quantifying heterogeneity. En: Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (eds.). *Introduction to meta-analysis.* John Wiley & Sons Ltd; 2009. p 107-26.
8. Pereda MA, Chavez MA, Hooper-Miele CC, Gilman RH, Steinhoff MC, Ellington LE, *et al.* Lung ultrasound for the diagnosis of pneumonia in children: a meta-analysis. *Pediatrics.* 2015;135:714-22.
9. Kulkarni M, Gokulakrishnan G, Price J, Fernandes CJ, Leeflang M, Pammi M. Diagnosis significant PDA using natriuretic peptides in preterm neonates: a systematic review. *Pediatrics.* 2015;135:e510-25.