# Síntesis de las evidencias científicas

# Revisión sistemática y meta-análisis

**Roberto Luis Lede**

Maestría en Investigación Clínica Farmacológica (UAI)

Profesor Adjunto de Obstetricia (UBA)

**María Teresa Rosanova**

Maestría en Investigación Clínica Farmacológica (UAI)

Universidad de Buenos Aires (UBA)

**Norberto Osvaldo Barabini**

Maestría en Investigación Clínica Farmacológica (UAI)

Universidad de Buenos Aires (UBA)

## Resumen

En el presente capítulo expondremos los conceptos que sustentan la revisión sistemática como una herramienta de gran valor para responder a las preguntas clínicas. No es nuestro objetivo capacitar al lector para realizar una revisión sistemática sino que pretendemos comentar los conceptos generales del tema bajo la premisa de que lo primero es saber evaluar la calidad del producto. En las últimas dos décadas, las revisiones sistemáticas han mostrado una evolución sorprendente, tanto en la cantidad publicada como en el desarrollo metodológico que aún sigue en franca evolución. Mencionarlas ya no es inusual pero no son muchos los profesionales de la

salud que conocen sus ventajas y desventajas. El contenido del capítulo está basado en las innúmeras publicaciones existentes sobre el tema, lo que ha convertido este conocimiento en un bien público, haciendo compleja la pretensión de la citación bibliográfica detallada. Al final del capítulo, recomendamos algunas publicaciones para quienes deseen ahondar el tema, muchas de las cuales han nutrido este documento.

## Palabras clave

Revisión sistemática; meta-análisis.

## Objetivo

El objetivo primordial es exponer la metodología y las características fundamentales de estas piezas de investigación sobre temas de cuidados médicos a fin de permitir su adecuada interpretación por parte del usuario de la información. Se busca por lo tanto, responder a la pregunta general: *¿Cuáles son las condiciones que debe reunir una revisión sistemática (RS) o un meta-análisis (MA) de investigaciones clínicas experimentales para otorgar verosimilitud a sus resultados y cuál es su aplicación en la práctica asistencial?*

Esta pregunta general incluye preguntas específicas que se intentará responder, como:

- ¿Qué son las RS y el MA?
- ¿Cómo es el diseño de una RS y su MA?
- ¿Cómo se hace el análisis de la calidad de una RS y su MA?
- ¿Cómo se analizan sus resultados?
- ¿Cómo se interpreta la representación gráfica de los resultados?
- ¿Qué otros aspectos se pueden considerar?

# Introducción

En más de una ocasión, ante la necesidad de adoptar una conducta para el tratamiento de un paciente, el profesional se encuentra con que la información sobre la eficacia y seguridad de la intervención a aplicar, presenta discrepancias entre sí a partir de resultados contradictorios o no definitorios de los estudios primarios disponibles. Definimos como «estudios primarios» a la publicación original de una investigación.

## ¿Cómo determinar si el tratamiento a indicar es una terapéutica efectiva?

Disponer de una RS sobre el tema es el camino ideal para resolver el dilema, ya que estos estudios (llamados «estudios secundarios») se construyen mediante una estrategia de investigación desarrollada explícitamente para responder a ese tipo de preguntas.

El gran número de publicaciones existentes en ciencias de la salud, dificulta que los profesionales se mantengan actualizados. Destacamos que en ese ámbito se publican anualmente varios cientos de miles de artículos, lo que significa que para mantenerse actualizado, se necesitaría leer, como mínimo, varias decenas de artículos originales por día, lo cual sería prácticamente imposible y es por ello que las RS de la literatura científica emergen como herramientas esenciales de consulta, ya que en un solo documento resumen criteriosamente los hallazgos de los estudios primarios.

## ¿Qué tipo de revisiones existen?

Existen dos tipos diferentes de revisiones: las *no sistemáticas* (también denominadas revisiones narrativas) y las *sistemáticas*. La diferencia fundamental entre ambas está dada porque las primeras carecen de un protocolo que defina los pasos que seguirá el investigador para responder la pregunta científica formulada y ello expone a la intromisión descontrolada de sesgos, restando credibilidad a la conclusión. Las segundas (sistemáticas), se adscriben a pautas metodológicas estrictas para su realización, permitiendo así su reproducibilidad, exigencia fundamental del método científico.

Se observa que en las revisiones narrativas, los autores utilizan métodos informales y subjetivos para seleccionar e interpretar los estudios, teniendo tendencia a reforzar sus ideas preconcebidas o a promover sus propios puntos de vista. Esto lleva a que los resultados y las conclusiones presenten sesgos y que las recomendaciones puedan ser inapropiadas.

Resaltamos que las revisiones narrativas no tienen protocolo, mientras que las sistemáticas deben ajustarse estrictamente a un protocolo preestablecido, lo que les otorga credibilidad al hacerlas evaluables y reproducibles. De ellas vamos a ocuparnos.

Dentro de las RS, se reconocen dos partes: la *cualitativa* y la *cuantitativa* (conocido como MA).La primera (la cualitativa) analiza y expone la calidad metodológica de los artículos incluidos y justifica la realización o la omisión de efectuar la síntesis cuantitativa (o MA).

# Revisión sistemática y meta-ánalisis

## ¿Qué entendemos por revisión sistemática y por meta-ánalisis?

Las RS se definen como la síntesis formal, cualitativa y cuantitativa de investigaciones de calidad metodológica comparable que poseen en común la misma intervención y el mismo punto final de resultado, las cuales mediante el seguimiento de un protocolo minucioso, son agrupadas con la intención de sintetizar la evidencia científica.

Su valor y utilidad radica en que permiten responder a preguntas clínicas concretas reuniendo estudios realizados de manera independiente (estudios primarios), a veces con resultados opuestos y sintetizar sus resultados.

Si la síntesis cualitativa de esa investigación bibliográfica reúne determinadas exigencias metodológicas, se puede analizar la información disponible mediante la aplicación de procedimientos estadísticos especialmente diseñados que permiten la síntesis cuantitativa de los datos (MA). Por lo tanto, MA no es sinónimo de RS sino sólo una parte de ella.

A las RS, con o sin MA, se las denomina «estudios secundarios» pues se realizan sobre estudios primarios (las investigaciones individuales). No confundir «secundarios» con un orden de calidad.

Las RS más frecuentemente publicadas son las dirigidas a efectuar la síntesis de estudios primarios experimentales. Es decir, de aquellos que prueban la fuerza de asociación entre una intervención y de un efecto beneficioso para la salud de los participantes.

## ¿Cuáles son las características generales de una RS sobre cuidados para la salud?

La RS condensa la información devolviendo el conocimiento sobre la calidad de la evidencia disponible y si es suficiente, sobre la dirección, magnitud y precisión del efecto.

Una RS sobre cuidados médicos también puede valerse de estudios observacionales que consideren a la misma intervención y punto final. Dado que el diseño de estos estudios no ofrece una protección relevante frente a los sesgos, la calidad de la evidencia será menor que si proviene de estudios experimentales adecuadamente aleatorizados y conducidos. Es de buena práctica que los resultados de las diferentes categorías de estudios, se presenten claramente agrupados a fin de que el lector pueda realizar su propia evaluación.

## ¿Cuál es la utilidad de las revisiones sistemáticas y meta-análisis?

- Resolver conflictos por resultados diferentes obtenidos por los estudios primarios.
- Aumentar la confiabilidad en las conclusiones al estrechar los intervalos de confianza de la medida del efecto, como producto del incremento del tamaño muestral producido por la sumatoria de estudios.
- Identificar sub-grupos de pacientes que podrían verse más beneficiados o perjudicados por la intervención estudiada.
- Identificar efectos adversos infrecuentes, favorecido por el incremento del tamaño muestral.
- Sugerir nuevas hipótesis sobre futuras investigaciones, aportando información para la planificación de ensayos clínicos

aleatorizados; para la estimación del costo-efectividad de las intervenciones; para estudios de evaluación de tecnologías sanitarias o para el desarrollo de guías de práctica clínica.

- Agilizar la lectura de la evidencia, ya que un solo artículo condensa el conocimiento existente sobre el tema.

## ¿Cuáles son limitaciones de las RS?

- Su calidad está determinada por la calidad de los artículos primarios incluidos.
- Los MA son muy sensibles a la metodología utilizada en su realización, como el diseño de los estudios incluidos, la selección de las variables, el análisis estadístico empleado para obtener la medida común de resultado.
- Los resultados pueden ser desviados del verdadero valor por la «intromisión» de sesgos de selección o publicación. Los efectos de los sesgos no pueden «verse» pero sí suponerse.
- La interpretación del MA en caso de existir notoria divergencia (heterogeneidad) entre los estudios, es difícil y controvertida.

Resumiendo:

Una RS (llamados estudios secundarios) es un estudio de revisión para responder a una pregunta concreta, en el que se utiliza una metodología científica claramente explicitada para la identificación, selección y valoración de estudios comparables, sintetizando sus resultados.

El MA es una parte de la RS en la cual el resultado individual de los estudios incluidos se combinan estadísticamente para calcular la

estimación global del efecto (fuerza de asociación) sobre la variable dependiente de interés.

## Aspectos del diseño de una RS y su MA

Las RS y el MA deben considerarse como verdaderas investigaciones clínicas, en las que las unidades de observación son los estudios primarios y requieren de una adecuada planificación, una considerable dedicación de recursos (especialmente humanos) y la elaboración de un protocolo de trabajo que formalice las decisiones tomadas durante la fase de diseño para conseguir los objetivos.

### ¿Cuáles son las etapas del proceso de realización de una RS?

Las etapas del proceso de planificación y realización de una RS tienen como finalidad facilitar la eventual reproducibilidad de las conclusiones. Podemos enumerar dichas etapas de la siguiente manera:

1. Formulación de la pregunta.
2. Formulación de los objetivos específicos del estudio.
3. Definir los criterios de inclusión y exclusión de los estudios primarios.
4. Si es pertinente, definir los subgrupos a estudiar.
5. Diseñar la estrategia de búsqueda de la información y aplicarla.
6. Valorar la calidad de los estudios hallados.
7. Extraer los datos y elaborar las tablas de evidencia.
8. Aplicar los procedimientos estadísticos para la combinación de los resultados.
9. Evaluar la heterogeneidad de los resultados individuales.
10. Efectuar el análisis de sensibilidad.

11. Evaluar la probabilidad de la presencia del sesgo de publicación.
12. Si es pertinente, realizar los análisis de resultados en los subgrupos.
13. Discutir los hallazgos
14. Elaborar la conclusión y recomendaciones.
15. Publicación del estudio.

En síntesis:

Las RS y su eventual MA constituyen las revisiones de mayor rigor científico. Poseen una serie de particularidades que las hacen especialmente atractivas dado que, en primer lugar, permiten una mayor generalización de sus resultados respecto de los estudios primarios, habida cuenta, entre otras, que las muestras no provienen de la misma población.

En segundo lugar, al incorporar varios estudios, se aumenta el tamaño de la muestra y por ende, se incrementa la precisión de la estimación del efecto pues se generan intervalos de confianza más estrechos, promoviendo la reducción de la incertidumbre sobre el resultado final.

Además, su realización es menos costosa que efectuar un estudio primario de similar cantidad total de participantes, planteando entonces menos problemas logísticos que los que genera el desarrollo de un estudio primario.


# Análisis e interpretación de una revisión sistemática y su meta-análisis

## ¿Qué se debe evaluar al leer una RS?

Como hemos mencionado, una RS se asemeja a una investigación clínica, en la cual los «pacientes» o unidades de observación son los artículos primarios publicados y, como en toda investigación, al leerla, se deben evaluar su validez y la utilidad de sus resultados.

Al igual que en un estudio primario, la RS será válida en la medida que haya sido efectuada siguiendo un protocolo riguroso, en el cual estén claramente definidos los objetivos, aclarando cuál es la intervención a estudiar y los puntos finales de resultados y los criterios de inclusión y exclusión de los artículos.

La definición de los criterios de inclusión es una de las decisiones más importantes de la RS. Dentro de ellos se encuentran, el diseño de los estudios primarios, las características de los participantes, el comparador, la intervención, y además que los resultados de los estudios incluidos coincidan con el enfoque de la RS.

También debe especificarse, con detalle, la exclusión de estudios siendo preciso que se informe la cantidad de exclusiones y sus motivos.

Un punto clave en la validez de una RS es la manera en la que se rastrearon los artículos o los datos de las investigaciones realizadas sobre el tema. Es decir, cuál fue la estrategia de búsqueda bibliográfica aplicada. No es difícil darse cuenta de que si esta fue realizada en forma insuficiente, las conclusiones de la RS estarán sesgadas hacia lo que surja del conjunto de artículos incluidos y no representará la síntesis verdadera de la experiencia general con respecto al tema en análisis.

A su vez, la búsqueda bibliográfica tiene dos aspectos fundamentales: las fuentes de información y las palabras clave empleadas para realizarla.

La búsqueda debe incluir todas las fuentes posibles de información incluyendo las fuentes informales, como los archivos personales, la literatura gris (tesis doctorales, comunicaciones a congresos, informes hechos para administraciones públicas, trabajos no publicados), la búsqueda manual (a partir de las citas bibliográficas de los artículos que ya se tienen) y el contacto directo con investigadores dedicados al tema, requiriéndoles información no publicada o aclaraciones sobre la que ya lo está.

No es infrecuente hallar RS que refieran que la búsqueda ha sido realizada exclusivamente sobre una base bibliográfica (típicamente, Medline). Esta estrategia de búsqueda es insuficiente pues deja afuera una enorme cantidad de medios de publicación. De hecho, no existe una base bibliográfica que incluya a la totalidad de las publicaciones médicas.

También es frecuente leer que ha sido incluida solamente bibliografía publicada en idioma inglés. Esto impone un sesgo (llamado *sesgo de lenguaje*) y el lector deberá evaluar si es posible que alguna información

de interés se haya publicado en otras lenguas. Por ejemplo, si el tema aborda algún problema de salud prevalente en determinadas regiones, será relevante que se haya buscado en la bibliografía regional. En otras ocasiones, la búsqueda aparece limitada a un determinado período de tiempo, lo cual puede ser apropiado si la intervención apareció en ese período pero en otras ocasiones, esto puede resultar excluyente de algunos estudios.

## ¿Cómo deben ser las palabras clave?

La pregunta clínica a la que se quiere dar respuesta mediante la RS debe tener un objetivo claro y ser concreta, de lo contrario, puede resultar difícil conseguir la información relevante o encontrarle sentido a la información, habida cuenta de que a partir de su enunciado rescataremos las palabras clave, necesarias para llevar a cabo una búsqueda eficiente.

Las palabras clave deben estar especificadas en la RS para permitir su evaluación y, eventualmente, la replicación de la búsqueda. Es por ello que se recomienda formular la pregunta siguiendo la premisa de que deben definirse adecuadamente al paciente, a la intervención a probar, al comparador y al resultado —*outcome*— esperado. Esta estrategia se la conoce como «pregunta PICO», acrónimo conformado con la primera letra de los conceptos: *Paciente*, *Intervencion*, *Comparador* y *Outcome* (Resultado).

## ¿Cómo es la recuperación de la bibliografía?

La recuperación de la bibliografía es una tarea larga y laboriosa. Requiere la recopilación de todos los trabajos localizados, lo cual puede ser más o menos complejo ya que los artículos pueden ser lenta pero

eficazmente recuperados, tanto en bibliotecas nacionales, extranjeras o a través de peticiones al autor. La literatura gris es más difícil de obtener.

No toda limitación en la búsqueda invalida una RS. Obviamente, la deficiencia no debería existir pero si la hay, debe ser evaluada por quien lee la RS y aplicará sus resultados.

## ¿Cómo hacer el análisis de los artículos hallados?

Una vez obtenido el material bibliográfico, comienza la etapa de analizar si ese material responde a las exigencias metodológicas de calidad que los autores se plantearon en el protocolo de la RS.

Al evaluar artículos para una RS sobre un cuidado médico (experimentales), lo habitual es exigir que:

1. Se trate de estudios experimentales en los cuales las intervenciones estudiadas se correspondan con el objetivo.
2. Que la asignación de las intervenciones haya sido adecuadamente aleatorizada.
3. Que se haya intentado un enmascaramiento apropiado de las intervenciones.
4. Que los resultados hayan sido analizados siguiendo el principio de intención de tratamiento.
5. Que la pérdida de seguimiento de pacientes haya sido mínima.
6. Que el análisis de las características de los artículos se haya efectuado por más de un investigador en forma separada y que, ante discrepancias, otro dirima la situación. Es de buena práctica que se informe la concordancia de las opiniones habida entre los evaluadores. Si fue elevada, quedarán menos dudas sobre la calidad de los artículos incluidos.

Se han propuesto varias guías orientativas destinadas a calificar la calidad metodológica de los artículos. Una de ellas es la escala propuesta por Alejandro Jadad, que se basa en determinar el control del sesgo en la asignación, el control del sesgo de evaluación del punto final y en el control del sesgo en el análisis de resultados, calificándolos cuantitativamente.

Dicha herramienta, plantea las siguientes preguntas:

- ¿El estudio fue descripto como aleatorizado?
- ¿Se describe el método para generar la secuencia de la aleatorización y fue adecuado?
- ¿El estudio se describe como doble ciego?
- ¿Se describe el método de enmascaramiento y fue un método adecuado?
- ¿Existió una descripción de las pérdidas y los retiros?

Las respuestas y puntuación posibles, son: Sí (un punto) o No (cero punto). La puntuación máxima puede llegar a 5 puntos. Si es < 3, la calidad se considera pobre.

Esta escala no es de aplicación obligada. Los criterios aplicados pueden ser estos u otros, pero siempre deben estar claramente definidos en la sección metodológica del artículo para que el lector pueda evaluar su pertinencia.

## ¿Cuáles son los elementos para considerar válida una RS?

Ya entrando en la sección Resultados, una RS bien realizada e informada incluye el detalle de las razones de las eventuales exclusiones

de artículos. Recordar que una exclusión tendenciosa compromete seriamente la confiabilidad de la RS.

Así también, debe presentar el detalle de las características fundamentales de los artículos incluidos. Ambos análisis suelen exponerse en tablas *ad hoc* (llamadas *tablas de evidencia*), herramientas que facilitan al lector la evaluación de la calidad y confiabilidad de la RS.

Si se incluyeron artículos con diferentes diseños, es conveniente presentar una tabla de evidencia por cada tipo de diseño. Esto facilita la apreciación de la calidad y resultados individuales.

Si las exigencias mencionadas se cumplen, admitiremos que la RS puede ser válida y llegará la etapa de analizar los resultados e interpretarlos cautelosamente. Si no satisface los criterios, su lectura debe ser obviada.

A fin de reducir las falencias en la publicación de una RS, un grupo de editores de las revistas médicas más reconocidas del mundo acordaron cuales serían sus exigencias mínimas para que una RS sea publicada en sus medios. El primer acuerdo se llamó «Acuerdo Quorum» y fue reemplazado en 2010 por el «Acuerdo Prisma». El acuerdo contiene un listado de exigencias que son de gran utilidad para los autores a fin de corroborar que todo lo importante esté consignado adecuadamente. Son exigencias, por ejemplo, exponer la estrategia de búsqueda bibliográfica y un flujograma que resuma los hallazgos y cómo se llegó a la cantidad de artículos incluidos. Ahondar en sus detalles excede al alcance de este capítulo.

## Análisis e interpretación de los resultados

Dos preguntas resultan relevantes a este tópico:

- ¿Cuáles son los resultados (dirección y magnitud)?
- ¿Cuán precisos son esos resultados?

### ¿Cómo se expresan los resultados?

Estos se expresan mediante el resultado de cálculos que arrojan una medida de resumen elaborada desde las medidas individuales de los resultados de los estudios incluidos, mediante una técnica estadística específica ofreciendo un indicador del efecto, llamado común o típico. Estos son, el *riesgo relativo (RR) típico*, *odds ratio (OR) típico*, *diferencia absoluta de riesgo (DAR) típico* o la *cantidad necesaria a tratar (CNT) típica*, si las variables de resultados son dicotómicas. Si son continuas, se las expresa como la diferencia ponderada de medias o la diferencia estandarizada de medias.

Se leen de igual manera que al interpretar una medida del efecto en un estudio individual. Ellas informan la dirección del efecto de la intervención y la magnitud, mientras que sus intervalos de confianza lo hacen sobre la precisión de la medición.

Atención, su cálculo no es el promedio surgido de la simple sumatoria de los casos del grupo intervención y del grupo control, a fin de obtener una medida global, sino que considera la variabilidad del resultado de cada uno de ellos. El resultado de los estudios con escasa cantidad de participantes, está más sujetos al azar y por lo tanto, se les otorga una influencia menor (denominada peso) en los resultados finales. Es así que los estudios con gran cantidad de participantes, tienen mayor influencia en los resultados finales que los estudios pequeños.

### ¿Qué modelos matemáticos se aplican para su cálculo?

Se han descrito dos maneras de calcular la medida de resumen común según se presuma la existencia o no de heterogeneidad. Son los llamados modelo de «efecto fijo» o de «efecto aleatorio».

- **Modelo de efecto fijo** («*fixed-effect model*», método de Mantel-Haenzel y método de Richard Peto —éste sólo es aplicable cuando se utiliza el indicador *odds ratio*—): asume que no existe heterogeneidad entre los estudios incluidos en la revisión, de modo que todos ellos estiman el mismo efecto y las diferencias observadas se deben únicamente al azar. Arroja intervalos de confianza más estrechos, por lo que es menos probable que el intervalo de confianza no cruce la línea del valor 1.
- **Modelo de efecto aleatorio** («*random-effect model*», propuesto por Der Simmonian-Laird): asume que los estudios incluidos en la revisión constituyen una muestra aleatoria de todos los estudios existentes. Arroja IC más amplios, por lo que resulta más cauto para admitir diferencias.

## Heterogeneidad, sensibilidad, subgrupos, publicación

Al tratarse de una medida de síntesis de los resultados de experiencias primarias, se deberán considerar otros aspectos que orientarán sobre la robustez o fiabilidad de las conclusiones obtenidas. La presencia de heterogeneidad entre los ensayos clínicos incluidos y el sesgo de publicación, constituyen dos de los principales problemas metodológicos que podrían afectar la validez de la RS.

## ¿Qué es y cómo se evalúa la existencia de heterogeneidad?

Se habla de «heterogeneidad» cuando los resultados de los estudios primarios no resultan razonablemente coincidentes, introduciendo la duda sobre la utilidad de una medida de resumen del efecto. Las causas de la heterogeneidad son varias, dentro de las cuales las más relevantes, son que existan diferencias en la intervención experimental o en la de comparación o diferencias en los criterios de elegibilidad de los participantes. Quien realiza o evalúa una RS, debe examinar con detalle esos ítems para integrar los artículos a la RS.

La evaluación de la heterogeneidad es la medida que expresa si las diferencias del efecto observadas entre los estudios podrían ser debidas a causas diferentes del azar, ya que es muy poco probable que todos los estudios incluidos muestren un efecto exactamente igual, tanto en dirección como en magnitud.

Interesará evaluar si tales diferencias podrían ser debidas a la simple variabilidad que impone el azar o a diferencias objetivas entre los estudios. Si la evaluación presume que hay heterogeneidad (es decir, que las diferencias observadas se deban a otros factores distintos del mero azar), se cuestiona si correspondería elaborar el MA o remitirse a discutir las probables fuentes de la discrepancia hallada. Estas pueden estar vinculadas a diferencias en las poblaciones participantes; a la medición y definición de las variables estudiadas; a que algunos de los aspectos de la intervención (por ejemplo, dosis, duración del tratamiento) no sean equiparables o al diseño del estudio y su calidad metodológica general. Existen varias alternativas para evaluar la presencia o ausencia de heterogeneidad.

## ¿Cuáles son las alternativas para evaluar la heterogeneidad?

Una manera simple e intuitiva de evaluar la heterogeneidad, es apreciar visualmente la representación gráfica de los resultados («*forest plot*») y analizar si la dirección del efecto de los estudios incluidos apunta en el mismo sentido y si la magnitud y precisión (barras indicativas de los intervalos de confianza) están superpuestas en su mayor parte. Si así ocurre, se podría suponer que no hay heterogeneidad entre los estudios (se dice que los estudios son homogéneos). Si esto no ocurre, es decir, si alguna o todas las apreciaciones están ausentes, se debe admitir la disparidad entre ellas y, por ende, suponer la presencia de heterogeneidad.

Además de la apreciación visual, existen métodos estadísticos para determinarla con mayor seguridad, de los que hablaremos más adelante.

Si hay heterogeneidad, está cuestionado si corresponde calcular una medida de resumen ya que su confiabilidad se vería reducida. Si se la calcula a pesar de la heterogeneidad detectada, es recomendable aplicar el modelo de efecto aleatorio ya que es más cauto para admitir diferencias pues su intervalo de confianza es más amplio.

## ¿Qué es la evaluación de la sensibilidad?

Consiste en una estrategia que permite apreciar si los resultados se mantienen al modificar el conjunto de las publicaciones incluidas en el MA. Es decir, determinar en cuánto se modifica la medida común de resultado al recalcularla sin incluir un determinado estudio. Si la proporción de modificación es pequeña (no modifica la dirección), sustenta la idea de que el valor obtenido para la medida común no es dependiente de un estudio en particular.

## ¿Qué es el análisis de subgrupos?

Al sintetizar los resultados de varios estudios primarios, podría ocurrir que el indicador común resulte particularmente influenciado por alguno de ellos en particular o que los investigadores decidan explorar ese indicador en determinados grupos de participantes. Para ello, se seleccionan los estudios que contienen los participantes de interés y se realiza un nuevo MA. De esta manera, el indicador común de resultado será el propio de ese conjunto de participantes permitiendo la elucubración científica con respecto al efecto del cuidado médico evaluado en la población seleccionada. Es decir, se obvia el efecto que podría imponer la existencia de una característica determinada en la población analizada.

Atención: no debe confundirse el análisis de subgrupos con el análisis de sensibilidad. En el primero, se excluyen los participantes o estudios que contienen una variable determinada. En el segundo, se excluye un estudio primario completo por vez para medir su influencia en el resultado final del conjunto.

## ¿Qué es el sesgo de publicación?

Se denomina así a un desvío potencial de la conclusión de un MA relacionado a que no se han incorporado la totalidad de los estudios realizados.

Anteriormente comentamos que las conclusiones de una RS, pueden verse afectadas por una búsqueda bibliográfica insuficiente. Pero, por otro lado, la búsqueda puede haber sido bien ejecutada cubriendo todas las posibles fuentes de información y rastreando los artículos con las palabras

claves apropiadas y sin embargo, estar presente el llamado «sesgo de publicación».

Este sesgo se produce cuando lo que se publica no representa el total de las investigaciones realizadas acerca de un tema y constituye una amenaza potencial para la validez de las conclusiones de un MA.

Es un hecho bien conocido que muchos ensayos terminados no llegan a publicarse. Esto es más frecuente cuando el resultado del ensayo es «negativo», es decir, cuando no se demuestran diferencias significativas entre los grupos comparados.

También influye tendenciosamente, el hecho de no publicar un estudio cuando su resultado es desfavorable a un nuevo fármaco presentado por el financiador. Por otra parte, los editores de revistas médicas tienden más a rechazar una publicación con resultados «negativos» ya que no son «noticia» (el reconocimiento de la existencia del sesgo de publicación comenzó en 1956, cuando el director de la revista *Journal of Abnormal Social Psychology*, señaló que los estudios negativos tenían menos probabilidades de publicarse en su revista).

## ¿Qué circunstancias favorecen la presencia del sesgo de publicación?

Hasta el momento se han identificado diferentes circunstancias que influyen en la presencia del sesgo de publicación, como la financiación, el conflicto de intereses, el prejuicio, el prestigio de la institución, la duplicación de los estudios publicados y el idioma, pero dos son los factores que más claramente se han relacionado y que guardan una estrecha relación entre ellos: la *significación estadística* y el *tamaño de la muestra*.

El impacto de este sesgo, independientemente de su causa, puede generar un desvío ficticio de la dirección y magnitud del efecto ya que faltarán en el MA, por ejemplo, las experiencias «negativas» que ejercerían una acción contraria al efecto sugerido por las experiencias «positivas».

El riesgo de este sesgo es mayor cuando la revisión dispone de gran cantidad de estudios de tamaño muestral reducido ya que los autores suelen apurarse por publicar sus hallazgos iniciales, particularmente cuando son «positivos».

### ¿Cómo se evalúa el sesgo de publicación?

La evaluación de la existencia de sesgo de publicación se basa en la suposición empírica de que los resultados de los estudios se distribuyen armónicamente. Es decir, que se supone que por cada estudio con un efecto determinado, debe existir otro con el efecto opuesto. Existen varios métodos disponibles para la estimación del sesgo de publicación.

Suele utilizarse el denominado gráfico de embudo o «*funnel plot*», cuya descripción e interpretación se desarrollan más adelante en este capítulo.

## Representación gráfica de los resultados

### ¿Cómo se representan gráficamente los resultados?

Se utiliza el llamado gráfico «*forest plot*». Este nombre surge pues la distribución de los puntos estimados de los estudios primarios se asemejaría a la imagen de la distribución de los árboles en un bosque. Conservando la estructura básica, se pueden hallar en diferentes formatos y componentes. De todas maneras, la técnica de lectura es siempre semejante.

La siguiente figura presenta un gráfico «*forest plot*» con su diagramación y componentes habituales. Puede observarse que está compuesto por un eje horizontal (se lo puede colocar en la parte superior o inferior), reglado con una escala logarítmica que se extiende a ambos lados del valor de la unidad (valor 1) y un eje vertical, que se ubica en la posición que corresponde a ella. Esta expresa que el riesgo de adquirir el efecto analizado, es similar entre la intervención en prueba y la comparación. Lo que se ubique a la izquierda del observador estará sugiriendo que el riesgo es menor al recibir la intervención en prueba y mayor, si se ubica a la derecha.

La significación clínica de esa posición dependerá de cuál sea el punto final de medición. Si, por ejemplo, es muerte, será deseable que se ubique sobre el campo de la izquierda, pero si es mejoría, será deseable que se ubique en el campo de la derecha. Por ello, al leer estos gráficos, el lector debe cerciorarse de cuál es el punto final adoptado.

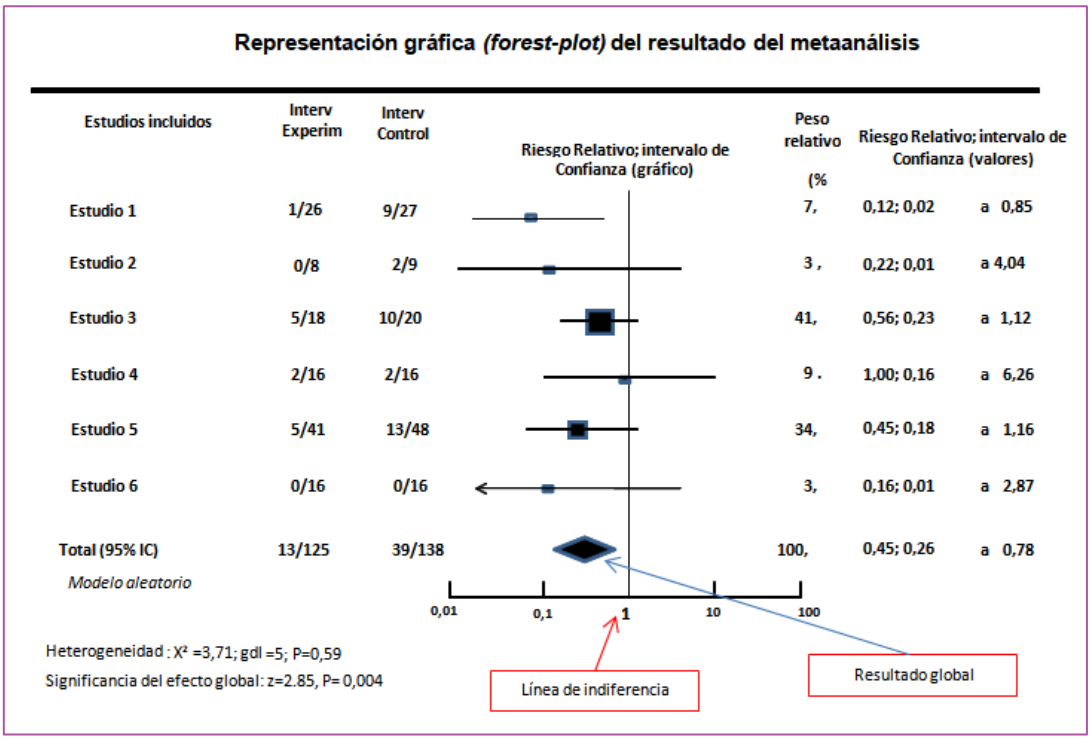
El gráfico presenta el punto estimado del efecto de cada estudio (aquí representado con un cuadrado cuyo tamaño guarda relación directa con el tamaño de la muestra y la precisión del mismo) y su correspondiente intervalo de confianza, representado por una línea horizontal extendiéndose a cada lado y que va desde el menor valor del intervalo de confianza al mayor. En el gráfico, el valor numérico del peso relativo está expuesto en la columna correspondiente.

Si la línea horizontal que representa al intervalo de confianza no contacta la línea vertical, admitimos que el efecto es decididamente el expresado, es decir que reduce o aumenta la probabilidad de que ocurra el punto final, acorde a su posición. Si la contacta, persistirá la incertidumbre sobre cuál es la verdadera ubicación del punto estimado, si a la derecha, a la izquierda o sobre la línea.

Suele estimarse la probabilidad de que exista una diferencia estadísticamente significativa en el efecto de las intervenciones en comparación, mediante el cálculo del valor del estadístico Z y la probabilidad representada. Decimos entonces que el resultado no resulta estadísticamente significativo.

El valor de la medida común o típica suele mostrarse por un rombo (inicialmente aplicado en las revisiones de la Biblioteca Cochrane), cuyo eje vertical debe estar en la posición del punto estimado, mientras que sus extremos laterales abarcan el intervalo de confianza. En la columna de la izquierda se listan los estudios y su año de publicación. En la columna de la derecha se consignan los tamaños muestrales de cada uno de los estudios primarios y al final de la misma, el valor de la medida común (ver figura siguiente).

La medida común de resultado pretende ser más fiable que la obtenida en cada uno los estudios individuales, entre otras cosas porque es extraída desde una muestra poblacional más amplia ya que participan los participantes incluidos en todos los estudios considerados. Con ello, mejora la precisión (recordemos que la precisión mantiene una relación directa con la cantidad de participantes).




## ¿Qué muestra el ejemplo de la figura?

El ejemplo muestra los resultados imaginarios de un MA llevado a cabo para evaluar el efecto de la intervención experimental *versus* la intervención control o comparativa, sobre el punto final XXXX.

El Estudio 1, si bien tiene un tamaño muestral reducido, es el único cuyo intervalo de confianza no atraviesa el punto de indiferencia. El Estudio 3 es el de mayor tamaño muestral y mayor precisión (el cuadrado que representa al punto estimado es el de mayor superficie y su intervalo de confianza estrecho). El resto de los estudios están representados por cuadrados menores en función del tamaño de su muestra. Sus intervalos de confianza son más extensos y transponen la línea vertical. El Estudio 4 ubica su punto estimado sobre la línea de indiferencia.

Finalmente, está representada la medida común del efecto (rombo), el riesgo relativo típico (RRt), en tanto que se aclara que para su obtención se aplicó el modelo de efecto aleatorio. El punto estimado (eje vertical del

rombo) dice que la intervención experimental aumenta la probabilidad de mejorar el punto final de resultado al recibirla en comparación con el grupo de control en un 55 % (RRt = 0,45). El extremo derecho del rombo no contacta con la línea vertical (IC95 % superior, 0,78), indicando que la intervención experimental tiene al menos un 95 % de probabilidades de que su efecto sea al menos de 22 %, en tanto que su límite inferior (IC95 % inferior, 0,26) indica que esa intervención tiene al menos un 95 % de probabilidades de que su efecto sea de hasta de un 78 %.

El análisis de heterogeneidad muestra que no hay heterogeneidad relevante ya que el valor del $chi^2$ es menor que los grados de libertad y la probabilidad de que no exista heterogeneidad es del 59 % (que luego comentaremos en mayor detalle). También se agrega la medida de la significancia del efecto que dice que la probabilidad de que la diferencia del efecto no tenga significación estadística, es del 4 por mil.

## ¿Qué otras cosas muestra el ejemplo de la figura?

Volviendo al ejemplo, la dirección del efecto de la intervención sobre el punto final de resultado muestra un claro efecto, ya que la intervención experimental disminuye la probabilidad de ocurrencia del punto final con mayor efectividad que la intervención comparadora. Observe que los puntos estimados (representados por cuadrados) se hallan casi todos (menos uno) ubicados sobre el mismo lado del gráfico, la dirección del efecto es similar y los límites superior e inferior de sus intervalos, se superponen.

Bajo estas circunstancias, puede suponerse mediante la apreciación visual que existe homogeneidad entre los estudios incluidos en el MA, lo

cual permitiría aumentar la confianza en la verosimilitud de la medida de resumen común del efecto.

También es posible evaluarla por procedimientos estadísticos probando la hipótesis de homogeneidad. Para ello, la más empleada es la prueba de $chi^2$ partiendo de la hipótesis nula de que los estudios son homogéneos.

## ¿Qué mostró el resultado de la evaluación de la heterogeneidad en el ejemplo?

El resultado del test mostró: Heterogeneidad: $chi^2$ = 3,71, gdl (grados de libertad) = 5; P = 0,59 (un valor p no significativo y por ende, la prueba confirma la ausencia de heterogeneidad, ya que en éste caso no es posible rechazar la hipótesis nula).

Otra manera de evaluar la presencia de heterogeneidad, es considerando el valor de $chi^2$. Se admite que si el valor del $chi^2$ es menor a la sumatoria de los grados de libertad + 1, la p será siempre mayor a 0,05 y por lo tanto, no se rechazaría la hipótesis de homogeneidad entre los estudios. En el ejemplo, $chi^2$ = 3,71, lo que resulta menor que 6 (es decir, es menor que la sumatoria de los gdel + 1). Bajo esta premisa, p debería ser mayor a 0,05 y lo es, p = 0,59.

A su vez, el estadístico inglés Douglas Altman propuso medir la inconsistencia (en lugar de hablar de heterogeneidad) de los resultados mediante un índice denominado $I^2$ (I cuadrado o Índice de inconsistencia) que expresa la proporción de la variabilidad de los resultados que sería debida a heterogeneidad y no al azar. Se sugiere que si el valor de $I^2$ es menor a 40 %, puede aceptarse que no hay heterogeneidad relevante; si es de 40 a 60 %, la situación es dudosa y si es mayor a 60 %, se debe considerar que la probabilidad de heterogeneidad, es marcada.

Si hay heterogeneidad, está cuestionado si corresponde calcular una medida de resumen ya que su confiabilidad se vería reducida. Si se la calcula a pesar de la heterogeneidad detectada, es recomendable aplicar el modelo de efecto aleatorio ya que es más cauto para admitir diferencias pues su intervalo de confianza es más amplio.

## ¿Cómo se hace la evaluación y representación gráfica del sesgo de publicación?

Existen varios métodos disponibles para la estimación del sesgo de publicación.

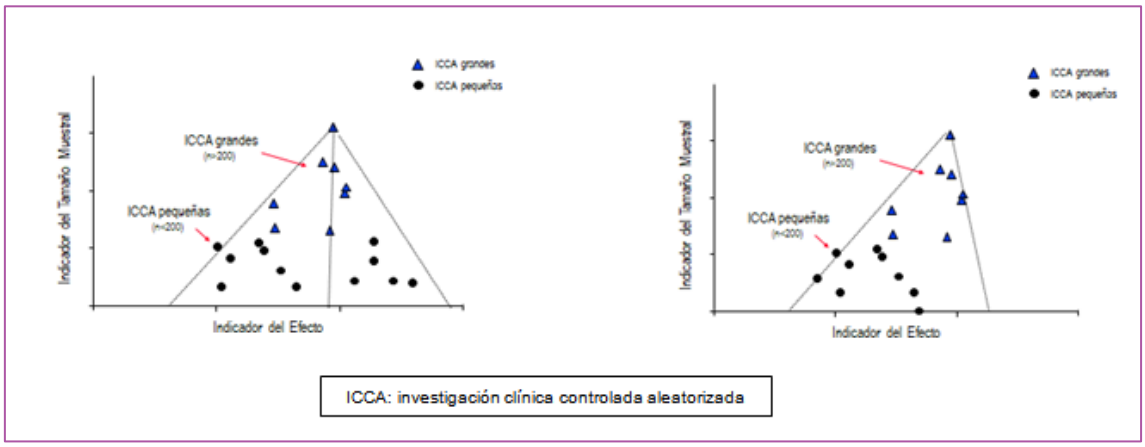
Es frecuentemente utilizado el gráfico de embudo o «*funnel plot*». Éste consiste en un diagrama de puntos donde en el eje horizontal se ubica cada estudio en relación con su medida del tamaño del efecto y en el vertical según el tamaño de su muestra. La distribución suele adoptar la forma de un triángulo (embudo), de allí su nombre.

El primer paso es la evaluación visual del gráfico, mediante la cual se estima si la distribución de los puntos que representan a cada estudio, es asimétrico respecto a la línea que representa la altura del triángulo y cuyo vértice va a estar determinado por el estudio de mayor tamaño muestral. Los catetos se diagraman desde el vértice pasando por los extremos de intervalo de confianza del estudio de mayor tamaño muestral. Si esa distribución resulta asimétrica, se presume que faltan estudios y, en consecuencia, se asume que es posible que exista un sesgo de publicación. La detección del sesgo de publicación disminuye la confianza en las conclusiones de la revisión.

Existen otras técnicas estadísticas probabilísticas, tal como la prueba de Egger, disponible en la mayoría de los softwares para la realización de MA

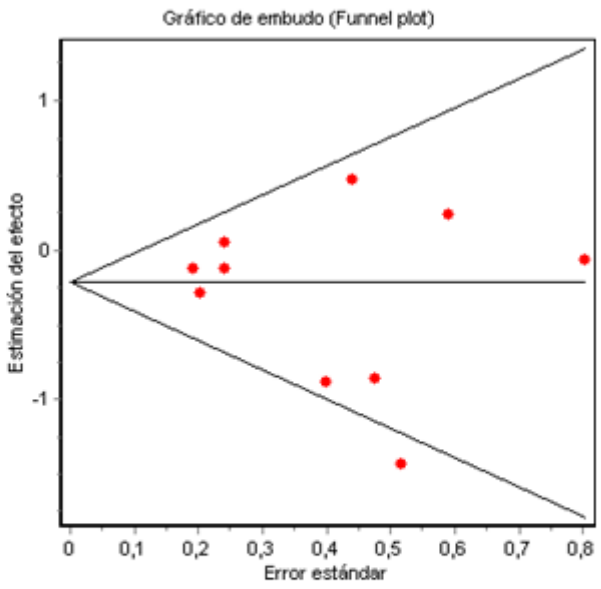
que permiten evaluar de manera más objetiva la existencia de un potencial sesgo de publicación, arrojando un valor de probabilidad de que exista ese sesgo. Convencionalmente, se admite que ese sesgo es muy improbable si la probabilidad resultante es menor al 5 % ($p < 0,05$).

Las siguientes figuras muestran ejemplos de simetría del embudo (sugerencia de que no existiría sesgo) y asimetría del embudo (sugerencia de que existiría sesgo). La ubicación de cada punto, es acorde al efecto hallado en cada estudio en relación con su tamaño muestral.




ICCA: investigación clínica controlada aleatorizada

El siguiente ejemplo, el gráfico «*funnel plot*» (realizado con el software Epi Data vers 3.1) se presenta orientado de manera horizontal en el que se aprecia una distribución razonablemente simétrica de los puntos (es la ubicación de los puntos estimados de cada uno de los estudios incluidos en el meta-análisis). Aplicada la técnica estadística de la prueba de Egger, esta dice que la probabilidad de que exista sesgo de publicación es de 45 % por lo que se rechaza la existencia del mismo (nivel convencional habitual para el rechazo de la hipótesis nula, es $p > 5$ %).




Prueba de Egger

| Estadístico t | gdl | Valor p |
|---|---|---|
| -0,7894 8 | 0,4527 | (no hay sesgo de publicación) |

Recuérdese que la evaluación es siempre empírica, pues nunca se podría aseverar que existen esos artículos no publicados.

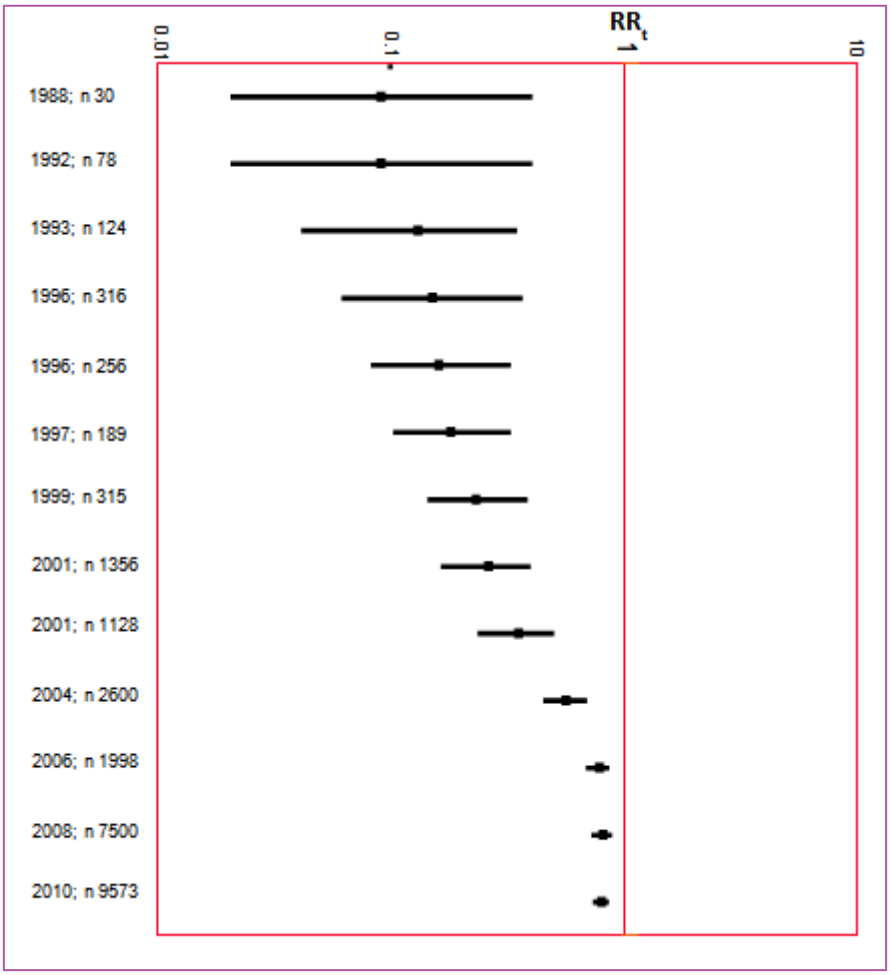
## Meta-análisis acumulativo

Se denomina así al procedimiento estadístico en que se calcula una nueva medida de resumen cada vez que se adiciona un nuevo estudio primario. La medida final es coincidente con la medida que arrojó el MA realizado de manera tradicional.

Este procedimiento aporta una visión de la evolución de la medida de resumen y permite presumir si tiene probabilidad de modificarse sustancialmente por un nuevo estudio. Esto permite decidir si es razonable la realización de nuevos estudios primarios sobre el tema. Debe tomarse en cuenta que realizar estudios cuando es remota la probabilidad de

generar un aporte de significación a la evidencia disponible, es inapropiado.

En el ejemplo siguiente puede verse como fue modificándose la posición del punto estimado y la amplitud del intervalo de confianza, en la medida que se adicionaron los estudios en orden cronológico (según año de publicación) e incrementó el tamaño muestral, hasta llegar a un punto en el que las modificaciones no produjeron un efecto evidente sobre la medida de la fuerza de asociación típica adoptada. También se lo presenta como una variante utilizada para analizar la tendencia de la evidencia a lo largo del tiempo (ver la figura siguiente).



## Consideraciones finales

Las RS y su subproducto, el MA, constituyen procedimientos que incrementan la calidad de las revisiones bibliográficas dado que obligan a una ejecución metódica, haciendo sus resultados más confiables e interpretables.

Aún persisten controversias y cuestiones de investigación pendientes en torno a sus aspectos metodológicos, ya que su implementación es una técnica empírica y no experimental.

A pesar de ello en la actualidad son aceptadas como las herramientas más apropiadas para producir evidencias de alta calidad, a pesar de su carácter de estudios observacionales retrospectivos.

## Lecturas recomendadas


CHALMERS I, ALTMAN D. *Systematic Reviews*. BMJ Publishing Group 1995.

CHALMERS I, HAYNES B. Reporting, updating, and correcting systematic reviews of the effects of health care. *BMJ*. 1994;309(6958):862-865. <doi:10.1136/bmj.309.6958.862>

CLARKE M, BRICE A, CHALMERS I. Accumulating Research: A Systematic Account of How Cumulative Meta-Analyses Would Have Provided Knowledge, Improved Health, Reduced Harm and Saved Resources. *PLOS One*. 2014;9:e102670. <doi:10.1371/journal.pone.0102670>

EGGER M, SMITH GD, SCHNEIDER M, MINDER C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997;315:629-634. <doi:10.1136/bmj.315.7109.629>

EYSENCK HJ. Meta-analysis and its problems. *BMJ*. 1994;309:789-792. <doi:10.1136/bmj.309.6957.789>

GLASS GV. Primary, secondary and meta-analysis of research. *Educational Researcher*. 1976;5:3-8. <https://doi.org/10.3102/0013189X005010003>

HIGGINS JP, THOMPSON SG, DEEKS JJ, ALTMAN DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327:557-560. <doi:10.1136/bmj.327.7414.557>

HIGGINS JPT, GREEN S (editors). Manual Cochrane de revisiones sistemáticas de intervenciones. Version 5.1.0. The Cochrane Collaboration, 2011. <https://es.cochrane.org/sites/es.cochrane.org/files/public/uploads/manual_cochrane_510_web.pdf>

JADAD AR, MOORE RA, CARROL D, JENKINSON C, REYNOLDS DJM, GAVAGHAN DJ, ET AL. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Controlled Clinical Trials*. 1996;17:1-12. <doi:10.1016/0197-2456(95)00134-4>

JAMES LIND LIBRARY. Cómo reducir la obra de la casualidad a través del metanálisis. <https://www.jameslindlibrary.org/essays/>

JUNI P, ALTMAN DG, EGGER M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ*. 2001;323:42-46. <doi:10.1136/bmj.323.7303.42>

KHAN K, DAYA S, JADAD A. The importance of quality of primary studies in producing unbiased systematic reviews. *Arch Intern Med*. 1996;156:661-666.

MOHER D, JADAD AR, COOK DJ, JONES A, ET AL. Does the poor quality of reports of randomized trials exaggerate estimates of intervention effectiveness reported in meta-analysis? *Lancet*. 1998;352:609-613. <doi:10.1016/S0140-6736(98)01085-X>

MULROW C, LANGHORNE P, GRIMSHAW J. Integrating heterogeneous pieces of evidence in systematic reviews. *Ann Intern Med*. 1997;127:989-995. <doi:10.7326/0003-4819-127-11-199712010-00008>

SACKETT DL, STRAUSS SE, RICHARDSON WS, ROSEMBERG W, ET AL. *Evidence-based medicine: how to practice and teach EBM*. 2nd. Ed. London: Churchill-Livingstone; 2000.

SACKS H, BERRIR J, REITMAN D, ANCONA-BERK VA, CHALMERS T. Meta-analyses of randomized controlled trials. *N Engl J Med*. 1987;316:450-455. <doi:10.1056/NEJM198702193160806>

SCHULZ KF, CHALMERS I, HAYES RJ, ALTMAN DG. Empirical evidence of bias. *JAMA*. 1995;273:408-412. <doi:10.1001/jama.273.5.408>

STERNE JAC, EGGER M, SMITH GD. Investigating and dealing with publication and other biases in meta-analysis. *BMJ*. 2001;323:101-105. <doi:10.1136/bmj.323.7304.101>

THE COCHRANE COLLABORATION. *Manual Cochrane de revisiones sistemáticas de intervenciones, 2011.*

<https://es.cochrane.org/sites/es.cochrane.org/files/uploads/Manual_Cochrane_510_reduit.pdf>

THOMPSON SG. Why sources of heterogeneity in meta-analysis should be investigated? *BMJ*. 1994;309:1351-1355. <doi:10.1136/bmj.309.6965.1351>

YOUNG HL. An overview of meta-analysis for clinicians. *Korean J Intern Med*. 2018;33:277-283. <doi:10.3904/kjim.2016.195>