DOI: https://doi.org/10.11144/Javeriana.umed63-1.aphl

Abordaje práctico de la heterogeneidad en la lectura crítica de revisiones sistemáticas y metanálisis

Practical Approach to Heterogeneity in Critical Appraisal of Systematic Reviews and Meta-analyses

Recibido: 23/08/2021 | Aceptado: 11/10/2021

Mateo Pineda-Álvarez[a]
Médico y cirujano, Facultad de Medicina, Universidad de Antioquia, Medellín, Colombia
ORCID: https://orcid.org/0000-0002-2552-9117
Juan Pablo Zapata-Ospina
Médico psiquiatra. Magíster en Epidemiología Clínica. Profesor del Instituto de Investigaciones Médicas, Facultad de Medicina, Universidad de Antioquia, Medellín, Colombia. Grupo Académico de Epidemiología Clínica (GRAEPIC)., Colombia
ORCID: https://orcid.org/0000-0002-1815-5583

RESUMEN

Las revisiones sistemáticas de la literatura (RSL) y los metanálisis (MA) constituyen una herramienta confiable para la toma de decisiones dentro de la práctica de la medicina basada en la evidencia, por su alto poder y precisión, al combinar los resultados provenientes de estudios primarios. Uno de los problemas más frecuentes es que los resultados de los estudios varíen de manera significativa entre sí, lo que se conoce como heterogeneidad, esto es, la variabilidad existente entre los resultados de estudios primarios sobre un desenlace, que puede deberse a las diferencias inherentes a cada estudio. Esta se puede identificar a través de métodos estadísticos y gráficos. Y una vez identificada, se debe ahondar en las causas que la expliquen para comprender cómo pueden aplicarse los resultados. También es posible incorporar la heterogeneidad en el MA y minimizarla con modelos estadísticos para combinar los efectos. En la presente revisión narrativa de la literatura, se pretende definir la heterogeneidad en las RSL y MA y ofrecer un abordaje práctico para la lectura crítica de estos estudios secundarios.

Palabras clave
revisión sistemática; metanálisis; métodos; medicina basada en la evidencia; sesgo.

[a] Autor de correspondencia: mateo.pinedaa@udea.edu.co

Cómo citar: Pineda Álvarez M, Zapata-Ospina JP. Abordaje práctico de la heterogeneidad en la lectura crítica de revisiones sistemáticas y metanálisis. Univ. Med. 2022;63(1). https://doi.org/10.11144/Javeriana.umed63-1.aphl

ABSTRACT

Systematic reviews (SR) and meta-analyses (MA) are studies that constitute a highly confident tool for decision-making within the practice of evidence-based medicine due to their high power and precision when combining the results of primary studies. One common issue in this type of study appears when there is a significant variation in the results of the primary studies, which is known as heterogeneity. Heterogeneity can be defined as the existing variability in measured outcomes between primary studies that can be explained by the inherent differences across individual studies. It can be detected through statistical and graphic methods. Once identified, the causes that explain such heterogeneity must be recognized

to apply accurately the results. Heterogeneity can also be minimized in a MA by changing the statistical models used to combine the effects. The present narrative review aims to define heterogeneity in SR and MA and offer a practical approach to the critical appraising of these studies. Systematic literature reviews (RSL) and meta-analyzes (MA) due to their high power and precision when combining the results from primary studies

Keywords
systematic review; meta-analysis; methods; evidence-based medicine; bias.

Introducción

El médico enfrenta problemas clínicos que lo harán plantearse preguntas que posiblemente deba responder con artículos científicos. Sin embargo, resulta poco práctico buscar todos los disponibles, pues la cantidad es exorbitante e incrementa continuamente. Para el 2017, por ejemplo, se registraron más de 800.000 artículos en Medline y para el año siguiente fueron 900.000 (1). Por eso, las revisiones sistemáticas de la literatura (RSL) cumplen un papel fundamental, pues intentan resumir todos los artículos primarios sobre una pregunta clínica estructurada. A diferencia de las revisiones narrativas generales, la búsqueda de la literatura es explícita, para que sea reproducible y exhaustiva y para encontrar la mayor cantidad de artículos, de manera que se seleccionen luego los más apropiados.

Una RSL puede incluir diferentes tipos de estudios primarios, desde experimentales (como ensayos clínicos aleatorizados) hasta observacionales (como los de cohortes) (2). Si los resultados son consistentes en todos los estudios, se pueden agrupar y sintetizar de forma cuantitativa o resumir estadísticamente, que es lo que conocemos como metanálisis (MA) (2,3). Muchos de estos estudios miden sus resultados con estimadores puntuales —por ejemplo, el riesgo relativo (RR), la razón de posibilidades u *odds ratio* (OR) o la diferencia de medias —, que se resumen en el MA. Así, se reúne información de distintas muestras para tener una mejor idea del universo de pacientes de la pregunta y alcanzar estimaciones más precisas. De hacerse de forma creíble, una RSL con MA opera como una lente que enfoca otros estudios y podría ser una herramienta muy confiable para la toma de decisiones dentro de la práctica de la medicina basada en la evidencia (2,4).

Uno de los problemas más frecuentes en las RSL es que los resultados de los estudios varíen de manera significativa entre sí, lo que conduce a una alta heterogeneidad al realizar el MA (2). Debido a que las RSL siempre agruparán estudios de diferentes autores y características distintas, la heterogeneidad siempre estará presente en menor o mayor medida; de ahí la importancia de identificarla y analizarla para decidir si es tolerable, al punto de hacer confiable un MA. El objetivo de esta revisión narrativa es ofrecer una guía práctica dirigida a usuarios de la literatura médica para la comprensión de la heterogeneidad en las RSL.

¿En qué consiste la heterogeneidad?

La heterogeneidad es la variabilidad existente entre los resultados de los estudios primarios acerca de un desenlace y que puede explicarse por las diferencias inherentes a cada estudio (5,6). Por ello, se divide en tres tipos: clínica, que se refiere a diferencias en los pacientes, el escenario en que fue realizado o las intervenciones; metodológica, por diferencias en el diseño del estudio, su calidad o análisis realizado, y estadística, la variación de la estimación del efecto como tal, consecuencia de la heterogeneidad clínica, de la heterogeneidad metodológica o por azar (7). Al combinar en un MA los resultados de varios estudios, es de esperarse que se obtenga un estimador puntual con intervalos de confianza mucho más estrechos y puedan aplicarse a una población más amplia. Pero si los estudios difieren mucho entre sí, estos cálculos pueden estar sesgados y llevar a conclusiones erróneas (2).

¿Cómo determinar que existe heterogeneidad?

Debe comenzarse por la evaluación desde el punto de visto clínico. Para eso el lector

debe saber sobre la enfermedad en estudio, para que al ver los estudios incluidos, pueda analizar las características de los participantes, las intervenciones (en estudio y cointervenciones), así como la medición de los desenlaces, y así determinar si alguno de esos factores plantea *a priori* un problema de estar combinando "peras y manzanas". Luego puede evaluarse por medio de pruebas estadísticas y de gráficos.

Pruebas estadísticas

Q de Cochran

Las pruebas de hipótesis para calcular un valor de *p* son una forma estadística que se encuentra frecuentemente en la literatura médica. En este caso, esta prueba relaciona el aporte o peso de cada estudio con la medida resumen del MA. De esta forma se ve qué tan alejado está cada estudio de ese resumen. Con esa lógica, los autores usan este método para probar si hay homogeneidad o no, con la proposición de una hipótesis nula (que afirma que hay homogeneidad: los resultados son similares entre sí) y una hipótesis alterna (*no* hay homogeneidad: los resultados no son similares). Se estima una probabilidad sigue una distribución χ^2 (chi cuadrado), que no es más que una descripción de la forma abstracta que toman todos los valores posibles de esa probabilidad y con la que se define un punto crítico (o valor de *p*).

Lo que evalúan los autores al calcular ese valor de *p* es la probabilidad de que con los datos de todos los estudios se rechace o no la hipótesis nula. Como toda probabilidad, va de 0 a 1, y entre más cercano se encuentre del 0, es más probable que la hipótesis nula sea falsa y se rechace, es decir, es muy probable que no exista homogeneidad y los resultados sean heterogéneos. En cambio, entre más cercano se encuentre el valor de *p* al 1, más probabilidad hay de que con los datos la hipótesis nula sea verdadera y no se pueda rechazar. Ello indicará que los resultados son homogéneos (8). En los MA, usualmente, se toma un valor crítico de *p* de 0,1 para mostrar que si es menor de ese valor, no hay homogeneidad; en tanto que si es mayor, existe homogeneidad (9). Indica únicamente si hay o no heterogeneidad, y ante un número pequeño de estudios su poder estadístico es limitado (10).

Test de Breslow-Day

Se utiliza específicamente para examinar la homogeneidad en las OR. Se calcula a partir del número de eventos observados frente a los esperados y cómo fueron en el grupo de intervención. Es una prueba de hipótesis donde la nula es que las OR son similares entre estudios y el valor de *p* también tiene una distribución χ^2. Tiene un bajo poder cuando el número de estudios es pequeño; lo ideal es más de 20 (11). Por lo general, los autores de RSL anotan en la sección de métodos que usaron este test y el nivel de significancia que tomaron. Por ejemplo, en una RSL se comparó el uso de griseofulvina versus terbinafina para tratar la *tinea capitis* y se calculó una OR combinada de 0,86 en favor de la terbinafina, aunque no fue significativa (IC95 %: 0,57 a 1,27). El test de Breslow-Day dio un valor de *p* de 0,015, que fue considerado significativo, por ser menor de 0,05 (12), e indica que los datos apoyan rechazar la hipótesis nula de homogeneidad, es decir, rechazar que las OR son similares, y habría heterogeneidad.

Estadístico de inconsistencia (I2)

Aunque toma como base la Q de Cochran, el I^2 permite cuantificar la heterogeneidad en lugar de probar si está o no presente. De ahí que sea más interpretable y la forma más frecuente en la que los autores reportan la heterogeneidad. Se calcula a partir de la resta del valor de la Q de Cochran y el conjunto de observaciones (estudios) con los que se estimó (llamado grados de libertad, que sería número de estudios menos 1) dividida por el valor de Q, lo cual se multiplica por 100 para calcular un porcentaje (13). Lo que subyace en esta operación matemática es la variabilidad que se encuentra entre los estudios, que se debe a

una verdadera variabilidad (es decir, debida a una verdadera heterogeneidad). Un porcentaje del 0 % indica que de la variabilidad que hay entre los estudios, ninguna se debe a una verdadera heterogeneidad. Por otro lado, un porcentaje del 100 % indica que toda la variabilidad entre estudios es por heterogeneidad. Se ha convenido utilizar la calificación de baja, moderada y alta para valores del I^2 de 25 %, 50 % y 75 %, respectivamente (14).

A pesar de ser la medida más popular, se ha visto que el I^2 tiende a subestimar la heterogeneidad ante un número pequeño de estudios; por ejemplo, al combinar menos de siete estudios, el I^2 disminuye hasta 28 puntos (15). También es necesario disminuir la incertidumbre sobre el verdadero valor del I^2, por lo que idealmente los autores deben reportar el intervalo de confianza, que sería el rango en el que oscila ese valor, según los datos de los estudios incluidos (16).

Métodos gráficos

La inspección de gráficos complementa a la evaluación estadística.

Forest plot

Consta de un eje vertical (*y*) en el que están los estudios incluidos, ordenados por algún criterio, como año de publicación o tamaño de muestra. En el eje horizontal (*x*) está la medida de efecto, que para el caso de los desenlaces dicotómicos (ocurrió/no ocurrió) sería el RR o la OR, y para los desenlaces cuantitativos (números), la diferencia de medias estandarizada (DME). El no efecto se representa con una línea sólida, que para el caso de medidas calculadas por división (RR u OR) sería el 1 o por sustracción (DME) sería el 0.

Cada estudio está representado por un cuadrado, cuyo tamaño depende del tamaño de la muestra y se ubicará en el eje horizontal según el efecto que reporten los autores. Cada cuadrado está inmerso en una línea horizontal que representa el intervalo de confianza de dicho estudio. Al final, se hace un acumulado del efecto total o, en otras palabras, la sumatoria del efecto de todos los estudios. Este acumulado está representado por un diamante, cuyo ancho representa su intervalo de confianza. Usualmente, hay una línea punteada que atraviesa el diamante por la mitad, es decir, pasa por el estimado del efecto acumulado. Si esta línea punteada atraviesa los intervalos de confianza de todos los estudios primarios, es posible que exista homogeneidad. Por el contrario, si hay estudios cuyos intervalos de confianza no son atravesados por la línea punteada, probablemente dichos estudios sean responsables de la heterogeneidad.

En la figura 1a se presenta el *forest plot* a partir de un ejemplo hipotético de los estudios que evaluaron el efecto sobre la mortalidad de una intervención, el cual es un desenlace dicotómico evaluado con el RR. Al ser el RR producto de una división, la línea sólida de este *forest plot* corresponde a la unidad. Se puede ver cómo el tamaño del estudio afecta su precisión; entonces, los estudios más grandes tienen intervalos de confianza más estrechos y el resumen final muestra un aumento del riesgo de morir en la intervención, que no es estadísticamente significativo (RR = 1,1; IC95 %: 0,98 a 1,32). Nótese cómo la línea punteada atraviesa los intervalos de confianza de todos los estudios, excepto del estudio 4, posible fuente de heterogeneidad para el MA.

Existe una variación conocida como *forest plot acumulativo*. Aquí se ordenan los estudios por un criterio establecido por los autores, y en lugar de mostrar el efecto de los estudios individuales, se muestra el efecto combinado a medida que se va sumando cada estudio (17). En la figura 1b se muestra un ejemplo en el que se ven los estudios ordenados por año de publicación con su respectivo efecto y cómo, con cada estudio adicional, se calcula el nuevo estimado del efecto, por lo que, en lugar de un cuadrado, cada estudio se representa con diamantes, pues es un acumulado en sumatoria. Nótese que el intervalo de confianza se va volviendo cada vez más estrecho, debido a que se va sumando la

muestra de los estudios anteriores, y el resultado es cada vez más preciso.

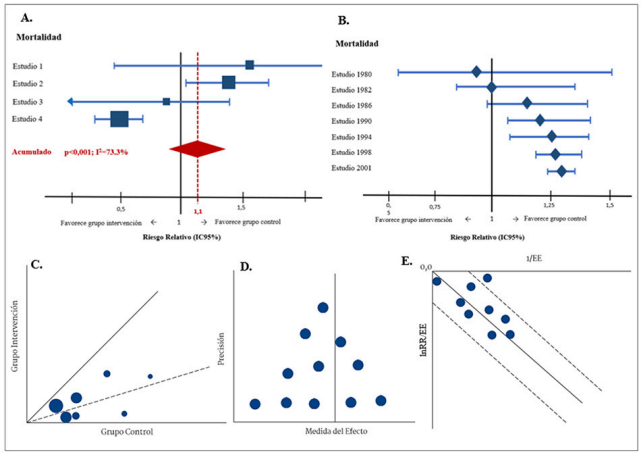

Figura 1
Métodos gráficos para evaluar la heterogeneidad

Gráfico de L'Abbé

En el eje *y* se ubica el grupo de intervención, y en el eje *x*, el de control (figura 1c). La medida que se usa en este gráfico es RR u OR. Los estudios primarios están representados por círculos, cuyo tamaño dependerá de su tamaño muestral. Usualmente, encontraremos dos líneas en esta gráfica: una línea sólida que representa el no efecto (en el caso de RR y OR sería el número uno por tratarse de divisiones) y una línea punteada que representa el acumulado del efecto. Si los círculos se encuentran por debajo de la línea sólida, significa que el desenlace se presentó con menor frecuencia que en el grupo de control. Por el contrario, si los círculos están por encima de la línea sólida, el desenlace se dio con mayor frecuencia con la intervención.

Así, si el desenlace fuera mortalidad, los círculos por debajo de la línea sólida indicarían que la intervención es eficaz, puesto que representó menos muertes. O si el desenlace es curación, los círculos por encima de la línea sólida indicarían mayor eficacia de la intervención, porque la frecuencia del desenlace fue mayor. Es necesario analizar la cercanía de los círculos a la línea punteada: entre más cerca estén de la línea punteada y agrupados entre sí, significa poca heterogeneidad; y si los círculos están muy dispersos, la heterogeneidad es mayor. Se ha determinado, sin embargo, que el tamaño de la muestra afecta la distancia entre los círculos y la línea punteada, por lo que se debería realizar un ajuste por el tamaño de la muestra para evitar resultados espurios (7,11,18). En la figura puede verse cómo los círculos más pequeños suelen estar más dispersos, y de ahí la importancia de hacer el ajuste por el tamaño de muestra.

Gráfico del embudo (funnel plot)

Pretende detectar un sesgo de publicación al comparar la precisión de los estudios en el eje *y* (que está relacionado con el tamaño de la muestra) y la magnitud del efecto en el eje *x* (medido por el estimador puntual, por ejemplo: RR u OR). Una línea sólida representa el no efecto y se espera que los estudios más grandes sean más precisos; mientras que los pequeños, más imprecisos y con mayor riesgo de error aleatorio. Así, los estudios más pequeños estarán más dispersos y a ambos lados de la línea, lo que da una apariencia de embudo invertido (figura 1d). En la figura podemos observar que los estudios que están más abajo, en la base, son los estudios más pequeños y, por ende, más imprecisos. Los estudios que están en la parte superior son más precisos, probablemente por un mayor tamaño de muestra.

El objetivo con este gráfico es observar si hay simetría a simple vista o, idealmente, con pruebas estadísticas que lo confirmen. Una asimetría podría indicar sesgo de publicación y, al tener solo estudios tendientes a beneficiar la intervención, habría una homogeneidad artificial o sesgada. Este método requiere, por lo menos, 30 estudios para ser lo más confiable posible, aunque con 10 es factible. Resulta inconveniente que no existe una estandarización en la medida que se va a usar en el eje *x* e *y*, pues esto varía según el autor. Al cambiar la forma de medir el efecto, se

altera la distribución del gráfico y, por tanto, su interpretación.

Conviene advertir que el sesgo de publicación no es la única explicación para la asimetría del embudo, así que su interpretación debe ser cuidadosa, pues puede darse por una verdadera heterogeneidad entre los estudios (2,11,18). En el ejemplo presentado en la figura 1d, podemos notar que hay una aparente simetría visual, con más de 10 estudios incluidos. Sin embargo, sería ideal realizar pruebas estadísticas para verificar esta simetría aparente.

Gráfico de Galbraith

Se construye con una medida estandarizada del efecto en el eje ., que puede calcularse dividiendo el logaritmo natural de RR u OR sobre el error estándar (EE), es decir, *lnRR/EE*. El eje *x* representa una medida de la precisión de ese efecto, que generalmente es el inverso del error estándar, es decir, 1/EE. Así, la relación es entre el efecto obtenido y su precisión (figura 1e). Una línea sólida representa la medición acumulada del efecto y unas líneas punteadas representan el intervalo de confianza de dicho acumulado. Los estudios más pequeños e imprecisos se ubican muy cerca del origen de la gráfica, hacia la coordenada 0,0, y los más precisos se encuentran más alejados del origen.

Evaluamos la heterogeneidad viendo qué tan cerca están los estudios de la línea sólida: entre más equidistantes estén a la línea sólida, más homogéneos son (7,11,18). Tomando el ejemplo de la figura 1e, se aprecia que la mayoría de los círculos se encuentran equidistantes a la línea sólida, pero hay un estudio impreciso que se localiza cerca del origen y otro más que se sale del límite de las líneas punteadas. Estos dos estudios podrían representar una fuente de heterogeneidad para la RSL.

¿Qué hacer si se detecta heterogeneidad?

Al encontrar evidencia de heterogeneidad, podemos decir que los resultados del efecto evaluado y que se pretenden combinar en el MA difieren mucho entre sí, más allá de lo esperado por errores del azar. Eso nos traería problemas si quisiéramos aplicar un tratamiento a nuevos pacientes, porque no estaríamos seguros de alcanzar el efecto visto en unos u otros estudios. La idea, entonces, es que los autores de RSL traten de explicar por qué se puede estar dando ese grado de inconsistencia y minimizarlo. Si los autores consideran que no es posible minimizar dicha inconsistencia, debería abandonarse la idea de combinar los resultados en el MA y únicamente describir los resultados de los estudios (síntesis cualitativa). Sabemos de casos en los que, aun en presencia de heterogeneidad alta, los autores de RSL presentan un MA, el cual debe criticarse y omitirse, porque estaría dando una combinación del efecto sesgada.

Investigación de fuentes de heterogeneidad

Análisis clínico

Como dijimos, el primer paso es evaluar la heterogeneidad que parte de los elementos de la pregunta PICO (población, intervención, comparador y *outcomes* o desenlaces). Se deben considerar las características de los participantes, de los estudios primarios incluidos y evaluar la edad, el sexo o comorbilidades y determinar qué tanta variación hay. En la intervención, evaluar dosificación, vía de administración o adherencia. También las comparaciones realizadas: ¿se usó placebo? Si se usó el tratamiento usual, ¿fue definido de la misma forma? ¿Se usó el mismo comparador? Si el comparador es otro medicamento, ¿cómo eran las dosis, vía de administración y adherencia? En cuanto a los desenlaces, se debe identificar si usaron instrumentos o definiciones operativas muy diferentes (19,20). El segundo paso es evaluar los aspectos metodológicos de los estudios. Si se están evaluando ensayos clínicos aleatorizados, se debe determinar si fueron adecuados la aleatorización, el cegamiento, el ocultamiento, el seguimiento de los pacientes, y si se presentaron conflictos de interés (11,19).

Análisis de sensibilidad

El objetivo es medir hasta qué punto las variables responsables de heterogeneidad repercuten en la medida del efecto del MA. Uno de ellos es la exclusión de *outliers* o resultados extremos, es decir, aquellos que se alejan tanto del promedio que afectan el resultado. Estos *outliers* pueden ser pacientes o estudios completos. El análisis de sensibilidad excluirá aquellos participantes o estudios que tengan resultados más extremos, para ver si se obtiene un resumen más homogéneo. Se pueden excluir estudios completos por problemas en su diseño. Así, un ensayo clínico aleatorizado, por ejemplo, podría ser excluido si no se realizó una adecuada aleatorización y ver cómo cambia el MA. Sin embargo, estas exclusiones pueden comprometer el rigor metodológico de la RSL (11,21) y lo que debemos analizar es cómo cambian los resultados con la exclusión y sin esta.

Análisis por subgrupos

La diversidad en las características de los pacientes de cada estudio puede afectar los resultados de un MA. Así, por ejemplo, si los pacientes de mayor edad tienden a morir más con la intervención que los de menor edad, puede haber un resultado sesgado en cuanto a mortalidad, por la diferencia que se presenta entre ambos subgrupos de edad, más que por efecto de la intervención. Los autores de una RSL deben identificar *a priori* aquellas variables que puedan explicar una posible heterogeneidad en los resultados, desde su conocimiento clínico y fisiopatológico. Esto se ve en un *forest plot* en el que se combina el efecto para un subgrupo y para el otro y se calcula el valor de *p* de la interacción o de la modificación del efecto. De encontrar significancia estadística, se sugiere que la variable que forma los subgrupos influye en el efecto de la intervención y los resultados pueden ser más homogéneos y aplicables a uno de los subgrupos. Se debe ser cuidadoso con este análisis, ya que puede dar resultados espurios y, por eso, las variables definidas deben estar bien sustentadas. Un análisis por subgrupos *post hoc* (después de tener los datos) no es recomendable (2,22).

Metarregresión

Se usa con variables cuantitativas que puedan explicar la heterogeneidad (como edad, peso corporal, concentraciones séricas de colesterol o año de publicación). Estas variables se denominan *variables independientes* y el efecto resumido es la variable dependiente. Se realiza un modelo estadístico de regresión que relaciona la variable dependiente y las independientes y se ve la significancia estadística de dicha relación. Entre más lineal sea la relación entre la variable independiente con la dependiente, mayor probabilidad hay de que la covariable esté contribuyendo con la heterogeneidad.

Por lo general, se informa el valor de *p* de dicha variable en la regresión. Por ejemplo, en una RSL sobre el efecto de la suplementación con vitamina D para prevención cardiovascular en jóvenes (23), no se observó influencia de las concentraciones plasmáticas previas de S-25-hidroxivitamina D en el desenlace cardiovascular (valor de $p > 0,29$). De todas formas, se debe tener en cuenta que la metarregresión es propensa a falsos positivos (19,24).

Minimización de la heterogeneidad: modelos estadísticos para medir el efecto

Es posible para los autores incorporar la heterogeneidad en el MA a través de los modelos estadísticos para combinar los efectos. Por un lado, el modelo de efectos aleatorios se usa cuando la heterogeneidad es grande y el estimado del efecto tiene gran variabilidad entre estudios. Este modelo permite evaluar la variabilidad que hay dentro de cada estudio por sí solo y en todos los estudios entre sí. Los intervalos de confianza que se generan suelen ser más amplios y los estudios grandes tienen un peso menor (25,26).

Por otro lado, el modelo de efectos fijos se usa cuando la heterogeneidad no es muy grande, por lo que se considera que la variabilidad del

estimador del efecto entre estudios no es tan alta. No se evalúa la variabilidad que hay entre los estudios, pero sí la variabilidad que hay dentro de cada estudio. Los intervalos de confianza en este modelo son más pequeños y se les da un mayor peso a los estudios que sean más grandes (25).

Conclusiones

La heterogeneidad puede ser un problema al realizar las RSL, si los autores no la identifican y manejan adecuadamente para evitar presentar resultados sesgados. Los usuarios de la literatura médica deben tener en cuenta la heterogeneidad en las RSL y verificar que los autores sí la hayan explorado y minimizado, para poder creer y confiar en los resultados en el momento de aplicarlos en la práctica clínica. Con este fin se propone, a modo de resumen, un algoritmo para la identificación y evaluación de la heterogeneidad (figura 2).

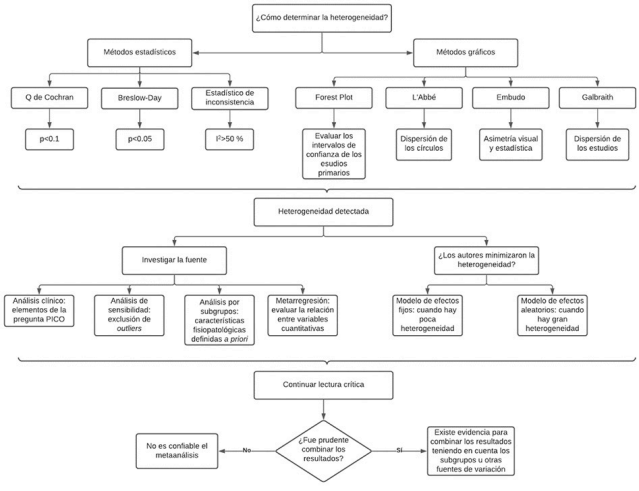

Figura 2
Cómo identificar y evaluar la heterogeneidad

Financiación

No se recibió financiación de ningún ente público o privado para la realización de este artículo.

Conflicto de intereses

Ninguno por declarar.

Referencias

1. Citations Added to MEDLINE by Fiscal Year. Bethesda: U.S. National Library of Medicine; c2019 [citado 2020 abr 15]. Disponible en: https://www.nlm.nih.gov/bsd/stats/cit_added.html

2. Guyatt GH, Rennie D, Meade MO, Cook DJ, editores. Users# guide to the medical literature: a manual for evidence-based clinical practice. 3.ª ed. New York: McGraw-Hill Education; 2015. p. 459-89.

3. García-Perdomo HA. Conceptos fundamentales de las revisiones sistemáticas/metanálisis. Rev Urol Colomb. 2015;XXIV(1):28-34.

4. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. BMJ Evidence-Based Medicine. 2016;21:125-127. https://doi.org/10.1136/ebmed-2016-110401

5. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Stat Med. 1999;18(20):2693-708. https://doi.org/10.1002/(sici)1097-0258(19991030)18:20<2693::aid-sim235>3.0.co;2-v

6. Engels EA, Schmid CH, Terrin N, Olkin I, Joseph L. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. Stat Med. 2000;19(13):1707-28. https://doi.org/10.1002/1097-0258(20000715)19:13<1707::aid-sim491>3.0.co;2-p

7. Xu H, Platt RW, Luo ZC, Wei S, Fraser WD. Exploring heterogeneity in meta-analyses: needs, resources and challenges. Paediatr Perinat Epidemiol.

2008;22(Suppl 1):18-28. https://doi.org/10.1111/j.1365-3016.2007.00908.x

8. Dahiru T. P-value, a true test of statistical significance? a cautionary note. Ann Ib Postgrad Med. 2008;6(1):21-26.

9. Centro Cochrane Iberoamericano, traductores. Manual Cochrane de revisiones sistemáticas de intervenciones, versión 5.1.0 [actualizada en marzo de 2011] [internet]. Barcelona: Centro Cochrane Iberoamericano; 2012. Disponible en: http://www.cochrane.es/?q=es/node/269

10. Viechtbauer W. Hypothesis tests for population heterogeneity in meta-analysis. Br J Math Stat Psychol. 2007;60(1):29-60. https://doi.org/10.1348/000711005X64042

11. Song F, Sheldon TA, Sutton AJ, Abrams KR, Jones DR. Methods for exploring heterogeneity in meta-analysis. Eval Health Prof. 2001;24(2):126-51.

12. Fleece D, Gaughan JP, Aronoff SC. Griseofulvin versus terbinafine in the treatment of tinea capitis:a meta-analysis of randomized, clinical trials. Pediatrics. 2004;114(5):1312-5. https://doi.org/10.1177/016327870102400203

13. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539-58. https://doi.org/10.1002/sim.1186

14. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557-60. https://doi.org/10.1136/bmj.327.7414.557

15. Von Hippel PT. The heterogeneity statistic I(2) can be biased in small meta-analyses. BMC Med Res Methodol. 2015;15:35. https://doi.org/10.1186/s12874-015-0024-z

16. Ioannidis JPA, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. BMJ. 2007;335(7626):914-6. https://doi.org/10.1136/bmj.39343.408449.80

17. Lau J, Schmid CH, Chalmers TC. Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. J Clin Epidemiol. 1995;48(1):45-57. https://doi.org/10.1016/0895-4356(94)00106-z

18. Bax L, Ikeda N, Fukui N, Yaju Y, Tsuruta H, Moons KGM. More than numbers: the power of graphs in meta-analysis. Am J Epidemiol. 2009;169(2):249-55. https://doi.org/10.1093/aje/kwn340

19. Petitti DB. Approaches to heterogeneity in meta-analysis. Stat Med. 2001;20(23):3625-33. https://doi.org/10.1002/sim.1091

20. West SL, Gartlehner G, Mansfield AJ, Poole C, Tant E, Lenfestey N, et al. Comparative effectiveness review methods: clinical heterogeneity [internet]. Agency for Healthcare Research and Quality; September 2010. Disponible en: http://effectivehealthcare.ahrq.gov/

21. De Souza RJ, Eisen RB, Perera S, Bantoto B, Bawor M, Dennis BB, et al. Best (but oft-forgotten) practices: sensitivity analyses in randomized controlled trials. Am J of Clin Nutr. 2016;103(1):5-17. https://doi.org/10.3945/ajcn.115.121848

22. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Ann Intern Med. 1992;116(1):78-84. https://doi.org/10.7326/0003-4819-116-1-78

23. Hauger H, Laursen RP, Ritz C, Mølgaard C, Lind MV,

Damsgaard CT. Effects of vitamin D supplementation on cardiometabolic outcomes in children and adolescents:a systematic review and meta-analysis of randomized controlled trials. Eur J Nutr. 2020;59(3):873-84. https://doi.org/10.1007/s00394-019-02150-x

24. Higgins JPT, Thompson SG. Controlling the risk of spurious findings from meta-regression. Stat Med. 2004;23(11):1663-82. https://doi.org/10.1002/sim.1752

25. Hedges LV, Vevea JL. Fixed- and random-effects models in meta-analysis. Psychol Methods. 1998;3(4):486-504. https://doi.org/10.1037/1082-989X.3.4.486

26. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. Contemp Clin Trials. 2007;28(2):105-14. https://doi.org/10.1016/j.cct.2006.04.004