Universitat Autònoma de Barcelona

# Evaluación e incorporación del riesgo de sesgo de estudios no experimentales en revisiones sistemáticas y metaanálisis

**Isabel Oliveras Boté**

Universitat Autònoma de Barcelona

# Evaluación e incorporación del riesgo de sesgo de estudios no experimentales en revisiones sistemáticas y metaanálisis

Isabel Oliveras Boté

**Tesis doctoral**

Directores

Josep Maria Losilla Vidal

Jaume Vives Brosa

Estudis de Doctorat en Psicologia de la Salut i Psicologia de l'Esport

Departament de Psicobiologia i Metodologia de les Ciències de la Salut

Facultat de Psicologia - Universitat Autònoma de Barcelona

**2018**

**Fuentes de financiación**

Esta tesis se ha llevado a cabo gracias a la ayuda para la contratación de personal investigador novel (FI-DGR) de la Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) de la Generalitat de Catalunya, de la ayuda de Formación de Profesorado Universitario (FPU) de la Secretaría de Estado de Investigación, Desarrollo e Innovación del Ministerio de Educación, Cultura y Deporte, y del proyecto I+D PSI2014-52962-P, financiado por el Ministerio de Economía y Competitividad.

# AGRADECIMIENTOS

En primer lugar, como no podía ser de otro modo, agradecer a mis directores, los doctores Josep Maria Losilla y Jaume Vives, toda la implicación y horas de trabajo depositadas en esta tesis doctoral. Desde el inicio de nuestra relación, me contagiaron su entusiasmo por la metodología y supieron infundirme confianza en los momentos en que más la necesitaba. Gracias por ser excelentes profesores, directores, tutores y, por encima de todo, maravillosas personas.

Agradecer especialmente a mi madre y a mi hermana Montse su acompañamiento en cada pequeña victoria y su paciencia en los momentos más duros de este trayecto. Os quiero.

Gracias a mis compañeras y amigas Laura y Marta, por todo lo que hemos compartido a lo largo de estos años. Nadie como Laura para entender las alegrías y angustias de las que hemos escogido la carrera académica.

Finalmente, dar las gracias a todo aquel y aquella de mi entorno que ha creído en mí y me ha apoyado desde aquel día lejano, hace ya trece años, en que decidí emprender esta emocionante experiencia desde cero.

# AGRAÏMENTS

En primer lloc, com no podia ser d'altra manera, agrair als meus directors, els doctors Josep Maria Losilla i Jaume Vives, tota la implicació i hores de feina dipositades en aquesta tesi doctoral. Des de l'inici de la nostra relació, em van encomanar el seu entusiasme per la metodologia i van saber infondre'm confiança en els moments que més la necessitava. Gràcies per ser excel·lents professors, directors, tutors i, per damunt de tot, meravelloses persones.

Agrair especialment a la meva mare i a la meva germana Montse el seu acompanyament en cada petita victòria i la seva paciència en els moments més durs d'aquest trajecte. Us estimo.

Gràcies a les meves companyes i amigues Laura i Marta, per tot el que hem compartit al llarg d'aquests anys. Ningú com la Laura per entendre les alegries i angoixes de les que hem escollit la carrera acadèmica.

Finalment, donar les gràcies a tot aquell i aquella del meu entorn que ha cregut en mi i m'ha recolzat des d'aquell dia llunyà, fa ja tretze anys, en que vaig decidir emprendre aquesta emocionant experiència des de zero.

# ÍNDICE

# 1. INTRODUCCIÓN

## 1.1. Práctica basada en la evidencia y síntesis de investigación

A lo largo de los años, el enfoque de la práctica basada en la evidencia (PBE) ha adquirido una gran relevancia para que tanto profesionales como políticos puedan tomar mejores decisiones sobre la aplicación de programas, tratamientos e intervenciones basadas en las mejores evidencias científicas (Sánchez-Meca, Marín-Martínez, & López-López, 2011; Sánchez Meca & Botella, 2010). La definición más común de la PBE procede del campo de la medicina y es la propuesta por Sackett, Rosenberg, Gray, Haynes y Richardson (1996):

> *"La medicina basada en la evidencia es el uso consciente, explícito y juicioso de la mejor evidencia actual para tomar decisiones sobre el cuidado del paciente individual. La práctica de la medicina basada en la evidencia significa integrar la experiencia clínica con la mejor evidencia clínica externa disponible procedente de la investigación sistemática."* (p. 71)

También el ámbito de la psicología se ha situado a la vanguardia de la PBE durante décadas (APA Presidential Task Force on Evidence-Based Practice, 2006), a partir del momento en que la American Psychological Association (APA) decidió incluir entre sus políticas la integración de la ciencia en la práctica de la psicología, introduciendo en la formación de doctores en psicología tanto la vertiente científica como la profesional (Hilgard et al., 1947; Thorne, 1947).

No obstante, a pesar de que la PBE ha proporcionado grandes avances y éxitos durante más de dos décadas (Greenhalgh et al., 2014), estos solo parecen darse en áreas específicas y de forma anecdótica (Every-Palmer & Howick, 2014). Así, en la actualidad, la PBE se enfrenta a serias dificultades que incluso han llevado a autores como John P.A. Ioannidis, importante impulsor del movimiento desde sus inicios, a criticar duramente el modo en que este movimiento ha sido "secuestrado" para beneficio de la industria y otros intereses creados (Ioannidis, 2016a). Entre otros problemas que también amenazan la correcta

implementación de la PBE se encuentran el gran volumen de evidencia disponible (no siempre válida y fiable), el inmanejable número de directrices clínicas (que a menudo entran en conflicto) y los sesgos de publicación, que a menudo favorecen la publicación de estudios con resultados positivos (Every-Palmer & Howick, 2014; Greenhalgh et al., 2014; Ioannidis, 2005, 2016a).

Por otra parte, la gran cantidad de información disponible y el elevado ritmo de publicación han convertido a las revisiones sistemáticas (RS) y su síntesis cuantitativa, el metaanálisis (MA), en herramientas fundamentales para la práctica de la PBE. Con origen en los ámbitos de la educación y la psicología (Chalmers, Hedges, & Cooper, 2002; Glass, 1976) y, a diferencia de las revisiones narrativas, las RSs y los MAs se presentan como metodologías rigurosas y sistemáticas para la síntesis de la evidencia, ofreciendo una aproximación científica para la identificación, análisis y síntesis de datos cuantitativos de estudios previos (Botella & Sánchez-Meca, 2015; Littell, Corcoran, & Pillai, 2008). Es así como RSs y MAs se han convertido en una importante fuente de información para investigadores, profesionales, gestores científicos y políticos en diversas disciplinas, en las cuales el sesgo ha sido reducido gracias a la sistematización de la identificación, evaluación, síntesis y agrupación estadística de los estudios relevantes sobre una materia (Manchikanti, Benyamin, Helm, & Hirsch, 2009; Sánchez Meca & Botella, 2010; Wells & Littell, 2009).

No obstante, a pesar de las grandes aportaciones y ventajas que ofrece, especialmente el MA, esta metodología tampoco está exenta de debate y de críticas. Por ejemplo, la controversia que genera el extendido uso del modelo de efectos aleatorios (DerSimonian & Laird, 1986), modelo que según varios autores puede llevar a estimaciones sesgadas (Al Khalaf, Thalib, & Doi, 2011; Cornell et al., 2014; Doi, 2015), lo que ha conducido a algunos de ellos a proponer modelos alternativos (Doi, Barendregt, Khan, Thalib, & Williams, 2015a, 2015b; Stanley & Doucouliagos, 2015), aunque ninguno con una aceptación generalizada. Otras importantes críticas a RSs i MAs son las que se dirigen a los sesgos de reporte y publicación (Hopewell, Loudon, Clarke, Oxman, & Dickersin, 2009; Ioannidis, Munafò, Fusar-Poli, Nosek, & David, 2014; Page et al., 2016), así como a "la producción masiva de RSs y MAs redundantes, engañosos y conflictivos" (Ioannidis, 2016b, p. 485). Según este autor, tan sólo aproximadamente un 3% de los MAs producidos se pueden considerar válidos, verdaderamente informativos y clínicamente útiles. Resultados similares han sido puestos de manifiesto por otros autores (e.g. Page et al., 2016).

## 1.2. Calidad metodológica/riesgo de sesgo en el contexto de la síntesis de investigación

Entre las limitaciones de RSs y MAs destaca la variable calidad metodológica de los estudios primarios que se integran y sintetizan. La calidad metodológica de los estudios primarios es una fuente potencial de sesgo que es necesario controlar a la hora de sintetizar los resultados, puesto que si los estudios primarios no tienen una adecuada calidad metodológica entonces las conclusiones de la RS/MA pueden no ser válidas (Higgins & Green, 2011; Jüni, Altman, & Egger, 2001; Voss & Rehfuess, 2013). Sin embargo, la evaluación de la calidad metodológica no es una tarea fácil debido, por un lado, a la gran variedad de conceptuaciones e interpretaciones que coexisten de este constructo y, por otro, a la falta de evidencia empírica sobre qué dimensiones de la calidad afectan en mayor medida a la validez de los resultados (Jarde, Losilla, Vives, & Rodrigo, 2013).

Si tenemos en cuenta la variedad de significados atribuidos en la literatura al constructo "calidad metodológica" o "calidad", que incluyen validez interna, riesgo de sesgo (RdS), limitaciones de los estudios, precisión, etc. (Balshem et al., 2011; Viswanathan & Berkman, 2012), es importante asumir una definición precisa de este constructo, especialmente en el desarrollo de herramientas para su evaluación. A menudo, el término "evaluación de la calidad metodológica" sugiere una evaluación sobre en qué magnitud los investigadores han llevado a cabo su investigación siguiendo los estándares más elevados posibles (Higgins & Green, 2011), pero habitualmente se refiere a un juicio sobre el RdS de un estudio individual (Balshem et al., 2011).

En consecuencia, las más importantes guías y colaboraciones (e.g. Cochrane Collaboration, Grading of Recommendations Assessment, Development and Evaluation – GRADE -) plantean una distinción entre la evaluación de la calidad metodológica y la evaluación del RdS (i.e. el riesgo de que los resultados de los estudios sobreestimen o subestimen los efectos reales de la intervención), y recomiendan centrarse en esta última (Higgins & Green, 2011). También según Wells y Littell (2009) el marco del RdS parece ser la aproximación ideal para la evaluación de la calidad de los estudios ya que permite evaluar las mayores fuentes de sesgo que constituyen amenazas a la validez interna y de constructo. La ventaja de asumir la aproximación del RdS es que clarifica qué debe ser evaluado y qué

indicadores adicionales de calidad deben ser evaluados aparte (validez externa, precisión, estándares de reporte y criterios éticos) porque todos ellos pueden no estar asociados con el RdS (Higgins & Green, 2011). Pese a que la incorporación de la dirección y magnitud de los sesgos parecen ser un elemento clave para la evaluación del RdS (Higgins & Green, 2011), los esfuerzos de algunos autores (Thompson et al., 2011; Turner, Spiegelhalter, Smith, & Thompson, 2009) para establecer la dirección y magnitud de los sesgos han dado como fruto métodos que, cuando menos, se muestran poco viables.

Por otra parte, mientras la incorporación de estudios experimentales en RSs y MAs está bien consolidada, la consideración de los estudios no experimentales en este ámbito es aún un debate abierto (Ijaz, Verbeek, Mischke, & Ruotsalainen, 2014; Jarde, Losilla, & Vives, 2012a, 2012b). Esto ocurre a pesar de que gran parte del conocimiento clínico y de la salud proviene de estudios no manipulativos, como se manifiesta en el hecho de que entre el 80-90% de los artículos publicados en revistas del ámbito clínico y entre el 70-80% de las publicaciones en el ámbito biomédico correspondan a estudios con diseños no experimentales y descriptivos (Funai, Rosenbush, Lee, & Del Priore, 2001; Manterola & Otzen, 2017; Primo, Gazzola, Primo, Tovo, & Faraco Junior, 2014; Scales, Norris, Peterson, Preminger, & Dahm, 2005). Asimismo, a menudo los diseños no experimentales son los más eficientes para responder algunas preguntas e incluso pueden representar la única manera de estudiar un determinado problema (Glasziou, Vandenbroucke, & Chalmers, 2004; Harder et al., 2014; Mann, 2003).

## 1.3. Herramientas de evaluación del riesgo de sesgo para estudios primarios no experimentales

Una cuestión esencial a la hora de evaluar el RdS de los estudios primarios son las herramientas utilizadas. En este sentido, la Colaboración Cochrane propone una herramienta basada en el enfoque de RdS que, teniendo en cuenta su extensión de uso, se puede considerar hoy en día un estándar de facto para la evaluación del RdS en el caso de estudios experimentales (Higgins & Green, 2011). La situación es, sin embargo, contraria en el caso de los estudios de tipo no experimental. Diversas RSs de herramientas para evaluar el RdS de estudios no experimentales (Deeks et al., 2003; Jarde et al., 2012a;

Sanderson, Tatt, & Higgins, 2007), muestran que no existe todavía un consenso acerca de cuál o cuáles son las herramientas más apropiadas para evaluar el RdS de este tipo de estudios, ni tampoco sobre las dimensiones que deben ser tenidas en cuenta a la hora de valorar dicho RdS. Es más, la propia evaluación del RdS de los estudios no experimentales presenta complicaciones añadidas atribuibles a la mayor variedad de diseños de investigación y a su mayor susceptibilidad al sesgo (Centre for Reviews and Dissemination., 2009; Hootman, Driban, Sitler, Harris, & Cattano, 2011; Margulis et al., 2014). Pese a que se han propuesto y publicado docenas de herramientas para evaluar el RdS en estudios no experimentales, muy pocas han sido desarrolladas siguiendo los criterios que se esperarían de cualquier instrumento de medida (Armijo-Olivo, Stiles, Hagen, Biondo, & Cummings, 2012; Deeks et al., 2003) y, aún en los pocos casos en que se ha seguido un procedimiento estandarizado en su desarrollo, son muy pocas las ocasiones en que se han obtenido resultados aceptables de fiabilidad inter-jueces (Downs & Black, 1998; Shamliyan, Kane, & Dickinson, 2010; Stang, 2010; Viswanathan & Berkman, 2012).

Por otra parte, la utilización de escalas que proporcionan una puntuación global ha sido fuertemente criticada (Herbison, Hay-Smith, & Gillespie, 2006; Jüni, Witschi, Bloch, & Egger, 1999; Whiting, Harbord, & Kleijnen, 2005) porque implica la ponderación de diferentes componentes que, además, pueden no estar relacionados con el RdS (Higgins & Green, 2011; Sanderson et al., 2007). La alternativa al uso de puntuaciones globales apunta hacia las herramientas basadas en dominios de sesgo (Higgins et al., 2011; Jarde, Losilla, Vives, & Rodrigo, 2013; Sterne et al., 2016; Whiting et al., 2011), que parecen ofrecer un mejor marco para detectar las potenciales fuentes de sesgo (O'Connor et al., 2015).

Como consecuencia de todo lo descrito, los resultados de la evaluación del RdS pueden diferir en gran medida dependiendo de la herramienta escogida y, por tanto, llegar a conclusiones diferentes en cuanto a los estudios evaluados, lo que parece ocurrir tanto en los estudios experimentales (Armijo-Olivo et al., 2012; Colle, Rannou, Revel, Fermanian, & Poiraudeau, 2002; Hartling et al., 2009) como en los no experimentales (Hootman et al., 2011; Jarde et al., 2012b; Margulis et al., 2014).

Como ya se ha comentado, frente a la falta de herramientas de consenso para la evaluación del RdS de los estudios de tipo no experimental, nuestro grupo desarrolló una propuesta concreta basada en dominios de sesgo, y materializada en una herramienta denominada Q-

Coh (Quality of Cohort studies) aplicable en estudios con diseño de cohortes comparativos (Jarde et al., 2013). Esta herramienta fundamenta el concepto de calidad metodológica en iniciativas ampliamente aceptadas, como son la iniciativa STROBE (Strengthening the Reporting of Observational studies in Epidemiology; Von Elm et al., 2007), los Estándares de Comunicación de Artículos (JARS, en sus siglas en inglés) de la American Psychological Association (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008) y las cuatro dimensiones de validez interna ampliamente reconocidas en las Ciencias Sociales (Shadish, Cook, & Campbell, 2002; Valentine & Cooper, 2008). El desarrollo de Q-Coh se llevó a cabo atendiendo a los estándares psicométricos y cuenta con buenos índices de fiabilidad entre evaluadores y de validez concurrente (Jarde et al., 2013).

## 1.4. Incorporación del riesgo de sesgo en la síntesis de investigación

Tal y como señalan diferentes autores (Ahn & Becker, 2011; Ioannidis, 2011), además de enfrentar la carencia antes mencionada de herramientas consensuadas para medir el RdS de los estudios no experimentales, es necesario profundizar en el estudio de la relación entre dichas medidas del RdS y la estimación de los efectos en los MAs, así como investigar sobre los procedimientos y técnicas estadísticas más apropiadas para llevar a cabo dicho estudio.

En esta línea, en una exploración preliminar de las aproximaciones propuestas para incluir los resultados de la evaluación del RdS en el MA (Oliveras, Jarde, Losilla, & Vives, 2013) se clasificaron las principales propuestas en tres grupos generales, no excluyentes entre ellos, en función del grado de manipulación ejercido sobre los resultados del MA. El primer grupo está constituido por las aproximaciones de carácter descriptivo en las que no se realiza ninguna modificación sobre los efectos. Los métodos incluidos en este grupo se limitan a representar gráficamente la relación entre el tamaño del efecto y puntuaciones de RdS (Ahn & Becker, 2011; Conn & Rantz, 2003; Detsky, Naylor, O'Rourke, McGeer, & L'Abbe, 1992), o bien utilizan métodos numéricos para filtrar u ordenar los resultados de los estudios antes de realizar el MA, como la inclusión/exclusión de los estudios en función

de los resultados de la evaluación del RdS (Ahn & Becker, 2011; Detsky et al., 1992; Linde et al., 1999; Moher et al., 1998), el MA acumulativo en orden de calidad (Conn & Rantz, 2003; Detsky et al., 1992) y el análisis de subgrupos según la calidad (Ioannidis, 2011; Jüni et al., 2001).

El segundo grupo lo forman aquellas aproximaciones en las que se estudia el resultado del MA (i.e., la estimación de los efectos) en función del RdS de los estudios primarios. El procedimiento de este tipo más utilizado es el que lleva a cabo un análisis de variables moderadoras con el RdS como variable predictiva del tamaño del efecto (Greenland & O'Rourke, 2001; Jüni et al., 2001; Littell et al., 2008), ya sea a través de un análisis de la varianza ponderado o bien de una meta-regresión.

Por último, un tercer grupo lo constituyen los procedimientos y técnicas cuyo enfoque se basa en la modificación del tamaño del efecto obtenido en los estudios primarios como paso previo a su incorporación en el MA, ya sea utilizando puntuaciones de calidad como factor de ponderación en los análisis estadísticos –método éste desaconsejado por un gran número de expertos (e.g., Ahn & Becker, 2011; Detsky et al., 1992; Jüni et al., 2001; Moher et al., 1998)-, o bien mediante el ajuste de los resultados de los estudios primarios con base en sesgos específicos (Ioannidis, 2011; Salanti & Ioannidis, 2009; Thompson et al., 2011; Turner et al., 2009). Este último enfoque presenta la dificultad añadida de requerir la estimación no sólo de la presencia, sino también de la magnitud y la dirección de los sesgos de los estudios primarios.

En este sentido, además de la problemática relativa a la elección del procedimiento estadístico más adecuado, hay que enfrentar también la controversia existente acerca de cuál debe ser el tipo de medida del RdS que se debe utilizar en los MAs: valoración global del RdS (cuantitativa o categórica ordinal), valoración separada para cada dominio de sesgo, aspectos particulares de la valoración del diseño medidos mediante ítems individuales, estimaciones directas de la magnitud y dirección de sesgos específicos, etc. (Greenland, 2005; Greenland & O'Rourke, 2001; Herbison et al., 2006; Ioannidis, 2011). Se pone de manifiesto, por tanto, la íntima relación entre la investigación sobre la medida del RdS y la investigación sobre los procedimientos para su incorporación en el MA, con la dificultad añadida que conlleva para su estudio la relación circular existente entre ambos objetivos de investigación. Hasta la fecha, los resultados de la investigación sobre esta

relación no parecen concluyentes (Ahn & Becker, 2011; Conn & Rantz, 2003; Verhagen et al., 2002).

## 1.5. Objetivos de esta tesis doctoral

### *1.5.1. Objetivo general*

El objetivo general del proyecto en el que se enmarca esta tesis doctoral es avanzar en la respuesta a la ya justificada necesidad de tomar en consideración la calidad metodológica de los estudios primarios a la hora de llevar a cabo RSs y MAs.

Para ello, es necesario el trabajo coordinado en dos frentes:

1. La revisión y ampliación de la herramienta de evaluación del RdS, Q-Coh (Jarde et al., 2013), de modo que pueda ser aplicada para la valoración de estudios con diseños no experimentales de cohortes y de casos y controles.

2. El análisis comparativo de los procedimientos propuestos hasta la fecha para la incorporación en el MA de las medidas de RdS de los estudios primarios, y a su vez, la utilización de estos procedimientos para dar soporte empírico a las dimensiones de calidad evaluadas.

El segundo de estos objetivos de la línea de investigación es el que se corresponde con el objetivo principal de esta tesis doctoral. La consecución de este objetivo es fundamental para dar soporte empírico a la propia medida de RdS, tanto de estudios experimentales como no experimentales, puesto que el efecto que esta medida pueda tener sobre los resultados del MA constituye a su vez una evidencia de su validez. A su vez, es necesario conocer el alcance y limitaciones de los diferentes procedimientos estadísticos de incorporación del RdS en el MA para poder elegir el procedimiento más adecuado en cada caso, así como para decidir el tipo de medida del RdS que se debe utilizar (puntuación global del RdS, puntuación separada para cada dimensión de la calidad o tipo de sesgo, estimaciones directas de la magnitud y dirección de sesgos específicos, etc.).

### 1.5.2. Objetivos específicos

Para llevar a cabo el objetivo general de esta tesis, fue necesario establecer una serie de objetivos específicos:

- *Objetivo 1*. Revisión y sistematización de las diferentes técnicas propuestas en la literatura para incorporar los indicadores de RdS en una síntesis de investigación.
- *Objetivo 2*. Análisis exploratorio de los efectos diferenciales que supone la aplicación de estos procedimientos de incorporación de los resultados de la evaluación del RdS sobre los resultados de MAs publicados.
- *Objetivo 3*. Meta-revisión sobre cómo se está llevando a cabo la evaluación e incorporación del RdS de los estudios primarios en la síntesis de investigación de psicología sanitaria.
- *Objetivo 4*. Estudio comparativo de fiabilidad, validez e idoneidad de tres herramientas de evaluación del RdS para estudios primarios no experimentales, así como determinar el efecto de los resultados de la evaluación del RdS sobre los resultados del MA.

### 1.5.3. Desarrollo y estructura de esta tesis doctoral

Después de llevar a cabo la revisión y clasificación de las técnicas propuestas para incorporar los indicadores de RdS en la síntesis de investigación (objetivo 1), se aplicaron algunas de estas técnicas a un MA con diseño de cohortes utilizando diferentes herramientas de RdS y diferentes tipos de medidas del RdS (objetivo 2). El análisis en profundidad de los resultados obtenidos, sin efectos significativos, puso de manifiesto que era necesaria información más detallada sobre los factores que pueden influir en la relación entre RdS y resultados de la síntesis de investigación. Con este fin se llevó a cabo una meta-revisión que aportara información sobre cómo se está evaluando e incorporando el RdS de los estudios primarios no experimentales en la síntesis de investigación del ámbito de la psicología sanitaria (objetivo 3).

El trabajo llevado a cabo con estos tres objetivos específicos se materializó en la primera publicación de este compendio. Aunque los resultados obtenidos a partir del segundo objetivo fueron descartados para su publicación (Anexo 1), fueron presentados en el XIV Congreso de Metodología de las Ciencias Sociales y de la Salud, celebrado en Palma de Mallorca (2015).

A continuación, la información obtenida en el desarrollo del primer artículo nos llevó hasta uno de los factores clave en la evaluación e incorporación del RdS en la síntesis de investigación: la influencia de la elección de una determinada herramienta de evaluación en los resultados. Así, se llevó a cabo un estudio empírico en el que se compararon tres herramientas de evaluación del RdS, aplicándolas a los estudios primarios de un MA publicado, con tal de valorar su fiabilidad, validez e idoneidad (objetivo 4). También se intentó, en este mismo estudio, determinar el efecto de los resultados de la evaluación del RdS sobre los resultados del MA. Este amplio objetivo se ha materializado en la segunda publicación que conforma este compendio.

Aunque parte del trabajo en el primer objetivo del proyecto de investigación (revisión y ampliación de la herramienta Q-Coh) ha transcurrido de forma paralela al desarrollo de esta tesis, por cuestiones temporales no ha sido posible integrarlo como parte de la misma. No obstante, gracias a los resultados aportados por los dos artículos, el trabajo avanza en este último objetivo, lo que esperamos que, en breve, nos permita obtener una herramienta de consenso y que pueda ser referente a la hora de evaluar el RdS de los estudios primarios no experimentales de una amplia variedad de diseños.

# 2. ARTÍCULOS QUE CONFORMAN ESTE COMPENDIO

## 2.1. Methodological quality is underrated in systematic reviews and meta-analyses in health psychology (Artículo 1)

El primero de los dos objetivos específicos que reúne este artículo es la revisión y sistematización de las técnicas propuestas para incorporar los indicadores de RdS en la síntesis de investigación. La revisión bibliográfica permitió reexaminar las aproximaciones identificadas en una exploración previa (Oliveras et al., 2013) y que, atendiendo a nuevos criterios, fueron clasificadas en cuatro grandes grupos: 1) RdS como criterio de inclusión en la RS o MA; 2) estrategias descriptivas, incluyendo las narrativas y las representaciones gráficas de los resultados de la evaluación del RdS; 3) exploración de las asociaciones entre RdS y el tamaño del efecto de los estudios primarios, como los análisis de sensibilidad, los MAs acumulativos, análisis de subgrupos y meta-regresiones; y 4) las estrategias que implican un ajuste del tamaño del efecto en base al RdS detectado como, por ejemplo, su uso como factor de ponderación del MA.

El segundo de los objetivos cubiertos por este artículo era llevar a cabo una meta-revisión que aportara una imagen actual de cómo se está evaluando e incorporando el RdS de los estudios primarios no experimentales en la síntesis de investigación en el ámbito de la psicología sanitaria. Para ello, se revisaron cuatro bases de datos (PsycINFO, Medline, Web of Science y Scopus) con el propósito de seleccionar RSs y MAs de estudios primarios con diseños de cohortes y casos y controles del ámbito mencionado. Se creó un manual de codificación para extraer las variables de interés de una muestra aleatoria de 90 estudios de los 1378 potencialmente relevantes identificados a través de la estrategia de búsqueda. Los resultados obtenidos mostraron que tan solo un 11% de las revisiones analizadas utilizaron una herramienta estándar para evaluar el RdS y utilizaron, al mismo tiempo, alguna estrategia para poner de manifiesto la influencia que el RdS de los estudios primarios tuvo

en sus resultados. Estos preocupantes resultados revelan, una vez más, la falta de unas directrices claras para incorporar los resultados del RdS en la síntesis de investigación.

Al mismo tiempo, la exploración de las diversas técnicas para incorporar los resultados del RdS en RSs y MAs, nos llevó a recomendar tan sólo dos de las propuestas, y únicamente cuando se cuenta con suficiente potencia estadística: el análisis de subgrupos y la meta-regresión.

Los resultados de la revisión las técnicas propuestas para incorporar los indicadores de RdS en la síntesis de investigación fueron presentados en el XIV Congreso de Metodología de las Ciencias Sociales y de la Salud, celebrado en Palma de Mallorca (2015), mientras que los resultados preliminares de la meta-revisión fueron presentados en el VII European Congress of Methodology, celebrado en Palma de Mallorca (2016).

El artículo en su versión final puede obtenerse en el siguiente vínculo:

Oliveras, I., Losilla, J.-M., & Vives, J. (2017). Methodological quality is underrated in systematic reviews and meta-analyses in health psychology. *Journal of Clinical Epidemiology*, *86*, 59–70. http://doi.org/10.1016/j.jclinepi.2017.05.002

Factor de impacto 2017 JCR: 4.245, Q1

Factor de impacto 2017 SJR: 2.862, Q1

# Methodological quality is underrated in systematic reviews and meta-analyses in health psychology

Isabel Oliveras, Josep-Maria Losilla and Jaume Vives.

Department of Psychobiology and Methodology of Health Sciences, Psychology Faculty, Universitat Autònoma de Barcelona,

Carrer de la Fortuna, Edifici B, 08193 Bellaterra, Barcelona, Spain.

Corresponding author: Jaume Vives, Department of Psychobiology and Methodology of Health Sciences, Psychology Faculty, Universitat Autònoma de Barcelona, Carrer de la Fortuna, Edifici B, 08193 Bellaterra, Barcelona, Spain.

Tel.: 0034-93 5812331; fax: 0034-93 5812125.

E-mail address: Jaume.Vives@uab.cat

Conflict of interest: All authors declare no conflict of interest.

Funding: This work was supported by the Spanish Ministry of Science and Innovation (grant number: PSI2014-52962-P). I.O. was supported by funding from two predoctoral grants from the Government of Catalonia and from the Ministry of Education, Culture and Sport of Spanish Government (grant numbers 2016 FI_B2 00115 and FPU14/04514). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Objectives:** In this paper, we compile and describe the main approaches proposed in the literature to include methodological quality or risk of bias into research synthesis. We also meta-review how the risk of bias of observational primary studies is being assessed and to what extent the results are incorporated in the conclusions of research synthesis.

**Study Design and Setting:** Electronic databases were searched for systematic reviews or meta-analyses related to health and clinical psychology. A random sample of 90 reviews published between January 2010 and May 2016 was examined.

**Results:** A total of 46 reviews (51%) performed a formal assessment of the risk of bias of primary studies. Only 17 reviews (19%) linked the outcomes of quality assessment with the results of the review.

**Conclusion:** According to previous literature, our results corroborate the lack of guidance to incorporate the risk of bias assessment in the results of systematic reviews and meta-analyses. Our recommendation is to appraise methodological quality according to domains of risk of bias to rate the degree of credibility of the results of a research synthesis, as well as subgroup analysis or meta-regression as analytical methods to incorporate the quality assessment.

Keywords: methodological quality; risk of bias; research synthesis; systematic review; meta-analysis.

Running title: Methodological Quality in Research Synthesis

Word count: 4,135

**What is new?**

**Key findings**

- Only 11% of the reviews analysed used a standard tool that assess the different domains of risk of bias, and stated the influence of the methodological quality of primary studies on their results.
- Only two of the proposed analytical methods to include methodological quality into research synthesis can be recommended without reservation, and only when there is enough statistical power: subgroup analysis and meta-regression.
- Our results pointed out the lack of specific guidance to incorporate the risk of bias assessment in the results of systematic reviews and meta-analysis.

**What this adds to what is known?**

- This article sets out the most common ways to manage methodological quality in a research synthesis, as well as the implications of each of these alternatives.
- This paper provides a thorough meta-review that appraise in depth how methodological quality is being assessed and incorporated into research synthesis based on primary studies with cohort and case-control designs.

**What is the implication and what should change now?**

- It is necessary to work to generate solid, well-defined and replicable procedures that guide the incorporation of the methodological quality of primary studies into systematic reviews and meta-analyses. This may substantially improve the decisions taken according to evidence based practice.
- Our proposal is to use the assessment of methodological quality preferably in two ways: (1) to rate the degree of credibility of the results of a systematic review or a meta-analysis, and (2) to improve the quality of research in a particular area, and to reduce the heterogeneity attributable to the risk of bias.

## 1. Introduction

Nowadays, there is broad consensus among the scientific community on the relevance of assessing the methodological quality (MQ) of the studies, especially when carrying out a research synthesis [1,2]. Assessing the MQ of the primary studies in the

context of a systematic review or meta-analysis (MA) is often a challenging process [3,4], especially when the synthesis is based on observational studies [5].

However, the debate continues about how we should define, assess and, especially, incorporate MQ into research synthesis [6,7]. Regarding the latter, several studies have explored the role of MQ in systematic reviews and MA [8–11], but to date guidelines on how to incorporate quality into the conclusions of a research synthesis remain scarce and vague [10].

On the other hand, previous research findings do not seem conclusive about the influence of the MQ of primary studies on the MA results [12–14]. Furthermore, certain methods proposed to incorporate the MQ, as for instance weighting effect sizes (ES) on the basis of MQ appraisal, could be introducing bias in the results of MA [12,15].

Considering the large number of questions which remain unanswered about the inclusion of MQ in research synthesis, this review aims to identify the strengths and weaknesses of the different approaches proposed to date, as well as to find out if there is a consensus procedure to carry out the inclusion of MQ into a MA. This general objective is based on the following specific objectives:

- Review the approaches to the assessment of the MQ of primary studies.
- Review the main strategies to include MQ into systematic reviews and MA.
- Meta-review how published systematic reviews and MA (a) assess the MQ of primary studies, (b) incorporate the MQ of primary studies, and (c) take into account the influence of the MQ of primary studies on the conclusions of the research syntheses.

## 2. Approaches to the assessment of the methodological quality of primary studies

Although there is no absolute consensus on what is and what should encompass the definition of MQ, in recent years many authors and organizations (e.g., the Cochrane Collaboration [16] and GRADE guidelines [17]) have adopted the *risk of bias* (RoB) framework. According to the Cochrane Collaboration, RoB in a systematic review may be defined as the risk that the results overestimate or underestimate the true effect of the intervention [18]. Generalizing to other causal effects, Viswanathan and Berkman [19] consider that a central goal is the assessment of the believability of the findings, which

entails evaluating the degree to which the effects reported by the study represent the true causal relationship between exposure and outcome. The RoB framework allows for a more accurate assessment of the main sources of bias that undermine the validity of a study [2]. Moreover, this perspective of MQ allows us to contextualize the importance of the different sources of bias depending on the study design and the field of the review [18]. In this paper, we refer to MQ within the RoB framework.

Moreover, and despite it not being the main purpose of this paper, it is essential to note that the assessment of MQ within the RoB framework presents some critical challenges, which clearly influence the use of MQ in research synthesis:

- The lack of validation of many of the assessment tools available make it difficult to appraise the MQ of the studies in a valid and reliable way [20–24]. This becomes more complex when the research synthesis includes observational research, which encompasses more diverse study designs than experimental research, and in which authors often create their own ad-hoc assessment tools [20,25].
- Although the use of global quality scores has been widely criticized [26–29], many quality assessment tools are still reducing the set of MQ domains to a single numerical value. This approach completely overlooks the fact that the relative importance of each of these domains can vary depending on the study design, the research field, or the research aim itself [13].
- Last but not least, it is the problem caused by poor reporting of primary studies [7,10,13] that, despite the existence of many standards about this issue (e.g., CONSORT [30] or STROBE [5]), results in incomplete information in many studies, thus making it impossible to obtain a proper assessment of MQ.

## 3. Strategies to incorporate methodological quality into research synthesis

To date, several strategies have been proposed to include the MQ component in a research synthesis [8,12,13,31]. Table 1 shows a classification of the methods available in four general approaches which do not exclude each other. Below are described the main features and criticisms of each method.

*3.1. Methodological quality as inclusion criteria in research synthesis*

This approach uses a quality-based threshold to decide the inclusion of primary studies in the review or in primary analyses. This threshold can be defined, for example, by classifying studies into two or more quality levels, or by using a measure of central tendency of an overall quality score as a cut-off point. However, the decisions made about excluding studies on the basis of MQ are often somewhat arbitrary [13]. Furthermore, the exclusion of studies when the available number of studies is small limits the power of the statistical analyses, or may even lead to the loss of valuable information or publication bias [3,13].

*3.2. Descriptive strategies*

The results of quality assessment may be reported as a narrative discussion [11]. Narrative strategies often constitute the only suitable alternative to deal with quality, especially when a statistical approach is not appropriate. Nevertheless, as Ioannidis [31] argues: “the process of recording biases and interpreting results can be fragmented, idiosyncratic, non-standardized and potentially conflicted” [op. cit. p778]. Descriptive strategies may also include graphical representations such as the RoB bar chart or the summary figure used in the Cochrane Handbook [18].

*3.3. Exploration of associations between effect sizes and methodological quality*

Exploratory methods allow to analyse the associations between the ES of the primary studies and their MQ. When global quality scores are used for this purpose, analyses may be affected by the problems related to these types of scores, as indicated above.

*3.3.1. Methodological quality assessment to conduct sensitivity analyses.*

A standard procedure for dealing with quality assessment using sensitivity analysis is to exclude studies under a certain quality threshold from some parts of the analyses [12], and thus to compare this outcome with the results from all the studies included. However, sensitivity analyses are not restricted to that purpose. Indeed, sensitivity analyses are used to assess the potential impact of any assumption or decision made during the review process, for example, the inclusion/exclusion criteria of primary studies, or how missing data are addressed [32,33].

### *3.3.2. Cumulative meta-analysis in order of quality.*

In cumulative MA studies are sequentially added to the MA in a specified order. Studies are often sorted chronologically, but they may also be sorted by other variables such as study quality, starting with the highest quality study and entering studies consecutively according to quality scores [7,8]. Therefore, cumulative MA is not an analysis procedure, but a graphical method to display results from a series of separate MA in one plot [32].

### *3.3.3. Quality-based subgroup analyses.*

Subgroup analyses entail distributing participants or studies into subsets to make comparisons between them. Thus, MQ may be treated as an effect modifier to build the subgroups of studies [18]. To compare the mean effect across subgroups, different methods can be applied, such as a z-test, an analysis of variance or a Q-test of homogeneity [32]. On the other hand, subgroup analyses are prone to provide misleading results due to multiple testing, inadequate power or the lack of *a priori* specification of the analyses [34,35]. However, these issues are not inherent in subgroup analyses whether they are carefully planned, justified and powered [36].

### *3.3.4. Including methodological quality in a meta-regression.*

Meta-regression is used to assess the potential impact of one or more continuous or categorical covariates on the effect estimate [33]. Thus, quantitative or categorical quality assessment can be included as covariates in a meta-regression. However, performing a meta-regression with a reduced number of studies is not a recommended option because it lacks statistical power [32].

## *3.4. Bias adjustment methods*

This group of methods involve applying some sort of adjustment to MA, either modifying the ES of primary studies or modifying the pooled ES estimate.

### *3.4.1. Methodological quality as a factor to weight a meta-analysis.*

Quality scores can be incorporated as weights into a MA (e.g. multiplying scores by inverse-variance weights), thus giving studies with high MQ a greater impact in pooled results and decreasing the impact of poorer ones [37]. On the other hand, weighting the ES estimates by quality lacks statistical or empirical justification [8]. In

fact, quality weighting adds uncertainty to the pooled ES estimator, as well as it can add bias in many cases [12].

*3.4.2. Quality effects model meta-analysis.*

Proposed by Doi and Thalib [38], this MA model uses a ranking of quality scores to weight the studies. The quality effects model proposes adjusting inter-studies variance by using an estimate of bias variance based on the overall quality score of the studies. Despite the recent improvements of this approach [39], overall quality scores remain at the basis of this meta-analytic model, with the implications discussed above.

*3.4.3. Elicitation from experts of primary studies biases.*

The approach proposed by Turner *et al*. [40] involves adjusting the ES of primary studies through quantifying the magnitude and direction of potential biases, as identified by a group of experts. This is a complex strategy which requires a large number of experts and a long time to carry it out [31,41], not to mention its subjectivity and difficulty in replication.

## 4. Methods

A meta-review was performed to obtain a current overview of how the MQ of observational studies in systematic reviews and MA is being assessed and taken into account, and how the MQ influences the conclusions of the research syntheses. To narrow the results to a manageable number of publications, we restrict the scope of the search to recent systematic reviews and MA related to health and clinical psychology area.

*4.1. Search strategy*

To find relevant publications, electronic databases including PsycINFO, Medline, Web of Science, and Scopus were systematically searched from January 2010 to May 2016. The search strategy was performed combining terms related to the research field (psycho*, clinic*, health), methodology (systematic review*, meta-analys*, meta analys*) and study designs (cohort, case-control, follow-up stud*, follow up stud*, observational, non-experimental, non experimental, prospective, retrospective, epidemiologic stud*). The search was restricted to articles published in peer-reviewed journals, including only human population, and written in English or Spanish. The detailed search strategies can be found in Appendix A at www.jclinepi.com.

*4.2. Selection criteria*

Studies were considered eligible if they were systematic reviews or MA (1) including primary studies clearly identifiable as cohort or case-control designs, (2) assessing the effects of an intervention or an exposure, and (3) related to health or clinical psychology fields. Systematic reviews or MA of descriptive studies (i.e. prevalence or incidence studies), measurement tools, genetic association studies, and other papers than systematic reviews or MA (e.g. protocols, letters, meta-reviews) were excluded from this meta-review.

*4.3. Sample selection*

For practical reasons, only a random sample of the retrieved studies was reviewed. The number of studies to include was calculated to ensure a maximum margin of error of 10% (assuming a 95% confidence level), leaving a sample of $n = 90$ studies. To create a representative random sample, the retrieved records (once duplicates were removed) were ordered alphabetically according to the surname of the first author. Then, in the order specified by a sequence of random numbers, the selection criteria were applied to the records (title/abstract or full text when necessary) until 90 eligible systematic reviews or MA were reached.

*4.4. Data extraction*

A draft of the data extraction form was piloted on five of the selected studies by the three authors. The extraction form was revised and problematic questions were reworked. Finally, a coding manual was developed to incorporate the improvements discussed (Appendix B at www.jclinepi.com). Data extraction was performed by one of the researchers (I.O.), following the coding manual and using a MS Access database created for this purpose. This process was supervised by the remaining authors (J.M.L. and J.V.) and any disagreement was discussed until consensus was achieved.

*4.5. Data analysis*

Extracted variables were divided into two main blocks: description of the variables common to all the studies, and description of the variables for the studies that assessed RoB. Descriptive analyses were performed using frequencies and percentages for categorical variables, and medians (with IQRs and ranges) for continuous variables.

Data were analysed using Statistical Package for Social Sciences (SPSS) Version 20.0 (SPSS, Inc., 2009, Chicago, Illinois, USA).

## 5. Results

### *5.1. Search results*

A total of 1378 potentially relevant systematic reviews and MAs related to clinical or health psychology were identified from the reviewed databases. Of these, a random sample of 230 was reviewed following the procedure described above until reaching $n$ = 90 eligible studies [42–131]. Some relevant features of the included studies can be found in Appendix C at www.jclinepi.com. Reasons for the exclusion of 140 publications are shown in Figure 1.

### *5.2. General description of the reviews*

Figure 2 provides an overview on the distribution by year of publication of both the retrieved records and the selected studies. In both cases there is a steadily increase in the number of published systematic reviews and MAs. Most of the selected reviews were published in journals ranked in the first quartile of Journal Citation Reports (61.1% of 72 reviews in Science Edition and 76.1% of 46 reviews in Social Sciences Edition) and in the first quartile of Scimago Journal Rank (88.9% of 90 reviews). Of the included reviews, only 10 of 90 (11.1%) assessed the effects of some kind of intervention. The median of primary studies included in the reviews was 18.50 (IQR 11-28.5; range 5-80); for the reviews including cohort studies ($n$ = 75) the median of cohort studies was 13 (IQR 5-22; range 1-76); and for the reviews including case-control studies ($n$ = 25) the median of case-control studies was 6 (IQR 2.5-12.5; range 1-41).

Details related to the quality of included reviews are outlined in Table 2. In summary, 48 reviews (53.3%) did not follow any guideline related to the quality of reporting or the quality of evidence; 87 reviews (96.7%) did not register the protocol of the review; 78 reviews (86.7%) mentioned a particular domain of bias as a methodological limitation; only 30 reviews (33.3%) clearly reported the set of confounders that could threaten the results; and finally, only 46 reviews (51.1%) performed a formal assessment of the RoB of the primary studies.

*5.3. Risk of bias assessment*

Only 31 (34.4%) of the total number of included reviews ($n$ = 90) appraised the RoB of the primary studies using standard tools (modified or not). As can be seen from the Table 3, this percentage increases to 67.4% when only reviews that assessed RoB ($n$ = 46) were taken into account. The most used MQ assessment tool in these 31 reviews was the Newcastle-Ottawa Scale (NOS) [132], which was applied in 16 reviews (51.6%). Of these 46 reviews that assessed RoB, 31 (67.4%) used a quantitative score and 25 (54.3%) provided both an overall RoB outcome and an outcome by RoB domains.

*5.4. Incorporation of risk of bias assessment and its influence on the results*

As Table 4 shows, most of the reviews that assessed RoB ($n$ = 46) incorporated the results only in a descriptive way (29 reviews, 63.0%), while 11 reviews (23.9%) used an analytical approach. Analytical strategies included sensitivity analysis, subgroup analysis and meta-regression. It is noteworthy that none of the reviewed studies used a method of bias adjustment to incorporate the results of RoB assessment in a MA. On the other hand, 29 reviews (63.0%) did not clearly define or did not define at all the influence of the MQ of primary studies on the review results, despite of having assessed their RoB. Of the 17 reviews (36.9%) that incorporated the RoB into the interpretation of the conclusions, only 9 reviews (19.5%) found some influence of MQ on the review results.

## 6. Discussion

Our study provides a current overview of how the MQ of observational primary studies is being assessed when a systematic review or MA is carried out, which are the main strategies used to incorporate these assessments into the reviews, and to what extent the MQ is taken into account in the interpretation of the results of the review. Overall, our findings show that, although about half of the reviews formally assessed the RoB of primary studies, only 10 (11.1%) used a standard or a modified standard tool that considers different domains of bias, and stated the influence of the MQ of primary studies on the results of the review.

The present findings seem to be consistent with other studies with similar purpose, although focusing on different fields or study designs. For example, Deeks et al. [20] examined 511 systematic reviews of non-randomised intervention studies, of which only 169 (33%) assessed study quality, and only 69 (14%) incorporated the results in a

quantitative manner. Whiting and colleagues [11] identified 114 diagnostic systematic reviews, of which 58 (51%) performed an explicit quality assessment. In this case, the main strategies used to incorporate the quality assessment in the reviews were a table or narrative description only (19%), sensitivity analysis (11%), and recommendations for future research (10%) among others. In 2005, Moja et al. [9] examined 965 systematic reviews (809 Cochrane reviews and 156 paper based reviews), and they found that MQ of primary studies was assessed in 94% of Cochrane reviews and 60% of paper based reviews. Overall, 496 (51%) of the reviews used the quality assessment to analyse or interpret the results. A more recent study of diagnostic accuracy reviews, carried out by Ochodo and colleagues [10], found that 60 reviews (92%) of the 65 reviews analysed had formally assessed the MQ of included studies, but only 6 reviews (9%) linked the results to their conclusions.

In the light of the aforementioned, it seems obvious that the MQ of the primary studies is increasingly assessed. However, gaps remain regarding the incorporation of RoB assessment in the results of reviews, probably because the lack of specific guidance to carry it out [133].

*6.1. Strengths and limitations*

To the best of our knowledge, this is the first study that appraise in depth how MQ is being assessed and incorporated into research synthesis based on primary studies with cohort and case-control designs. The analysis of the MQ of observational studies becomes relevant if we take into account the large amount of research and fields that must resort to such designs for efficiency or ethical reasons. Moreover, in addition to updating the strategies used to incorporate MQ into research synthesis, this paper provides a thorough meta-review of numerous quality-related elements that provide a broad vision of how MQ is being used in a particular field. These quality-related elements have been addressed using a coding manual, which was widely discussed by the authors until consensus was reached.

On the other hand, this paper has limitations that should be considered. In many cases, the design of the primary studies included in systematic reviews or MA are difficult to distinguish because of the different terms used by the authors to refer to them, or even

occasionally because of the total lack of reporting of the designs included. This may have led to the exclusion of some publications relevant to our meta-review.

The other major limitation to consider relates to data collection because some items required certain degree of judgement, since information available in the publications was often incomplete.

*6.2. Recommendations*

Although there are ongoing efforts to improve MQ assessment tools [134,135], it is necessary to continue the development of these tools according to a proper quality framework and following rigorous psychometric standards. Without valid and reliable tools, and without the appropriate procedures for the integration of quality into the MA, it is virtually impossible to obtain results that can be interpreted in the sense of whether or not MQ has an influence on the results. Theoretical background suggests that this influence exists and could vary depending on the magnitude and direction of bias, that seems to be key in MQ assessment [18,40,136]. This difficulty in obtaining clear results on the influence of MQ on the research synthesis results could be one of the main reasons for the lack of inclusion of MQ into the conclusions of the reviews.

The lack of guidance to incorporate MQ assessment into research synthesis, as pointed out by other authors [10,11,133,137] is an important handicap in the research synthesis and that should be addressed urgently as this may alter decisions taken according to evidence based practice. It is necessary to work to generate solid, well-defined and replicable procedures that guide the incorporation of the methodological quality of primary studies into systematic reviews and meta-analyses.

Regarding the strategies to include MQ into systematic reviews or MA, there is a relevant question that it is worth considering. Even if we had the ability to properly estimate the risk of bias of a study this does not mean that the study is biased. For this reason, the question that arises is whether the mere possibility of the existence (i.e., risk) of bias should be used to correct the ES of a MA. Just as the ES of a given study is not corrected on the basis of the poor validity or reliability of the tool that measures an exposure or a response variable, correcting the ES on the basis of unconfirmed biases may be a reckless decision. This might become especially critical when a global quality score is used to make the adjustment of the ES, as it is very likely that this combination

will lead to an increase of bias in the MA results. In our meta-review, we found no study that used this kind of correction of ES.

According to the results of our meta-review, the only analytical methods that may be recommended without the uncertainty associated with bias adjustment methods, are the subgroup analysis and meta-regression, but always restricted by the statistical power that these analyses require.

*6.3. Conclusion*

Our results show that only 11% of the reviews analysed used a standard tool that assess the different domains of risk of bias, and stated the influence of the MQ of primary studies on their results.

This situation highlights the need to significantly improve both MQ assessment tools and procedures for the inclusion of MQ in research synthesis. Until this occurs, instead of using MQ to adjust the ES of primary studies or to adjust the pooled ES in MA results, our proposal is to use the assessment of MQ preferably in two ways: (1) to rate the degree of credibility of the results of systematic reviews or MA, and (2) to evaluate the methodological limitations and the main sources of risk of bias in a particular research area, so that it may help to improve their methodological quality, and to reduce the heterogeneity in the results of primary studies that is attributable to risk of bias.

**References**

[1] Johnson BT, Low RE, MacDonald H V. Panning for the gold in health research: Incorporating studies' methodological quality in meta-analysis. Psychol Health 2014;30:135–52. doi:10.1080/08870446.2014.953533.

[2] Wells K, Littell JH. Study Quality Assessment in Systematic Reviews of Research on Intervention Effects. Res Soc Work Pract 2009;19:52–62. doi:10.1177/1049731508317278.

[3] Jüni P, Altman DG, Egger M. Systematic reviews in health care - Assessing the quality of controlled clinical trials. Br Med J 2001;323:42–6. doi:10.1136/bmj.323.7303.42.

[4] Whittemore R, Chao A, Jang M, Minges KE, Park C. Methods for knowledge synthesis: an overview. Heart Lung 2014;43:453–61. doi:10.1016/j.hrtlng.2014.05.014.

[5] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008;61:344–9. doi:10.1016/j.jclinepi.2007.11.008.

[6] Botella J, Sánchez-Meca J. Meta-análisis en ciencias sociales y de la salud. Madrid: Editorial Síntesis; 2015.

[7] Linde K, Scholz M, Ramirez G, Clausius N, Melchart D, Jonas WB. Impact of study quality on outcome in placebo-controlled trials of homeopathy. J Clin Epidemiol 1999;52:631–6. doi:10.1016/S0895-4356(99)00048-7.

[8] Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating Variations in the Quality of Individual Randomized Trials into Metaanalysis. J Clin Epidemiol 1992;45:255–65. doi:10.1016/0895-4356(92)90085-2.

[9] Moja LP, Telaro E, D'Amico R, Moschetti I, Coe L, Liberati A. Assessment of methodological quality of primary studies by systematic reviews: results of the metaquality cross sectional study. BMJ 2005;330:1053. doi:10.1136/bmj.38414.515938.8F.

[10] Ochodo EA, van Enst WA, Naaktgeboren CA, de Groot JAH, Hooft L, Moons KGM, et al. Incorporating quality assessments of primary studies in the conclusions of diagnostic accuracy reviews: a cross-sectional study. BMC Med Res Methodol 2014;14:33. doi:10.1186/1471-2288-14-33.

[11] Whiting P, Rutjes AWS, Dinnes J, Reitsma JB, Bossuyt PMM, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. J Clin Epidemiol 2005;58:1–12. doi:10.1016/j.jclinepi.2004.04.008.

[12] Ahn S, Becker BJ. Incorporating Quality Scores in Meta-Analysis. J Educ Behav Stat 2011;36:555–85. doi:10.3102/1076998610393968.

[13] Conn VS, Rantz MJ. Focus on research methods: Research methods: Managing primary study quality in meta-analyses. Res Nurs Health 2003;26:322–33. doi:10.1002/nur.10092.

[14] Verhagen AP, de Vet HCW, Vermeer F, Widdershoven JWMG, de Bie R a, Kessels AGH, et al. The influence of methodologic quality on the conclusion of a landmark meta-analysis on thrombolytic therapy. Int J Technol Assess Health Care 2002;18:11–23. doi:10.1017/S0266462309091016.

[15] Greenland S, O'Rourke K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. Biostatistics 2001;2:463–71. doi:10.1093/biostatistics/2.4.463.

[16] Higgins J, Green S, editors. Cochrane handbook for systematic reviews of interventions. 2008.

[17] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924–6. doi:10.1136/bmj.39489.470347.AD.

[18] Higgins J, Green S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011 2011.

[19] Viswanathan M, Berkman ND. Development of the RTI item bank on risk of bias and precision of observational studies. J Clin Epidemiol 2012;65:163–78. doi:10.1016/j.jclinepi.2011.05.008.

[20] Deeks JJ, Dinnes J, D'Amico R, Sowden a J, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7:iii–x, 1-173.

[21] Sanderson S, Tatt ID, Higgins JPT. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol 2007;36:666–76. doi:10.1093/ije/dym018.

[22] Bai A, Shukla VK, Bak G, Wells G. Quality Assessment Tools Project Report. Ottawa: Canadian Agency for Drugs and Technologies in Health: 2012.

[23] Jarde A, Losilla JM, Vives J. Methodological quality assessment tools of non-experimental studies: a systematic review. An Psicol 2012;28:617–28. doi:10.6018/analesps.28.2.148911.

[24] Jarde A, Losilla JM, Vives J. Suitability of three different tools for the assessment of methodological quality in ex post facto studies. Int J Clin Heal Psychol 2012;12:97–108.

[25] Hootman JM, Driban JB, Sitler MR, Harris KP, Cattano NM. Reliability and validity of three quality rating instruments for systematic reviews of observational studies. Res Synth Methods 2011;2:110–8. doi:10.1002/jrsm.41.

[26] Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. Jama-Journal Am Med Assoc 1999;282:1054–60. doi:10.1001/jama.282.11.1054.

[27] Herbison P, Hay-Smith J, Gillespie WJ. Adjustment of meta-analyses on the basis of quality scores should be abandoned. J Clin Epidemiol 2006;59:1249–56. doi:10.1016/j.jclinepi.2006.03.008.

[28] Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol 2005;5:19. doi:10.1186/1471-2288-5-19.

[29] Cooper H, Hedges L V, Valentine JC, editors. The Handbook of Research Synthesis and Meta-Analysis. New York: Russell Sage Foundation; 2009.

[30] Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. J Clin Epidemiol 2010;63:834–40. doi:10.1016/j.jclinepi.2010.02.005.

[31] Ioannidis JPA. Commentary: Adjusting for bias: a user's guide to performing plastic surgery on meta-analyses of observational studies. Int J Epidemiol 2011;40:777–9. doi:10.1093/ije/dyq265.

[32] Borenstein M, Hedges L V., Higgins JPT, Rothstein HR. Introduction to Meta-Analysis. Chichester, West Sussex, UK: John Wiley & Sons Ltd; 2009. doi:10.1002/9780470743386.

[33] Littell JH, Corcoran J, Pillai VK. Systematic reviews and meta-analysis. Oxford: Oxford University Press; 2008.

[34] Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005;365:176–86. doi:10.1016/S0140-6736(05)17709-5.

[35] Tanniou J, van der Tweel I, Teerenstra S, Roes KCB. Subgroup analyses in confirmatory clinical trials: time to be specific about their purposes. BMC Med Res Methodol 2016;16:20. doi:10.1186/s12874-016-0122-6.

[36] Burke JF, Sussman JB, Kent DM, Hayward RA. Three simple rules to ensure reasonably credible subgroup analyses. BMJ 2015;351:h5651. doi:10.1136/bmj.h5651.

[37] Berard A, Bravo G. Combining studies using effect sizes and quality scores: Application to bone loss in postmenopausal women. J Clin Epidemiol 1998;51:801–7. doi:10.1016/S0895-4356(98)00073-0.

[38] Doi SAR, Thalib L. A quality-effects model for meta-analysis. Epidemiology 2008;19:94–100. doi:10.1097/EDE.0b013e31815c24e7.

[39] Doi SAR, Barendregt JJ, Khan S, Thalib L, Williams GM. Advances in the meta-analysis of heterogeneous clinical trials II: The quality effects model. Contemp Clin Trials 2015;45:123–129. doi:10.1016/j.cct.2015.05.010.

[40] Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. Bias modelling in evidence synthesis. J R Stat Soc Ser A-Statistics Soc 2009;172:21–47. doi:10.1111/j.1467-985X.2008.00547.x.

[41] Dias S, Sutton AJ, Welton NJ, Ades AE. Evidence synthesis for decision making 3: heterogeneity--subgroups, meta-regression, bias, and bias-adjustment. Med Decis Making 2013;33:618–40. doi:10.1177/0272989X13485157.

[42] AlAqeel B, Margolese HC. Remission in schizophrenia: Critical and systematic review. Harv Rev Psychiatry 2012;20:281–97. doi:10.3109/10673229.2012.747804.

[43] Alisic E, Krishna RN, Groot A, Frederick JW. Children's Mental Health and Well-Being After Parental Intimate Partner Homicide: A Systematic Review. Clin Child Fam Psychol Rev 2015;18:328–45. doi:10.1007/s10567-015-0193-7.

[44] Allott K, Liu P, Proffitt TM, Killackey E. Cognition at illness onset as a predictor of later functional outcome in early psychosis: Systematic review and methodological critique. Schizophr Res 2011;125:221–35. doi:10.1016/j.schres.2010.11.001.

[45] Boden JM, Fergusson DM. Alcohol and depression. Addiction 2011;106:906–14. doi:10.1111/j.1360-0443.2010.03351.x.

[46] Borges G, Bagge CL, Orozco R. A literature review and meta-analyses of cannabis use and suicidality. J Affect Disord 2016;195:63–74. doi:10.1016/j.jad.2016.02.007.

[47] Brennan ME, Spillane AJ. Uptake and predictors of post-mastectomy reconstruction in women with breast malignancy - Systematic review. Ejso 2013;39:527–41. doi:10.1016/j.ejso.2013.02.021.

[48] Brito K, Edirimanne S, Eslick GD. The extent of improvement of health-related quality of life as assessed by the SF36 and Paseika scales after parathyroidectomy in patients with primary hyperparathyroidism - A systematic review and meta-analysis. Int J Surg 2015;13:245–9. doi:10.1016/j.ijsu.2014.12.004.

[49] Burton BK, Hjorthoj C, Jepsen JR, Thorup A, Nordentoft M, Plessen KJ. Research Review: Do motor deficits during development represent an endophenotype for schizophrenia? A meta-analysis. J Child Psychol Psychiatry 2016;57:446–56. doi:10.1111/jcpp.12479.

[50] Cerimele JM, Katon WJ. Associations between health risk behaviors and symptoms of schizophrenia and bipolar disorder: a systematic review. Gen Hosp Psychiatry 2013;35:16–22. doi:10.1016/j.genhosppsych.2012.08.001.

[51] Chen S, Zhong X, Jiang L, Zheng X, Xiong Y, Ma S, et al. Maternal autoimmune diseases and the risk of autism spectrum disorders in offspring: A systematic review and meta-analysis. Behav Brain Res 2016;296:61–9. doi:10.1016/j.bbr.2015.08.035.

[52] da Silva J, Goncalves-Pereira M, Xavier M, Mukaetova-Ladinska EB. Affective disorders and risk of developing dementia: systematic review. Br J Psychiatry 2013;202:177–86. doi:10.1192/bjp.bp.111.101931.

[53] de Maat S, de Jonghe F, de Kraker R, Leichsenring F, Abbass A, Luyten P, et al. The current state of the empirical evidence for psychoanalysis: A meta-analytic approach. Harv Rev Psychiatry 2013;21:107–37.

[54] Diaz-Piedra C, Di Stasi LL, Baldwin CM, Buela-Casal G, Catena A. Sleep disturbances of adult women suffering from fibromyalgia: A systematic review of observational studies. Sleep Med Rev 2015;21:86–99. doi:10.1016/j.smrv.2014.09.001.

[55] Dickens C, Katon W, Blakemore A, Khara A, McGowan L, Tomenson B, et al. Does depression predict the use of urgent and unscheduled care by people with long term conditions? A systematic review with meta-analysis. J Psychosom Res 2012;73:334–42. doi:10.1016/j.jpsychores.2012.08.018.

[56] DiGangi JA, Gomez D, Mendoza L, Jason LA, Keys CB, Koenen KC. Pretrauma risk factors for posttraumatic stress disorder: A systematic review of the literature. Clin Psychol Rev 2013;33:728–44. doi:10.1016/j.cpr.2013.05.002.

[57] Diniz BS, Butters MA, Albert SM, Dew MA, Reynolds CF 3rd. Late-life depression and risk of vascular dementia and Alzheimer's disease: systematic review and meta-analysis of community-based cohort studies. Br J Psychiatry 2013;202:329–35. doi:10.1192/bjp.bp.112.118307.

[58] Elbers NA, Hulst L, Cuijpers P, Akkermans AJ, Bruinvels DJ. Do compensation processes impair mental health? A meta-analysis. Injury 2013;44:674–83. doi:10.1016/j.injury.2011.11.025.

[59] El-Nashar SA, Hopkins MR, Barnes SA, Pruthi RK, Gebhart JB, Cliby WA, et al. Health-related quality of life and patient satisfaction after global endometrial ablation for menorrhagia in women with bleeding disorders: a follow-up survey and systematic review. Am J Obstet Gynecol 2010;202:348.e1-7. doi:10.1016/j.ajog.2009.11.032.

[60] Faedda GL, Marangoni C, Serra G, Salvatore P, Sani G, Vazquez GH, et al. Precursors of bipolar disorders: A systematic literature review of prospective studies. J Clin Psychiatry 2015;76:614–24. doi:10.4088/JCP.13r08900.

[61] Fan Z, Wu Y, Shen J, Ji T, Zhan R. Schizophrenia and the risk of cardiovascular diseases: A meta-analysis of thirteen cohort studies. J Psychiatr Res 2013;47:1549–56. doi:10.1016/j.jpsychires.2013.07.011.

[62] Flak AL, Su S, Bertrand J, Denny CH, Kesmodel US, Cogswell ME. The association of mild, moderate, and binge prenatal alcohol exposure and child neuropsychological outcomes: a meta-analysis. Alcohol Clin Exp Res 2014;38:214–26. doi:10.1111/acer.12214.

[63] Geulayov G, Gunnell D, Holmen TL, Metcalfe C. The association of parental fatal and non-fatal suicidal behaviour with offspring suicidal behaviour and depression: A systematic review and meta-analysis. Psychol Med 2012;42:1567–80. doi:10.1017/S0033291711002753.

[64] Goldfarb SS, Tarver WL, Locher JL, Preskitt J, Sen B. A systematic review of the association between family meals and adolescent risk outcomes. J Adolesc 2015;44:134–49. doi:10.1016/j.adolescence.2015.07.008.

[65] Heerde JA, Scholes-Balog KE, Hemphill SA. Associations between youth homelessness, sexual offenses, sexual victimization, and sexual risk behaviors: A systematic literature review. Arch Sex Behav 2015;44:181–212. doi:10.1007/s10508-014-0375-2.

[66] Heikkilä K, Madsen IEH, Nyberg ST, Fransson EI, Ahola K, Alfredsson L, et al. Job strain and the risk of inflammatory bowel diseases: individual-participant meta-analysis of 95,000 men and women. PLoS One 2014;9:e88711. doi:10.1371/journal.pone.0088711.

[67] Hemmi MH, Wolke D, Schneider S. Associations between problems with crying, sleeping and/or feeding in infancy and long-term behavioural outcomes in childhood: a meta-analysis. Arch Dis Child 2011;96:622–9. doi:10.1136/adc.2010.191312.

[68] Hu R, Li Y, Zhang Z, Yan W. Antenatal depressive symptoms and the risk of preeclampsia or operative deliveries: a meta-analysis. PLoS One 2015;10:e0119018. doi:10.1371/journal.pone.0119018.

[69] Hughes K, Bellis MA, Jones L, Wood S, Bates G, Eckley L, et al. Prevalence and risk of violence against adults with disabilities: A systematic review and meta-analysis of observational studies. Lancet 2012;379:1621–9. doi:10.1016/S0140-6736(11)61851-5.

[70] Jokela M, Batty GD, Hintsa T, Elovainio M, Hakulinen C, Kivimaki M. Is personality associated with cancer incidence and mortality? An individual-participant meta-analysis of 2156 incident cancer cases among 42,843 men and women. Br J Cancer 2014;110:1820–4. doi:10.1038/bjc.2014.58.

[71] Jokela M, Batty GD, Nyberg ST, Virtanen M, Nabi H, Singh-Manoux A, et al. Personality and all-cause mortality: individual-participant meta-analysis of 3,947 deaths in 76,150 adults. Am J Epidemiol 2013;178:667–75. doi:10.1093/aje/kwt170.

[72] Käkelä J, Panula J, Oinas E, Hirvonen N, Jääskeläinen E, Miettunen J, et al. Family history of psychosis and social, occupational and global outcome in schizophrenia: A meta-analysis. Acta Psychiatr Scand 2014;130:269–78. doi:10.1111/acps.12317.

[73] Kaymaz N, Drukker M, Lieb R, Wittchen H-U, Werbeloff N, Weiser M, et al. Do subthreshold psychotic experiences predict clinical outcomes in unselected non-help-seeking population-based samples? A systematic review and meta-analysis, enriched with new results. Psychol Med 2012;42:2239–53. doi:10.1017/S0033291711002911.

[74] Keall RM, Clayton JM, Butow PN. Therapeutic life review in palliative care: a systematic review of quantitative evaluations. J Pain Symptom Manage 2015;49:747–61. doi:10.1016/j.jpainsymman.2014.08.015.

[75] Keshaviah A, Edkins K, Hastings ER, Krishna M, Franko DL, Herzog DB, et al. Re-examining premature mortality in anorexia nervosa: A meta-analysis redux. Compr Psychiatry 2014;55:1773–84. doi:10.1016/j.comppsych.2014.07.017.

[76] Kropelin TF, Neyens JCL, Halfens RJG, Kempen GIJM, Hamers JPH, Kröpelin TF, et al. Fall determinants in older long-term care residents with dementia: A systematic review. Int Psychogeriatrics 2013;25:549–63. doi:10.1017/S1041610212001937.

[77] Kuijpers KF, van der Knaap LM, Lodewijks IAJ. Victims' influence on intimate partner violence revictimization: A systematic review of prospective evidence. Trauma, Violence, Abus 2011;12:198–219. doi:10.1177/1524838011416378.

[78] Laisné F, Lecomte C, Corbière M. Biopsychosocial predictors of prognosis in musculoskeletal disorders: A systematic review of the literature (corrected and republished)*. Disabil Rehabil An Int Multidiscip J 2012;34:1912–41. doi:10.3109/09638288.2012.729362.

[79] LeBlanc AG, Spence JC, Carson V, Gorber SC, Dillman C, Janssen I, et al. Systematic review of sedentary behaviour and health indicators in the early years (aged 0-4 years). Appl Physiol Nutr Metab Appl Nutr Metab 2012;37:753–72. doi:10.1139/h2012-063.

[80] Lehmann V, Hagedoorn M, Tuinman MA. Body image in cancer survivors: a systematic review of case-control studies. J Cancer Surviv 2014;9:339–48. doi:10.1007/s11764-014-0414-y.

[81] Leung YW, Flora DB, Gravely S, Irvine J, Carney RM, Grace SL. The impact of premorbid and postmorbid depression onset on mortality and cardiac morbidity among patients with coronary heart disease: meta-analysis. Psychosom Med 2012;74:786–801. doi:10.1097/PSY.0b013e31826ddbed.

[82] Linsell L, Malouf R, Johnson S, Morris J, Kurinczuk JJ, Marlow N. Prognostic factors for behavioral problems and psychiatric disorders in children born very preterm or very low birth weight: A systematic review. J Dev Behav Pediatr 2016;37:88–102. doi:10.1097/DBP.0000000000000238.

[83] Liu B, Tarigan LH, Bromet EJ, Kim H. World Trade Center disaster exposure-related probable posttraumatic stress disorder among responders and civilians: a meta-analysis. PLoS One 2014;9:e101491. doi:10.1371/journal.pone.0101491.

[84] Liu RT, Alloy LB. Stress generation in depression: A systematic review of the empirical literature and recommendations for future study. Clin Psychol Rev 2010;30:582–93. doi:10.1016/j.cpr.2010.04.010.

[85] Liu S, Zhu Y, Chen W, Sun T, Cheng J, Zhang Y. Risk factors for the second contralateral hip fracture in elderly patients: A systematic review and meta-analysis. Clin Rehabil 2015;29:285–94. doi:10.1177/0269215514542358.

[86] Long MH, Johnston V, Bogossian F. Work-related upper quadrant musculoskeletal disorders in midwives, nurses and physicians: A systematic review of risk factors and functional consequences. Appl Ergon 2012;43:455–67. doi:10.1016/j.apergo.2011.07.002.

[87] Lovato C, Watts A, Stead LF. Impact of tobacco advertising and promotion on increasing adolescent smoking behaviours. Cochrane Database Syst Rev 2011:CD003439. doi:10.1002/14651858.CD003439.pub2.

[88] Mäkikangas A, Kinnunen U, Feldt T, Schaufeli W. The longitudinal development of employee well-being: a systematic review. Work Stress 2016;8373:1–25. doi:10.1080/02678373.2015.1126870.

[89] Makin SDJ, Turpin S, Dennis MS, Wardlaw JM. Cognitive impairment after lacunar stroke: systematic review and meta-analysis of incidence, prevalence and comparison with other stroke subtypes. J Neurol Neurosurg Psychiatry 2013;84:893–900. doi:10.1136/jnnp-2012-303645.

[90] Malik S, Kanwar A, Sim LA, Prokop LJ, Wang Z, Benkhadra K, et al. The association between sleep disturbances and suicidal behaviors in patients with psychiatric diagnoses: a systematic review and meta-analysis. Syst Rev 2014;3:18. doi:10.1186/2046-4053-3-18.

[91] Masson M, East-Richard C, Cellard C. A meta-analysis on the impact of psychiatric disorders and maltreatment on cognition. Neuropsychology 2016;30:143–56. doi:10.1037/neu0000228.

[92] Modabbernia A, Mollon J, Boffetta P, Reichenberg A. Impaired Gas Exchange at Birth and Risk of Intellectual Disability and Autism: A Meta-analysis. J Autism Dev Disord 2016;46:1847–59. doi:10.1007/s10803-016-2717-5.

[93] Monette MCE, Baird A, Jackson DL. A Meta-Analysis of Cognitive Functioning in Nondemented Adults with Type 2 Diabetes Mellitus. Can J Diabetes 2014;38:401–8. doi:10.1016/j.jcjd.2014.01.014.

[94] Moulton CD, Koychev I. The effect of penicillin therapy on cognitive outcomes in neurosyphilis: A systematic review of the literature. Gen Hosp Psychiatry 2015;37:49–52. doi:10.1016/j.genhosppsych.2014.10.008.

[95] Murri MB, Prestia D, Mondelli V, Pariante C, Patti S, Olivieri B, et al. The HPA axis in bipolar disorder: Systematic review and meta-analysis. Psychoneuroendocrinology 2016;63:327–42. doi:10.1016/j.psyneuen.2015.10.014.

[96] Nieuwenhuijsen K, Bruinvels D, Frings-Dresen M. Psychosocial work environment and stress-related disorders, a systematic review. Occup Med (Chic Ill) 2010;60:277–86. doi:10.1093/occmed/kqq081.

[97] Nilaweera I, Doran F, Fisher J. Prevalence, nature and determinants of postpartum mental health problems among women who have migrated from South Asian to high-income countries: A systematic review of the evidence. J Affect Disord 2014;166:213–26. doi:10.1016/j.jad.2014.05.021.

[98] Okun MA, Yeung EWH, Brown S. Volunteering by older adults and risk of mortality: A meta-analysis. Psychol Aging 2013;28:564–77. doi:10.1037/a0031519.

[99] Palmisano GL, Innamorati M, Vanderlinden J. Life adverse experiences in relation with obesity and binge eating disorder: A systematic review. J Behav Addict 2016;5:11–31. doi:10.1556/2006.5.2016.018.

[100] Picorelli AMA, Pereira LSM, Pereira DS, Felicio D, Sherrington C, Assumpcao Picorelli AM, et al. Adherence to exercise programs for older people is influenced by program characteristics and personal factors: a systematic review. J Physiother 2014;60:151–6. doi:10.1016/j.jphys.2014.06.012.

[101] Pinquart M, Duberstein PR. Depression and cancer mortality: a meta-analysis. Psychol Med 2010;40:1797–810. doi:10.1017/S0033291709992285.

[102] Pompili M, Gonda X, Serafini G, Innamorati M, Sher L, Amore M, et al. Epidemiology of suicide in bipolar disorders: a systematic review of the literature. Bipolar Disord 2013;15:457–90. doi:10.1111/bdi.12087.

[103] Prang K-H, Newnam S, Berecki-Gisolf J. The impact of family and work-related social support on musculoskeletal injury outcomes: a systematic review. J Occup Rehabil 2015;25:207–19. doi:10.1007/s10926-014-9523-8.

[104] Prieto ML, Cuéllar-Barboza AB, Bobo W V., Roger VL, Bellivier F, Leboyer M, et al. Risk of myocardial infarction and stroke in bipolar disorder: A systematic review and exploratory meta-analysis. Acta Psychiatr Scand 2014;130:342–53. doi:10.1111/acps.12293.

[105] Proper KI, Singh AS, van Mechelen W, Chinapaw MJM. Sedentary behaviors and health outcomes among adults: A systematic review of prospective studies. Am J Prev Med 2011;40:174–82. doi:10.1016/j.amepre.2010.10.015.

[106] Raggi A, Giovannetti AM, Quintas R, D'Amico D, Cieza A, Sabariego C, et al. A systematic review of the psychosocial difficulties relevant to patients with migraine. J Headache Pain 2012;13:595–606. doi:10.1007/s10194-012-0482-1.

[107] Ramond A, Bouton C, Richard I, Roquelaure Y, Baufreton C, Legrand E, et al. Psychosocial risk factors for chronic low back pain in primary care--a systematic review. Fam Pract 2011;28:12–21. doi:10.1093/fampra/cmq072.

[108] Rashid M, Goez HR, Mabood N, Damanhoury S, Yager JY, Joyce AS, et al. The impact of pediatric traumatic brain injury (TBI) on family functioning: a systematic review. J Pediatr Rehabil Med 2014;7:241–54. doi:10.3233/PRM-140293.

[109] Rehm J, Taylor B, Mohapatra S, Irving H, Baliunas D, Patra J, et al. Alcohol as a risk factor for liver cirrhosis: A systematic review and meta-analysis. Drug Alcohol Rev 2010;29:437–45. doi:10.1111/j.1465-3362.2009.00153.x.

[110] Rossi G, Frediani S, Rossi R, Rossi A. Long-acting antipsychotic drugs for the treatment of schizophrenia: Use in daily practice from naturalistic observations. BMC Psychiatry 2012;12.

[111] Sbarra DA, Law RW, Portley RM. Divorce and Death: A Meta-Analysis and Research Agenda for Clinical, Social, and Health Psychology. Perspect Psychol Sci 2011;6:454–74. doi:10.1177/1745691611414724.

[112] Schmidt U, Willmund G-D, Holsboer F, Wotjak CT, Gallinat J, Kowalski JT, et al. Searching for non-genetic molecular and imaging PTSD risk and resilience markers: Systematic review of literature and design of the German Armed Forces PTSD biomarker study. Psychoneuroendocrinology 2015;51:444–58. doi:10.1016/j.psyneuen.2014.08.020.

[113] Sømhovd MJ, Hansen BM, Brok J, Esbjørn BH, Greisen G. Anxiety in adolescents born preterm or with very low birthweight: A meta-analysis of case-control studies. Dev Med Child Neurol 2012;54:988–94. doi:10.1111/j.1469-8749.2012.04407.x.

[114] Song J, Viggiano A, Monda M, De Luca V. Peripheral glutamate levels in schizophrenia: Evidence from a meta-analysis. Neuropsychobiology 2014;70:133–41. doi:10.1159/000364828.

[115] Stehr MD, Von Lengerke T. Preventing weight gain through exercise and physical activity in the elderly: A systematic review. Maturitas 2012;72:13–22. doi:10.1016/j.maturitas.2012.01.022.

[116] Sturmberg C, Marquez J, Heneghan N, Snodgrass S, van Vliet P. Attentional focus of feedback and instructions in the treatment of musculoskeletal dysfunction: a systematic review. Man Ther 2013;18:458–67. doi:10.1016/j.math.2013.07.002.

[117] Surkan PJ, Kennedy CE, Hurley KM, Black MM. Maternal depression and early childhood growth in developing countries: systematic review and meta-analysis. WHO Bull 2011;89:608–15. doi:10.2471/BLT.11.088187.

[118] Taylor G, McNeill A, Girling A, Farley A, Lindson-Hawley N, Aveyard P. Change in mental health after smoking cessation: systematic review and meta-analysis. BMJ 2014;348:g1151.

[119] Tetley A, Moghaddam NG, Dawson DL, Rennoldson M. Parental bonding and eating disorders: A systematic review. Eat Behav 2014;15:49–59. doi:10.1016/j.eatbeh.2013.10.008.

[120] Thangaratinam S, Rogozinska E, Jolly K, Glinkowski S, Duda W, Borowiack E, et al. Interventions to reduce or prevent obesity in pregnant women: a systematic review. Health Technol Assess 2012;16:iii–iv, 1-191. doi:10.3310/hta16310.

[121] Theorell T, Hammarstrom A, Aronsson G, Traskman Bendz L, Grape T, Hogstedt C, et al. A systematic review including meta-analysis of work environment and depressive symptoms. BMC Public Health 2015;15:738. doi:10.1186/s12889-015-1954-4.

[122] Trenchard SO, Rust S, Bunton P. A systematic review of psychosocial outcomes within 2 years of paediatric traumatic brain injury in a school-aged population. Brain Inj 2013;27:1217–37. doi:10.3109/02699052.2013.812240.

[123] Vangeli E, Stapleton J, Smit ES, Borland R, West R. Predictors of attempts to stop smoking and their success in adult general population samples: A systematic review. Addiction 2011;106:2110–21. doi:10.1111/j.1360-0443.2011.03565.x.

[124] Vu JA, Babikian T, Asarnow RF. Academic and language outcomes in children after traumatic brain injury: A meta-analysis. Except Child 2011;77:263–81.

[125] Wang Y-M, Shu B-C, Fetzer S, Chang Y-J. Parenting style of women who conceived using in vitro fertilization: a meta-analysis. J Nurs Res 2014;22:69–80. doi:10.1097/JNR.0000000000000025.

[126] Weierink L, Vermeulen RJ, Boyd RN. Brain structure and executive functions in children with cerebral palsy: A systematic review. Res Dev Disabil 2013;34:1678–88. doi:10.1016/j.ridd.2013.01.035.

[127] Wilson LC. Mass Shootings: A Meta-Analysis of the Dose-Response Relationship. J Trauma Stress 2014;27:631–8. doi:10.1002/jts.21964.

[128] Winters BD, Eberlein M, Leung J, Needham DM, Pronovost PJ, Sevransky JE. Long-term mortality and quality of life in sepsis: a systematic review. Crit Care Med 2010;38:1276–83. doi:10.1097/CCM.0b013e3181d8cc1d.

[129] Xenaki L-A, Pehlivanidis A. Clinical, neuropsychological and structural convergences and divergences between Attention Deficit/Hyperactivity Disorder and Borderline Personality Disorder: A systematic review. Pers Individ Dif 2015;86:438–49. doi:10.1016/j.paid.2015.06.049.

[130] Zhu T, Ye X, Zhang T, Lin Z, Shi W, Wei X, et al. Association between alcohol consumption and multiple sclerosis: A meta-analysis of observational studies. Neurol Sci 2015;36:1543–50. doi:10.1007/s10072-015-2326-7.

[131] Zijlmans MAC, Riksen-Walraven JM, de Weerth C. Associations between maternal prenatal cortisol concentrations and child outcomes: A systematic review. Neurosci Biobehav Rev 2015;53:1–24. doi:10.1016/j.neubiorev.2015.02.015.

[132] Wells G, Shea B, O'connell D, Peterson J. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses n.d. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp (accessed November 24, 2016).

[133] Hopewell S, Boutron I, Altman DG, Ravaud P. Incorporation of assessments of risk of bias of primary studies in systematic reviews of randomised trials: a cross-sectional study. BMJ Open 2013;3:e003342. doi:10.1136/bmjopen-2013-003342.

[134] Jarde A, Losilla JM, Vives J, Rodrigo MF. Q-Coh: A tool to screen the methodological quality of cohort studies in systematic reviews and meta-analyses. Int J Clin Heal Psychol 2013;13:138–46.

[135] Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomized studies of interventions. Bmj 2016:4–10. doi:10.1136/bmj.i4919.

[136] Thompson S, Ekelund U, Jebb S, Lindroos AK, Mander A, Sharp S, et al. A proposed method of bias adjustment for meta-analyses of published observational studies. Int J Epidemiol 2011;40:765–77. doi:10.1093/ije/dyq248.

[137] Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. Ann Intern Med 2013;158:280–6. doi:10.7326/0003-4819-158-4-201302190-00009.

[138] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6:e1000097. doi:10.1371/journal.pmed.1000097.

[139] Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of Observational Studies in Epidemiology - A Proposal for Reporting. JAMA 2000;283:2008–12. doi:10.1001/jama.283.15.2008.

[140] Centre for Reviews and Dissemination U of Y. Systematic reviews: CRD's guidance for undertaking reviews in health care. CRD, University of York; 2009. doi:10.1016/S1473-3099(10)70065-7.

[141] Effective Public Health Practice Project: Quality assessment Tool for Quantitative Studies n.d. http://www.ephpp.ca/tools.html (accessed November 24, 2016).

[142] Stalenhoef PA, Crebolder HFJM, Knottnerus JA, Van Der Horst FGEM. Incidence, risk factors and consequences of falls among elderly subjects living in the community: a criteria-based analysis. Eur J Public Health 1997;7:328–34.

[143] Sherehiy B, Karwowski W, Marek T. Relationship between risk factors and musculoskeletal disorders in the nursing profession: a systematic review. Occup Ergon 2004;4:241–279 39p.

[144] Tak LM, Cleare AJ, Ormel J, Manoharan A, Kok IC, Wessely S, et al. Meta-analysis and meta-regression of hypothalamic-pituitary-adrenal axis activity in functional somatic disorders. Biol Psychol 2011;87:183–94. doi:10.1016/j.biopsycho.2011.02.002.

[145] Mirza I, Jenkins R. Risk factors, prevalence, and treatment of anxiety and depressive disorders in Pakistan: systematic review. BMJ 2004;328:794. doi:10.1136/bmj.328.7443.794.

[146] Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health 1998;52:377–84.

[147] Kmet LM, Lee RC, Cook LS. Standard Quality Assessment Criteria for Evaluating Primary Research Papers. Alberta Herit Found Med Res 2004;13:1–22.

http://www.ihe.ca/publications/standard-quality-assessment-criteria-for-evaluating-primary-research-papers-from-a-variety-of-fields (accessed November 25, 2016).

[148] National Institute for Mental Health and Clinical Excellence. The guidelines manual. London: NICE; 2009.

[149] Ariëns GAM, Van Mechelen W, Bongers PM, Bouter LM, Van Der Wal G. Physical risk factors for neck pain. Scand J Work Environ Heal 2000;26:7–19. doi:10.5271/sjweh.504.

**Table 1.** Approaches to inclusion of methodological quality assessment in research synthesis

| | |
|---|---|
| Criteria for inclusion/exclusion | Quality assessment as criteria for inclusion/exclusion in the review |
| Descriptive strategies | Narrative strategies: discussion of quality, results of quality assessment presented in a table or study quality as a basis for recommendations for future research |
| | Plot/Figure |
| Exploration of associations between effect sizes and quality | Quality assessment to conduct sensitivity analyses |
| | Cumulative meta-analysis in order of quality |
| | Quality-based subgroup analyses |
| | Including quality as a variable in a regression analysis: meta-regression models |
| Bias adjustment methods | Quality as a factor to weight a meta-analysis |
| | Quality effects model |
| | Elicitation of bias distributions from experts |

**Table 2.** Description of the methodological quality elements of the reviews

| Attribute assessed | *n* | Category | Frequency | Percent |
|---|---|---|---|---|
| Was any guideline followed? | 90 | No | 48 | 53.3 |
| | | Yes | 33 | 36.7 |
| | | Partially | 9 | 10.0 |
| Which guideline was followed? * | 42 | PRISMA [138] | 28 | 66.7 |
| | | MOOSE [139] | 12 | 28.6 |
| | | CRD [140] | 3 | 7.1 |
| | | GRADE [17] | 5 | 11.9 |
| | | Other | 4 | 9.5 |
| Was the protocol of the review registered? | 90 | No | 87 | 96.7 |
| | | Yes (PROSPERO) | 3 | 3.3 |
| Was any element related to RoB used as eligibility criteria? | 90 | No | 66 | 73.3 |
| | | Yes | 22 | 24.4 |
| | | Unclear | 2 | 2.2 |
| Domains of RoB used as eligibility criteria * | 22 | Selection | 3 | 13.6 |
| | | Confounding | 2 | 9.1 |
| | | Exposure | 7 | 31.8 |
| | | Performance | 14 | 63.6 |
| | | Outcome | 3 | 13.6 |
| Was a particular domain of bias mentioned as methodological limitation? | 90 | No | 12 | 13.3 |
| | | Yes | 78 | 86.7 |
| Domain of bias mentioned as methodological limitation * | 78 | Selection | 25 | 32.1 |
| | | Confounding | 46 | 59.0 |
| | | Exposure | 46 | 59.0 |
| | | Performance | 7 | 9.0 |
| | | Outcome | 42 | 53.8 |
| | | Attrition | 10 | 12.8 |
| | | Reported results | 2 | 2.6 |
| Were the main confounding variables that could threaten the results reported? | 90 | No | 30 | 33.3 |
| | | Yes | 30 | 33.3 |
| | | Unclear | 30 | 33.3 |
| Was the RoB of primary studies assessed? | 90 | No | 41 | 45.6 |
| | | Yes | 46 | 51.1 |
| | | Unclear | 3 | 3.3 |
| Was any meta-analysis performed? | 90 | No | 44 | 48.9 |
| | | Yes | 46 | 51.1 |

PRISMA = Preferred Reporting Items for Systematic reviews and Meta-Analyses; MOOSE = Meta-analysis Of Observational Studies in Epidemiology; CRD = Centre for Reviews and Dissemination; GRADE = Grading of Recommendations Assessment, Development and Evaluation; PROSPERO = International prospective register of systematic reviews; RoB = risk of bias.

* Percentage of cases (multi-response items).

**Table 3.** Risk of bias assessment of the primary studies included on the reviews

| Attribute assessed | *n* | Category | Frequency | Percent |
|---|---|---|---|---|
| Type of tool used to assess RoB | 46 | Specific biases | 10 | 21.7 |
| | | Standard tool | 18 | 39.1 |
| | | Standard modified tool | 13 | 28.3 |
| | | New tool | 2 | 4.3 |
| | | Mixture | 3 | 6.5 |
| Standard scale/check-list used to assess RoB (modified or not) | 31 | NOS [132] | 16 | 51.6 |
| | | EPHPP [141] | 3 | 9.7 |
| | | GRADE [17] | 2 | 6.5 |
| | | Stalenhoef et al. [142] | 1 | 3.2 |
| | | Sherehiy et al. [143] | 1 | 3.2 |
| | | Tak et al. [144] | 1 | 3.2 |
| | | Mirza and Jenkins [145] | 1 | 3.2 |
| | | Downs and Black [146] | 1 | 3.2 |
| | | QualSyst tool [147] | 1 | 3.2 |
| | | NICE [148] | 1 | 3.2 |
| | | Ariens et al. [149] | 1 | 3.2 |
| | | STROBE [5] | 1 | 3.2 |
| | | QUIPS [137] | 1 | 3.2 |
| Was the RoB assessment tool reported when no standard tool was used? | 28 | No | 7 | 25.0 |
| | | Yes | 21 | 75.0 |
| Type of index of the RoB assessment tool | 46 | Quantitative | 31 | 67.4 |
| | | Categories | 8 | 17.4 |
| | | Not reported | 7 | 15.2 |
| Type of outcome provided by the RoB assessment tool | 46 | Overall RoB | 14 | 30.4 |
| | | Overall RoB and RoB domains | 25 | 54.3 |
| | | Not reported | 7 | 15.2 |
| Were the results of RoB assessment used as exclusion criteria? | 46 | No | 41 | 89.1 |
| | | Yes, overall RoB | 5 | 10.9 |
| Reporting of the results of RoB assessment | 46 | Detailed results for each primary study | 30 | 65.2 |
| | | Overall results | 15 | 32.6 |
| | | Not reported | 1 | 2.2 |
| Variability in RoB assessment results | 46 | Low | 13 | 28,3 |
| | | Moderate to high | 25 | 54,3 |
| | | Not reported | 8 | 17.4 |

RoB = risk of bias; NOS = Newcastle-Ottawa Scale; EPHPP = Effective Public Health Practice Project; GRADE = Grading of Recommendations Assessment, Development and Evaluation; NICE = National Institute for Mental Health and Clinical Excellence; STROBE = STrengthening the Reporting of OBservational studies in Epidemiology; QUIPS = Quality In Prognostic Studies.

**Table 4.** Incorporation of the risk of bias assessment results into the research syntheses

| Attribute assessed | *n* | Category | Frequency | Percent |
|---|---|---|---|---|
| How were RoB assessment results incorporated into the review results? | 46 | Only descriptive | 29 | 63.0 |
| | | Analytical | 11 | 23.9 |
| | | Not incorporated | 6 | 13.0 |
| Type of descriptive incorporation | 39 | Only narrative | 36 | 92.3 |
| | | Plot | 3 | 7.7 |
| RoB narrative incorporation: abstract | 38 | Specifically mentioned | 8 | 21.1 |
| | | General comment | 6 | 15.8 |
| | | Not reported | 24 | 63.2 |
| RoB narrative incorporation: discussion | 38 | Specifically mentioned | 25 | 65.8 |
| | | General comment | 7 | 18.4 |
| | | Not reported | 6 | 15.8 |
| RoB narrative incorporation: conclusions | 38 | Specifically mentioned | 5 | 13.2 |
| | | General comment | 6 | 15.8 |
| | | Not reported | 27 | 71.1 |
| RoB narrative incorporation: recommendations | 38 | Specifically mentioned | 21 | 55.3 |
| | | General comment | 8 | 21.1 |
| | | Not reported | 9 | 23.7 |
| RoB analytical incorporation* | 11 | Sensitivity analysis | 5 | 45.5 |
| | | Subgroup analysis | 4 | 36.4 |
| | | Meta-regression | 2 | 18.2 |
| Was there some significant influence of RoB on the results? | 46 | No | 8 | 17.4 |
| | | Yes, overall RoB | 6 | 13.0 |
| | | Yes, RoB domains (confounding) | 3 | 6.5 |
| | | Unclear | 10 | 21.7 |
| | | Not reported | 19 | 41.3 |

RoB = risk of bias.

* Percentage of cases (multi-response items).

**Figure 1.** Flowchart of the selection process.

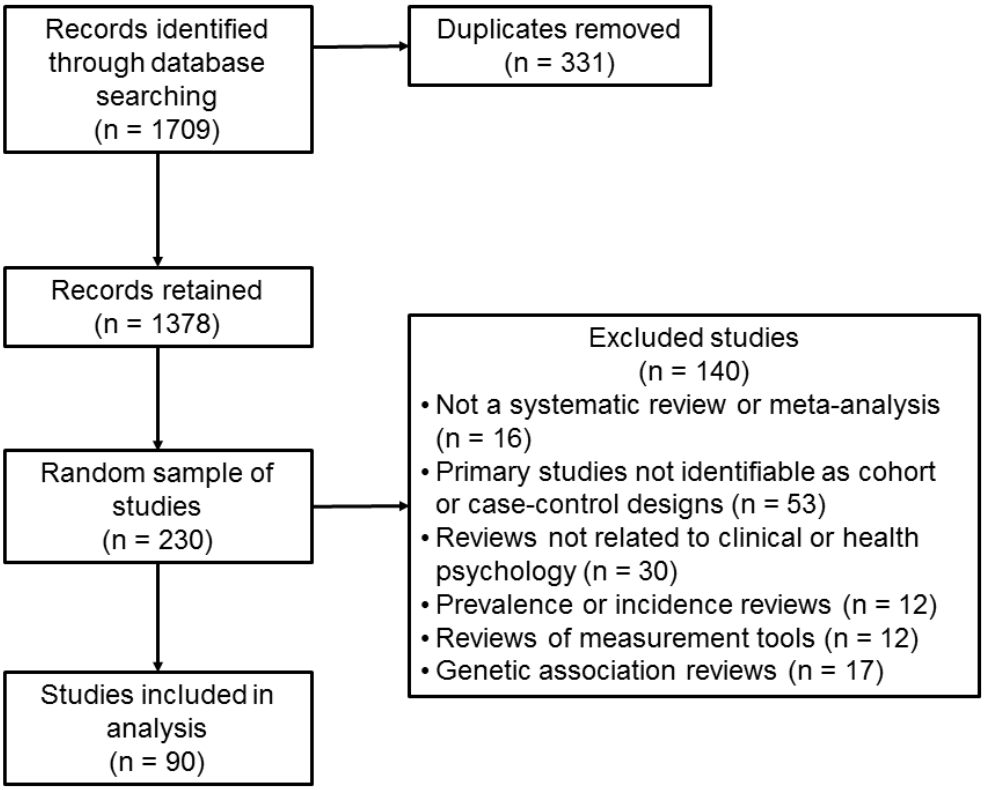
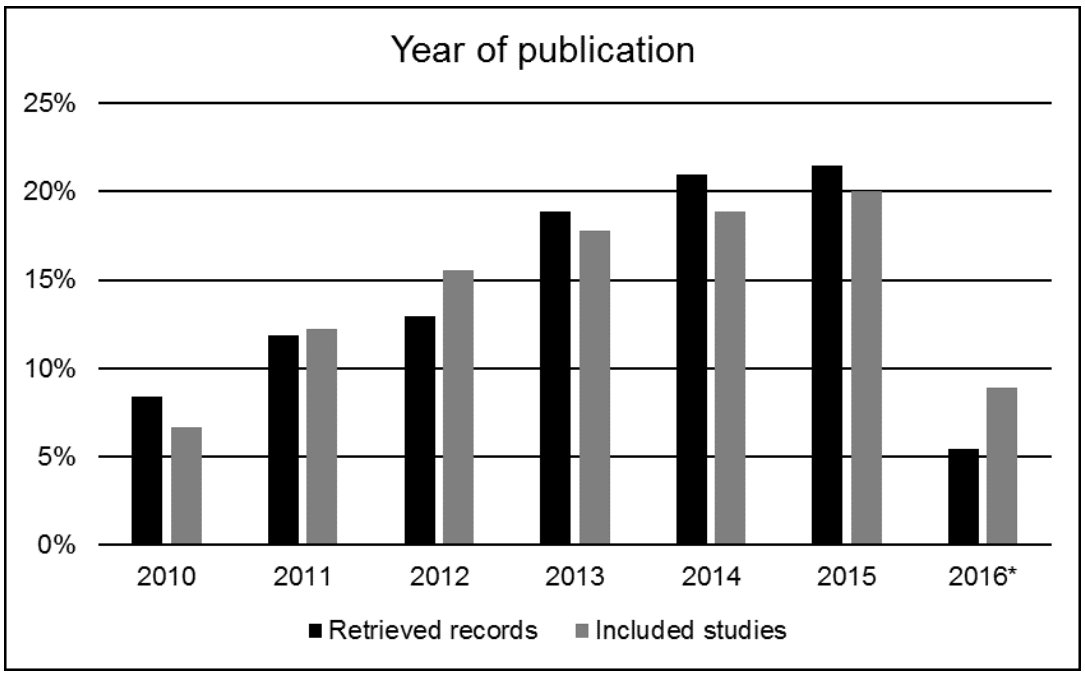



**Figure 2.** Distribution by year of publication of the retrieved records (once duplicates were removed, *n* = 1378) and the included studies (*n* = 90).

* Data until May 2016.



## 2.2. Three risk of bias tools lead to opposite conclusions in observational research synthesis (Artículo 2)

El primer artículo puso de manifiesto una serie de importantes carencias respecto a la evaluación y la incorporación del RdS en la síntesis de investigación. El siguiente paso se centró en uno de los factores clave de estas carencias: la repercusión de la elección de una determinada herramienta de evaluación del RdS. Así, el segundo artículo publicado es un estudio empírico comparativo de la fiabilidad, validez e idoneidad de tres herramientas de evaluación para estudios de diseños no experimentales, que además se propone determinar el efecto de los resultados de la evaluación del RdS sobre los resultados del MA.

Para ello, las tres herramientas seleccionadas fueron las siguientes: 1) la Newcastle-Ottawa Scale (NOS; Wells, Shea, O'Connell, & Peterson, 2000), como herramienta más utilizada para evaluar el RdS de estudios de cohortes y casos y controles (Oliveras, Losilla, & Vives, 2017); 2) una versión piloto del nuevo Q-Coh (Jarde et al., 2013), herramienta desarrollada por nuestro grupo de investigación, basada en dominios de sesgo y con buenas propiedades psicométricas; y 3) Risk Of Bias In Nonrandomized Studies of Interventions (ROBINS-I; Sterne et al., 2016) una nueva herramienta, también basada en dominios de sesgo, propuesta por Cochrane y destinada principalmente a estudios no aleatorizados de intervención, pero también aplicable a una amplia variedad de estudios no experimentales. Estas tres herramientas fueron testadas en un estudio piloto por dos evaluadores independientes con el fin de poder acordar algunos elementos previos, necesarios para su aplicación (e.g. factores de confusión relevantes).

Las herramientas fueron aplicadas por estos mismos evaluadores a 28 estudios de cohortes incluidos en un MA del ámbito de la psicología sanitaria, siguiendo un esquema de aleatorización del orden de los estudios y de las herramientas. También la facilidad de uso de las herramientas fue evaluada para varios aspectos, utilizando una escala de cinco puntos. La posible influencia del RdS sobre los resultados del MA se evaluó a través de análisis de sensibilidad, análisis de subgrupos y meta-regresiones.

Los resultados de la evaluación del RdS mostraron gran variabilidad en función de la herramienta utilizada: los estudios fueron calificados como bajo RdS en un 75% de los

casos con la NOS, en un 11% con Q-Coh y en ninguno con ROBINS-I. Respecto al acuerdo entre evaluadores, la NOS mostró un acuerdo entre moderado y bueno, mientras que Q-Coh y ROBINS-I mostraron unos niveles de acuerdo más discretos. La correlación entre Q-Coh y ROBINS-I fue buena para la mayoría de los dominios de sesgo, mientras que la correlación entre estas dos herramientas y la NOS fue pobre en general. También los resultados de la evaluación de la facilidad de uso de las tres herramientas mostraron las similitudes entre Q-Coh y ROBINS-I. Por último, los resultados de los análisis de subgrupos y meta-regresiones utilizados para determinar la influencia del RdS sobre los resultados del MA no mostraron asociaciones significativas entre calidad y tamaño del efecto.

Los resultados de la comparación de las tres herramientas fueron presentados en el XV Congreso de Metodología de las Ciencias Sociales y de la Salud, celebrado en Barcelona (2017).

El artículo en su versión final puede obtenerse en el siguiente vínculo:

Losilla, J.-M., Oliveras, I., Marin-Garcia, J. A., & Vives, J. (2018). Three risk of bias tools lead to opposite conclusions in observational research synthesis. *Journal of Clinical Epidemiology*, *101*, 61–72. http://doi.org/10.1016/j.jclinepi.2018.05.021

Factor de impacto 2017 JCR: 4.245, Q1

Factor de impacto 2017 SJR: 2.862, Q1

# Three risk of bias tools lead to opposite conclusions in observational research synthesis

Josep-Maria Losilla[a], Isabel Oliveras[a], Juan A. Marin-Garcia[b] and Jaume Vives[a].

[a] Department of Psychobiology and Methodology of Health Sciences, Psychology Faculty, Universitat Autònoma de Barcelona,

Carrer de la Fortuna, Edifici B, 08193 Bellaterra, Barcelona, Spain.

[b] Department of Business Management, School of Industrial Engineering,

Universitat Politècnica de València,

Dept. Organización de Empresas, Edificio 7D, Camino de Vera s/n, 46022 Valencia, Spain.

Corresponding author: Jaume Vives, Department of Psychobiology and Methodology of Health Sciences, Psychology Faculty, Universitat Autònoma de Barcelona, Carrer de la Fortuna, Edifici B, 08193 Bellaterra, Barcelona, Spain.

Tel.: 0034-93 5812331; fax: 0034-93 5812001.

E-mail address: Jaume.Vives@uab.cat

Conflict of interest: All authors declare no conflict of interest.

Funding: This work was supported by the Spanish Ministry of Science and Innovation (grant number: PSI2014-52962-P). I.O. was supported by funding from a predoctoral grant from the Ministry of Education, Culture and Sport of Spanish Government (grant number: FPU14/04514). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Objectives:** The aim of this study is to assess the agreement and compare the performance of three different instruments in assessing risk of bias (RoB) of comparative cohort studies included in a health psychology meta-analysis.

**Study Design and Setting:** Three tools were applied to 28 primary studies included in the selected meta-analysis: the Newcastle-Ottawa Scale (NOS), Quality of Cohort studies (Q-Coh), and Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I).

**Results:** Inter-rater agreement varied greatly from tool to tool. For overall RoB, 75% of the studies were rated as low RoB with the NOS, 11% of the studies with Q-Coh, and no study was found to be at low RoB using ROBINS-I. No influence of quality ratings on the meta-analysis results was found for any of the tools.

**Conclusion:** Assessing RoB using the three tools may lead opposite conclusions, especially at low and high levels of RoB. Domain-based tools (Q-Coh and ROBINS-I) provide a more comprehensive framework for identifying potential sources of bias, which is essential to improving the quality of future research. Both further guidance on the application of RoB tools and improvements in the reporting of primary studies are necessary.

Keywords: risk of bias; methodological quality; systematic review; meta-analysis; domain of bias; quality tool.

Running title: NOS, Q-Coh and ROBINS-I may lead to opposite conclusions

Word count: 4990

**What is new?**

**Key findings**

- Assessing RoB using the three tools may lead opposite conclusions, especially at low and high levels of RoB, where most of the studies were rated as low RoB with the NOS, contrary to ROBINS-I with which most of the studies were rated as high RoB, while Q-Coh showed greater variability. Therefore, both the NOS and ROBINS-I showed low capability in grading RoB in observational studies.
- Correlation between Q-Coh and ROBINS-I was good for most of the domains of bias, while correlations between these two tools and the NOS showed poorer agreement. Raters' assessments of the usability of the tools also reveal the similarities between Q-Coh and ROBINS-I.
- The results of subgroup and meta-regression analyses showed no clear association between RoB and combined effect sizes when a meta-analysis is performed.

**What this adds to what is known?**

- Although this study has found that Q-Coh and ROBINS-I are comprehensive and valid tools compared to the NOS, their reliability need to be improved.
- This paper provides empirical evidence that the NOS assessment of RoB is overly positive.
- To our knowledge, this is the first time that the properties of ROBINS-I have been tested. When applying ROBINS-I, the use of a target trial makes it difficult to discriminate levels of RoB between observational studies and hinders the understanding of some items.

**What is the implication and what should change now?**

- To improve the reliability of the tools, two conditions must be met: (1) the development of detailed guidance and training in the application of RoB assessment tools, and (2) improvements in the reporting of primary studies.
- In the context of systematic reviews and meta-analysis, RoB assessments make it possible to identify weaknesses in research designs and should guide the improvement of the quality of future studies, which is especially relevant to synthesize the results of non-experimental research.

## 7. Introduction

Assessing the methodological quality or RoB of primary studies is an essential component of any systematic review or meta-analysis [1,2], and should play a relevant role in interpreting the results of the review [3]. Moreover, the inclusion of poor quality studies in a review may lead to invalid conclusions [3,4]. In fact, the results of such quality assessments often exert an important influence on some decisions made in the review process, such as whether to exclude studies not meeting certain quality standards, to perform sensitivity analyses, to determine the strength of evidence or to guide recommendations for future research and clinical practice [5,6].

Compared to clinical trials, the quality assessment of observational studies is often more demanding due to the variety of designs comprised and their increased susceptibility to bias [5,7,8]. These difficulties are probably the reason why in some areas such as health psychology, only about half of all reviews that include cohort and case-control studies assessed the RoB of the primary studies [9]. Although a wide range of tools suitable for observational studies have been reviewed by several authors [10–12], there is no consensus on which is the best procedure or tool to assess RoB in observational designs, despite observational studies are usually included in systematic reviews including those of Cochrane [13]. Moreover, most of these tools were poorly developed, and their developers often failed to follow standard methodological procedures or to test their tools' validity and reliability [10,14]. Thus, RoB assessments of a single study using different tools may lead to different conclusions [4,15,16], both in randomized controlled trials [1,14,17] and in observational studies [7,8,18].

Meanwhile, the use of scales that provide a single summary score is strongly discouraged [4,15,19], because it involves the weighting of component items, even though some of them may be not related to RoB [3,11]. The alternative seems to perform a RoB assessment based in domains [20–23], which is increasingly applied and apparently provides a more structured framework within which to make qualitative decisions on the overall quality of studies and to detect potential sources of bias [16].

The general purpose of this study is to assess the agreement and compare the performance of three different instruments in assessing the RoB of comparative cohort studies included in a meta-analysis related to health psychology. The selected tools were:

(a) the Newcastle-Ottawa Scale (NOS) [24], the most frequently used scale to assess the quality of cohort and case-control studies [9], which provides a summary score; (b) Quality of Cohort studies (Q-Coh) [21], a specific domain-based tool to assess the RoB of cohort studies with good psychometric properties; and (c) Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I) [22], a new domain-based tool proposed by Cochrane which is intended to assess RoB in non-randomized studies of interventions but is also applicable to a wide variety of observational designs [25]. To be more precise, the specific objectives are:

- To estimate, for each tool, the degree of inter-rater agreement when examining items, domains of RoB and overall quality rating.
- To estimate the level of agreement between tools for specific biases, domains of RoB and overall quality rating.
- To appraise the qualitative aspects of the tools related to their usability: the average time spent, clarity of instructions and items, coverage and validity.
- To determine the effect of quality ratings on the results of a meta-analysis.

## 8. Methods

### *8.1. Risk of bias assessment tools*

The NOS [24] was developed to assess the quality of observational studies included in systematic reviews. This tool exists in separate versions for cohort and case-control designs, although only the scale for cohort studies was applied here. Studies are assessed using eight items broken down into three dimensions: selection (four items), comparability (one item) and exposure for case-control studies or outcome for cohort studies (three items). A study can be awarded a maximum of nine stars. Although the tool's developers have said that the validity and reliability of the tool have been established, no further specific information has been published. Nevertheless, subsequent studies that have tested the NOS have come to quite varied results in terms of inter-rater reliability and validity [7,26–28].

Q-Coh [21] is a bias domain-based tool specifically intended to assess the RoB of cohort studies, and developed by two of the authors of the present paper. A pilot of the second version of the tool was applied in this study (Appendix A). This version of Q-Coh is structured around four domains of bias: (a) selection (3 items), (b) confounding (4

items), (c) exposure measures (3 items), and (d) outcome measures (5 items). Each domain of RoB is evaluated as "yes" (potential bias) and "no" (no bias), and the overall assessment of RoB based on the four domains is rated as "low RoB", "moderate RoB" or "high RoB". The tool also includes several previous considerations about the most important confounding factors in the research field under study, the acceptable percentage of missing data, and the exposure and outcome variables of interest. The reliability and validity of the original tool were established by the developers, with inter-rater agreement kappa values ranging from 0.60 to 0.87 for the different domains (0.75 for overall assessment), and a weighted kappa value equal to 0.41 ($p = 0.003$) for validity analyzed by studying the agreement of the ratings of the overall RoB of the studies with an external rating.

The ROBINS-I tool [22] is a new published tool from Cochrane for assessing RoB in NRSI (observational and quasi-randomized studies). The authors define bias as "systematic difference between the results of the NRSI and the results expected from the target trial" [22], where the "target trial" is a hypothetical randomized trial without threat of bias. Although ROBINS-I was designed to compare the effects of two or more interventions, the term "intervention" refers here to either treatment or exposure, thus including studies in which no intervention was carried out by the investigators [25]. The tool is structured around seven domains of bias: (a) bias due to confounding, (b) bias in selection of participants into the study, (c) bias in classification of interventions, (d) bias due to deviations from intended interventions, (e) bias due to missing data, (f) bias in measurement of outcomes, and (g) bias in selection of the reported results. Every domain includes a series of signaling questions to help the reviewers judge the RoB in the same four categories as in the overall RoB (low, moderate, serious and critical RoB), based on the responses given to the signaling questions. The developers also provide detailed guidance on the use of ROBINS-I [25] but, to the best of our knowledge, to date there is no published data on the reliability and validity of this tool.

*8.2. Selected studies*

A systematic review with meta-analysis [29] in the field of health psychology, including 28 comparative cohort studies (the study design which is common to the three tools), was selected to test the properties of the tools. The references of the 28 primary studies are provided in Appendix B. The selected review explores the prospective

association between depression and the risk of developing stroke in adults. No intervention was applied to the participants of the studies. The main meta-analysis, including 31 effect sizes, was performed using a random-effects model, and its summary effect shows an increased risk of stroke morbidity and mortality in depressed participants compared with non-depressed individuals, with a hazard ratio (HR) estimate of 1.45 (95% CI, 1.29-1.63) and moderate to high heterogeneity ($I2 = 66.0\%$; Cochrane Q statistic = 88.1, $p < 0.001$).

*8.3. Procedure*

The three tools were pilot tested by two independent raters (IO and JAM) on a study not included in the meta-analysis. The study team discussed and agreed upon some previous considerations that were necessary for the application of the tools (i.e. exposure and outcome of interest, the most important confounding factors, acceptable percentage of missing data and minimum duration of follow-up). A randomization scheme was then developed to randomize the order of studies and the order of the tools applied by each rater. Following this randomization scheme, the two raters independently applied each of the three tools to the 28 selected papers over the same time frame (approximately two months) to reduce both intra-rater and inter-rater variation in quality assessment.

Consensus on quality ratings was reached via the following process: (1) if the scores of both raters were the same, then that score was used as a consensus score; (2) if the scores differed, a consensus score was agreed upon by both raters; and (3) if the raters were unable to reach a consensus, then the score provided by a third rater (JML) was used for discussion and final agreement.

Finally, the ease of use of the tools was rated on a five-point scale for each of the following aspects (Appendix C): (a) coverage of the tool (the extent to which the tool covers all the relevant domains of bias and features or if it includes non-relevant items), (b) clarity of instructions, (c) clarity of the items, and (d) ability to discriminate between studies of high and low RoB.

*8.4. Analysis*

For the purposes of comparison and using a strategy similar to that applied by other authors (e.g. [30–33]), three categories were created using the NOS scores: studies scoring 0-3, 4-6 and 7-9 were deemed to be of low, moderate and high quality,

respectively. When at least one item in an NOS domain reflected bias, RoB was considered to be present in that domain.

Traditionally, the most commonly used statistics to evaluate inter-rater agreement are Cohen's kappa coefficient [34] or its generalization for multiple raters proposed by Fleiss [35]. However, when the prevalence of a given response category is notably high or low these statistics are not advisable. Under these circumstances, the "kappa paradox" appears, meaning that the value of the kappa statistic is low even when the observed proportion of agreement is significantly high [36]. A second kappa paradox arises when the extent to which raters disagree on the proportion of cases in each response category is large, resulting in kappa values higher than when this bias is low or absent. Given that kappa is difficult to interpret in the presence of different prevalence or bias, several studies have recommended including detailed information about the proportions of specific agreement between raters for each response category in order to make it possible to evaluate the possible effects of prevalence or bias [37–39]. Additionally, in the presence of different prevalence or bias, a widely used alternative to Cohen's kappa is the Prevalence-Adjusted and Bias-Adjusted Kappa (PABAK) proposed by Byrt, Bishop and Carlin [40].

Therefore, several statistics are provided for each item, domain and overall RoB: the proportion of inter-rater agreement, the proportion of choices for each response category, and the Fleiss kappa statistic (or PABAK when necessary). Following the recommendations of Q-Coh and ROBINS-I authors', inter-rater reliability was calculated once the equivalent categories of response had been merged (i.e. yes/probably yes, no/probably no for ROBINS-I, yes/presumably and yes/not necessary for Q-Coh). Landis & Koch's [41] criteria were used to define the levels of agreement for kappa (or PABAK): no agreement (< 0), slight (0-0.20), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80) and almost perfect agreement (0.81-1). All these analyses were performed using the irr package (v.0.84) [42] for R version 3.3.2.

In order to determine agreement between tools, the items of each tool were classified into common domains and aspects of bias. The comparison of the groups of related items was performed applying the non-parametric Kendall's tau-b correlation coefficient [43], using the Statistical Package for Social Sciences (SPSS) version 20.0

(SPSS, Inc., 2009, Chicago, IL, USA). Good agreement between the tools indicate that they estimate comparable constructs, offering an indirect measure of concurrent validity.

Beyond the qualitative assessment of the tools, some indicators of their usability were also calculated using Microsoft Excel 2016. The median time spent on the application of the tools was calculated, counting only the scoring time in the cases where each tool was applied first. Each tool was also measured in terms of the mode number of items in which the answer was "not reported" or "no information" and the mode number of items that required a third rater to reach a consensus.

Finally, subgroup, meta-regression and sensitivity analyses were performed to test the effects of quality ratings on the results of the meta-analysis for every tool, both in terms of each individual domain of bias and of overall RoB, following the classification previously created for the comparison between tools. These analyses were performed using Comprehensive Meta-analysis (CMA) version 3 (Biostat, Inc., 2014, Englewood, NJ, USA), following a mixed-effects model for subgroup analyses and a random-effects model with the Knapp Hartung adjustment for meta-regressions [44].

## 9. Results

### *9.1. Risk of bias assessment*

Figure 1 shows a summary of the consensus results of RoB assessment for each tool, for overall RoB and by domain. The NOS scores ranged from 5 to 9, with a median and mode of 8 (25th percentile [p25] = 6; p75 = 9). Once the studies were classified into categories, there were 21 studies with low RoB and 7 studies with moderate RoB. None of the studies were placed in the category of high RoB. According to the Q-Coh results, 3 studies were classified as low RoB, 11 studies as moderate RoB and 14 studies as high RoB. Finally, when ROBINS-I was applied, 4 studies were deemed to have moderate RoB and 24 to have serious RoB, with no studies classified as having critical RoB or no information categories. The fourth domain of RoB in ROBINS-I (bias due to deviations from intended interventions) was judged inapplicable due to the non-experimental design of the studies.

### *9.2. Inter-rater reliability*

Table 1 presents the proportion of agreement between raters for items (all the tools), individual domains and overall RoB (Q-Coh and ROBINS-I) and the kappa (or

PABAK) values, once equivalent response categories were joined as specified in the methods section above. It was not possible to synthesize Q-Coh results for items corresponding to the confusion domain (items from 4 to 7), as they were applied to all individual confounding variables within the same study.

Inter-rater agreement for the NOS items ranged from fair to almost perfect (kappa/PABAK: .20 to 1), with a proportion of agreement between raters ranging from 46% to 100%. Agreement between raters for Q-Coh ranged from no agreement to almost perfect agreement (kappa/PABAK: -.05, .93) for items and no agreement to substantial agreement (kappa/PABAK: -.20 to .64) for domains, registering no agreement for overall RoB (kappa = -.16). The percentage of agreement ranged from 39% to 96% for items and from 50% to 82% for domains, and the figure was 25% for overall RoB. The results for ROBINS-I ranged between slight and almost perfect agreement for items (kappa/PABAK: .00 to 1) and domains (kappa/PABAK: .08 to .93), and showed moderate agreement for overall RoB (PABAK = .57). The proportion of agreement between raters ranged from 39% to 100% for items, from 36% to 96% for domains, and was 79% for overall RoB. Note that the lowest agreement values were those corresponding to items or domains related to missing data for Q-Coh and ROBINS-I.

Additionally (see Appendix D), over 90% of the responses of the raters were concentrated in a single response category in three items of the NOS (two of them corresponding to comparability domain), only in one item of Q-Coh regarding outcome domain, and finally, in 18 items of ROBINS-I belonging to domains of confounding, selection, intervention, outcome and reported results. The results for the outcome domain for ROBINS-I also showed a significant lack of variability as 95%, of the responses were concentrated in a single category.

*9.3. Agreement between tools*

In order to make comparisons between the tools possible, the first step needed was an analysis of the content of the tools and classify related items into 19 aspects of bias, all while preserving the original domains of each of the tools (Table 2). Items related to missing data are distributed among several domains; these items were also grouped together in a single domain and compared.

The agreement coefficients between tools for overall RoB and the different domains are shown in Table 3. Regarding the NOS, only the associations of the missing data domain for Q-Coh and ROBINS-I and the outcome domain for Q-Coh were statistically significant. In contrast, the correlations between Q-Coh and ROBINS-I were significant for overall RoB and for most of the domains, with the exceptions of the selection and outcome domains.

*9.4. Usability of the tools*

The raters gave generally poor scores to the NOS for coverage and discriminative ability, while the results for the clarity of instructions were very good and the scores for the clarity of the items were moderate to very good (Table 4). The Q-Coh tool obtained moderate to very good rates for all the aspects assessed. The ratings of ROBINS-I were very poor for all the aspects, except for the coverage of the tool, which was considered moderate to good.

Table 4 also shows the median of time spent to rate the studies for each tool. Since the procedure used by the raters consisted of a prior identification of the necessary data for the application of the three RoB tools, the time spent in data extraction could not be accounted for individually for each tool. However, it must be pointed out that the time invested in ROBINS-I training was significantly higher compared to Q-Coh, while the time spent in the NOS was significantly lower compared to both. Results as to the number of items that were answered with a "not reported" or "no information" option, as well as items that required a third rater judgment to achieve consensus, highlighted the similarities between Q-Coh and ROBINS-I, as well as the differences between the NOS and the other two tools.

Additional raters' comments were centered in the poor quality of reporting of most primary studies, especially regarding data loss, and in the difficulties involved in applying the ROBINS-I tool to non-experimental studies. Another comment was on the need to come to a more detailed agreement on the criteria to be used prior to the application of the tools, thus allowing the raters to make quality assessment decisions with greater confidence.

*9.5. Effect of quality rating on meta-analysis results*

Only the subgroup analysis for outcome domain of ROBINS-I ($p < .001$) and the meta-regression analyses for selection and outcome domains of ROBINS-I ($p = .023$ and

p = .001, respectively) led to statistically significant results (Appendix E). However, these results cannot be considered relevant because of the low number of studies included in some of the analysis categories. In any case, it is worth commenting on some trends observed in the results of these analyses.

Figure 2 shows the results of the subgroup analyses for each tool and each domain of bias, and for overall RoB. Equivalent levels of RoB of each tool were grouped and classified by domains to facilitate comparison between tools. No significant differences were found between the tools for overall RoB. Moreover, no association between the level of bias and estimated effect sizes was found for any of the tools. Nevertheless, despite the absence of statistically significant results, the confounding domain in both Q-Coh and ROBINS-I shows a trend in the sense that the smaller the bias the smaller the effect size. The same does not happen with the NOS. It should also be noted that in the exposure and response domains the results are more homogeneous for the three tools, with the trend in the opposite direction to that of the confounding domain (i.e. the lower the RoB, the greater the effect size). Finally, sensitivity analyses excluding the studies at high RoB also provided non-relevant results (NOS HR 95% CI: 1.29-1.63; Q-Coh HR 95% CI: 1.19-1.90; ROBINS-I HR 95% CI: 0.93-2.02). However, the number of studies included showed that the least demanding tool was the NOS (n=31; i.e., all studies) compared to Q-Coh (n=15) and ROBINS-I (n=4).

**10. Discussion**

Our comparison of three tools for RoB assessment of non-experimental studies suggests that we are dealing here with three different approaches to RoB assessment, each of which could lead to different conclusions about the final quality grade assigned to each study. In this study, no agreement between tools was found for overall RoB. While 75% of the studies can be considered to be at low RoB when the NOS is applied, 86% of the studies would be at serious RoB according to ROBINS-I. Overall RoB measured with Q-Coh showed greater variability (11% low, 39% moderate and 50% high RoB). This lack of agreement corroborates the findings of a great deal of the previous work comparing quality tools for both experimental and non-experimental studies [1,4,14,17].

The findings on inter-rater agreement of the NOS are consistent with those of Hootman et al. [7], showing moderate to good inter-rater reliability and good usability (i.e. clarity of items, short scoring time and ease of consensus). In contrast, the lesser

degree of agreement between raters found in Q-Coh and ROBINS-I can be attributed to the broader scope of these tools, which implies a more comprehensive analysis of the primary studies than the NOS requires. Furthermore, Q-Coh and ROBINS-I are more demanding in terms of the amount of information collected and the level of detail used in assessing RoB, but, unfortunately, the quality of the reporting of primary studies was not always up to the standards set by these demands. In fact, among the main causes that have been pointed out in the literature to explain low inter-rater agreement is the difficulty in extracting certain specific information from poorly reported studies [1,14]. The fact that Q-Coh and ROBINS-I obtained the lowest agreements in items related to missing data would also point in this direction. It should be noted that, although agreement between raters in RoB domains for ROBINS-I and Q-Coh have a similar range, it is not the case of overall RoB, where ROBINS-I offers better results compared to Q-Coh. This might be explained by the detailed algorithm for the overall RoB judgment in ROBINS-I, which leaves little margin to rater's decision.

On the other hand, in our opinion, some items are hard to understand and may have negatively affected the inter-rater agreement. This could be due to the fact that ROBINS-I identifies a target trial as the gold standard against which all observational studies are assessed, as well as the fact that ROBINS-I "*to keep the analogy with the target trial (...) uses the term 'intervention' groups to refer to 'treatment' or 'exposure' groups in observational studies even though in such studies no actual intervention was implemented by the investigators"* [25, p.4]. Furthermore, the use in ROBINS-I of a target trial as a reference also makes it difficult to discriminate between the distinct levels of RoB in different observational studies. For example, compared to a target trial, no observational study can achieve low RoB in the confusion domain meaning that no observational study can be given the grade of low overall RoB. In this context, the difficulties we found in understanding the ROBINS-I items may have led to an over-agreement in the previous phase of the application of the tool, as discussed later in our study limitations.

Regarding agreement between the tools, Q-Coh and ROBINS-I showed good correlation for overall RoB and for three out of five domains of bias, while correlations between these two tools and the NOS showed poorer agreement. In this sense, these results are within the realm of what is to be expected, especially if we consider that Q-

Coh and ROBINS are tools based on bias domains, while the NOS is a global scoring scale. Only the selection domain showed no significant correlation between any of the tools, which is probably due to the different definition and conceptualization of this domain in the three tools. Otherwise, it is somewhat surprising that no significant correlation was found in the outcome domain for ROBINS-I. This discordance of ROBINS-I could be explained by the lack of direct assessment of the validity and reliability of data collection methods.

The raters' assessments of the usability of the tools reveals the similarities between Q-Coh and ROBINS-I, especially regarding the coverage of the tool, scoring time, loss of information and ease of consensus. However, the clarity of instructions and items of ROBINS-I, as well as its discriminative ability, were given very low scores by both raters. In addition, as expected, both tools clearly differ from the NOS in most aspects of usability evaluated. In this sense, the shorter scoring time required by the NOS may be one of the reasons for its greater generalization of use.

Finally, although the results of subgroup and meta-regression analyses showed some clear trends when Q-Coh or ROBINS-I are applied, almost all the estimates were small and non-significant. Our results are consistent with some previous literature that had found no association between quality rating and combined effect sizes [45–47]. Although other studies reported significant effects [16,48–50], there seems to be no clear patterns of associations [51,52]. It might be the case that low variability in RoB is hindering the emergence of an association between RoB and effect size estimates, since only moderate to high-quality studies tend to be included in meta-analyses.

### *10.1. Strengths and limitations*

To the best of our knowledge, this is the first time that the reliability and validity of ROBINS-I have been tested. Furthermore, this is the first study to our knowledge to compare the performance of two domain-based tools and a composite scoring scale applied to observational research.

However, these findings are subject to several limitations. First, the confusing instructions of ROBINS-I and its use of a target trial raised serious doubts among the raters during the pilot phase. Having a trial as reference study forced us to agree to very specific criteria not covered by the tool itself. These specific criteria allowed us to apply

the tool in a non-interventional context, where ROBINS-I is hardly applicable. We expect ROBINS-E [53], currently under development, to overcome this shortcoming. Unlike the NOS and Q-Coh, this pre-agreement in ROBINS-I entailed an ad-hoc tailoring for which this tool provides no guidelines. This non-standardized adaptation of ROBINS-I could have been properly evaluated if two teams of raters had been included. Second, there is no gold standard to adequately test concurrent validity, although good correlations between Q-Coh and ROBINS-I indicate that they are measuring similar constructs. Third, the relatively small number of studies considered in some categories of RoB limited the power of subgroup analyses and meta-regressions, leading to wide confidence intervals in those subgroups with few studies. Finally, RoB categories for the NOS were obtained from the overall quantitative score by setting cutoff points, which may be somewhat arbitrary.

### *10.2. Recommendations and future research*

There are some questions arising from our findings that should be explored. Firstly, it is not clear how reviewers should handle quality assessment in observational research, whether they should take as a reference a target trial or should assess studies against the best available evidence [8]. Although it seems that the tendency is to choose the first option [25,54,55], this seems to be in detriment of the ability to discriminate between different levels of quality when only observational studies are being assessed.

Moreover, although it seems domain-based tools showed better attributes and properties than composite scores [16], it is essential to find new methods or procedures that allow for improving the reliability of these tools. This improvement seems to depend on two essential conditions: detailed guidance and training in applying RoB assessment tools, and clear and complete reporting of primary studies. Specific guidance for RoB tools should include clear decision rules to reduce the sort of discrepancies that arose from differing interpretations of the tool [47]. Moreover, Faggion [56] suggested that researchers have to make accessible the rationale used for supporting their judgements to the end-users of systematic reviews. Regarding the quality of the reporting, it has proven to be crucial to the carrying out of a proper RoB assessment [14,47,48]. In our experience, it is too often difficult or even impossible to gather the necessary information to assess certain domains of bias (e.g. missing data). This situation is likely to improve considerably if scientific journals systematically include the results of the implementation

of reporting guidelines in its publications, as some journals have already done (e.g. BJUI International [57]).

## 11. Conclusions

The current study, comparing the performance of three different tools when assessing the RoB of 28 cohort studies, shows that assessing RoB on the same study using different tools may lead to opposite conclusions, especially at low and high levels of RoB, where most of the studies were rated as low RoB with the NOS, contrary to ROBINS-I with which most of the studies were rated as high RoB. Therefore, both the NOS and ROBINS-I showed low capability in grading RoB in observational studies. Our results showed also lower inter-rater agreement for the most comprehensive tools (Q-Coh and ROBINS-I), as well as lack of association between RoB and combined effect sizes when a meta-analysis is performed.

In the light of the results found, we must emphasize the important role of RoB assessment in systematic reviews and in the context of meta-analyses. In this context, RoB assessment provides invaluable information to describe the strength of the evidence found, beyond the usual tests of association between the levels of RoB and the effect estimates of primary studies as potential explanation for part of the observed heterogeneity. The analysis of the results of RoB assessment makes it possible to identify weaknesses in research designs and the most common deficiencies in reporting. This information plays an essential role in guiding the improvement of the quality of studies in a research area, which in turn is a basic objective of research synthesis, especially in non-experimental research.

## References

[1] Hartling L, Ospina M, Liang Y, Dryden DM, Hooton N, Krebs Seida J, et al. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. Bmj 2009;339:b4012–b4012. doi:10.1136/bmj.b4012.

[2] Johnson BT, Low RE, MacDonald H V. Panning for the gold in health research: Incorporating studies' methodological quality in meta-analysis. Psychol Health 2014;30:135–52. doi:10.1080/08870446.2014.953533.

[3] Higgins J, Green S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available at: http://training.cochrane.org/handbook. Last accessed December 12, 2017.

[4] Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. Jama-Journal Am Med Assoc 1999;282:1054–60. doi:10.1001/jama.282.11.1054.

[5] Centre for Reviews and Dissemination. Systematic reviews : CRD's guidance for undertaking reviews in health care. CRD, University of York; 2009. Available at: https://www.york.ac.uk/crd/guidance/. Last accessed May 30, 2017.

[6] Jüni P, Altman DG, Egger M. Systematic reviews in health care - Assessing the quality of controlled clinical trials. Br Med J 2001;323:42–6. doi:10.1136/bmj.323.7303.42.

[7] Hootman JM, Driban JB, Sitler MR, Harris KP, Cattano NM. Reliability and validity of three quality rating instruments for systematic reviews of observational studies. Res Synth Methods 2011;2:110–8. doi:10.1002/jrsm.41.

[8] Margulis A V, Pladevall M, Riera-Guardia N, Varas-Lorenzo C, Hazell L, Berkman ND, et al. Quality assessment of observational studies in a drug-safety systematic review, Comparison of two tools: The Newcastle-Ottawa scale and the RTI item bank. Clin Epidemiol 2014;6:981–93. doi:10.2147/CLEP.S66677.

[9] Oliveras I, Losilla J-M, Vives J. Methodological quality is underrated in systematic reviews and meta-analyses in health psychology. J Clin Epidemiol 2017;86:59–70. doi:10.1016/j.jclinepi.2017.05.002.

[10] Deeks JJ, Dinnes J, D'Amico R, Sowden a J, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7:iii–x, 1-173.

[11] Sanderson S, Tatt ID, Higgins JPT. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol 2007;36:666–76. doi:10.1093/ije/dym018.

[12] Jarde A, Losilla JM, Vives J. Methodological quality assessment tools of non-experimental studies: a systematic review. An Psicol 2012;28:617–28. doi:10.6018/analesps.28.2.148911.

[13] Ijaz S, Verbeek JH, Mischke C, Ruotsalainen J. Inclusion of nonrandomized studies in Cochrane systematic reviews was found to be in need of improvement. J Clin Epidemiol 2014;67:645–53. doi:10.1016/j.jclinepi.2014.01.001.

[14] Armijo-Olivo S, Stiles CR, Hagen NA, Biondo PD, Cummings GG. Assessment of study quality for systematic reviews: a comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: methodological research. J Eval Clin Pract 2012;18:12–8. doi:10.1111/j.1365-2753.2010.01516.x.

[15] Herbison P, Hay-Smith J, Gillespie WJ. Adjustment of meta-analyses on the basis of quality scores should be abandoned. J Clin Epidemiol 2006;59:1249–56. doi:10.1016/j.jclinepi.2006.03.008.

[16] O'Connor SR, Tully MA, Ryan B, Bradley JM, Baxter GD, McDonough SM. Failure of a numerical quality assessment scale to identify potential risk of bias in a systematic review: a comparison study. BMC Res Notes 2015;8:224. doi:10.1186/s13104-015-1181-1.

[17] Colle F, Rannou F, Revel M, Fermanian J, Poiraudeau S. Impact of quality scales on levels of evidence inferred from a systematic review of exercise therapy and low back pain. Arch Phys Med Rehabil 2002;83:1745–52. doi:10.1053/apmr.2002.35657.

[18] Jarde A, Losilla JM, Vives J. Suitability of three different tools for the assessment of methodological quality in ex post facto studies. Int J Clin Heal Psychol 2012;12:97–108.

[19] Whiting P, Rutjes AWS, Dinnes J, Reitsma JB, Bossuyt PMM, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. J Clin Epidemiol 2005;58:1–12. doi:10.1016/j.jclinepi.2004.04.008.

[20] Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ Br Med J 2011;343:1–9.

[21] Jarde A, Losilla JM, Vives J, Rodrigo MF. Q-Coh: A tool to screen the methodological quality of cohort studies in systematic reviews and meta-analyses. Int J Clin Heal Psychol 2013;13:138–46.

[22] Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomized studies of interventions. Bmj 2016:4–10. doi:10.1136/bmj.i4919.

[23] Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529–36. doi:10.7326/0003-4819-155-8-201110180-00009.

[24] Wells G, Shea B, O'Connell D, Peterson J. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses 2000. Available at: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. Last accessed May 30, 2017.

[25] Sterne JA, Higgins JPT, Elbers RG, Reeves BC, the development group for ROBINS-I. Risk Of Bias In Non-randomized Studies of Interventions ( ROBINS-I ): detailed guidance. 2016. Available at: http:/www.riskofbias.info. Last accessed May 30, 2017.

[26] Hartling L, Milne A, Hamm MP, Vandermeer B, Ansari M, Tsertsvadze A, et al. Testing the Newcastle Ottawa Scale showed low reliability between individual reviewers. J Clin Epidemiol 2013;66:982–93. doi:10.1016/j.jclinepi.2013.03.003.

[27] Lo CK-L, Mertz D, Loeb M. Newcastle-Ottawa Scale: comparing reviewers' to authors' assessments. BMC Med Res Methodol 2014;14:45. doi:10.1186/1471-2288-14-45.

[28] Oremus M, Oremus C, Hall GBC, McKinnon MC. Inter-rater and test–retest reliability of quality assessments by novice student raters using the Jadad and

Newcastle–Ottawa Scales. BMJ Open 2012;2:e001368. doi:10.1136/bmjopen-2012-001368.

[29] Pan A, Sun Q, Okereke OI, Rexrode KM, Hu FB. Depression and risk of stroke morbidity and mortality: A meta-analysis and systematic review. JAMA J Am Med Assoc 2011;306:1241–9. doi:10.1001/jama.2011.1282.

[30] Jike M, Itani O, Watanabe N, Buysse DJ, Kaneita Y. Long sleep duration and health outcomes: A systematic review, meta-analysis and meta-regression. Sleep Med Rev 2017. doi:10.1016/j.smrv.2017.06.011.

[31] Porcelli B, Pozza A, Bizzaro N, Fagiolini A, Costantini MC, Terzuoli L, et al. Association between stressful life events and autoimmune diseases: A systematic review and meta-analysis of retrospective case-control studies. Autoimmun Rev 2016;15:325–34. doi:10.1016/j.autrev.2015.12.005.

[32] Xue J, Chen W, Chen L, Gaudet L, Moher D, Walker M, et al. Significant discrepancies were found in pooled estimates of searching with Chinese indexes versus searching with English indexes. J Clin Epidemiol 2016;70:246–53. doi:10.1016/j.jclinepi.2015.09.014.

[33] Zheng Y, Wu X, Lin X, Lin H. The Prevalence of Depression and Depressive Symptoms among Eye Disease Patients: A Systematic Review and Meta-analysis. Sci Rep 2017;7:1–9. doi:10.1038/srep46453.

[34] Cohen J. A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas 1960;20:37–46. doi:10.1177/001316446002000104.

[35] Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76:378–82. doi:10.1037/h0031619.

[36] Feinstein AR, Cicchetti D V. High agreement but low Kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43:543–9. doi:10.1016/0895-4356(90)90158-L.

[37] Lantz CA, Nebenzahl E. Behavior and interpretation of the κ statistic: Resolution of the two paradoxes. J Clin Epidemiol 1996;49:431–4. doi:10.1016/0895-4356(95)00571-4.

[38] Cicchetti D V., Feinstein AR. High Agreement but Low Kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990;43:551–8. doi:10.1016/0895-4356(90)90159-M.

[39] Uebersax J. Statistical methods for diagnostic agreement, http://www.john-uebersax.com/stat/agree.htm; 2015. Last accessed October 25, 2017.

[40] Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46:423–9. doi:10.1016/0895-4356(93)90018-V.

[41] Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics 1977;33:159. doi:10.2307/2529310.

[42] Gamer M, Lemon J, Fellows I, Sing P. irr: Various coefficients of interrater reliability and agreement (Version 0.84) [software] 2012.

[43] Kendall MG. A New Measure of Rank Correlation. Biometrika 1938;30:81. doi:10.2307/2332226.

[44] Hartung J, Knapp G, Sinha BK. Statistical Meta-Analysis with Applications. Wiley; 2008. doi:10.1002/9780470386347.

[45] Verhagen AP, de Vet HCW, Vermeer F, Widdershoven JWMG, de Bie R a, Kessels AGH, et al. The influence of methodologic quality on the conclusion of a landmark meta-analysis on thrombolytic therapy. Int J Technol Assess Health Care 2002;18:11–23. doi:10.1017/S0266462309091016.

[46] Balk EM, Bonis PAL, Moskowitz H, Schmid CH, Ioannidis JPA, Wang C, et al. Correlation of Quality Measures With Estimates of Treatment Effect in Meta-analyses of Randomized Controlled Trials. JAMA 2002;287:2973–82. doi:10.1001/jama.287.22.2973.

[47] Hartling L, Hamm M, Milne A, Vandermeer B, Santaguida PL, Ansari M, et al. Validity and Inter-rater Reliability Testing of Quality Assessment Instruments. Agency for Healthcare Research and Quality (US); 2012.

[48] Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998;352:609–13. doi:10.1016/S0140-6736(98)01085-X.

[49] Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273:408–12. doi:10.1097/00043764-199603000-00007.

[50] Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. vol. 7. 2003. doi:10.3310/hta7010.

[51] Ahn S, Becker BJ. Incorporating Quality Scores in Meta-Analysis. J Educ Behav Stat 2011;36:555–85. doi:10.3102/1076998610393968.

[52] Conn VS, Rantz MJ. Focus on research methods: Research methods: Managing primary study quality in meta-analyses. Res Nurs Health 2003;26:322–33. doi:10.1002/nur.10092.

[53] Morgan R, Sterne J, Higgins J, Thayer K, Schunemann H, Rooney A TK. A new instrument to assess Risk of Bias in Non-randomised Studies of Exposures (ROBINS-E): Application to studies of environmental exposure. Abstr. Glob. Evid. Summit, Cape Town: Cochrane Database of Systematic Reviews; 2017. doi:10.1002/14651858.CD201702.

[54] Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. Bias modelling in evidence synthesis. J R Stat Soc Ser A-Statistics Soc 2009;172:21–47. doi:10.1111/j.1467-985X.2008.00547.x.

[55] Hernán MA. With Great Data Comes Great Responsibility. Epidemiology 2011;22:290–1. doi:10.1097/EDE.0b013e3182114039.

[56] Faggion CM. The rationale for rating risk of bias should be fully reported. J Clin Epidemiol 2016;76:238. doi:10.1016/j.jclinepi.2016.03.007.

[57] Bhindi B, Wallis CJD, Boorjian SA, Thompson RH, Farrell A, Kim SP, et al. The Role of Lymph Node Dissection in the Management of Renal Cell Carcinoma: A Systematic Review and Meta-Analysis. BJU Int 2018. doi:10.1111/bju.14127.

Table 1. Results of inter-rater agreement

| NOS | | | Q-Coh | | | ROBINS-I[a] | | |
|---|---|---|---|---|---|---|---|---|
| Items | P. Overall agreement | Kappa/ PABAK | Items and domains | P. Overall agreement | Kappa/ PABAK | Items and domains | P. Overall agreement | Kappa/ PABAK |
| Selection 1 | .46 | .24 | Item 1 | .79 | .57[b] | 1.1 | 1.00 | 1.00[b] |
| Selection 2 | .96 | .93[b] | Item 2 | .96 | .93[b] | 1.2 | .93 | .86[b] |
| Selection 3 | .96 | .93[b] | Item 3 | .46 | .23 | 1.3 | .93 | .86[b] |
| Selection 4 | .75 | .50[b] | Selection | .61 | .32 | 1.4 | .79 | .57[b] |
| Comparability a | .93 | .86[b] | Confounding[c] | .82 | .64[b] | 1.5 | .68 | .36[b] |
| Comparability b | 1.00 | 1.00[b] | Item 8 | .79 | .57[b] | 1.6 | .96 | .93[b] |
| Outcome 1 | .68 | .36[b] | Item 9 | .64 | .29[b] | 1.7 | 1.00 | 1.00[b] |
| Outcome 2 | .96 | .93[b] | Item 10 | .39 | .02 | 1.8 | 1.00 | 1.00[b] |
| Outcome 3 | .61 | .21[b] | Exposure | .50 | -.20 | Confounding | .75 | .50[b] |
| | | | Item 11 | .71 | .43[b] | 2.1 | .82 | .64[b] |
| | | | Item 12 | .89 | .79[b] | 2.2 | .82 | .64[b] |
| | | | Item 13 | .93 | .86[b] | 2.3 | .82 | .64[b] |
| | | | Item 14 | .79 | .57[b] | 2.4 | .96 | .93[b] |
| | | | Item 15 | .39 | -.05 | 2.5 | .89 | .79[b] |
| | | | Outcome | .64 | .29[b] | Selection | .82 | .64[b] |
| | | | Overall RoB | .25 | -.16 | 3.1 | .82 | .64[b] |
| | | | | | | 3.2 | .96 | .93[b] |
| | | | | | | 3.3 | 1.00 | 1.00[b] |
| | | | | | | Intervention | .79 | .57[b] |
| | | | | | | 5.1 | .75 | .50[b] |
| | | | | | | 5.2 | .50 | .14 |
| | | | | | | 5.3 | .43 | .00 |
| | | | | | | 5.4 | .39 | .11 |
| | | | | | | 5.5 | .54 | .08 |
| | | | | | | Missing data | .36 | .08 |
| | | | | | | 6.1 | 1.00 | 1.00[b] |
| | | | | | | 6.2 | .86 | .71[b] |
| | | | | | | 6.3 | 1.00 | 1.00 |
| | | | | | | 6.4 | 1.00 | 1.00 |
| | | | | | | Outcome | .96 | .93[b] |
| | | | | | | 7.1 | .75 | .50[b] |
| | | | | | | 7.2 | .57 | .14[b] |
| | | | | | | 7.3 | .89 | .79[b] |
| | | | | | | Reported result | .64 | .29[b] |
| | | | | | | Overall RoB | .79 | .57[b] |

Note: P. Overall agreement = proportion of agreement between raters; RoB = Risk of Bias.
[a]Domain 4 of ROBINS-I "Bias due to deviations from intended interventions" was considered not applicable.
[b]PABAK
[c]Agreement for items of confounding domain (4 to 7) could not be calculated.

Table 2. Items from the three tools classified into common domains and aspects of bias

| Domains and aspects of bias | NOS Items | Q-Coh Items | ROBINS-I Signaling questions |
|---|---|---|---|
| Confounding / comparability | | | |
| Potential of confounding | | | 1.1 |
| Baseline confounding factors | C1a, C1b | 4, 7 | 1.4, 1.5, 1.6 |
| Confounding during follow-up | | 4, 7 | 1.2, 1.3, 1.7, 1.8 |
| Missing data on confounders | | 5, 6 | 5.3 |
| Selection | | | |
| Representativeness | S1 | | |
| Exclusion of participants or different criteria | S2 | 3 | |
| Selection based on variables after start | | | 2.1, 2.2, 2.3, 2.5 |
| Outcome not present at start | S4 | 1 | |
| Coincidence of intervention and follow-up start | | 2 | 2.4, 2.5 |
| Exposure | | | |
| Exposure measure | S3 | 8 | |
| Classification of participants | | | 3.1, 3.2, 3.3 |
| Missing data on exposure | | 9, 10 | 5.2 |
| Outcome | | | |
| Blinding of assessors | | 12 | 6.1, 6.2 |
| Outcome measure | O1 | 11 | 6.3, 6.4 |
| Length of follow-up | O2 | 13 | |
| Attrition/lost to follow-up | O3 | 14, 15 | 5.1 |
| Missing data | O3 | 5, 6, 9, 10, 14, 15 | 5.1, 5.2, 5.3, 5.4, 5.5 |
| Selective reporting of results | | | 7.1, 7.2, 7.3 |

Table 3. Results of agreement between tools

| | | NOS | Q-Coh |
|---|---|---|---|
| Overall risk of bias | ROBINS-I | -.058 | .580** |
| | NOS | - | -.200 |
| Confounding/Comparability | ROBINS-I | -.160 | -.913** |
| | NOS | - | .175 |
| Selection | ROBINS-I | .250 | -.258 |
| | NOS | - | .167 |
| Exposure/Classification of interventions | ROBINS-I | .000 | -.595* |
| | NOS | - | .093 |
| Outcome | ROBINS-I | .167 | .132 |
| | NOS | - | .640** |
| Missing data | ROBINS-I | -.546** | -.691** |
| | NOS | - | .683** |

Note: Kendall's tau-b correlation coefficient and its significance test were used.
* $p < .05$, ** $p < .01$

Table 4. Summary of the usability of the tools

| Attribute | NOS | Q-Coh | ROBINS-I |
|---|---|---|---|
| Coverage of the tool[a] | 2-3 | 4-5 | 2-3 |
| Clarity of instructions[a] | 5 | 3-4 | 1 |
| Clarity of items[a] | 3-5 | 4 | 1 |
| Discriminative ability[a] | 2 | 3-4 | 1 |
| Time[b] | 4 (2, 6) | 13 (11, 20) | 17 (14, 20) |
| Number of items answered with NR/NI option[c] | 0 (0-1) | 2 (0-6) | 1 (1-7) |
| Number of items requiring a third rater for consensus[c] | 0 (0-1) | 0 (0-6) | 0 (0-10) |

Note: NR = not reported; NI = no information.
[a]Range of scores. Each of these attributes was rated from 1 (*poor*) to 5 (*excellent*).
[b]Median of minutes in the cases where each tool was first applied ($25^{th}$ and $75^{th}$ percentiles in parentheses). Only accounted for scoring time. Time spent identifying relevant information before data extraction varies considerably depending on the tool.
[c]Mode number (range in parentheses).

Figure 1. Consensus results of risk of bias assessment for overall risk of bias and domains of bias for each tool. Categories for overall risk of bias were: low, moderate and high risk of bias for NOS and Q-Coh; low, moderate, serious, critical risk of bias and no information for ROBINS-I. Categories for domains of bias were: risk of bias "yes" or "no" for NOS and Q-Coh; for ROBINS-I the same categories as overall risk of bias were applied.

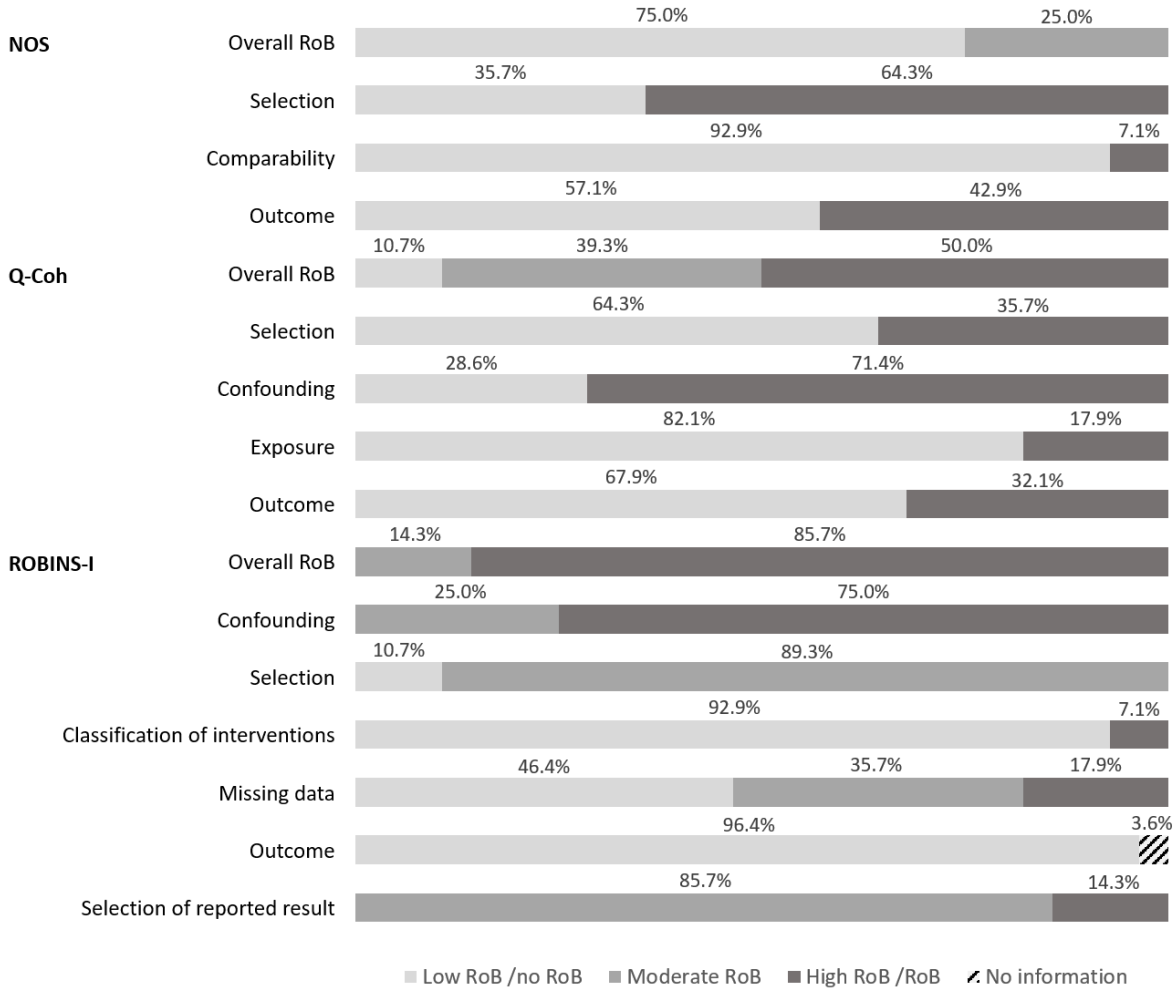
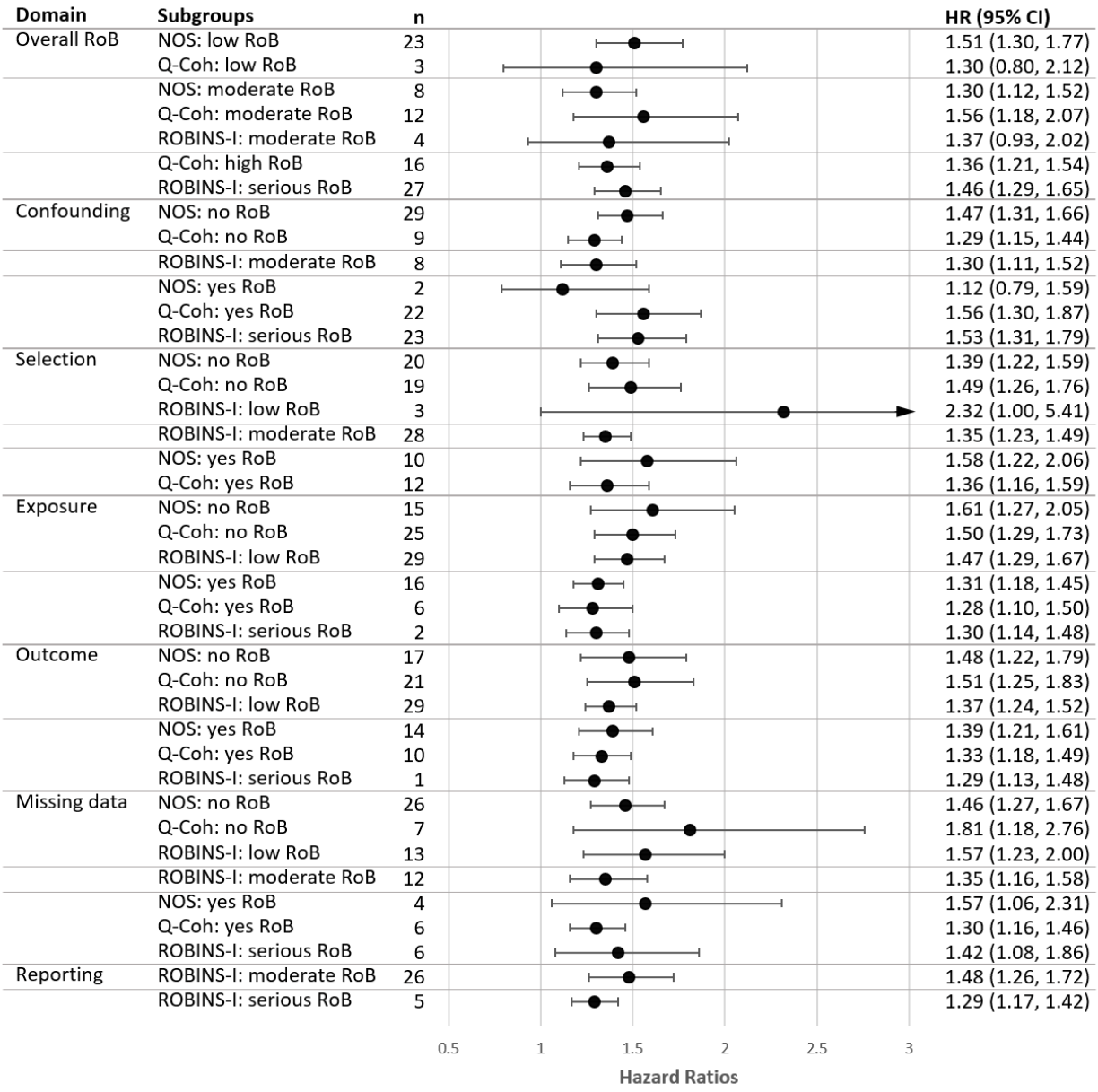


Figure 2. Forest plot of subgroup analyses results for the NOS, Q-Coh and ROBINS-I grouped by domain and RoB level.



# 3. DISCUSIÓN

A lo largo de los dos estudios que conforman esta tesis hemos mostrado que, a pesar de la existencia de una gran cantidad de herramientas para evaluar el RdS de los estudios no experimentales y de propuestas para incorporarlo en la síntesis de investigación, siguen existiendo importantes lagunas sobre qué herramienta y procedimiento son la mejor opción para llevar a cabo una RS o MA con las mejores garantías.

## 3.1. Evaluación e incorporación del riesgo de sesgo en psicología sanitaria

Según los resultados de nuestro primer estudio, la evaluación e incorporación del RdS sigue siendo una práctica poco habitual en la síntesis de investigación en el ámbito de la psicología sanitaria. A pesar de que el 51% de las 90 revisiones evaluadas llevaron a cabo una evaluación del RdS, tan sólo un 34% del total utilizó una herramienta estándar y únicamente un 11%, además, relacionó de algún modo los resultados de esta evaluación con las conclusiones de la revisión. Aunque en estos últimos años parece apreciarse un incremento de las revisiones que evalúan el RdS de los estudios primarios, sigue siendo infrecuente la utilización de esta evaluación para explicar los resultados de la revisión, más allá de alguna mención en el apartado de discusión de los artículos.

## 3.2. Evaluación del riesgo de sesgo de estudios no experimentales

Por otra parte, como ya se ha mencionado, a pesar de existir multitud de herramientas para la evaluación del RdS de estudios no experimentales, la gran mayoría de ellas no presenta garantías de validez ni de fiabilidad (Deeks et al., 2003; Jarde et al., 2012a; Sanderson et al., 2007). En cuanto al tipo de medida del RdS a utilizar, aunque las herramientas basadas en dominios parecen ofrecer mejores atributos y propiedades (O'Connor et al., 2015), es

necesaria una importante mejora en la fiabilidad entre evaluadores, lo que parece depender tanto de unas directrices claras en la aplicación de las herramientas (Faggion, 2016; Hartling et al., 2012), como de un reporte exhaustivo de los estudios primarios.

Respecto a lo primero, creemos que para que una herramienta destinada a los estudios no experimentales consiga un amplio consenso, además de seguir un riguroso proceso de elaboración de acuerdo con estándares psicométricos, debe cumplir otros requisitos destinados a mejorar su validez, fiabilidad y aplicabilidad:

- Que tenga en consideración las principales amenazas específicas a la validez de los diseños no experimentales.
- Que sea una herramienta basada en dominios de sesgo y no en puntuaciones globales.
- Que sea capaz de discriminar el nivel de RdS de los estudios evaluados.
- Que la fórmula para su aplicación sea clara y sencilla, también para investigadores no expertos en metodología.
- Que sea aplicable a los diseños no experimentales más habituales.
- Que ofrezca una fase previa de adiestramiento para mayor seguridad en su aplicación.
- Que su tiempo de aplicación sea razonable.

Como se ha señalado previamente, esperamos que el trabajo realizado en el nuevo Q-Coh, nos permita cumplir con estos requisitos. También tenemos constancia de que la Colaboración Cochrane se encuentra trabajando en una herramienta específica para evaluar el RdS de los estudios no aleatorizados en los que no se aplica ninguna intervención (ROBINS-E; Morgan et al., 2017). Aunque se trata de una propuesta aún en desarrollo, en su versión preliminar no parece que vaya a aportar demasiadas novedades respecto a ROBINS-I, al menos en lo que respecta a la complejidad de su aplicación.

En cuanto a las deficiencias que presenta el reporte de los estudios primarios son, posiblemente, una de las principales causas de la baja fiabilidad entre evaluadores en la aplicación de herramientas para valorar el RdS (Armijo-Olivo et al., 2012; Hartling et al., 2009; Moher et al., 1998). La extracción de datos para la evaluación del RdS puede resultar especialmente complicada cuando el reporte es pobre y, además, se aplican herramientas que requieren un análisis minucioso de los estudios primarios (e.g. Q-Coh, ROBINS-I).

Algunas de las medidas apuntadas por la literatura para la mejora de reporte y publicación de los estudios son fácilmente aplicables, como el registro de protocolos y la utilización de guías de reporte por parte de los autores (e.g. CONSORT, STROBE, PRISMA), o bien como propuestas de las publicaciones para poder discriminar claramente los trabajos con un buen reporte (e.g. BJU International). En cuanto al proceso de publicación, hay que pedir valentía a los editores para publicar más estudios independientes y más estudios con resultados negativos, que a menudo tienen menos posibilidades de ser publicados (Hopewell et al., 2009), pero que también constituyen evidencia. También animar a esos mismos editores a promover y publicar la replicación de estudios (Ioannidis, 2016b), fundamental en ciencia y muy necesaria para que la síntesis de investigación pueda basarse en estudios comparables, reduciendo la diversidad en los resultados finales.

## 3.3. Incorporación del riesgo de sesgo en la síntesis de investigación

Después de llevar a cabo nuestro primer estudio, a pesar de existir varias propuestas para la incorporación del RdS en la síntesis de investigación, tan sólo pudimos recomendar la aplicación de los análisis de subgrupos y las meta-regresiones, y únicamente cuando la potencia estadística fuera la adecuada. Los resultados de la aplicación de estas técnicas en nuestro segundo estudio no mostraron una asociación clara entre el RdS y los tamaños de efecto combinados en un MA. Estas conclusiones fueron las mismas para todas las herramientas aplicadas. Aunque el alcance de nuestros análisis publicados es muy limitado para afirmar que no existe relación, lo cierto es que en la literatura los resultados inconsistentes son habituales (Ahn & Becker, 2011; Conn & Rantz, 2003). A destacar, que a lo largo del desarrollo de esta tesis hemos llevado a cabo análisis similares, utilizando diversos MAs y diferentes herramientas, pero obteniendo siempre resultados no significativos (Oliveras et al., 2013; Oliveras, Losilla, & Vives, 2015). Es posible que, si existe una asociación entre RdS y tamaño del efecto, esta no se manifieste a causa de la poca variabilidad en el RdS de los estudios primarios, especialmente cuando en la evaluación del RdS de los estudios no experimentales se ha tomado como referente un

estudio experimental, como hemos constatado en la aplicación de la herramienta ROBINS-I (Losilla, Oliveras, Marin-Garcia, & Vives, 2018).

## 3.4. Líneas de investigación actuales

Además de los esfuerzos por promover mejoras en las herramientas de evaluación del RdS y en el reporte de los estudios primarios, en la actualidad existen interesantes líneas de investigación que van un nivel más allá y que destinan sus esfuerzos a la mejora de la síntesis de investigación y, en general, de la PBE. Creemos que vale la pena revisar algunas de estas propuestas y tenerlas en consideración a la hora de diseñar futuros proyectos.

### *3.4.1. Síntesis de investigación: revisiones sistemáticas vivas*

En relación a la mejora de síntesis de investigación, uno de los problemas que ya apuntábamos en la introducción, es la producción masiva y sin sentido de RSs y MAs (Ioannidis, 2016b). El desperdicio de recursos de investigación y, por tanto, de su financiación, es un tema que preocupa enormemente y que está a la orden del día (Chalmers & Glasziou, 2009; Moher et al., 2016). Aunque existen iniciativas para buscar un mayor aprovechamiento de los recursos de investigación (The Lancet Series "increasing value, reducing waste"; Al-Shahi Salman et al., 2014; Chalmers et al., 2014; Chan et al., 2014; Ioannidis et al., 2014), la implementación de mejoras se prevé larga y compleja, pues requiere de la implicación de todas las partes interesadas: financiadores, publicaciones, instituciones académicas e investigadores (Moher et al., 2016).

Para solventar la redundancia de RSs y MAs, Page y Moher (2016) en sus comentarios sobre el artículo de Ioannidis (2016b), proponen un modelo de "revisiones sistemáticas vivas" ("living systematic reviews", en inglés; Elliott et al., 2014) o de una evolución del mismo, los "metaanálisis en red acumulativos vivos" ("living cummulative network meta-analyses", en inglés; Créquit, Trinquart, Yavchitz, & Ravaud, 2016). Según la definición de Elliott et al. (2014), las revisiones sistemáticas vivas son resúmenes online actualizados y de alta calidad sobre investigación en salud, que se actualizan a medida que hay

disponibilidad de nuevas investigaciones. Se trata de una aproximación novedosa a las tradicionales RSs, que aprovecha las oportunidades que ofrecen las nuevas tecnologías y en las que el propósito es que las revisiones se mantengan siempre actualizadas sin perder rigor metodológico para conseguir síntesis de evidencia confiable (Elliott et al., 2017). A través de una serie de cuatro artículos en el Journal of Clinical Epidemiology (Akl, Meerpohl, et al., 2017; Elliott et al., 2017; Simmonds et al., 2017; Thomas et al., 2017), los investigadores que conforman Living Systematic Review Network proporcionan orientación para llevar a cabo este tipo de revisiones, que ya se están probando en diferentes proyectos Cochrane y no Cochrane (Akl et al., 2017; Cnossen et al., 2016; Hodder et al., 2018; Rahal, Badgett, & Hoffman, 2016).

### *3.4.2. Directrices para evaluar la calidad de la evidencia: GRADE y PRECEPT*

En cuanto al RdS, parece que por sí solo, sin contextualizar ni relacionar con los resultados de la síntesis de investigación, tiene poco sentido a la hora de dar el paso de la investigación a la práctica. La evaluación del RdS, aunque esencial, es tan solo uno de los elementos del concepto de evidencia científica y del proceso de toma de decisiones en la EBP. El sistema GRADE (Guyatt et al., 2008), además de proporcionar una guía para el desarrollo de directrices de intervención en salud, permite contextualizar el RdS de los estudios que forman parte de una síntesis de investigación en un marco más amplio, como es la calidad de un determinado cuerpo de evidencia. Este marco, además del RdS, tiene en cuenta otros elementos distintos que son cruciales a la hora de clasificar la evidencia, reduciendo su grado de credibilidad (riesgo de sesgo, inconsistencia, evidencia indirecta, imprecisión y sesgo de publicación) o bien aumentándolo (efecto de gran magnitud, gradiente dosis-respuesta y efecto de los potenciales factores de confusión residual).

No obstante, el sistema GRADE está diseñado para aplicarse, básicamente, en el ámbito de las intervenciones biomédicas, por lo que existen determinados problemas a la hora de aplicarlo en intervenciones complejas (Durrheim & Reingold, 2010; Rehfuess & Akl, 2013), que en muchas ocasiones se basan en diseños no experimentales, y que son muy frecuentes en ámbitos como la psicología, la educación o la salud pública (Craig et al., 2008; Movsisyan, Melendez-Torres, & Montgomery, 2016a). Uno de los principales

problemas que algunos autores apuntan (Movsisyan et al., 2016a; Rehfuess & Akl, 2013) en cuanto a la aplicación de GRADE en estos ámbitos es, precisamente, algo que hemos señalado también respecto a ROBINS-I en el segundo estudio: su incapacidad para discriminar adecuadamente la calidad de la evidencia cuando se dispone de resultados de diseños de investigación no experimental. Para intentar sortear estas dificultades asociadas a la aplicación de GRADE en los ámbitos señalados, surge el proyecto GRADE Guidance for Complex Social Interventions, que tiene como objetivo la adaptación de GRADE a las particularidades de las intervenciones complejas. Aunque esta línea de investigación es aún incipiente, el proyecto cuenta con tres artículos publicados (Movsisyan, Dennis, Rehfuess, Grant, & Montgomery, 2018; Movsisyan et al., 2016a; Movsisyan, Melendez-Torres, & Montgomery, 2016b) en que se empiezan a establecer las bases para la adaptación de GRADE a intervenciones complejas.

Por otra parte, el uso de jerarquías basadas únicamente en el diseño no tiene en cuenta cuál es el diseño de investigación más apropiado en función de cuál es la pregunta de investigación (Glasziou et al., 2004). Como se ha comentado previamente, es importante recordar que los estudios no experimentales son, en muchas ocasiones, la mejor o la única forma de responder a determinadas preguntas de investigación, como la prevalencia, incidencia, etiología o pronóstico de determinadas patologías o condiciones (Mann, 2003). En consecuencia, parece mucho más lógico establecer cuál debe ser el diseño de referencia en cada caso para cada pregunta de investigación.

En esta línea, Harder et al. (2014) propone un "enfoque basado en preguntas" con el propósito de definir una metodología para evaluar y clasificar la evidencia y solidez de las recomendaciones en el área de la salud pública y, particularmente, en la epidemiología, prevención y control de las enfermedades infecciosas. Este proyecto, llamado PRECEPT (Project on a Framework for Rating Evidence in Public Health), parte de una selección de preguntas de investigación relevantes para identificar los diseños de estudio más apropiados para responder a estas preguntas para, finalmente, identificar las herramientas de evaluación de RdS más apropiadas para cada diseño (Harder et al., 2014). Esta aproximación propone utilizar el sistema GRADE como uno de sus componentes clave (Harder et al., 2015), ofreciendo directrices metodológicas y sugiriendo adaptaciones para aplicarlo a cuatro dominios relevantes en el área de la salud pública y las enfermedades infecciosas: carga de la enfermedad (significancia del problema), factores de riesgo (causas

del problema), diagnósticos (detección del problema) e intervenciones (consecuencias de las acciones para solventar el problema) (Harder et al., 2017). Finalmente, esta propuesta define cuatro pasos para evaluar la evidencia: identificar las preguntas relevantes, llevar a cabo la RS, aplicar el sistema de clasificación de la evidencia y documentar los resultados y, finalmente, preparar un resumen narrativo de la evidencia (Harder et al., 2017).

### *3.4.3. Más allá de la práctica basada en la evidencia: la ciencia de la implementación*

A pesar de las mejoras propuestas a todos los niveles, la realidad es que sigue existiendo divorcio entre investigación y práctica. Son muchos los factores que pueden impedir la adopción de la PBE en la práctica diaria, incluyendo las diversas exigencias a los profesionales que trabajan en primera línea, la falta de conocimientos, habilidades y recursos, así como el desajuste entre la evidencia de la investigación y las prioridades operativas en el ámbito de la salud (Bauer, Damschroder, Hagedorn, Smith, & Kilbourne, 2015). Intentar superar estos retos a la hora de promover la PBE ha favorecido el auge que está experimentando la ciencia o investigación de la implementación ("implementation science/research", en inglés), una perspectiva procedente de varias disciplinas y tradiciones de investigación que, aunque no es nueva, se ha abierto paso en esta última década (Peters, Adam, Alonge, Agyepong, & Tran, 2013).

La investigación de la implementación, definida por Eccles y Mittman (2006) es "el estudio científico de métodos para promover la adopción sistemática de los resultados de investigación y otras prácticas basadas en la evidencia de la práctica habitual, y, por consiguiente, para mejorar la calidad y la eficacia de los servicios de salud". La investigación de la implementación intenta entender y trabajar en condiciones reales en lugar de intentar controlar estas condiciones, así como explicar qué intervenciones funcionan en el "mundo real", por qué funcionan y cómo funcionan (Peters et al., 2013). El contexto es un elemento clave para la investigación de la implementación, y puede incluir el entorno social, cultural, económico, político, legal y físico, así como el entorno institucional que, a su vez, comprende a varias partes interesadas y sus interacciones, así como las condiciones demográficas y epidemiológicas (Peters et al., 2013).

A pesar de que el desarrollo de la ciencia de la implementación es reciente, cuenta ya con algunos programas implementados con éxito (Glasgow et al., 2012), así como como con publicaciones específicas como *Implementation Science* o *BMC Health Services Research*.

# 4. CONCLUSIONES

Los dos artículos publicados ponen de manifiesto la urgente necesidad de mejora en varios ámbitos si se quiere conseguir que la PBE consiga su propósito, especialmente en lo que respecta a la investigación no experimental. En primer lugar, es imprescindible mejorar tanto las herramientas para evaluar el RdS de los estudios no experimentales, como el reporte de los estudios primarios. Estos dos factores tienen una influencia crítica sobre la síntesis de investigación y sobre los procedimientos de incorporación del RdS en esta síntesis, pues creemos que no es posible determinar si hay alguna relación entre el RdS y los tamaños de efecto si no contamos con herramientas válidas y fiables que nos proporcionen unos resultados creíbles.

El propósito final de la línea de investigación en la que se enmarca esta tesis intenta dar respuesta a la primera de las necesidades: la mejora de las herramientas de evaluación. Los dos estudios incluidos en la tesis han aportado información crítica para la consecución de este propósito, que se halla en una fase muy avanzada. Una vez conseguido este propósito, es absolutamente necesario establecer procedimientos sólidos, bien definidos y replicables para la incorporación del RdS en la síntesis de investigación, así como para la clasificación de la evidencia. Esto mejoraría sustancialmente las decisiones tomadas de acuerdo con la PBE.

Mientras se llevan a cabo las mejoras necesarias y se avanza en nuevas líneas de investigación relacionadas, las recomendaciones van en el sentido de utilizar la evaluación del RdS como:

- base para establecer el grado de credibilidad de los resultados de RS y MA;
- base para la identificación de las limitaciones metodológicas y las principales fuentes de sesgo de un área determinada y, de este modo, contribuir a la mejora de la calidad de la investigación en ese ámbito;
- parte del procedimiento de evaluación de la calidad de la evidencia.

Una vez contemos con una buena herramienta de evaluación del RdS, se abren prometedoras posibilidades para futuras líneas de investigación que vayan un paso más allá de la evaluación del RdS de los estudios primarios, algunas relacionadas con las líneas de

investigación mencionadas en la discusión. En primer lugar, el estudio de los procedimientos metodológicos y/o estadísticos para la incorporación del RdS en RSs y MAs dista mucho de estar agotado. El diseño y elaboración de nuevas estrategias que vayan más allá de las existentes en la actualidad debería permitir, en algún momento, establecer relaciones más consistentes entre RdS y resultados del MA, lo que se traduciría en un acercamiento más fiable de los resultados a los efectos reales de una investigación. En segundo lugar, las intervenciones complejas, tan frecuentes en el ámbito de la psicología, abren varias posibilidades como la adaptación del sistema GRADE para su aplicación en psicología sanitaria, o bien la creación de un nuevo marco de clasificación de la evidencia adaptado a este ámbito, así como su aplicación en un entorno real tomando en consideración, por ejemplo, el uso de medidas ecológicas momentáneas. Por último, otra línea que parece muy prometedora es la de las revisiones sistemáticas vivas o los metaanálisis en red acumulativos vivos, que deberían permitir el encaje de todas las piezas que componen las intervenciones complejas, al mismo tiempo que favorecerían un enfoque multi-disciplinario y de investigación colaborativa e independiente, tan necesaria para evitar interferencias de intereses creados en la investigación, así como el sinsentido de la producción masiva de RSs y MAs poco creíbles. Todas estas posibles líneas de investigación contribuirían muy positivamente a que la implementación de la PBE fuera, por fin, una realidad.

# 5. REFERENCIAS

Ahn, S., & Becker, B. J. (2011). Incorporating Quality Scores in Meta-Analysis. *Journal of Educational and Behavioral Statistics*, *36*, 555–585. http://doi.org/10.3102/1076998610393968

Akl, E. A., Kahale, L. A., Hakoum, M. B., Matar, C. F., Sperati, F., Barba, M., … Schünemann, H. (2017). Parenteral anticoagulation in ambulatory patients with cancer. *Cochrane Database of Systematic Reviews*, (9). http://doi.org/10.1002/14651858.CD006652.pub5

Akl, E. A., Meerpohl, J. J., Elliott, J., Kahale, L. A., Schünemann, H. J., Agoritsas, T., … Pearson, L. (2017). Living systematic reviews: 4. Living guideline recommendations. *Journal of Clinical Epidemiology*, *91*, 47–53. http://doi.org/10.1016/j.jclinepi.2017.08.009

Al-Shahi Salman, R., Beller, E., Kagan, J., Hemminki, E., Phillips, R. S., Savulescu, J., … Chalmers, I. (2014). Increasing value and reducing waste in biomedical research regulation and management. *The Lancet*, *383*(9912), 176–185. http://doi.org/10.1016/S0140-6736(13)62297-7

Al Khalaf, M. M., Thalib, L., & Doi, S. A. R. (2011). Combining heterogenous studies using the random-effects model is a mistake and leads to inconclusive meta-analyses. *Journal of Clinical Epidemiology*, *64*(2), 119–23. http://doi.org/10.1016/j.jclinepi.2010.01.009

APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist*, *61*(4), 271–285. http://doi.org/10.1037/0003-066X.61.4.271

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: why do we need them? What might they be? *The American Psychologist*, *63*(9), 839–51. http://doi.org/10.1037/0003-066X.63.9.839

Armijo-Olivo, S., Stiles, C. R., Hagen, N. A., Biondo, P. D., & Cummings, G. G. (2012). Assessment of study quality for systematic reviews: a comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: methodological research. *Journal of Evaluation in Clinical Practice*, *18*(1), 12–8. http://doi.org/10.1111/j.1365-2753.2010.01516.x

Balshem, H., Helfand, M., Schünemann, H. J., Oxman, A. D., Kunz, R., Brozek, J., … Guyatt, G. H. (2011). GRADE guidelines: 3. Rating the quality of evidence. *Journal of Clinical Epidemiology*, *64*(4), 401–6. http://doi.org/10.1016/j.jclinepi.2010.07.015

Barendregt, J., & Doi, S. (2014). *MetaXL version 2.2*. Epigear. Retrieved from http://www.epigear.com/

Bauer, M. S., Damschroder, L., Hagedorn, H., Smith, J., & Kilbourne, A. M. (2015). An introduction to implementation science for the non-specialist. *BMC Psychology*, *3*(1), 1–12. http://doi.org/10.1186/S40359-015-0089-9

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). Comprehensive meta-analysis version 2. Englewood, NJ: Biostat.

Botella, J., & Sánchez-Meca, J. (2015). *Meta-análisis en ciencias sociales y de la salud*. Madrid: Editorial Síntesis.

Centre for Reviews and Dissemination. (2009). *Systematic reviews : CRD's guidance for undertaking reviews in health care*. CRD, University of York. Retrieved from https://www.york.ac.uk/crd/guidance/

Chalmers, I., Bracken, M. B., Djulbegovic, B., Garattini, S., Grant, J., Gülmezoglu, A. M., … Oliver, S. (2014). How to increase value and reduce waste when research priorities are set. *The Lancet*, *383*(9912), 156–165. http://doi.org/10.1016/S0140-6736(13)62229-1

Chalmers, I., & Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *The Lancet*, *374*(9683), 86–89. http://doi.org/10.1016/S0140-6736(09)60329-9

Chalmers, I., Hedges, L., & Cooper, H. (2002). A brief history of research synthesis. *Evaluation & the Health …*, *25*(1), 12–37. http://doi.org/10.1177/0163278702025001003

Chan, A. W., Song, F., Vickers, A., Jefferson, T., Dickersin, K., Gøtzsche, P. C., … Van Der Worp, H. B. (2014). Increasing value and reducing waste: Addressing inaccessible research. *The Lancet*, *383*(9913), 257–266. http://doi.org/10.1016/S0140-6736(13)62296-5

Cnossen, M. C., Scholten, A. C., Lingsma, H. F., Synnot, A., Tavender, E., Gantner, D., … Polinder, S. (2016). Adherence to Guidelines in Adult Patients with Traumatic Brain Injury: A Living Systematic Review. *Journal of Neurotrauma*, neu.2015.4121. http://doi.org/10.1089/neu.2015.4121

Colle, F., Rannou, F., Revel, M., Fermanian, J., & Poiraudeau, S. (2002). Impact of quality scales on levels of evidence inferred from a systematic review of exercise therapy and low back pain. *Archives of Physical Medicine and Rehabilitation*, *83*(12), 1745–1752. http://doi.org/10.1053/apmr.2002.35657

Conn, V. S., & Rantz, M. J. (2003). Focus on research methods: Research methods: Managing primary study quality in meta-analyses. *Research in Nursing & Health*, *26*, 322–333. http://doi.org/10.1002/nur.10092

Cornell, J. E., Mulrow, C. D., Localio, R., Stack, C. B., Meibohm, A. R., Guallar, E., & Goodman, S. N. (2014). Random-effects meta-analysis of inconsistent effects: A time for change. *Annals of Internal Medicine*, *160*(4), 267–270. http://doi.org/10.7326/M13-2886

Craig, P., Dieppe, P., Macintyre, S., Mitchie, S., Nazareth, I., & Petticrew, M. (2008). Developing and evaluating complex interventions: The new Medical Research Council guidance. *Bmj*, *337*(7676), 979–983. http://doi.org/10.1136/bmj.a1655

Créquit, P., Trinquart, L., Yavchitz, A., & Ravaud, P. (2016). Wasted research when systematic reviews fail to provide a complete and up-to-date evidence synthesis: The example of lung cancer. *BMC Medicine*, *14*(1). http://doi.org/10.1186/s12916-016-0555-0

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, a J., Sakarovitch, C., Song, F., … Altman, D. G. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment (Winchester, England)*, *7*(27), iii–x, 1-173.

DerSimonian, R., & Laird, N. (1986). Metaanalysis in Clinical-Trials. *Controlled Clinical*

*Trials*, *7*(3), 177–188. http://doi.org/10.1016/0197-2456(86)90046-2

Detsky, A. S., Naylor, C. D., O'Rourke, K., McGeer, A. J., & L'Abbe, K. A. (1992). Incorporating Variations in the Quality of Individual Randomized Trials into Metaanalysis. *Journal of Clinical Epidemiology*, *45*(3), 255–265. http://doi.org/10.1016/0895-4356(92)90085-2

Doi, S. A. R. (2015). Meta-analysis and the problem of inconsistent effects. *International Journal of Evidence-Based Healthcare*. http://doi.org/10.1097/XEB.0000000000000058

Doi, S. A. R., Barendregt, J. J., Khan, S., Thalib, L., & Williams, G. M. (2015a). Advances in the Meta-analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model. *Contemporary Clinical Trials*. http://doi.org/10.1016/j.cct.2015.05.009

Doi, S. A. R., Barendregt, J. J., Khan, S., Thalib, L., & Williams, G. M. (2015b). Advances in the meta-analysis of heterogeneous clinical trials II: The quality effects model. *Contemporary Clinical Trials*, *45*(A), 123–129. http://doi.org/10.1016/j.cct.2015.05.010

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health*, *52*(6), 377–84.

Durrheim, D. N., & Reingold, A. (2010). Modifying the GRADE framework could benefit public health. *Journal of Epidemiology and Community Health*, *64*(5), 387. http://doi.org/10.1136/jech.2009.103226

Eccles, M. P., & Mittman, B. S. (2006). Welcome to implementation science. *Implementation Science*, *1*(1), 1–3. http://doi.org/10.1186/1748-5908-1-1

Elliott, J. H., Synnot, A., Turner, T., Simmonds, M., Akl, E. A., McDonald, S., … Pearson, L. (2017). Living systematic review: 1. Introduction—the why, what, when, and how. *Journal of Clinical Epidemiology*, *91*, 23–30. http://doi.org/10.1016/j.jclinepi.2017.08.010

Elliott, J. H., Turner, T., Clavisi, O., Thomas, J., Higgins, J. P. T., Mavergames, C., &

Gruen, R. L. (2014). Living Systematic Reviews: An Emerging Opportunity to Narrow the Evidence-Practice Gap. *PLoS Medicine*, *11*(2), 1–6. http://doi.org/10.1371/journal.pmed.1001603

Every-Palmer, S., & Howick, J. (2014). How evidence-based medicine is failing due to biased trials and selective publication. *Journal of Evaluation in Clinical Practice*, *20*(6), 908–914. http://doi.org/10.1111/jep.12147

Faggion, C. M. (2016). The rationale for rating risk of bias should be fully reported. *Journal of Clinical Epidemiology*. http://doi.org/10.1016/j.jclinepi.2016.03.007

Funai, E. F., Rosenbush, E. J., Lee, M. J., & Del Priore, G. (2001). Distribution of study designs in four major US journals of obstetrics and gynecology. *Gynecologic and Obstetric Investigation*, *51*(1), 8–11. http://doi.org/10.1159/000052882

Glasgow, R. E., Vinson, C., Chambers, D., Khoury, M. J., Kaplan, R. M., & Hunter, C. (2012). National institutes of health approaches to dissemination and implementation science: Current and future directions. *American Journal of Public Health*, *102*(7), 1274–1281. http://doi.org/10.2105/AJPH.2012.300755

Glass, G. V. (1976). Primary, Secondary, and Meta-Analysis of Research'. *American Educational Research Association*, *5*(10), 3–8. http://doi.org/10.3102/0013189X005010003

Glasziou, P., Vandenbroucke, J. P., & Chalmers, I. (2004). Assessing the quality of research. *British Medical Journal*, *328*(7430), 39–41. http://doi.org/10.1136/bmj.328.7430.39

Greenhalgh, T., Howick, J., Maskrey, N., Brassey, J., Burch, D., Burton, M., … Spence, D. (2014). Evidence based medicine: A movement in crisis? *BMJ (Online)*, *348*(June), 1–7. http://doi.org/10.1136/bmj.g3725

Greenland, S. (2005). Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society Series A-Statistics in Society*, *168*(2), 267–291. http://doi.org/10.1111/j.1467-985X.2004.00349.x

Greenland, S., & O'Rourke, K. (2001). On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics*, *2*(4), 463–471. http://doi.org/10.1093/biostatistics/2.4.463

Gurol-Urganci, I., Bou-Antoun, S., Lim, C. P., Cromwell, D. a, Mahmood, T. a, Templeton, A., & van der Meulen, J. H. (2013). Impact of Caesarean section on subsequent fertility: a systematic review and meta-analysis. *Human Reproduction (Oxford, England)*, *28*(7), 1943–52. http://doi.org/10.1093/humrep/det130

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ (Clinical Research Ed.)*, *336*(7650), 924–6. http://doi.org/10.1136/bmj.39489.470347.AD

Harder, T., Abu Sin, M., Bosch-Capblanch, X., Bruno Coignard, de Carvalho Gomes, H., Duclos, P., … Zuiderent-Jerak, T. (2015). Towards a framework for evaluating and grading evidence in public health. *Health Policy*, *119*(6), 732–736. http://doi.org/10.1016/j.healthpol.2015.02.010

Harder, T., Takla, A., Eckmanns, T., Ellis, S., Forland, F., James, R., … Wichmann, O. (2017). PRECEPT: an evidence assessment framework for infectious disease epidemiology, prevention and control. *Euro Surveillance : Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin*, *22*(40), 1–10. http://doi.org/10.2807/1560-7917.ES.2017.22.40.16-00620

Harder, T., Takla, A., Rehfuess, E., Sánchez-Vivar, A., Matysiak-Klose, D., Eckmanns, T., … Wichmann, O. (2014). Evidence-based decision-making in infectious diseases epidemiology, prevention and control: Matching research questions to study designs and quality appraisal tools. *BMC Medical Research Methodology*, *14*(1). http://doi.org/10.1186/1471-2288-14-69

Hartling, L., Hamm, M., Milne, A., Vandermeer, B., Santaguida, P. L., Ansari, M., … Dryden, D. M. (2012). *Validity and Inter-rater Reliability Testing of Quality Assessment Instruments*. *Validity and Inter-Rater Reliability Testing of Quality Assessment Instruments*. Agency for Healthcare Research and Quality (US). Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22536612

Hartling, L., Ospina, M., Liang, Y., Dryden, D. M., Hooton, N., Krebs Seida, J., & Klassen, T. P. (2009). Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. *Bmj*, *339*(oct19 1), b4012–b4012. http://doi.org/10.1136/bmj.b4012

Herbison, P., Hay-Smith, J., & Gillespie, W. J. (2006). Adjustment of meta-analyses on the basis of quality scores should be abandoned. *Journal of Clinical Epidemiology*, *59*, 1249–1256. http://doi.org/10.1016/j.jclinepi.2006.03.008

Higgins, J., & Green, S. (2011). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Retrieved from http://training.cochrane.org/handbook

Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., … Sterne, J. A. C. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ: British Medical Journal*, *343*, 1–9. Retrieved from http://www.bmj.com/content/343/bmj.d5928

Hilgard, E. R., Kelly, E. L., Luckey, B., Sanford, R. N., Shaffer, L. F., & Shakow, D. (1947). Recommended Graduate Training Program in Clinical Psychology. *American Psychologist*, *2*(9), 539–558. http://doi.org/10.1037/h0058236

Hodder, R. K., O'Brien, K. M., Stacey, F. G., Wyse, R. J., Clinton-McHarg, T., Tzelepis, F., … Wolfenden, L. (2018). Interventions for increasing fruit and vegetable consumption in children aged five years and under. *Cochrane Database of Systematic Reviews*, (5). http://doi.org/10.1002/14651858.CD008552.pub5

Hootman, J. M., Driban, J. B., Sitler, M. R., Harris, K. P., & Cattano, N. M. (2011). Reliability and validity of three quality rating instruments for systematic reviews of observational studies. *Research Synthesis Methods*, *2*(2), 110–118. http://doi.org/10.1002/jrsm.41

Hopewell, S., Loudon, K., Clarke, M. J., Oxman, A. D., & Dickersin, K. (2009). Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database of Systematic Reviews*, (1). http://doi.org/10.1002/14651858.MR000006.pub3

Ijaz, S., Verbeek, J. H., Mischke, C., & Ruotsalainen, J. (2014, June). Inclusion of nonrandomized studies in Cochrane systematic reviews was found to be in need of improvement. *Journal of Clinical Epidemiology*. http://doi.org/10.1016/j.jclinepi.2014.01.001

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS*

*Medicine*, *2*(8), 0696–0701. http://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2011). Commentary: Adjusting for bias: a user's guide to performing plastic surgery on meta-analyses of observational studies. *International Journal of Epidemiology*, *40*(3), 777–779. http://doi.org/10.1093/ije/dyq265

Ioannidis, J. P. A. (2016a). Evidence-based medicine has been hijacked: A report to David Sackett. *Journal of Clinical Epidemiology*, *73*, 82–86. http://doi.org/10.1016/j.jclinepi.2016.02.012

Ioannidis, J. P. A. (2016b, September 1). The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. *Milbank Quarterly*. http://doi.org/10.1111/1468-0009.12210

Ioannidis, J. P. A., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., … Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. *The Lancet*, *383*(9912), 166–175. http://doi.org/10.1016/S0140-6736(13)62227-8

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*(5), 235–241. http://doi.org/10.1016/j.tics.2014.02.010

Jarde, A., Losilla, J. M., & Vives, J. (2012a). Methodological quality assessment tools of non-experimental studies: a systematic review. *Anales De Psicologia*, *28*, 617–628. http://doi.org/10.6018/analesps.28.2.148911

Jarde, A., Losilla, J. M., & Vives, J. (2012b). Suitability of three different tools for the assessment of methodological quality in ex post facto studies. *International Journal of Clinical and Health Psychology*, *12*, 97–108.

Jarde, A., Losilla, J. M., Vives, J., & Rodrigo, M. F. (2013). Q-Coh: A tool to screen the methodological quality of cohort studies in systematic reviews and meta-analyses. *International Journal of Clinical and Health Psychology*, *13*, 138–146.

Jüni, P., Altman, D. G., & Egger, M. (2001). Systematic reviews in health care - Assessing the quality of controlled clinical trials. *British Medical Journal*, *323*(July), 42–46. http://doi.org/10.1136/bmj.323.7303.42

Jüni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *Jama-Journal of the American Medical Association*, *282*(11), 1054–1060. http://doi.org/10.1001/jama.282.11.1054

Linde, K., Scholz, M., Ramirez, G., Clausius, N., Melchart, D., & Jonas, W. B. (1999). Impact of study quality on outcome in placebo-controlled trials of homeopathy. *Journal of Clinical Epidemiology*, *52*, 631–636. http://doi.org/10.1016/S0895-4356(99)00048-7

Littell, J. H., Corcoran, J., & Pillai, V. K. (2008). *Systematic reviews and meta-analysis*. Oxford: Oxford University Press.

Losilla, J.-M., Oliveras, I., Marin-Garcia, J. A., & Vives, J. (2018). Three risk of bias tools lead to opposite conclusions in observational research synthesis. *Journal of Clinical Epidemiology*, *101*, 61–72. http://doi.org/10.1016/j.jclinepi.2018.05.021

Manchikanti, L., Benyamin, R. M., Helm, S., & Hirsch, J. A. (2009). Evidence-based medicine, systematic reviews, and guidelines in interventional pain management: part 3: systematic reviews and meta-analyses of randomized trials. *Pain Physician*, *12*, 35–72.

Mann, C. J. (2003). Observational research methods. Research design II: cohort, cross sectional, and case-control studies. *Emergency Medicine Journal*, *20*(1), 54–60. http://doi.org/10.1136/emj.20.1.54

Manterola, C., & Otzen, T. (2017). Checklist for Reporting Results Using Observational Descriptive\r\nStudies as Research Designs. The MInCir Initiative\r\n. *International Journal of Morphology*, *35*(1), 72–76. http://doi.org/10.4067/S0717-95022017000100013

Margulis, A. V, Pladevall, M., Riera-Guardia, N., Varas-Lorenzo, C., Hazell, L., Berkman, N. D., … Perez-Gutthann, S. (2014). Quality assessment of observational studies in a drug-safety systematic review, Comparison of two tools: The Newcastle-Ottawa scale and the RTI item bank. *Clinical Epidemiology*, *6*, 981–993. http://doi.org/10.2147/CLEP.S66677

Moher, D., Glasziou, P., Chalmers, I., Nasser, M., Bossuyt, P. M. M., Korevaar, D. A., … Boutron, I. (2016). Increasing value and reducing waste in biomedical research:

Who's listening? *The Lancet*, *387*(10027), 1573–1586. http://doi.org/10.1016/S0140-6736(15)00307-4

Moher, D., Pham, B., Jones, A., Cook, D. J., Jadad, A. R., Moher, M., … Deborah, J. (1998). Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet*, *352*, 609–613. http://doi.org/10.1016/S0140-6736(98)01085-X

Morgan, R., Sterne, J., Higgins, J., Thayer, K., Schunemann, H., Rooney, A., & Taylor, K. (2017). A new instrument to assess Risk of Bias in Non-randomised Studies of Exposures (ROBINS-E): Application to studies of environmental exposure. In *Abstracts of the Global Evidence Summit*. Cape Town: Cochrane Database of Systematic Reviews. http://doi.org/10.1002/14651858.CD201702

Movsisyan, A., Dennis, J., Rehfuess, E., Grant, S., & Montgomery, P. (2018). Rating the quality of a body of evidence on the effectiveness of health and social interventions: A systematic review and mapping of evidence domains. *Research Synthesis Methods*, *9*(2), 224–242. http://doi.org/10.1002/jrsm.1290

Movsisyan, A., Melendez-Torres, G. J., & Montgomery, P. (2016a). Outcomes in systematic reviews of complex interventions never reached "high" GRADE ratings when compared with those of simple interventions. *Journal of Clinical Epidemiology*, *78*, 22–33. http://doi.org/10.1016/j.jclinepi.2016.03.014

Movsisyan, A., Melendez-Torres, G. J., & Montgomery, P. (2016b). Users identified challenges in applying GRADE to complex interventions and suggested an extension to GRADE. *Journal of Clinical Epidemiology*, *70*, 191–199. http://doi.org/10.1016/j.jclinepi.2015.09.010

O'Connor, S. R., Tully, M. A., Ryan, B., Bradley, J. M., Baxter, G. D., & McDonough, S. M. (2015). Failure of a numerical quality assessment scale to identify potential risk of bias in a systematic review: a comparison study. *BMC Research Notes*, *8*(1), 224. http://doi.org/10.1186/s13104-015-1181-1

Oliveras, I., Jarde, A., Losilla, J.-M., & Vives, J. (2013). La incorporación en el meta-análisis de la valoración de la calidad metodológica de los estudios primarios. In *XIII Congreso de Metodología de las Ciencias Sociales y de la Salud* (pp. 90–91). San Cristóbal de La Laguna (España).

Oliveras, I., Losilla, J.-M., & Vives, J. (2015). Gestión de la calidad metodológica en la síntesis de investigación de estudios no experimentales. In *XIV Congreso de Metodología de las Ciencias Sociales y de la Salud*. Palma de Mallorca (España).

Oliveras, I., Losilla, J.-M., & Vives, J. (2017). Methodological quality is underrated in systematic reviews and meta-analyses in health psychology. *Journal of Clinical Epidemiology*, *86*, 59–70. http://doi.org/10.1016/j.jclinepi.2017.05.002

Page, M. J., & Moher, D. (2016). Mass Production of Systematic Reviews and Meta-analyses: An Exercise in Mega-silliness? *Milbank Quarterly*. http://doi.org/10.1111/1468-0009.12211

Page, M. J., Shamseer, L., Altman, D. G., Tetzlaff, J., Sampson, M., Tricco, A. C., … Moher, D. (2016). Epidemiology and Reporting Characteristics of Systematic Reviews of Biomedical Research: A Cross-Sectional Study. *PLoS Medicine*, *13*(5), e1002028. http://doi.org/10.1371/journal.pmed.1002028

Peters, D. H., Adam, T., Alonge, O., Agyepong, I. A., & Tran, N. (2013). Implementation research: what it is and how to do it. *BMJ (Clinical Research Ed.)*, *347*, f6753. http://doi.org/10.1136/BMJ.F6753

Primo, N. A., Gazzola, V. B., Primo, B. T., Tovo, M. F., & Faraco Junior, I. M. (2014). Bibliometric analysis of scientific articles published in Brazilian and international orthodontic journals over a 10-year period. *Dental Press Journal of Orthodontics*, *19*(2), 56–65. http://doi.org/10.1590/2176-9451.19.2.056-065.oar

Rahal, A. K., Badgett, R. G., & Hoffman, R. M. (2016). Screening coverage needed to reduce mortality from prostate cancer: A living systematic review. *PLoS ONE*, *11*(4), e0153417. http://doi.org/10.1371/journal.pone.0153417

Rehfuess, E. A., & Akl, E. A. (2013). Current experience with applying the GRADE approach to public health interventions: an empirical study. *BMC Public Health*, *13*, 9. http://doi.org/10.1186/1471-2458-13-9

*Review manager (RevMan)*. (2014). Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration.

Sackett, D. L., Rosenberg, W. M., Gray, J., Haynes, R. B., & Richardson, W. S. (1996). Evidence Based Medicine: What it is and what it isn't. *Bmj*, *312*(7023), 71–72.

http://doi.org/10.1136/bmj.312.7023.71

Salanti, G., & Ioannidis, J. P. A. (2009). Synthesis of observational studies should consider credibility ceilings. *Journal of Clinical Epidemiology*, *62*(2), 115–22. http://doi.org/10.1016/j.jclinepi.2008.05.014

Sánchez-Meca, J., Marín-Martínez, F., & López-López, J. A. (2011). Meta-análisis e intervención psicosocial basada en la evidencia. *Psychosocial Intervention*, *20*, 95–107. http://doi.org/10.5093/in2011v20n1a9

Sánchez Meca, J., & Botella, J. (2010). Revisiones sistemáticas y meta-análisis: Herramientas para la práctica profesional. *Papeles Del Psicólogo: Revista Del Colegio Oficial de Psicólogos*, *31*(1), 7–17.

Sanderson, S., Tatt, I. D., & Higgins, J. P. T. (2007). Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *International Journal of Epidemiology*, *36*(3), 666–76. http://doi.org/10.1093/ije/dym018

Scales, C. D., Norris, R. D., Peterson, B. L., Preminger, G. M., & Dahm, P. (2005). Clinical research and statistical methods in the urology literature. *Journal of Urology*, *174*(4 I), 1374–1379. http://doi.org/10.1097/01.ju.0000173640.91654.b5

Shadish, W. R., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin. Retrieved from http://post.queensu.ca/~hh11/assets/applets/Causal_inference_in_experimental_and_quasi-experimental_designs.pdf

Shamliyan, T., Kane, R. L., & Dickinson, S. (2010). A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. *Journal of Clinical Epidemiology*, *63*(10), 1061–70. http://doi.org/10.1016/j.jclinepi.2010.04.014

Simmonds, M., Salanti, G., McKenzie, J., Elliott, J., Agoritsas, T., Hilton, J., … Pearson, L. (2017). Living systematic reviews: 3. Statistical methods for updating meta-analyses. *Journal of Clinical Epidemiology*, *91*, 38–46. http://doi.org/10.1016/j.jclinepi.2017.08.008

Stang, A. (2010). Critical evaluation of the Newcastle-Ottawa scale for the assessment of

the quality of nonrandomized studies in meta-analyses. *European Journal of Epidemiology*, *25*(9), 603–5. http://doi.org/10.1007/s10654-010-9491-z

Stanley, T. D., & Doucouliagos, H. (2015). Neither fixed nor random: Weighted least squares meta-analysis. *Statistics in Medicine*, *34*(13), 2116–2127. http://doi.org/10.1002/sim.6481

Sterne, J. A., Hernán, M. A., Reeves, B. C., Savovi , J., Berkman, N. D., Viswanathan, M., … Higgins, J. P. (2016). ROBINS-I: a tool for assessing risk of bias in non-randomized studies of interventions. *Bmj*, 4–10. http://doi.org/10.1136/bmj.i4919

Thomas, J., Noel-Storr, A., Marshall, I., Wallace, B., McDonald, S., Mavergames, C., … Pearson, L. (2017). Living systematic reviews: 2. Combining human and machine effort. *Journal of Clinical Epidemiology*, *91*, 31–37. http://doi.org/10.1016/j.jclinepi.2017.08.011

Thompson, S., Ekelund, U., Jebb, S., Lindroos, A. K., Mander, A., Sharp, S., … Wilks, D. (2011). A proposed method of bias adjustment for meta-analyses of published observational studies. *International Journal of Epidemiology*, *40*, 765–777. http://doi.org/10.1093/ije/dyq248

Thorne, F. C. (1947). The clinical method in science. *American Psychologist*, *2*(5), 159–166. http://doi.org/10.1037/h0060157

Turner, R. M., Spiegelhalter, D. J., Smith, G. C. S., & Thompson, S. G. (2009). Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society Series A-Statistics in Society*, *172*, 21–47. http://doi.org/10.1111/j.1467-985X.2008.00547.x

Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: the Study Design and Implementation Assessment Device (Study DIAD). *Psychological Methods*, *13*(2), 130–49. http://doi.org/10.1037/1082-989X.13.2.130

Verhagen, A. P., de Vet, H. C. W., Vermeer, F., Widdershoven, J. W. M. G., de Bie, R. a, Kessels, A. G. H., … van den Brandt, P. a. (2002). The influence of methodologic quality on the conclusion of a landmark meta-analysis on thrombolytic therapy. *International Journal of Technology Assessment in Health Care*, *18*(1), 11–23. http://doi.org/10.1017/S0266462309091016

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. Retrieved from http://www.jstatsoft.org/v36/i03/

Viswanathan, M., & Berkman, N. D. (2012). Development of the RTI item bank on risk of bias and precision of observational studies. *Journal of Clinical Epidemiology*, *65*(2), 163–78. http://doi.org/10.1016/j.jclinepi.2011.05.008

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., Vandenbroucke, J. P., … Initiative, S. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Preventive Medicine*, *45*(4), 247–51. http://doi.org/10.1016/j.ypmed.2007.08.012

Voss, P. H., & Rehfuess, E. a. (2013). Quality appraisal in systematic reviews of public health interventions: an empirical study on the impact of choice of tool on meta-analysis. *Journal of Epidemiology and Community Health*, *67*(1), 98–104. http://doi.org/10.1136/jech-2011-200940

Wells, G., Shea, B., O'Connell, D., & Peterson, J. (2000). The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Retrieved November 24, 2016, from http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp

Wells, K., & Littell, J. H. (2009). Study Quality Assessment in Systematic Reviews of Research on Intervention Effects. *Research on Social Work Practice*, *19*(1), 52–62. http://doi.org/10.1177/1049731508317278

Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., … Bossuyt, P. M. M. (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine*, *155*(8), 529–36. http://doi.org/10.7326/0003-4819-155-8-201110180-00009

Whiting, P., Harbord, R., & Kleijnen, J. (2005). No role for quality scores in systematic reviews of diagnostic accuracy studies. *BMC Medical Research Methodology*, *5*(1), 19. http://doi.org/10.1186/1471-2288-5-19

# 6. ANEXOS

# Anexo 1: Resultados de los análisis exploratorios de la incorporación del riesgo de sesgo en un metaanálisis publicado de estudios de cohortes

## Methods

*Selection of the meta-analysis*

To illustrate the procedures for including the risk of bias (RoB) in a meta-analysis (MA), we chose a MA extracted from Gurol-Urganci *et al*. (Gurol-Urganci et al., 2013), which includes 16 cohort studies examining the impact of Caesarean section on subsequent birth rate, as compared with vaginal delivery (control group). This MA was performed using inverse variance random effects model, and its summary effect shows that Caesarean delivery and lower birth rate are associated, with a point risk ratio estimate of 0.89 and 95% CI of 0.87 to 0.92 ($p < .001$), and high heterogeneity, $Q = 79.44$, $p < .001$, $I^2 = 81\%$.

*Quality assessment*

The authors of the MA assessed the RoB of primary studies using the Newcastle-Ottawa scale (NOS) for cohort studies (G. Wells et al., 2000). Assessing RoB with the NOS, a study can be scored with a maximum of nine 'stars' on three broad domains: the selection of study groups (0-4 stars), the comparability of the groups (0-2 stars) and the ascertainment of outcome of interest (0-3 stars). From the overall score of the NOS, we established three levels of RoB: 0 to 3 stars, low quality; 4 to 6 stars, moderate quality; and 7 to 9 stars, high quality. The scores of the primary studies in the MA of Gurol-Urganci *et al*. (Gurol-Urganci et al., 2013) ranged from 2 to 9 stars (Table 1).

To check the possible influence of the choice of quality assessment tool on the results, we applied an adaptation of Q-Coh (Jarde et al., 2013) to the same primary studies. This tool provides a classification of studies into three quality levels (high-moderate-low), based on the assessment of RoB in five domains: selection, exposure measures, performance, outcome measures and attrition. Using Q-Coh, seven studies were classified as high quality, three studies as moderate quality, and six studies as low quality (Table 2).

*Data analysis*

All the meta-analyses follow the random-effects model, using the DerSimonian and Laird estimator (DerSimonian & Laird, 1986), except the fixed-effect model and the quality effects model (Doi et al., 2015b). Regarding the latter, the MA was conducted using the MetaXL software (v.2.2) (Barendregt & Doi, 2014). Sensitivity analyses and subgroup analyses were run using Review manager (RevMan) (v.5.3) (Review manager, 2014). Cumulative MA was performed using Comprehensive Meta-analysis (v.2.2.064)

(Borenstein, Hedges, Higgins, & Rothstein, 2005), and the mixed-effects model meta-regressions were carried out with the Metafor package (v.1.9-4) (Viechtbauer, 2010) for R. The remaining analyses were completed using Microsoft Excel.

**Results**

As shown in Table 3, the outcome of MA without RoB adjustment was quite similar between the fixed effect and the random effects model, despite the high heterogeneity of the studies ($Q = 78.44$, $p < .001$, $I^2 = 81\%$).

The results obtained from sensitivity analyses excluding low quality studies are presented in Table 4. When we used the RoB levels obtained from the NOS, only the smallest study was excluded, leading to virtually identical results as compared to the original random-effects MA. On the other hand, RoB assessed with Q-Coh left out six studies rated as low quality. As a result, the ES estimate was similar to the original MA, with a slight increase in the precision of 95% CI and a small reduction of heterogeneity.

Figure 1 shows the subgroup analysis according to the median of the NOS scores, whereas Figure 2 shows subgroup analyses according to (a) the previously defined quality levels from the NOS scores, and (b) the three levels of Q-Coh. In all of these cases, there were differences between the estimates of the different subgroups, in the sense that the higher the quality, the smaller the effect size. Nevertheless, none of the tests for subgroup differences was statistically significant, although we must highlight the low statistical power of the analysis due to the small number of studies in some subgroups.

The studies were sequentially added to cumulative MA (Figure 3), from highest to lowest NOS quality score, and from lowest to highest standard error in the case of equal scores. According to the data depicted in Figure 3, the effect size of the MA increased slightly as lower quality studies were introduced in the analysis.

The results of introducing RoB as a moderator variable in meta-regression models, are shown in Table 5. The test of moderators ($Q_M$) showed significant differences in mean ES depending on quality levels, for both the NOS and Q-Coh. When we carried out the same analyses categorizing the scores from the NOS for each of its three quality domains, none of them showed significant results in the test of moderators. On the other hand, the test for Q-Coh domains found significant differences in two of its five domains: selection and attrition. In this case, it was not possible to apply the same test to the remaining domains

(exposure, performance, and outcome) because all the studies scored the same. In addition, regarding the explained variance ($R^2$), only the NOS overall quality levels and the selection domain of Q-Coh resulted in significant results.

Finally, as Table 6 shows, the results of applying two methods of bias adjustment using the overall quality score from the NOS. As a result, both direct weighting and *Quality effects model* reported almost identical results as compared to the original random-effects MA.

**References**

Barendregt, J., & Doi, S. (2014). *MetaXL version 2.2*. Epigear. Retrieved from http://www.epigear.com/

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). Comprehensive meta-analysis version 2. Englewood, NJ: Biostat.

DerSimonian, R., & Laird, N. (1986). Metaanalysis in Clinical-Trials. *Controlled Clinical Trials*, *7*(3), 177–188. http://doi.org/10.1016/0197-2456(86)90046-2

Doi, S. A. R., Barendregt, J. J., Khan, S., Thalib, L., & Williams, G. M. (2015). Advances in the meta-analysis of heterogeneous clinical trials II: The quality effects model. *Contemporary Clinical Trials*, *45*(A), 123–129. http://doi.org/10.1016/j.cct.2015.05.010

Gurol-Urganci, I., Bou-Antoun, S., Lim, C. P., Cromwell, D. a, Mahmood, T. a, Templeton, A., & van der Meulen, J. H. (2013). Impact of Caesarean section on subsequent fertility: a systematic review and meta-analysis. *Human Reproduction (Oxford, England)*, *28*(7), 1943–52. http://doi.org/10.1093/humrep/det130

Jarde, A., Losilla, J. M., Vives, J., & Rodrigo, M. F. (2013). Q-Coh: A tool to screen the methodological quality of cohort studies in systematic reviews and meta-analyses. *International Journal of Clinical and Health Psychology*, *13*, 138–146.

*Review manager (RevMan)*. (2014). Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. Retrieved from http://www.jstatsoft.org/v36/i03/

Wells, G., Shea, B., O'Connell, D., & Peterson, J. (2000). The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Retrieved from http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp

Studies included in the meta-analysis of Gurol-Urganci et al. (2013):

Hemminki, E., Graubard, B. I., Hoffman, H. J., Mosher, W. D., & Fetterly, K. (1985). Cesarean section and subsequent fertility: results from the 1982 National Survey of Family Growth. *Fertility and Sterility*, *43*(4), 520–8. Retrieved from http://europepmc.org/abstract/MED/3872816

Hall, M. H., Campbell, D. M., Fraser, C., & Lemon, J. (1989). Mode of delivery and future fertility. *BJOG: An International Journal of Obstetrics and Gynaecology*, *96*(11), 1297–1303. http://doi.org/10.1111/j.1471-0528.1989.tb03227.x

Tollånes, M. C., Melve, K. K., Irgens, L. M., & Skjaerven, R. (2007). Reduced Fertility After Cesarean Delivery. *Obstetrics & Gynecology*, *110*(6), 1256–1263. http://doi.org/10.1097/01.AOG.0000292089.18717.9f

Hemminki, E. (1987). Pregnancy and Birth after Cesarean Section: A Survey Based on the Swedish Birth Register. *Birth*, *14*(1), 12–17. http://doi.org/10.1111/j.1523-536X.1987.tb01443.x

Zdeb, M. S., Therriault, G. D., & Logrillo, V. M. (1984). Frequency, spacing, and outcome of pregnancies subsequent to primary cesarean childbirth. *American Journal of Obstetrics and Gynecology*, *150*(2), 205–212. http://doi.org/10.1016/S0002-9378(84)80017-4

LaSala, A. P., & Berkeley, A. S. (1987). Primary cesarean section and subsequent fertility. *American Journal of Obstetrics and Gynecology*, *157*(2), 379–383. http://doi.org/10.1016/S0002-9378(87)80177-1

Smith, G. C. S., Wood, A. M., Pell, J. P., & Dobbie, R. (2006). First cesarean birth and subsequent fertility. *Fertility and Sterility*, *85*(1), 90–5. http://doi.org/10.1016/j.fertnstert.2005.07.1289

Mollison, J., Porter, M., Campbell, D., & Bhattacharya, S. (2005). Primary mode of delivery and subsequent pregnancy. *BJOG : An International Journal of Obstetrics and Gynaecology*, *112*(8), 1061–5. http://doi.org/10.1111/j.1471-0528.2005.00651.x

Hemminki, E., & Meriläinen, J. (1996). Long-term effects of cesarean sections: Ectopic pregnancies and placental problems. *American Journal of Obstetrics and Gynecology*, *174*(5), 1569–1574. http://doi.org/10.1016/S0002-9378(96)70608-7

Hemminki, E., Shelley, J., & Gissler, M. (2005). Mode of delivery and problems in subsequent births: a register-based study from Finland. *American Journal of Obstetrics and Gynecology*, *193*(1), 169–77. http://doi.org/10.1016/j.ajog.2004.11.007

Gottvall, K., & Waldenstrom, U. (2002). Does a traumatic birth experience have an impact on future reproduction? *BJOG: An International Journal of Obstetrics and Gynaecology*, *109*(3), 254–260. http://doi.org/10.1111/j.1471-0528.2002.01200.x

Jolly, J., Walker, J., & Bhabra, K. (1999). Subsequent obstetric performance related to primary mode of delivery. *BJOG: An International Journal of Obstetrics and Gynaecology*, *106*(3), 227–232. http://doi.org/10.1111/j.1471-0528.1999.tb08235.x

Salem, W., Flynn, P., Weaver, A., & Brost, B. (2011). Fertility after cesarean delivery among Somali-born women resident in the USA. *Journal of Immigrant and Minority Health / Center for Minority Public Health*, *13*(3), 494–9. http://doi.org/10.1007/s10903-010-9362-4

Six L, Nather A, Grimm C, Joura EA. Fertility after cesarean section and vaginal delivery. [German]. *Journal Für Fertilität Und Reproduktion*, *15*(4), 16–18.

Eijsink, J. J. H., van der Leeuw-Harmsen, L., & van der Linden, P. J. Q. (2008). Pregnancy after Caesarean section: fewer or later? *Human Reproduction (Oxford, England)*, *23*(3), 543–7. http://doi.org/10.1093/humrep/dem428

Bahl, R., Strachan, B., & Murphy, D. J. (2004). Outcome of subsequent pregnancy three years after previous operative delivery in the second stage of labour: cohort study. *BMJ (Clinical Research Ed.)*, *328*(7435), 311. http://doi.org/10.1136/bmj.37942.546076.44

**Table 1.** Results of risk of bias assessment with the NOS

| Study | RR (95% ci) | NOS | | | | |
|---|---|---|---|---|---|---|
| | | Sel | Comp | Out | Total (quantitative) | Categorical[a] |
| Hemminki (1985) | 1,00 (0,88, 1,14) | 4 | 2 | 2 | 8 | 1 |
| Hall (1989) | 0,76 (0,72, 0,80) | 3 | 1 | 2 | 6 | 2 |
| Tollanes (2007) | 0,90 (0,90, 0,91) | 4 | 1 | 3 | 7 | 2 |
| Hemminki (1987) | 0,91 (0,88, 0,94) | 4 | 2 | 3 | 9 | 1 |
| Zdeb (1984) | 0,95 (0,92, 0,98) | 3 | 2 | 2 | 7 | 2 |
| LaSala (1987) | 0,95 (0,87, 1,05) | 2 | 1 | 2 | 5 | 3 |
| Smith (2006) | 0,93 (0,87, 1,00) | 4 | 2 | 3 | 9 | 1 |
| Mollison (2005) | 0,91 (0,89, 0,94) | 3 | 2 | 3 | 8 | 1 |
| Hemminki (1996) | 0,84 (0,82, 0,87) | 3 | 1 | 3 | 7 | 2 |
| Hemminki (2005) | 0,90 (0,89, 0,91) | 4 | 0 | 3 | 7 | 2 |
| Gottvall (2002) | 0,92 (0,79, 1,07) | 2 | 0 | 3 | 5 | 3 |
| Jolly (1999) | 0,82 (0,71, 0,95) | 2 | 1 | 1 | 4 | 3 |
| Salem (2011) | 0,65 (0,38, 1,09) | 2 | 0 | 0 | 2 | 3 |
| Six (2005) | 0,88 (0,70, 1,09) | 2 | 1 | 1 | 4 | 3 |
| Eijsink (2008) | 1,02 (0,80, 1,30) | 3 | 1 | 3 | 7 | 2 |
| Bahl (2004) | 0,90 (0,71, 1,15) | 2 | 2 | 1 | 5 | 3 |

RR = Risk ratio; CI = Confidence interval; Sel = Selection; Comp = Comparability; Out = Outcome.
[a]Categorical rank based on quantitative score: 1-High (8-9), 2-Acceptable (6-7), 3-Low (<6).

**Table 2.** Results of risk of bias assessment with Q-Coh

| Study | RR (95% IC) | Categorical[a] |
|---|---|---|
| Hemminki (1985) | 1,00 (0,88, 1,14) | 1 |
| Hall (1989) | 0,76 (0,72, 0,80) | 3 |
| Tollanes (2007) | 0,90 (0,90, 0,91) | 2 |
| Hemminki (1987) | 0,91 (0,88, 0,94) | 1 |
| Zdeb (1984) | 0,95 (0,92, 0,98) | 1 |
| LaSala (1987) | 0,95 (0,87, 1,05) | 3 |
| Smith (2006) | 0,93 (0,87, 1,00) | 1 |
| Mollison (2005) | 0,91 (0,89, 0,94) | 1 |
| Hemminki (1996) | 0,84 (0,82, 0,87) | 2 |
| Hemminki (2005) | 0,90 (0,89, 0,91) | 1 |
| Gottvall (2002) | 0,92 (0,79, 1,07) | 3 |
| Jolly (1999) | 0,82 (0,71, 0,95) | 3 |
| Salem (2011) | 0,65 (0,38, 1,09) | 3 |
| Six (2005) | 0,88 (0,70, 1,09) | 2 |
| Eijsink (2008) | 1,02 (0,80, 1,30) | 1 |
| Bahl (2004) | 0,90 (0,71, 1,15) | 3 |

RR = Risk ratio; CI = Confidence interval
[a]Categorical rank: 1-High, 2-Acceptable, 3-Low.

**Table 3.** Results of the fixed effect and the random effects model for the meta-analysis of Gurol-Urganci et al. (2013)

| | n | RR (95% CI) | Q (*p*) | $I^2$ (%) |
|---|---|---|---|---|
| Fixed effect model | 16 | 0.90 (0.89, 0.91) | 78.44 (< .001) | 81 |
| Random effects model | 16 | 0.89 (0.87, 0.92) | | |

Note: RR = Risk ratio; CI = Confidence interval; Q = Cochran's statistic of heterogeneity; $I^2$ = heterogeneity statistic.

**Table 4.** Results of the sensitivity analyses excluding the studies rated as low quality

| Quality assessment | n | RR (95% CI) | Q (*p*) | $I^2$ (%) |
|---|---|---|---|---|
| NOS | 15 | 0.90 (0.87, 0.92) | 76.94 (< .001) | 82 |
| Q-Coh | 10 | 0.91 (0.89, 0.92) | 38.38 (< .001) | 77 |

Note: RR = Risk ratio; CI = Confidence interval; Q = Cochran's statistic of heterogeneity; $I^2$ = heterogeneity statistic.

**Figure 1.** Subgroup analysis stratified by the median of NOS scores (Med = 7).

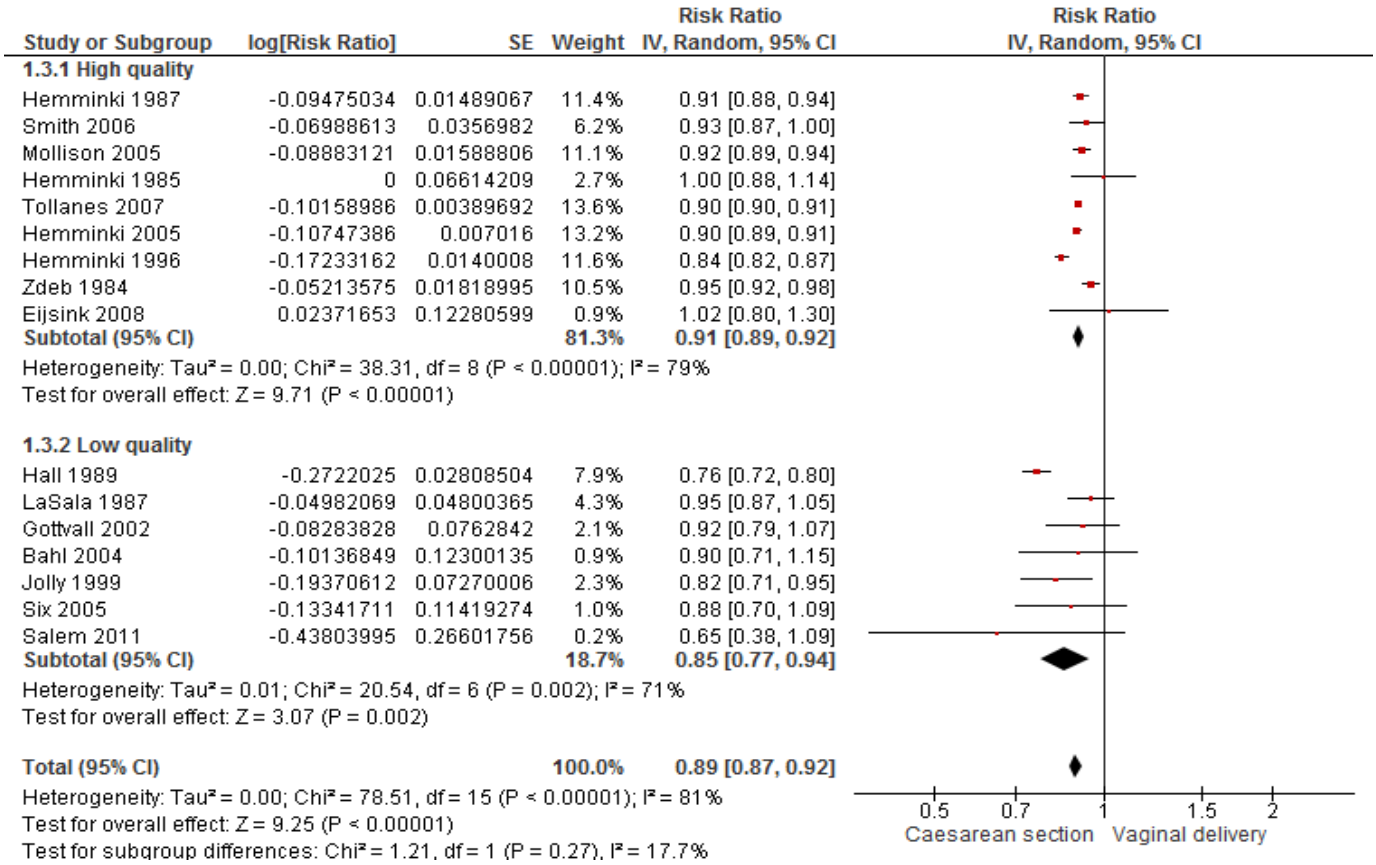
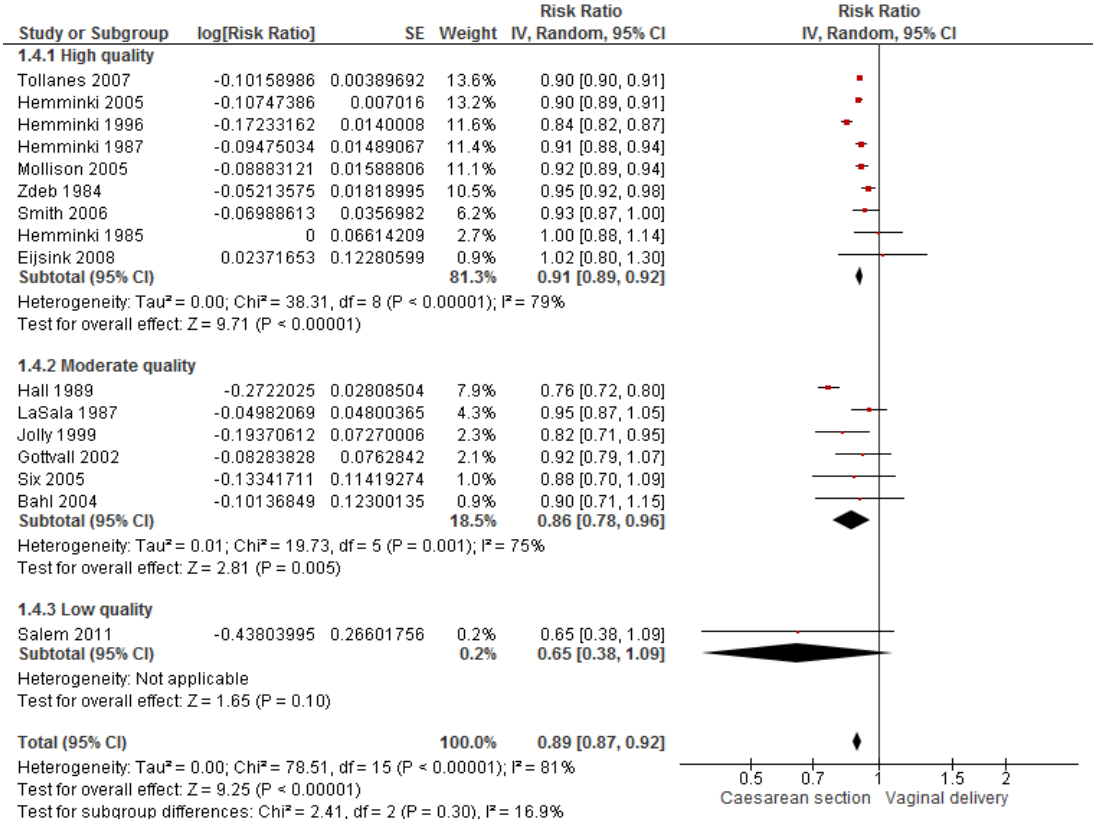
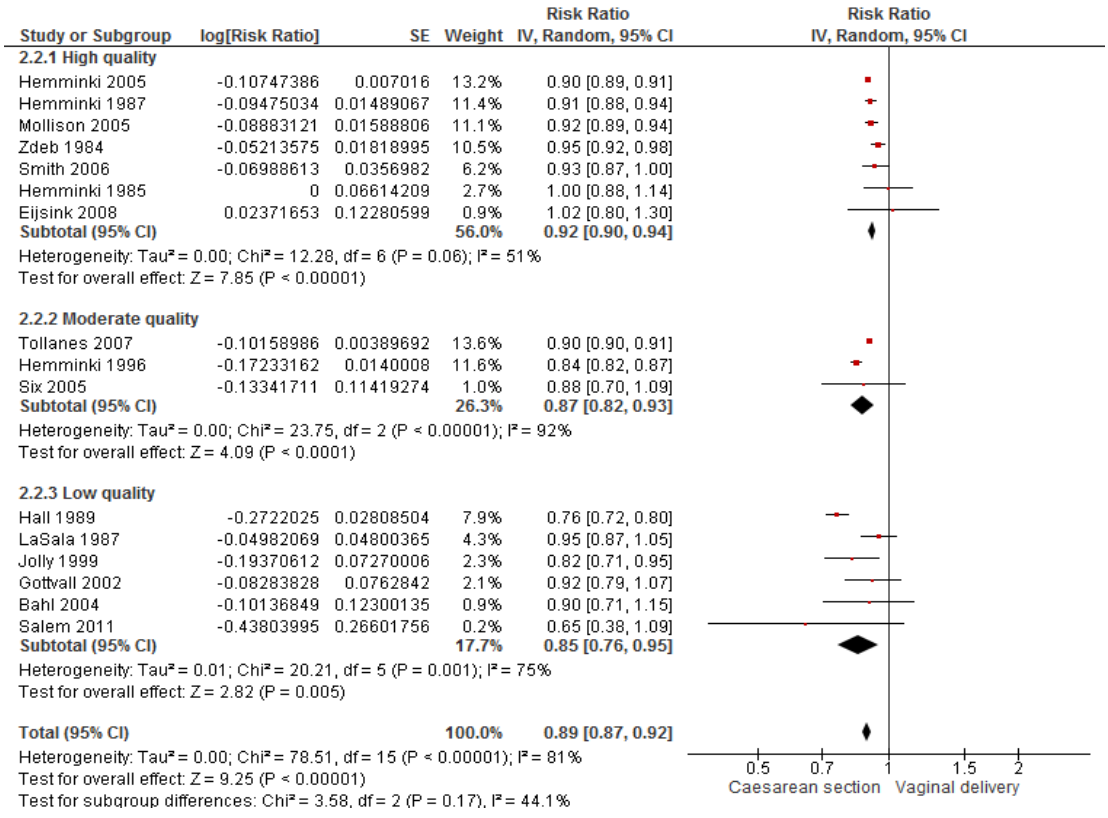


**Figure 2.** Subgroup analysis stratified by (a) the three quality levels derived from NOS scores (0-3 = low, 4-6 = moderate, 7-9 = high) and (b) the three quality levels of Q-Coh (low, moderate, high).

**(a)**




**(b)**



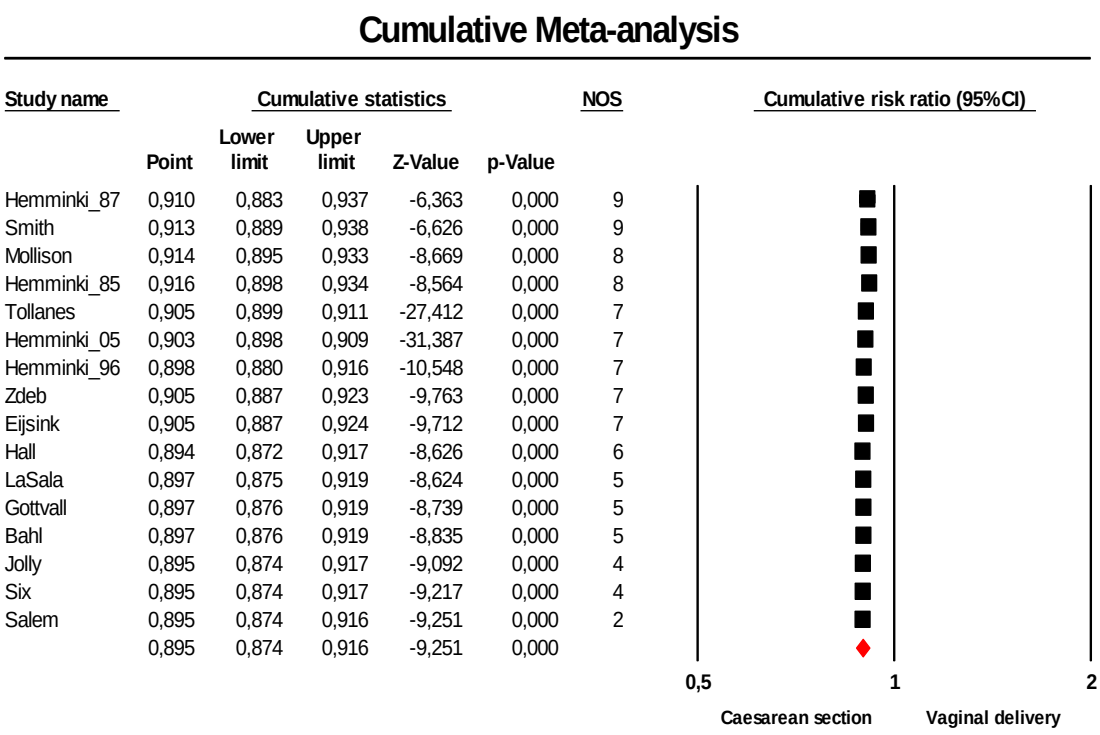



**Table 5.** Results of the test of moderators (meta-regression)

| Quality assessment | Moderator | $Q_M$ (*p*) | $R^2$ (%) |
|---|---|---|---|
| NOS | Overall quality | 9.42 (.009) | 28 |
| | Selection domain | 2.77 (.251) | 0 |
| | Comparability domain | 4.87 (.088) | 0 |
| | Outcome domain | 2.29 (.515) | 0 |
| Q-Coh | Overall quality | 8.44 (.015) | 0 |
| | Selection domain | 8.06 (.005) | 27 |
| | Attrition domain | 6.24 (.013) | 0 |

Note: $Q_M$ = Test of moderators; $R^2$ = Proportion of explained variance

**Table 6.** Results of bias-adjusted meta-analysis using the NOS quality scores

| Bias-adjustment method | n | RR (95% CI) | Q (*p*) | $I^2$ (%) |
|---|---|---|---|---|
| Weighting | 16 | 0.90 (0.87, 0.92) | 78.44 (< .001) | 81 |
| Quality effects model | 16 | 0.90 (0.87, 0.93) | | |

Note: RR = Risk ratio; CI = Confidence interval; Q = Cochran's statistic of heterogeneity; $I^2$ = heterogeneity statistic.

# Anexo 2: Estrategias de búsqueda

## Apéndice A del Artículo 1

## **Appendix A:** Search strategies

| **PsycINFO** search strategy from January 1st, 2010 to May 22th, 2016 (via APA PsycNET) | |
|---|---|
| Research field | |
| 1 | DE "health care psychology" OR DE "clinical psychology" |
| 2 | "psycho*" AND ("health" OR "clinic*") |
| Design | |
| 3 | DE "Followup Studies" OR DE "Prospective Studies" OR DE "Retrospective Studies" |
| 4 | "cohort" OR "case control" OR "follow-up stud*" OR "follow up stud*" OR "observational" OR "non-experimental" OR "non experimental" OR "prospective" OR "retrospective" OR "epidemiologic stud*" |
| 5 | DE "Clinical Trials" OR DE "Experimental Methods" OR DE "Quasi Experimental Methods" |
| 6 | "trial*" OR "RCT*" OR "experimental stud*" OR "quasi experiment" OR "quasi-experiment" OR "CCT*" OR "q-RCT*" OR "cross-sectional" OR "cross sectional" |
| Combined sets | |
| 7 | (#1 AND #3 NOT #5) OR (#2 AND #4 NOT #6) |
| Filters | |
| 8 | #7 AND Language: ("English" OR "Spanish") AND Methodology: Systematic Review OR Meta Analysis AND Population Group: Human AND Year: 2010 To 2016 AND Peer-Reviewed Journals only |
| Records retrieved: **606** | |

| **Pubmed** search strategy from January 1st, 2010 to May 22th, 2016 (via NLM) | |
|---|---|
| Research field | |
| 1 | "Behavioral Medicine"[MeSH Terms] OR "Psychology, Clinical"[MeSH Terms] |
| 2 | "psychology"[All Fields] AND ("health"[All Fields] OR "clinical"[All Fields]) |
| Design | |
| 3 | "cohort studies"[MeSH Terms] OR "case control studies"[MeSH Terms] |
| 4 | "cohort"[All Fields] OR "case control"[All Fields] OR "follow-up study"[All Fields] OR "follow-up studies"[All Fields] OR "follow up study"[All Fields] OR "follow up studies"[All Fields] OR "observational"[All Fields] OR "non-experimental"[All Fields] OR "non experimental"[All Fields] OR "prospective"[All Fields] OR "retrospective"[All Fields] OR "epidemiologic study" [All Fields] OR "epidemiologic studies"[All Fields] |
| 5 | "Clinical Studies as Topic"[ MeSH Terms] OR "Cross-Sectional Studies"[ MeSH Terms] |
| 6 | "trial*"[All Fields] OR "RCT*"[All Fields] OR "experimental study"[All Fields] OR "experimental studies"[All Fields] OR "quasi experiment"[All Fields] OR "quasi-experiment"[All Fields] OR "CCT*"[All Fields] OR "q-RCT*"[All Fields] OR "cross-sectional"[All Fields] OR "cross sectional"[All Fields] |
| Publication type | |
| 7 | "Meta-Analysis"[Publication Type] |
| 8 | "systematic review"[Title] OR "meta analysis"[Title] OR "meta-analysis"[Title] |
| Combined sets | |
| 9 | (#1 OR #2) AND (#3 OR #4) AND (#7 OR #8) NOT (#5 OR #6) |
| Filters | |
| 10 | #9 AND ("English"[Language] OR "Spanish"[Language]) AND ("2010/01/01"[PDAT]: "3000"[PDAT]) AND "humans"[MeSH Terms] |
| Records retrieved: **436** | |

| **Web of Science** search strategy from January 1st, 2010 to May 22th, 2016 (via FECYT) | |
|---|---|
| Research field | |
| 1 | TS=("psycho*" AND ("health" OR "clinic*")) |
| Design | |
| 2 | TS=("cohort" OR "case control" OR "follow-up stud*" OR "follow up stud*" OR "observational" OR "non-experimental" OR "non experimental" OR "prospective" OR "retrospective" OR "epidemiologic stud*") |
| 3 | TS= ("trial*" OR "RCT*" OR "experimental stud*" OR "quasi experiment" OR "quasi-experiment" OR "CCT*" OR "q-RCT*" OR "cross-sectional" OR "cross sectional") |
| Publication type | |
| 4 | TI=("systematic review*" OR "meta analys*" OR "meta-analys*") |
| Combined sets | |
| 5 | #1 AND #2 AND #4 NOT #3 |
| Filters | |
| 6 | #5 AND RESEARCH AREAS: (PSYCHOLOGY) AND LANGUAGES: (ENGLISH OR SPANISH) AND DOCUMENT TYPES: (ARTICLE OR REVIEW) AND Timespan: 2010-2016. |
| Records retrieved: **199** | |

| | |
|---|---|
| **Scopus** search strategy from January 1st, 2010 to May 22th, 2016 | |
| Research field | |
| 1 | ALL (psychology AND (health OR clinical)) |
| Design | |
| 2 | ALL (cohort OR "case control" OR "follow-up study" OR observational OR non-experimental OR prospective OR retrospective OR "epidemiologic study") |
| 3 | ALL (trial OR rct OR "experimental study" OR quasi-experiment OR cct OR q-rct OR cross-sectional) |
| Publication type | |
| 4 | TITLE ("systematic review" OR meta-analysis) |
| Combined sets | |
| 5 | #1 AND #2 AND #4 NOT #3 |
| Filters | |
| | #5 AND (LIMIT-TO (DOCTYPE, "ar") OR LIMIT-TO (DOCTYPE, "re")) AND (LIMIT-TO (SUBJAREA, "PSYC")) AND (LIMIT-TO (LANGUAGE, "English") OR LIMIT-TO (LANGUAGE, "Spanish")) AND (LIMIT- TO (PUBYEAR, 2016) OR LIMIT-TO (PUBYEAR, 2015) OR LIMIT-TO (PUBYEAR, 2014) OR LIMIT-TO (PUBYEAR, 2013) OR LIMIT-TO (PUBYEAR, 2012) OR LIMIT-TO (PUBYEAR, 2011) OR LIMIT-TO (PUBYEAR, 2010)) |
| Records retrieved: **468** | |

# Anexo 3: Manual de codificación de la extracción de datos

## Apéndice B del Artículo 1

## **Appendix B:** Coding manual for data extraction

### Header variables coding

Common database to all the reviews, including or not risk of bias assessment, and including or not a meta-analysis.

| Variable | Description |
|---|---|
| [STUDYID] | Study identification number: _____ |
| [TITLE] | Title of the study:<br>_____ |
| [AUTHOR] | First author of the study: _____ |
| [YEARPUB] | Year of publication:<br>1. 2010<br>2. 2011<br>3. 2012<br>4. 2013<br>5. 2014<br>6. 2015<br>7. 2016 |
| [JOURNAL] | Journal extended dictionary:<br>[JOURNALID]_____<br>[J_NAME] Complete name of the journal: _____<br>[J_RECOM] Journal recommendations (more than one response is possible):<br>0. None<br>1. Reporting standards<br>2. Risk of bias assessment<br>3. Registry of protocols (RCTs, reviews…) |
| [YEARIMP] | Extended information about the impact of journals:<br>[JOURNALID]: _____<br>[YEARIND] Index year: _____<br>[J_IMPACT_JCR] Impact factor JCR: _____<br>[J_CAT_SCIE] Number of categories indexing in JCR-SCIE: _____<br>[J_QUART_SCIE] Quartile JCR-SCIE*:<br>1. Quartile 1<br>2. Quartile 2<br>3. Quartile 3<br>4. Quartile 4<br>[J_CAT_SSCI] Number of categories indexing in JCR-SSCI: _____ |

| | |
|---|---|
| | [J_QUART_SSCI] Quartile JCR-SSCI*:<br>1. Quartile 1<br>2. Quartile 2<br>3. Quartile 3<br>4. Quartile 4<br>[J_IMPACT_SJR] Impact factor SJR: _____<br>[J_CAT_SJR] Number of categories indexing in SJR: _____<br>[J_QUART_JCR] Quartile SJR*<br>1. Quartile 1<br>2. Quartile 2<br>3. Quartile 3<br>4. Quartile 4<br>*When a journal is ranked in more than one category it will be chosen the highest quartile achieved |
| [NSTUDIES] | Number of studies included in the review (N): _____<br>Total number of primary studies included in qualitative synthesis, regardless of the number of studies included in quantitative synthesis when a meta-analysis was performed |
| [NCOHORT] | Number of cohort studies included: _____ |
| [NCASECON] | Number of case-control studies included: _____ |
| [OBJECTIVE] | Were the effects of some kind of intervention assessed in the review?<br>0. No<br>1. Yes<br>The included primary studies assessed the effect of an intervention, but the researchers did not allocate the participants to intervention levels. |
| [GUIDELINE] | Was any quality guideline followed?<br>0. No<br>There was no mention of quality guidelines<br>1. Yes<br>The authors declared that they followed the recommendations of some quality guideline/s related to reporting, review performance, quality of evidence, or others.<br>2. Partially<br>The authors incorporated some elements of quality guidelines (e.g. PRISMA flowchart) but they did not clearly follow a quality guideline in its whole. |
| [GUIDENAME] | Which quality guideline was followed? (only if the response to item GUIDELINE is 1 or 2). More than one response is possible:<br>1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)<br>2. MOOSE (Meta-analysis of Observational Studies in Epidemiology)<br>3. CRD (Centre for Reviews and Dissemination)<br>4. GRADE (Grading of Recommendations Assessment, Development and Evaluation)<br>5. Other*<br>*[GUIDENAMEother] Which? _____ |

| | |
|---|---|
| [PROTOCOL] | Was the protocol of the review registered?<br>0. No<br>There was no mention to protocol registry<br>1. Yes<br>The authors declared that they registered the review protocol |
| [PROTNAME] | Where was the protocol registered? (only if the response to item PROTOCOL is 1):<br>1. PROSPERO<br>2. Cochrane Collaboration<br>3. Other*<br>*[PROTNAMEother] Which? _____ |
| [ELEGIBIL] | Was any element related to risk of bias used as eligibility criteria?<br>0. No<br>1. Yes<br>Primary studies were included/excluded from review based on one or more elements related to risk of bias<br>2. Unclear |
| [ELEGDOMAIN] | What was the domain of bias to which the element used as eligibility criteria belong? (only if the response to item ELEGIBIL is 1). More than one response is possible:<br>1. Selection<br>Systematic differences in characteristics between the participants<br>2. Confounding<br>Presence of confounding variables that were not controlled neither in the design of the study nor in the analyses<br>3. Exposure<br>Inaccurate measurement or classification of subjects on study exposure/s<br>4. Performance<br>Systematic differences between participants appeared during the follow-up<br>5. Outcome<br>Inaccurate measurement or classification of subjects on study outcome/s<br>6. Attrition<br>Systematic differences between comparison groups as a result of differential withdrawals or exclusions of participants |
| [LIMITATION] | Was a particular domain of bias mentioned by the authors as a methodological limitation?<br>0. No<br>1. Yes<br>The authors reported that some risk of bias was possible in one or more domains<br>2. Unclear |

| | |
|---|---|
| [LIMITDOMAIN]* | Which was the domain at risk of bias according to the authors? (only if the response to item LIMITATION is 1). More than one response is possible:<br>1. Selection<br>2. Confounding<br>3. Exposure<br>4. Performance<br>5. Outcome<br>6. Attrition<br>7. Reported results<br>Selective reporting of results, analysis or subgroup of participants |
| [CONFOUND] | Were the main confounding variables that could threaten the review results reported?<br>0. No<br>1. Yes<br>2. Unclear<br>Some confounding/adjustment variables were reported but It is not clear if all the relevant variables were reported |
| [ROBASSESS] | Was the risk of bias of primary studies assessed?<br>0. No<br>There was no mention risk of bias assessment of the included studies<br>1. Yes<br>The authors declared that the risk of bias of the included studies was assessed (whether it is based on a previously published tool or not)<br>2. Unclear<br>The authors declared that the risk of bias of the included studies was assessed but there was no more information about it |
| [MA] | Was any meta-analysis performed?<br>0. No<br>1. Yes |
| [FEATUR] | Relevant review features (optional)<br>Note here any relevant/interesting feature about the review (e.g. significant flaws detected, comments about reporting, example of well-executed revision…)<br>_____ |

## Risk of bias variables coding

Database including only the reviews that assessed risk of bias

| Variable | Description |
|---|---|
| [STUDYID] | Study identification number: _____ |
| [TOOLTYPE] | Type of tool used to assess risk of bias:<br>1. Specific biases<br>Informal appraisal of specific biases<br>2. Standard scale/check-list<br>Previously published scale/check-list not modified<br>3. Modified standard scale/check-list<br>Previously published scale/checklist modified<br>4. New scale/check-list created by the authors<br>New structured scale/check-list<br>5. Mixture of two or more scales/check-lists<br>The authors used items/components extracted from two or more scales/check-lists<br>6. Not reported |
| [TOOL] | Scale/check-list used to assess risk of bias (only if the response to item TOOLTYPE is 2 or 3)<br>Tool dictionary:<br>[TOOLID] _____<br>[T_NAME] Name of the tool: _____ |
| [TOOLREP] | Was the tool to assess the risk of bias reported when no standard scale/check-list was used? (only if the response to item TYPETOOL is 1, 3, 4 or 5):<br>0. No<br>1. Yes<br>The tool or the specific biases assessed were reported |
| [TOOLIND] | Type of index used to assess risk of bias (except if the response to item TOOLTYPE is 6):<br>1. Quantitative<br>2. Categories/levels<br>3. Both<br>4. Unclear/not reported |
| [TOOLOUT] | Type of risk of bias outcome provided (except if the response to item TOOLTYPE is 6):<br>1. Overall risk of bias<br>2. Domains/dimensions/biases<br>3. Both<br>4. Unclear/not reported |

| | |
|---|---|
| [ROBEXCLUSION] | Were the results of risk of bias assessment used as exclusion criteria?<br>0. No<br>1. Yes, on the basis of the overall risk of bias<br>Studies that not met a certain overall quality standard were excluded from the review<br>2. Yes, on the basis of the domains of bias<br>Studies that not met a certain quality standard in one or more domains of bias were excluded from the review |
| [EXCLDOMAIN] | Which was the domain of bias used as exclusion criteria? (only if the response to item ROBEXCLUSION is 2). More than one response is possible:<br>1. Selection<br>2. Confounding<br>3. Exposure<br>4. Performance<br>5. Outcome<br>6. Attrition<br>7. Reported results |
| [ROBREP] | Reporting of the results of risk of bias assessment:<br>1. Detailed results reported<br>The results of risk of bias assessment were reported for each individual study<br>2. Overall results reported<br>Only overall results of risk of bias assessment were reported<br>3. Results not reported or not available |
| [ROBCAT] | Number of categories based on risk of bias assessment:<br>1. Two categories (high-low)<br>2. Three categories (high-moderate-low)<br>3. No categories reported |
| [ROBRESULT] | Results of risk of bias assessment (except if the response to item ROBCAT is 3):<br>[ROB_L] Number of studies at low risk of bias: _____<br>[ROB_M] Number of studies at moderate risk of bias: _____<br>[ROB_H] Number of studies at high risk of bias: _____ |
| [ROBVAR] | When sufficient data is available, was there variability in risk of bias between the included studies?<br>0. No<br>Most of the scores of assessed studies were concentrated on a single side of the central scores<br>1. Yes<br>The scores of assessed studies were distributed on both sides of the central scores<br>2. Not reported<br>For example, for the NOS scale (range 0-9), were the studies distributed across all score range or were concentrated above/below a score of 5? |

| | |
|---|---|
| [ROBINC] | How was risk of bias assessment incorporated into the results:<br>1. Only descriptive (qualitative)<br>Risk of bias assessment was linked to the interpretation of the results in a narrative way in some section/s of the review or a plot/figure was used<br>2. Analytical (quantitative)<br>Risk of bias assessment was used to test the differences between the studies at high and low risk of bias or used to adjust the results of the meta-analysis<br>3. Not incorporated<br>Risk of bias assessment was not linked to the interpretation of the results or used in calculations |
| [ROBDESC] | Descriptive incorporation:<br>1. Only narrative<br>2. Plot/figure |
| [ROBNARR]* | Section/s of the review where the risk of bias assessment was narratively linked to the interpretation of the review results:<br>[ROBNARR_ABS] Abstract:<br>1. Specifically mentioned<br>Risk of bias assessment results were specifically mentioned and linked to review results (e.g. X studies were at high risk of bias, were not controlled for confounding or had inadequate measurement of the variables)<br>2. General comment<br>General comment about the overall risk of bias of the evidence<br>3. Not reported<br>Risk of bias assessment results were not mentioned in this section<br>[ROBNARR_DISC] Discussion:<br>1. Specifically mentioned<br>2. General comment<br>3. Not reported<br>[ROBNARR_CONC] Conclusions:<br>1. Specifically mentioned<br>2. General comment<br>3. Not reported<br>[ROBNARR_REC] Recommendations:<br>1. Specifically mentioned<br>2. General comment<br>3. Not reported |

| | |
|---|---|
| [ROBANALYT] | Method used to incorporate risk of bias assessment into quantitative analysis (only if the response to item ROBINC is 2). More than one response is possible:<br>1. Sensitivity analysis<br>2. Cumulative meta-analysis<br>3. Subgroup analysis<br>4. Meta-regression<br>5. Quality weighting<br>6. Other*<br>*[ROBANALYTother] Which? _____ |
| [ROBSIGN] | Did the authors declare that there was some significant influence of risk of bias on the results?<br>0. No<br>The authors tested in some way the influence of the RoB of the included studies on the results but it was not found significant influence<br>1. Yes, on the basis of the overall risk of bias<br>The authors stated that overall RoB of the included studies had influence on the results, or they tested in some way that this influence existed<br>2. Yes, on the basis of the domains of bias<br>The authors stated that some domain/s of RoB had influence on the results, or they tested in some way that this influence existed<br>3. Unclear<br>It was speculated that the RoB of the included studies may had affected the results but it was not tested<br>4. Not reported<br>No comment about the RoB influence on the results |
| [SIGNDOMAIN] | Which was the domain of bias that influenced the results? (only if the response to item ROBSIGN is 2). More than one response is possible:<br>1. Selection<br>2. Confounding<br>3. Exposure<br>4. Performance<br>5. Outcome<br>6. Attrition<br>7. Reported results |

## Meta-analysis variables coding

Database including only the reviews that performed a meta-analysis

| Variable | Description |
|---|---|
| [STUDYID] | Study identification number:<br><br>_____ |
| [MAMODEL] | Meta-analysis model:<br>1. Fixed effect<br>2. Random effects<br>3. Both<br>4. Other*<br>*[MAMODELother] Which? _____ |
| [MARESULT] | Statistical significance of the meta-analysis result/s:<br>1. Significant<br>If the review contains two or more meta-analyses, at least one of the results can be considered significant<br>2. Non-significant<br>If the review contains two or more meta-analyses, none of the results can be considered significant<br>3. Unclear<br>4. Not reported |
| [HETEROG] | Statistical significance of heterogeneity statistic/analysis result/s:<br>1. Significant<br>2. Non-significant<br>3. Unclear<br>4. Not reported<br>*same response options as MARESULT |
| [PBASSESS] | Was publication bias assessed?<br>0. No<br>Publication bias was not assessed for this meta-analysis<br>1. Yes<br>The authors declared that publication bias was assessed.<br>2. Unclear |
| [PUBBIAS] | Statistical significance of publication bias assessment (only if response to item PBASSESS is 1):<br>1. Significant<br>2. Non-significant<br>3. Unclear<br>4. Not reported<br>*same response options as MARESULT |

# Anexo 4: Características de los estudios incluidos

Apéndice C del Artículo 1

**Appendix C**: Selected features of included studies *

| Study[a] | N studies[b] | N cohort[c] | N case-c[d] | Guideline[e] | Protocol[f] | RoB assessment[g] | Meta-analysis[h] | Type of tool[i] | RoB incorporation[j] | RoB influence[k] |
|---|---|---|---|---|---|---|---|---|---|---|
| **[41] AlAqeel, 2012** | 27 | 10 | 0 | N | N | N | N | - | - | - |
| **[42] Alisic, 2015** | 13 | 1 | 0 | N | N | N | N | - | - | - |
| **[43] Allott, 2011** | 22 | 22 | 0 | N | N | Y | N | 1 | OD | U |
| **[44] Boden, 2011** | 15 | NR | NR | N | N | N | Y | - | - | - |
| **[45] Borges, 2016** | 19 | 15 | 4 | N | N | N | Y | - | - | - |
| **[46] Brennan, 2013** | 28 | 28 | 0 | P | N | U | N | - | - | - |
| **[47] Brito, 2015** | 6 | 5 | 0 | Y | N | N | Y | - | - | - |
| **[48] Burton, 2016** | 23 | 1 | 1 | Y | Y | N | Y | - | - | - |
| **[49] Cerimele, 2013** | 8 | 8 | 0 | Y | N | N | N | - | - | - |
| **[50] Chen, 2016** | 10 | 1 | 9 | N | N | Y | Y | 2 | OD | NR |
| **[51] da Silva, 2013** | 51 | 36 | 15 | N | N | Y | N | 2 | OD | NR |
| **[52] de Maat, 2013** | 14 | 13 | 0 | N | N | Y | Y | 4 | A | U |
| **[53] Diaz-Piedra, 2015** | 34 | 0 | 34 | N | N | Y | N | 3 | OD | U |
| **[54] Dickens, 2012** | 16 | 16 | 0 | P | N | Y | Y | 3 | A | N |
| **[55] DiGangi, 2013** | 54 | 54 | 0 | N | N | N | N | - | - | - |
| **[56] Diniz, 2013** | 23 | 23 | 0 | Y | N | Y | Y | 2 | OD | N |
| **[57] Elbers, 2013** | 10 | 10 | 0 | Y | N | Y | Y | 3 | OD | YO |
| **[58] El-Nashar, 2010** | 12 | 4 | 0 | N | N | N | N | - | - | - |
| **[59] Faedda, 2015** | 26 | 26 | 0 | Y | N | N | N | - | - | - |
| **[60] Fan, 2013** | 13 | 13 | 0 | Y | N | N | Y | - | - | - |
| **[61] Flak, 2014** | 34 | 34 | 0 | N | N | Y | Y | 3 | A | YO |
| **[62] Geulayov, 2012** | 28 | 15 | 10 | N | N | N | Y | - | - | - |
| **[63] Goldfarb, 2015** | 26 | 10 | 0 | Y | N | N | N | - | - | - |
| **[64] Heerde, 2015** | 38 | 3 | 0 | Y | N | N | N | - | - | - |

| Study[a] | N studies[b] | N cohort[c] | N case-c[d] | Guideline[e] | Protocol[f] | RoB assessment[g] | Meta-analysis[h] | Type of tool[i] | RoB incorporation[j] | RoB influence[k] |
|---|---|---|---|---|---|---|---|---|---|---|
| **[65] Heikkila, 2014** | 11 | 11 | 0 | P | N | N | Y | - | - | - |
| **[66] Hemmi, 2011** | 22 | 22 | 0 | N | N | N | Y | - | - | - |
| **[67] Hu, 2015** | 12 | 10 | 2 | Y | N | Y | Y | 2 | OD | YD |
| **[68] Hughes, 2012** | 26 | 4 | 0 | N | N | Y | Y | 1 | A | U |
| **[69] Jokela, 2014** | 6 | 6 | 0 | N | N | N | Y | - | - | - |
| **[70] Jokela, 2013** | 7 | 7 | 0 | N | N | N | Y | - | - | - |
| **[71] Kakela, 2014** | 14 | 14 | 0 | Y | N | N | Y | - | - | - |
| **[72] Kaymaz, 2012** | 6 | 6 | 0 | N | N | U | Y | - | - | - |
| **[73] Keall, 2015** | 14 | 1 | 0 | Y | N | Y | N | 2 | OD | NR |
| **[74] Keshaviah, 2014** | 41 | 41 | 0 | N | N | N | Y | - | - | - |
| **[75] Kropelin, 2013** | 8 | 7 | 0 | N | N | Y | N | 3 | OD | NR |
| **[76] Kuijpers, 2011** | 15 | 15 | 0 | N | N | N | N | - | - | - |
| **[77] Laisné, 2012** | 68 | 68 | 0 | N | N | Y | N | 5 | OD | U |
| **[78] LeBlanc, 2012** | 23 | 21 | 1 | Y | Y | Y | N | 2 | OD | NR |
| **[79] Lehmann, 2015** | 25 | 0 | 25 | P | N | Y | N | 1 | OD | YO |
| **[80] Leung, 2012** | 22 | 22 | 0 | Y | N | Y | Y | 2 | A | YO |
| **[81] Linsell, 2016** | 15 | 14 | 0 | P | Y | Y | N | 3 | OD | NR |
| **[82] Liu, B, 2014** | 10 | 10 | 0 | P | N | N | Y | - | - | - |
| **[83] Liu, RT, 2010** | 57 | NR | NR | N | N | N | N | - | - | - |
| **[84] Liu, S, 2015** | 13 | 0 | 13 | N | N | Y | Y | 2 | NI | NR |
| **[85] Long, 2012** | 18 | 2 | 0 | Y | N | Y | N | 3 | OD | NR |
| **[86] Lovato, 2011** | 19 | 19 | 0 | Y | N | Y | N | 1 | OD | NR |
| **[87] Mäkikangas, 2016** | 40 | 40 | 0 | N | N | N | N | - | - | - |
| **[88] Makin, 2013** | 24 | 24 | 0 | Y | N | N | Y | - | - | - |
| **[89] Malik, 2014** | 19 | 14 | 0 | Y | N | Y | Y | 2 | OD | U |
| **[90] Masson, 2015** | 12 | 0 | 12 | Y | N | U | Y | - | - | - |
| **[91] Modabbernia, 2016** | 67 | NR | NR | Y | N | Y | Y | 2 | A | N |

| Study[a] | N studies[b] | N cohort[c] | N case-c[d] | Guideline[e] | Protocol[f] | RoB assessment[g] | Meta-analysis[h] | Type of tool[i] | RoB incorporation[j] | RoB influence[k] |
|---|---|---|---|---|---|---|---|---|---|---|
| **[92] Monette, 2014** | 25 | 5 | 2 | N | N | N | Y | - | - | - |
| **[93] Moulton, 2015** | 9 | 1 | 0 | Y | N | N | N | - | - | - |
| **[94] Murri, 2016** | 41 | 0 | 41 | N | N | Y | Y | 3 | A | N |
| **[95] Nieuwenhuijsen, 2010** | 7 | 7 | 0 | N | N | Y | Y | 3 | OD | NR |
| **[96] Nilaweera, 2014** | 15 | 3 | 0 | P | N | Y | N | 2 | OD | U |
| **[97] Okun, 2013** | 14 | 14 | 0 | N | N | N | Y | - | - | - |
| **[98] Palmisano, 2016** | 70 | 17 | 9 | N | N | N | N | - | - | - |
| **[99] Picorelli, 2014** | 9 | 1 | 0 | N | N | N | N | - | - | - |
| **[100] Pinquart, 2010** | 76 | 76 | 0 | N | N | N | Y | - | - | - |
| **[101] Pompili, 2013** | 34 | NR | NR | Y | N | Y | N | 1 | NI | NR |
| **[102] Prang, 2015** | 14 | 12 | 0 | N | N | Y | N | 2 | OD | U |
| **[103] Prieto, 2014** | 5 | 5 | 0 | Y | N | Y | Y | 1 | OD | YO |
| **[104] Proper, 2011** | 19 | 19 | 0 | N | N | Y | N | 5 | OD | YO |
| **[105] Raggi, 2012** | 51 | 12 | 2 | N | N | Y | N | 2 | NI | NR |
| **[106] Ramond, 2011** | 18 | 18 | 0 | P | N | Y | N | 5 | OD | NR |
| **[107] Rashid, 2014** | 9 | 8 | 0 | P | N | Y | N | 3 | OD | YD |
| **[108] Rehm, 2010** | 17 | 14 | 3 | N | N | N | Y | - | - | - |
| **[109] Rossi, 2012** | 80 | NR | NR | Y | N | N | N | - | - | - |
| **[110] Sbarra, 2011** | 32 | 32 | 0 | Y | N | N | Y | - | - | - |
| **[111] Schmidt, 2015** | 36 | 31 | 0 | Y | N | N | N | - | - | - |
| **[112] Sømhovd, 2012** | 6 | 0 | 6 | N | N | Y | Y | 2 | OD | NR |
| **[113] Song, 2014** | 10 | 0 | 10 | N | N | N | Y | - | - | - |
| **[114] Stehr, 2012** | 9 | 4 | 0 | N | N | N | N | - | - | - |
| **[115] Sturmberg, 2013** | 7 | 1 | 0 | Y | N | Y | N | 2 | OD | YD |
| **[116] Surkan, 2011** | 17 | 4 | 6 | Y | N | Y | Y | 3 | NI | NR |
| **[117] Taylor, 2014** | 26 | 11 | 0 | Y | N | Y | Y | 3 | A | N |
| **[118] Tetley, 2014** | 24 | 0 | 5 | N | N | Y | N | 4 | OD | NR |

| Study[a] | N studies[b] | N cohort[c] | N case-c[d] | Guideline[e] | Protocol[f] | RoB assessment[g] | Meta-analysis[h] | Type of tool[i] | RoB incorporation[j] | RoB influence[k] |
|---|---|---|---|---|---|---|---|---|---|---|
| **[119] Thangaratinam, 2012** | 26 | 19 | 5 | Y | N | Y | N | 2 | NI | NR |
| **[120] Theorell, 2015** | 59 | 55 | 4 | Y | N | Y | Y | 1 | A | N |
| **[121] Trenchard, 2013** | 44 | 36 | 0 | Y | N | Y | N | 1 | OD | U |
| **[122] Vangeli, 2011** | 8 | 8 | 0 | N | N | N | Y | - | - | - |
| **[123] Vu, 2011** | 18 | 13 | 0 | N | N | N | Y | - | - | - |
| **[124] Wang, 2014** | 14 | 0 | 14 | N | N | Y | Y | 2 | OD | NR |
| **[125] Weierink, 2013** | 6 | 2 | 1 | N | N | Y | N | 3 | NI | NR |
| **[126] Wilson, 2014** | 11 | 11 | 0 | N | N | N | Y | - | - | - |
| **[127] Winters, 2010** | 30 | 26 | 0 | N | N | Y | N | 1 | OD | U |
| **[128] Xenaki, 2015** | 45 | NR | NR | N | N | N | N | - | - | - |
| **[129] Zhu, 2015** | 10 | 1 | 9 | N | N | Y | Y | 2 | A | N |
| **[130] Zijlmans, 2015** | 28 | 28 | 0 | Y | N | Y | N | 1 | A | N |

*The complete database is available upon request to the authors.

[a] Name of the first author.

[b] Number of primary studies included.

[c] Number of cohort studies included. NR, not reported.

[d] Number of case-control studies included. NR, not reported.

[e] Was any quality guideline followed? N, no; Y, yes; P, partially.

[f] Was the protocol of the review registered? N, no; Y, yes.

[g] Was the risk of bias of primary studies assessed? N, no; Y, yes; U, unclear.

[h] Was any meta-analysis performed? N, no; Y, yes.

[i] Type of tool used to assess risk of bias: 1, specific biases; 2, standard scale/check-list; 3, modified standard scale/check-list; 4, new scale/check-list created by the authors; 5, mixture of two or more scales/check-lists.

[j] How was risk of bias assessment incorporated into the results? OD, only descriptive (qualitative); A, analytical (quantitative); NI, not incorporated.

[k] Did the authors declare that there was some significant influence of risk of bias on the results? N, no; YO, yes on the basis of the overall risk of bias; YD, yes on the basis of the domains of bias; U, unclear; NR, not reported.

# Anexo 5: Versión de Q-Coh aplicada en el estudio

## Apéndice A del Artículo 2

**PREVIOUS CONSIDERATIONS**

**Confounding**

In the research field of this study, the most important known confounding factors affecting the initial comparability of the participants are:

_____

_____

In the research field of this study, the most important known confounding factors affecting the comparability of the participants during follow-up are:

_____

_____

**Missing data**

The acceptable percentages of missing data in confounding, exposure and outcome variables are:

_____

**Exposure variables**

_____

**Outcome variables**

_____

**Selection**

**Detail below the different associations between exposure-outcome variables addressed in this study:**

Exposure-outcome variables: _____

(I1) Exposure preceded outcome? Yes / No / Not reported

(I2) Exposure and follow-up started at the same time? Yes / No / Not reported

(I3) Relevant groups excluded? Yes / No / Not reported

Exposure-outcome variables: _____

(I1) Exposure preceded outcome? Yes / No / Not reported

(I2) Exposure and follow-up started at the same time? Yes / No / Not reported

(I3) Relevant groups excluded? Yes / No / Not reported

…

**(I1) Is it guaranteed that the exposure temporally preceded the outcome of interest?**

Yes The exposure preceded the outcome of interest.

No There is no certainty that the exposure preceded the outcome of interest.

Not reported Not enough information was reported on this issue.

**(I2) Did the exposure and follow-up start at the same time?**

Yes The start of exposure and the start of follow-up coincided for most participants, or the time elapsed between the two moments was too short to exclude individuals who experienced the outcome before the start of follow-up.

No The start of exposure and the start of follow-up did not match for a significant number of participants, which may have caused the exclusion of individuals who had experienced the outcome before the start of follow-up.

Not reported Not enough information was reported on this issue.

**(I3) Were any relevant group of potential participants excluded from the study or different inclusion criteria applied?**

Yes The exclusion (or refusal to participate) of a relevant group of potential participants or the application of different inclusion criteria may have altered the results of the association between the exposure and the outcome variable.

No No relevant group of potential participants was excluded (or refused to participate) and the same inclusion criteria were applied.

Not reported Not enough information was reported on this issue.

**Inference1**

**Selection bias** occurs when the temporary precedence of the exposure was not guaranteed (Item1), or some outcome events (Item2) or participants (Item3) were excluded from the study, and this may have affected the association or associations of interest in a way that could not be adjusted for in the analysis or corrected by limiting the scope of study findings.

Do you consider that selection bias may have occurred in this study? **Yes / No**

**Number of items:**

Bias

No Bias

Not reported

Total

## Confounding

### Confounding affecting the initial comparability of the participants

If there are confounding variables that have been measured but have not been controlled for during the study design to enhance the comparability of participants (e.g. by matching or restriction), they still can be controlled for during the data analysis stage using, for instance, stratification, adjustment, propensity scores or regression models.

**Detail below the confounding factors affecting the initial comparability that have been accounted for in this study:**

Variable: _____

(I4) Properly measured? Yes / Presumably / Doubtfully

(I5) Acceptable percentage of missing data? Yes / No / Not reported

(I6) Selective missing data? Yes / No / Not reported

Variable: _____

(I4) Properly measured? Yes / Presumably / Doubtfully

(I5) Acceptable percentage of missing data? Yes / No / Not reported

(I6) Selective missing data? Yes / No / Not reported

…

### Confounding during follow-up

In some cases, the comparability of participants may have been altered during follow-up due to changes in some of the confounding factors measured at the baseline, and/or by the emergence of other confounding factors throughout follow-up. Therefore, it is important to take these events into consideration during the data analysis.

**Detail below the confounding factors that have been accounted for during follow-up in this study:**

Variable: _____

(I4) Properly measured? Yes / Presumably / Doubtfully

(I5) Acceptable percentage of missing data? Yes / No / Not reported

(I6) Selective missing data? Yes / No / Not reported

Variable: _____

(I4) Properly measured? Yes / Presumably / Doubtfully

(I5) Acceptable percentage of missing data? Yes / No / Not reported

(I6) Selective missing data? Yes / No / Not reported

…

**(I4) Was the confounding variable properly measured?**

Yes All the tools used to measure the confounding variable were validated and evidences of their validity were reported, as well as reliability indices with acceptable values were provided when necessary (e.g. inter-rater or test-retest reliability).

Presumably No evidences of validity or reliability indices were reported for all the measurement tools but they seem to be appropriate to measure the confounding variable, for example, because the variable is objective enough to make the validation of the tools unnecessary, or the measurement tools are well known and broadly used in the field (their suitability is assumed).

Doubtfully Not all the tools used to measure the confounding variable seem appropriate, or the construct measured by tools do not match the definition of the confounding variable.

Note: If a measurement tool was validated in a different study (with a different sample) it is important to consider how comparable the two samples are (the one used to validate the tool and the one under study). If the procedure to measure the confounding variable was not the same for all participants and this may have altered the results of the study, answer “Doubtfully”.

**(I5) Was the percentage of missing data in the confounding variable acceptable?**

Yes The percentage of missing data in the confounding variable was low enough to assume that it is unlikely that the study results based in this variable have been altered.

No The percentage of missing data in the confounding variable was too high to be acceptable and this probably altered the study results.

Not reported The percentage of missing data in the confounding variable was not assessed or not reported and cannot be calculated.

**(I6) Were those participants with registered value in the confounding variable different from those without value?**

Yes The differences between those participants with registered value in the confounding variable and those without value were important and this may have altered the study results.

No The differences between those participants with registered value in the confounding variable and those without value were not important enough to suppose that this may have altered the study results.

Not reported Not enough information is given to compare those participants with registered value in the confounding variable and those without value.

Note: In case that the participants were categorized into groups, it must be judged if their comparability may have been affected by selective missing data. While a low percentage of missing data is desired, it is important to make sure

that the underlying reasons are similar as they might point out systematic differences between the groups, and then this item should be answered “Yes”. If the reasons for missing data were not reported, differences in the missing percentages may be an indicator of systematic differences between the groups, and then this item should also be answered “Yes”.

**(I7) Is the measure of any of the most important known confounding variables in this research field missing?**

Yes

No

Write down the confounding variables that have not been measured:

_____

**Inference 2**

**Bias due to confounding** occurs when important factors related to the outcomes of interest were not properly measured (Item4), their missing data may have altered the results (Item5, Item 6), or were not controlled for by design or adjusted in the analysis, and this may have affected the comparability of the participants at baseline or during follow-up.

Do you consider that confounding bias may have occurred in this study? **Yes / No**

**Number of items:**

Bias

No Bias

Not reported

Total

**Exposure measures**

**Detail below the exposure variables in this study:**

Variable: _____

(I8) Properly measured? Yes / Presumably / Doubtfully

(I9) Acceptable percentage of missing data? Yes / No / Not reported

(I10) Selective missing data? Yes / No / Not reported

Variable: _____

(I8) Properly measured? Yes / Presumably / Doubtfully

(I9) Acceptable percentage of missing data? Yes / No / Not reported

(I10) Selective missing data? Yes / No / Not reported

…

**(I8) Was the exposure variable properly measured?**

Yes All the tools used to measure the exposure variable were validated and evidences of their validity were reported, as well as reliability indices with acceptable values were provided when necessary (e.g. inter-rater or test-retest reliability).

Presumably No evidences of validity or reliability indices were reported for all the measurement tools but they seem to be appropriate to measure the exposure, for example, because the exposure variable is objective enough to make the validation of the tools unnecessary, or the measurement tools are well known and broadly used in the field (their suitability is assumed).

Doubtfully Not all the tools used to measure the exposure variable seem appropriate, or the construct measured by the measuring tools do not match the definition of the exposure variable.

Note: If a measurement tool was validated in a different study (with a different sample) it is important to consider how comparable the two samples are (the one used to validate the tool and the one under study). If the procedure to measure the exposure variable was not the same for all participants and this may have altered the results of the study, answer “Doubtfully”.

**(I9) Was the percentage of missing data in the exposure variable acceptable?**

Yes The percentage of missing data in the exposure variable was low enough to assume that it is unlikely that the study results based in this variable have been altered.

No The percentage of missing data in the exposure variable was too high to be acceptable and this probably altered the study results.

Not reported The percentage of missing data in the exposure variable was not assessed or not reported and cannot be calculated.

**(I10) Were those participants with registered value in the exposure variable different from those without value?**

Yes The differences between those participants with registered value in the exposure variable and those without value were important and this may have the study results.

No The differences between those participants with registered value in the exposure variable and those without value were not important enough to suppose that this may have altered the study results.

Not reported Not enough information is given to compare those participants with registered value in the exposure variable and those without value.

Note: In case that the participants were categorized into groups, it must be judged if their comparability may have been affected by selective missing data. While a low percentage of missing data is desired, it is important to make sure that the underlying reasons are similar, as they might point out systematic differences between the groups, and then this item should be answered “Yes”. If the reasons for missing data were not reported, differences in the missing percentages may be an indicator of systematic differences between the groups, and then this item should also be answered “Yes”.

**Inference 3**

**Bias in the measurement of exposures** occurs when exposure variables were not properly measured (Item7) or missing data may have altered the results (Item8, Item9).

Do you consider that bias in the measurement of exposures may have occurred in this study? **Yes / No**

**Number of items:**

Bias

No Bias

Not reported

Total

**Outcome measures**

**Detail below the outcome variables in this study:**

Variable: _____

(I11) Properly measured? Yes / Presumably / Doubtfully

(I12) Blinding to the exposure status of the participants? Yes/ Not necessary/No

(I13) Long enough follow-up time? Yes / Presumably / Doubtfully

(I14) Acceptable percentage of missing data? Yes / No / Not reported

(I15) Selective missing data? Yes / No / Not reported

Variable: _____

(I11) Properly measured? Yes / Presumably / Doubtfully

(I12) Blinding to the exposure status of the participants? Yes/ Not necessary/No

(I13) Long enough follow-up time? Yes / Presumably / Doubtfully

(I14) Acceptable percentage of missing data? Yes / No / Not reported

(I15) Selective missing data? Yes / No / Not reported

…

**(I11) Was the outcome variable of interest properly measured?**

Yes All the tools used to measure the outcome variable were validated and evidences of their validity were reported, as well as reliability indices with acceptable values were provided when necessary (e.g. inter-rater or test-retest reliability).

Presumably No evidences of validity or reliability indices were reported for all the measurement tools but they seem to be appropriate to measure the outcome variable, for example, because the outcome variable is objective enough to make the validation of the tools unnecessary, or the measurement tools are well known and broadly used in the field (their suitability is assumed).

Doubtfully Not all the tools used to measure the outcome variable seem appropriate, or the construct measured by the measuring tools do not match the definitions of the outcome variable.

Note: If a measurement tool was validated in a different study (with a different sample) it is important to considerer how comparable the two samples are (the one used to validate the tool and the one under study). If the procedure to measure the outcome variable was not the same for all participants and this may have altered the results of the study, answer “Doubtfully”.

**(I12) Were those assessing the outcome successfully blinded to the exposure status of the participants?**

Yes Some procedure has been successfully used to blind those assessing the outcome to the exposure status of the participants.

Not necessary It is unlikely that the outcome may have been affected by the assessors' knowledge of the participant's condition (for instance, when the outcome is objective enough).

No Although necessary, no blinding procedure was successfully applied or reported.

**(I13) Was follow-up long enough to detect the effect of the exposure variables on the outcome?**

Yes The follow-up time was long enough to detect the effect of the exposure variables on the outcome.

Presumably The follow-up time seemed or was said to be long enough to detect the effect of the exposure variables on the outcome, but this was not verified.

Doubtfully The follow-up time could not be long enough to detect the effect of the exposure variables on the outcomes.

Note: If effects of more than one exposure variable on the outcome were measured, detail below in which of the exposure variables follow-up was not long enough to detect effects:

_____

**(I14) Was the percentage of missing data in the outcome variable acceptable?**

Yes The percentage of missing data in the outcome variable was low enough to assume that it is unlikely that the study results based in this variable have been altered.

No The percentage of missing data in the outcome variable was too high to be acceptable and this probably altered the study results.

Not reported The percentage of missing data in the outcome variable was not assessed or not reported and cannot be calculated.

Note: In the case that the dropout rates were not given for the different groups defined by the exposure, a low drop out is necessary to have enough confidence that the comparability of the groups remains at the end of the study.

**(I15) Were those participants with registered value in the outcome variable different from those without value?**

Yes The differences between those participants with registered value in the outcome variable and those without value were important and this may have altered the study results.

No The differences between those participants with registered value in the outcome variable and those without value were not important enough to suppose that this may have altered the study results.

Not reported Not enough information is given to compare those participants with registered value in the outcome variable and those without value.

Note: In case that the participants were categorized into groups, it must be judged if their comparability may have been affected by selective missing data or selective dropouts. While a low percentage of missing data or dropouts is desired, it is important to make sure that the underlying reasons are similar, as they might point out systematic differences between the groups, and then this item should be answered “Yes”. If the reasons for missing data or dropping out of the participants of the study are not reported, differences in the missing percentages or dropout rates may be an indicator of systematic differences between the groups, and then this item should also be answered “Yes”.

**Inference 4**

**Bias in the measurement of outcomes** occurs when outcome variables were not properly measured (Item10), when the lack of blinding of those assessing the outcomes may have affected the results (Item11), when follow-up was not long enough (Item12), or when missing data or dropouts may have altered the results (Item 13, Item14).

Do you consider that bias in the measurement of outcomes may have occurred in this study? **Yes / No**

**Number of items:**

Bias

No Bias

Not reported

Total

**Overall risk of bias**

Considering the responses given to the previous items and to the four potential biases (selection, confounding, exposure and outcome), the overall risk of bias for this study is:

**Low risk of bias** The study was well conducted and it seems to provide strong evidence.

**Moderate risk of bias** The study was generally well conducted but it shows some flaws that may have affected the results.

**High risk of bias** The study shows some major flaws and it may not provide strong evidence.

**Number of items:**

Bias

No Bias

Not reported

Total

# Anexo 6: Referencias de los estudios primarios incluidos en el metaanálisis

Apéndice B del Artículo 2

Appendix B: Studies included in Pan et al. (2011) meta-analysis

[1] Vogt T, Pope C, Mullooly J, Hollis J. Mental health status as a predictor of morbidity and mortality: a 15-year follow-up of members of a health maintenance organization. Am J Public Heal 1994;84:227–31.

[2] Wassertheil-Smoller S, Applegate WB, Berge K, Chang CJ, Davis BR, Grimm R, et al. Change in depression as a precursor of cardiovascular events. Arch Intern Med 1996;156:553–61. doi:10.1001/archinte.156.5.553.

[3] Everson SA, Roberts RE, Goldberg DE, Kaplan GA. Depressive symptoms and increased risk of stroke mortality over a 29-year period. Arch Intern Med 1998;158:1133–8. doi:10.1001/archinte.158.10.1133.

[4] Simons LA, McCallum J, Friedlander Y, Simons J. Risk Factor for Ishemic Stroke: Dubbo Study of the Elderly. J Am Hear Soc 1998;29. doi:http://dx.doi.org/10.1161/01.STR.29.7.1341.

[5] Whooley MA, Browner WS. Association between depressive symptoms and mortality in older women. Arch Intern Med 1998;158:2129–35.

[6] Jonas B, Mussolino M. Symptoms of depression as a prospective risk factor for stroke. Psychosom Med 2000;62:463–71.

[7] Larson SL, Owens PL, Ford D, Eaton W. Depressive disorder, dysthymia, and risk of stroke: thirteen-year follow-up from the Baltimore epidemiologic catchment area study. Stroke 2001;32:1979–83.

[8] Ohira T, Iso H, Satoh S, Sankai T, Tanigawa T, al et. Prospective study of depressive symptoms and risk of stroke among Japanese. Stroke 2001;32:903–8.

[9] Ostir G V, Markides KS, Peek MK, Goodwin JS. The Association Between Emotional Well-Being and the Incidence of Stroke in Older Adults. Psychosom Med 2001;63:210–5. doi:10.1097/00006842-200103000-00003.

[10] May M, McCarron P, Stansfeld S, Ben-Shlomo Y, Gallacher J, Yarnell J, et al. Does psychological distress predict the risk of ischemic stroke and transient ischemic attack? The Caerphilly study. Stroke 2002;33:7–12.

[11] Yasuda N, Mino Y, Koda S, Ohara H. Symptoms , as Assessed by the General Health Questionnaire , on Cause of Death in Older Persons Living in a Rural. J Am Geriatics Soc 2002;50:313–20.

[12] Wassertheil-Smoller S, Shumaker S, Ockene J, Talavera GA, Greenland P, Cochrane B, et al. Depression and Cardiovascular Sequelae in Postmenopausal Women. Arch Intern Med 2004;164:289. doi:10.1001/archinte.164.3.289.

[13] Gump BB, Matthews K a, Eberly LE, Chang Y. Depressive symptoms and mortality in men: results from the Multiple Risk Factor Intervention Trial. Stroke 2005;36:98–102. doi:10.1161/01.STR.0000149626.50127.d0.

[14] Avendano M, Kawachi I, Lenthe F Van, Boshuizen HC, Mackenbach JP, Van Den Bos GAM, et al. Socioeconomic status and stroke incidence in the US elderly: The role of risk factors in the EPESE study. Stroke 2006;37:1368–73. doi:10.1161/01.STR.0000221702.75002.66.

[15] Stürmer T, Hasselbach P, Amelang M. Personality, lifestyle, and risk of cardiovascular disease and cancer: follow-up of population based cohort. BMJ 2006;332:1359. doi:10.1136/bmj.38833.479560.80.

[16] Arbelaez JJ, Ariyo AA, Crum RM, Fried LP, Ford DE. Depressive symptoms, inflammation, and ischemic stroke in older adults: A prospective analysis in the cardiovascular health study. J Am Geriatr Soc 2007;55:1825–30. doi:10.1111/j.1532-5415.2007.01393.x.

[17] Kawamura T, Shioiri T, Takahashi K, Ozdemir V, Someya T. Survival Rate and Causes of Mortality in the Elderly with Depression. J Investig Med 2007;55:106–14. doi:10.2310/6650.2007.06040.

[18] Salaycik KJ, Kelly-Hayes M, Beiser A, Nguyen A-H, Brady SM, Kase CS, et al. Depressive Symptoms and Risk of Stroke: The Framingham Study. Stroke 2007;38:16–21. doi:10.1161/01.STR.0000251695.39877.ca.

[19] Bos MJ, Lindén T, Koudstaal PJ, Hofman a, Skoog I, Breteler MMB, et al. Depressive symptoms and risk of stroke: the Rotterdam Study. J Neurol Neurosurg Psychiatry 2008;79:997–1001. doi:10.1136/jnnp.2007.134965.

[20] Lee HC, Lin HC, Tsai SY. Severely Depressed Young Patients Have Over Five Times Increased Risk for Stroke: A 5-Year Follow-Up Study. Biol Psychiatry 2008;64:912–5. doi:10.1016/j.biopsych.2008.07.006.

[21] Liebetrau M, Steen B, Skoog I. Depression as a risk factor for the incidence of first-ever stroke in 85-year-olds. Stroke 2008;39:1960–5. doi:10.1161/STROKEAHA.107.490797.

[22] Surtees PG, Wainwright NWJ, Luben RN, Wareham NJ, Bingham SA, Khaw K-T. Psychological distress, major depressive disorder, and risk of stroke. Neurology 2008;70:788–94. doi:10.1212/01.wnl.0000304109.18563.81.

[23] Whooley MA, de Jonge P, Vittinghoff E, Otte C, Moos R, Carney RM, et al. Depressive Symptoms, Health Behaviors, and Risk of Cardiovascular Events in Patients With Coronary Heart Disease. JAMA 2008;300:2379. doi:10.1001/jama.2008.711.

[24] Wouts L, Voshaar RCO, Bremmer MA, Buitelaar JK, Penninx BWJH, Beekman ATF. Cardiac Disease, Depressive Symptoms, and Incident Stroke in an Elderly Population. Arch Gen Psychiatry 2008;65:596. doi:10.1001/archpsyc.65.5.596.

[25] Glymour MM, Maselko JM, Gilman SE, Patton KK, Avendano M. Depressive Symptoms Predict Incident Stroke Independently of Memory Impairments. Neurology 2010;75:2063–70.

[26] Nabi H, Kivimäki M, Suominen S, Koskenvuo M, Singh-Manoux A, Vahtera J. Does depression predict coronary heart disease and cerebrovascular disease equally well? The health and social support prospective cohort study. Int J Epidemiol 2010;39:1016–24. doi:10.1093/ije/dyq050.

[27] Peters R, Pinto E, Beckett N, Swift C, Potter J, McCormack T, et al. Association of depression with subsequent mortality, cardiovascular morbidity and incident dementia in people aged 80 and over and suffering from hypertension. Data from the Hypertension in the Very Elderly Trial (HYVET). Age Ageing 2010;39:439–45. doi:10.1093/ageing/afq042.

[28] Pan A, Okereke OI, Sun Q, Logroscino G, Manson JE, Willett WC, et al. Depression and incident stroke in women. Stroke 2011;42:2770–5. doi:10.1161/STROKEAHA.111.617043.

# Anexo 7: Escala de evaluación de la aplicabilidad de las herramientas

Apéndice C del Artículo 2

Appendix C: Usability of the tools

Rate in a five-point scale (where 1 is “very poor” and 5 is “very good”) the following aspects of every tool:

1- Coverage of the tool: Does the tool cover all important domains? Does the tool include items that are not relevant? (e.g. external validity items). Were any important points omitted?

2- Ease of use:
   a. Clarity of instructions
   b. Clarity of items

3- Discriminative ability: Does the tool distinguish between different levels of risk of bias?

| Item | NOS | Q-Coh | ROBINS-I |
|---|---|---|---|
| 1 | 1 2 3 4 5 | 1 2 3 4 5 | 1 2 3 4 5 |
| 2a | 1 2 3 4 5 | 1 2 3 4 5 | 1 2 3 4 5 |
| 2b | 1 2 3 4 5 | 1 2 3 4 5 | 1 2 3 4 5 |
| 3 | 1 2 3 4 5 | 1 2 3 4 5 | 1 2 3 4 5 |

# Anexo 8: Resultados del acuerdo entre evaluadores

## Apéndice D del Artículo 2

Appendix D

Table D1. Results of inter-rater agreement

| Tool | Items and domains | P. Overall agreement | Response category 1 | | | Response category 2 | | | Response category 3 | | | Response category 4 | | | Response category 5 | | | Kappa/ PABAK |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Resp. | P. cat. | P. agree. | Resp. | P. cat. | P. agree. | Resp. | P. cat. | P. agree. | Resp. | P. cat. | P. agree. | Resp. | P. cat. | P. agree. | |
| NOS[a] | Selection 1 | .46 | a | .30 | .47 | b | .38 | .38 | c | .25 | .71 | d | .07 | .00 | | | | .24 |
| | Selection 2 | .96 | a | .98 | .98 | b | .02 | .00 | c | .00 | | | | | | | | .93[b] |
| | Selection 3 | .96 | a | .02 | .00 | b | .48 | .96 | c | .50 | 1.00 | d | .00 | | | | | .93[b] |
| | Selection 4 | .75 | a | .73 | .83 | b | .27 | .53 | | | | | | | | | | .50[b] |
| | Comparability a | .93 | a | .96 | .96 | b | .04 | .00 | | | | | | | | | | .86[b] |
| | Comparability b | 1.00 | a | .96 | 1.00 | b | .04 | 1.00 | | | | | | | | | | 1.00[b] |
| | Outcome 1 | .68 | a | .13 | .57 | b | .70 | .77 | c | .16 | .44 | d | .02 | .00 | | | | .36[b] |
| | Outcome 2 | .96 | a | .80 | .98 | b | .20 | .91 | | | | | | | | | | .93[b] |
| | Outcome 3 | .61 | a | .27 | .40 | b | .64 | .72 | c | .07 | .50 | d | .02 | .00 | | | | .21[b] |
| Q-Coh | Item 1 | .79 | Y | .77 | .88 | N | .16 | .67 | NR | .07 | .00 | | | | | | | .57[b] |
| | Item 2 | .96 | Y | .23 | .92 | N | .00 | | NR | .77 | .98 | | | | | | | .93[b] |
| | Item 3 | .46 | Y | .30 | .35 | N | .39 | .55 | NR | .30 | .47 | | | | | | | .23 |
| | Selection | .61 | Y | .48 | .59 | N | .52 | .62 | | | | | | | | | | .32 |
| | Confounding[c] | .82 | Y | .70 | .87 | N | .30 | .71 | | | | | | | | | | .64[b] |
| | Item 8 | .79 | Y/P | .89 | .88 | D | .11 | .00 | | | | | | | | | | .57[b] |
| | Item 9 | .64 | Y | .79 | .77 | N | .05 | .00 | NR | .16 | .22 | | | | | | | .29[b] |
| | Item 10 | .39 | Y | .13 | .00 | N | .41 | .35 | NR | .46 | .54 | | | | | | | .02 |
| | Exposure | .50 | Y | .29 | .13 | N | .71 | .65 | | | | | | | | | | -.20 |
| | Item 11 | .71 | Y/P | .86 | .83 | D | .14 | .00 | | | | | | | | | | .43[b] |
| | Item 12 | .89 | Y/NN | .95 | .94 | N | .05 | .00 | | | | | | | | | | .79[b] |
| | Item 13 | .93 | Y/P | .86 | .96 | D | .14 | .75 | | | | | | | | | | .86[b] |
| | Item 14 | .79 | Y | .84 | .89 | N | .09 | .40 | NR | .07 | .00 | | | | | | | .57[b] |
| | Item 15 | .39 | Y | .09 | .00 | N | .43 | .42 | NR | .48 | .44 | | | | | | | -.05 |
| | Outcome | .64 | Y | .29 | .38 | N | .71 | .75 | | | | | | | | | | .29[b] |
| | Overall RoB | .25 | L | .16 | .22 | M | .41 | .17 | H | .43 | .33 | | | | | | | -.16 |
| ROBINS-I[d] | 1.1 | 1.00 | Y/PY | 1.00 | 1.00 | N/PN | .00 | | | | | | | | | | | 1.00[b] |
| | 1.2 | .93 | NA | .00 | | Y/PY | .04 | .00 | N/PN | .96 | .96 | NI | .00 | | | | | .86[b] |
| | 1.3 | .93 | NA | .96 | .96 | Y/PY | .00 | | N/PN | .04 | .00 | NI | .00 | | | | | .86[b] |
| | 1.4 | .79 | NA | .00 | | Y/PY | .32 | .67 | N/PN | .68 | .84 | NI | .00 | | | | | .57[b] |
| | 1.5 | .68 | NA | .00 | | Y/PY | .77 | .79 | N/PN | .20 | .36 | NI | .04 | .00 | | | | .36[b] |
| | 1.6 | .96 | NA | .00 | | Y/PY | .02 | .00 | N/PN | .98 | .98 | NI | .00 | | | | | .93[b] |
| | 1.7 | 1.00 | NA | 1.00 | 1.00 | Y/PY | .00 | | N/PN | .00 | | NI | .00 | | | | | 1.00[b] |
| | 1.8 | 1.00 | NA | 1.00 | 1.00 | Y/PY | .00 | | N/PN | .00 | | NI | .00 | | | | | 1.00[b] |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Confounding | .75 | L | .02 | .00 | M | .21 | .50 | S | .77 | .84 | C | .00 | | NI | .00 | | .50[b] |
| 2.1 | .82 | Y/PY | .09 | .00 | N/PN | .91 | .90 | NI | .00 | | | | | | | | .64[b] |
| 2.2 | .82 | NA | .91 | .90 | Y/PY | .05 | .00 | N/PN | .02 | .00 | NI | .02 | .00 | | | | .64[b] |
| 2.3 | .82 | NA | .91 | .90 | Y/PY | .05 | .00 | N/PN | .02 | .00 | NI | .02 | .00 | | | | .64[b] |
| 2.4 | .96 | Y/PY | .05 | .67 | N/PN | .00 | | NI | .95 | .98 | | | | | | | .93[b] |
| 2.5 | .89 | NA | .95 | .94 | Y/PY | .02 | .00 | N/PN | .04 | .00 | NI | .00 | | | | | .79[b] |
| Selection | .82 | L | .05 | .00 | M | .89 | .92 | S | .04 | .00 | C | .02 | .00 | NI | .00 | | .64[b] |
| 3.1 | .82 | Y/PY | .91 | .90 | N/PN | .09 | .00 | NI | .00 | | | | | | | | .64[b] |
| 3.2 | .96 | Y/PY | .98 | .98 | N/PN | .02 | .00 | NI | .00 | | | | | | | | .93[b] |
| 3.3 | 1.00 | Y/PY | 1.00 | 1.00 | N/PN | .00 | | NI | .00 | | | | | | | | 1.00[b] |
| Intervention | .79 | L | .88 | .90 | M | .04 | .00 | S | .09 | .00 | C | .00 | | NI | .00 | | .57[b] |
| 5.1 | .75 | Y/PY | .82 | .87 | N/PN | .16 | .22 | NI | .02 | .00 | | | | | | | .50[b] |
| 5.2 | .50 | Y/PY | .50 | .57 | N/PN | .43 | .50 | NI | .07 | .00 | | | | | | | .14 |
| 5.3 | .43 | Y/PY | .39 | .45 | N/PN | .52 | .48 | NI | .09 | .00 | | | | | | | .00 |
| 5.4 | .39 | NA | .36 | .40 | Y/PY | .09 | .40 | N/PN | .14 | .25 | NI | .41 | .43 | | | | .11 |
| 5.5 | .54 | NA | .36 | .40 | Y/PY | .04 | .00 | N/PN | .61 | .65 | NI | .00 | | | | | .08 |
| Missing data | .36 | L | .46 | .54 | M | .27 | .27 | S | .18 | .20 | C | .00 | | NI | .09 | .00 | .08 |
| 6.1 | 1.00 | Y/PY | .00 | | N/PN | 1.00 | 1.00 | NI | .00 | | | | | | | | 1.00[b] |
| 6.2 | .86 | Y/PY | .29 | .75 | N/PN | .68 | .89 | NI | .04 | 1.00 | | | | | | | .71[b] |
| 6.3 | 1.00 | Y/PY | .96 | 1.00 | N/PN | .00 | | NI | .04 | 1.00 | | | | | | | 1.00 |
| 6.4 | 1.00 | Y/PY | .96 | 1.00 | N/PN | .00 | | NI | .04 | 1.00 | | | | | | | 1.00 |
| Outcome | .96 | L | .95 | .98 | M | .00 | | S | .02 | .00 | C | .00 | | NI | .04 | 1.00 | .93[b] |
| 7.1 | .75 | Y/PY | .13 | .00 | N/PN | .88 | .86 | NI | .00 | | | | | | | | .50[b] |
| 7.2 | .57 | Y/PY | .21 | .00 | N/PN | .79 | .73 | NI | .00 | | | | | | | | .14[b] |
| 7.3 | .89 | Y/PY | .02 | .00 | N/PN | .95 | .94 | NI | .04 | .00 | | | | | | | .79[b] |
| Reported result | .64 | L | .00 | | M | .82 | .78 | S | .18 | .00 | C | .00 | | NI | .00 | | .29[b] |
| Overall RoB | .79 | L | .00 | | M | .18 | .40 | S | .82 | .87 | C | .00 | | NI | .00 | | .57[b] |

Note: P. Overall agreement = proportion of agreement between raters; Resp. = response; P. cat. = proportion of choices of the raters for a specific response category; P. agree. = proportion of specific agreement between raters for each response category; Kappa = Fleiss Kappa (or PABAK). Y = yes; PY = probably yes; N = no; PN = probably not; NA = not applicable; NI = no information; L = low; M = moderate; S = serious; C = critical; RoB = Risk of Bias; P = presumably; D = doubtfully; NN = not necessary; H = high.

[a]For the full text of response options, see Wells, G., Shea, B., O'Connell, D., & Peterson, J. (2000).

[b]PABAK

[c]Agreement for items of confounding domain (4 to 7) could not be calculated.

[d]Domain 4 of ROBINS-I "Bias due to deviations from intended interventions" was considered not applicable.

# Anexo 9: Resultados de los análisis de subgrupos y meta-regresiones

Apéndice E del Artículo 2

Appendix E

Table E1. Results of subgroup analyses for overall RoB and domains for each tool

| | Tool | Subgroup | n | HR (95% CI) | Z-value | P-value | Q-statistic | P-value for heterogeneity | $I^2$ value | P-value between subgroups |
|---|---|---|---|---|---|---|---|---|---|---|
| Overall RoB | NOS | Low RoB | 23 | 1.51 (1.30, 1.77) | 5.27 | < .001 | 76.30 | < .001 | 71% | .179 |
| | | Moderate RoB | 8 | 1.30 (1.12, 1.52) | 3.36 | .001 | 10.88 | .144 | 36% | |
| | Q-Coh | Low RoB | 3 | 1.30 (0.80, 2.12) | 1.07 | .286 | 7.03 | .030 | 72% | .658 |
| | | Moderate RoB | 12 | 1.56 (1.18, 2.07) | 3.13 | .002 | 48.26 | <.001 | 77% | |
| | | High RoB | 16 | 1.36 (1.21, 1.54) | 5.09 | < .001 | 29.49 | .014 | 49% | |
| | ROBINS-I | Moderate RoB | 4 | 1.37 (0.93, 2.02) | 1.58 | .114 | 7.28 | .063 | 59% | .757 |
| | | Serious RoB | 27 | 1.46 (1.29, 1.65) | 5.96 | < .001 | 80.53 | < .001 | 68% | |
| Confounding | NOS | No RoB | 29 | 1.47 (1.31, 1.66) | 6.31 | < .001 | 86.19 | < .001 | 68% | .147 |
| | | Yes RoB | 2 | 1.12 (0.79, 1.59) | 0.63 | .531 | 0.90 | .343 | 0% | |
| | Q-Coh | No Rob | 9 | 1.29 (1.15, 1.44) | 4.36 | < .001 | 12.82 | .118 | 38% | .073 |
| | | Yes RoB | 22 | 1.56 (1.30, 1.87) | 4.82 | < .001 | 70.36 | < .001 | 70% | |
| | ROBINS-I | Moderate RoB | 8 | 1.30 (1.11, 1.52) | 3.29 | .001 | 12.59 | .083 | 44% | .160 |
| | | Serious RoB | 23 | 1.53 (1.31, 1.79) | 5.29 | < .001 | 74.77 | < .001 | 71% | |
| Selection | NOS | No RoB | 20 | 1.39 (1.22, 1.59) | 4.99 | < .001 | 38.54 | .005 | 51% | .522 |
| | | Yes RoB | 10 | 1.58 (1.22, 2.06) | 3.43 | .001 | 48.47 | < .001 | 81% | |
| | | Missing | 1 | 1.21 (0.80, 1.83) | 0.90 | .367 | NA | NA | NA | |
| | Q-Coh | No RoB | 19 | 1.49 (1.26, 1.76) | 4.65 | < .001 | 66.77 | < .001 | 73% | .449 |
| | | Yes RoB | 12 | 1.36 (1.16, 1.59) | 3.85 | < .001 | 20.76 | .036 | 47% | |
| | ROBINS-I | Low RoB | 3 | 2.32 (1.00, 5.41) | 1.95 | .051 | 20.22 | < .001 | 90% | .212 |
| | | Moderate RoB | 28 | 1.35 (1.23, 1.49) | 6.14 | < .001 | 49.31 | .005 | 45% | |
| Exposure | NOS | No RoB | 15 | 1.61 (1.27, 2.05) | 3.95 | < .001 | 58.90 | < .001 | 76% | .113 |
| | | Yes RoB | 16 | 1.31 (1.18, 1.45) | 5.04 | < .001 | 23.16 | .081 | 35% | |
| | Q-Coh | No RoB | 25 | 1.50 (1.29, 1.73) | 5.39 | < .001 | 79.35 | < .001 | 70% | .157 |
| | | Yes RoB | 6 | 1.28 (1.10, 1.50) | 3.13 | .002 | 7.50 | .186 | 33% | |
| | ROBINS-I | Low RoB | 29 | 1.47 (1.29, 1.67) | 5.75 | < .001 | 87.58 | < .001 | 68% | .205 |
| | | Serious RoB | 2 | 1.30 (1.14, 1.48) | 3.91 | < .001 | 0.29 | .591 | 0% | |

| | Tool | Subgroup | n | HR (95% CI) | Z-value | P-value | Q-statistic | P-value for heterogeneity | $I^2$ value | P-value between subgroups |
|---|---|---|---|---|---|---|---|---|---|---|
| Outcome | NOS | No RoB | 17 | 1.48 (1.22, 1.79) | 3.97 | < .001 | 58.16 | < .001 | 72% | .637 |
| | | Yes RoB | 14 | 1.39 (1.21, 1.61) | 4.62 | < .001 | 28.08 | .009 | 54% | |
| | Q-Coh | No RoB | 21 | 1.51 (1.25, 1.83) | 4.32 | < .001 | 68.72 | < .001 | 71% | .244 |
| | | Yes RoB | 10 | 1.33 (1.18, 1.49) | 4.63 | < .001 | 15.90 | .069 | 43% | |
| | ROBINS-I | Low RoB | 29 | 1.37 (1.24, 1.52) | 6.12 | < .001 | 50.04 | .006 | 44% | <.001 |
| | | Serious RoB | 1 | 1.29 (1.13, 1.48) | 3.70 | < .001 | NA | NA | NA | |
| | | No information | 1 | 5.43 (3.47, 8.50) | 7.39 | < .001 | NA | NA | NA | |
| Missing data | NOS | No RoB | 26 | 1.46 (1.27, 1.67) | 5.39 | < .001 | 82.69 | < .001 | 70% | .376 |
| | | Yes RoB | 4 | 1.57 (1.06, 2.31) | 2.26 | .024 | 4.81 | .186 | 38% | |
| | | Missing | 1 | 1.29 (1.13, 1.48) | 3.70 | < .001 | NA | NA | NA | |
| | Q-Coh | No RoB | 7 | 1.81 (1.18, 2.76) | 2.74 | .006 | 39.79 | < .001 | 85% | .320 |
| | | Yes RoB | 6 | 1.30 (1.16, 1.46) | 4.41 | <. 001 | 3.82 | .575 | 0% | |
| | | Missing | 18 | 1.37 (1.19, 1.57) | 4.46 | < .001 | 38.80 | .002 | 56% | |
| | ROBINS-I | Low RoB | 13 | 1.57 (1.23, 2.00) | 3.62 | < .001 | 50.58 | < .001 | 76% | .603 |
| | | Moderate RoB | 12 | 1.35 (1.16, 1.58) | 3.91 | < .001 | 22.19 | .023 | 50% | |
| | | Serious RoB | 6 | 1.42 (1.08, 1.86) | 2.52 | .012 | 12.30 | .031 | 59% | |
| Selective reporting | ROBINS-I | Moderate RoB | 26 | 1.48 (1.26, 1.72) | 4.95 | < .001 | 81.47 | < .001 | 69% | .150 |
| | | Serious RoB | 5 | 1.29 (1.17, 1.42) | 5.19 | < .001 | 4.51 | .341 | 11% | |

Note: Risk of bias domains are based on the classification presented in the Table 2 of the main report. Analyses were performed following a mixed effects model. RoB = risk of bias; n = number of effect sizes in each category; HR = hazard ratio; CI = confidence interval; NA = not applicable.

Table E2. Results of meta-regression analyses for overall RoB and domains for each tool

| | Tool | Category | n | β (95% CI) | Test of the model | | Goodness of fit | | | $R^2$ analog |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | F | P-value | Q | P-value | $I^2$ value | |
| Overall RoB | NOS[a] | Intercept | - | 0.23 (-0.37, 0.84) | 0.07 | .796 | 49.83 | .007 | 44% | 0% |
| | | Overall score | - | 0.01 (-0.07, 0.09) | | | | | | |
| | NOS | Low RoB (intercept) | 23 | 0.33 (0.19, 0.46) | 0.22 | .644 | 49.93 | .007 | 44% | 0% |
| | | Moderate RoB | 8 | -0.06 (-0.30, 0.19) | | | | | | |
| | Q-Coh | Low RoB (intercept) | 3 | 0.27 (-0.21, 0.76) | 0.28 | .757 | 84.78 | < .001 | 67% | 0% |
| | | Moderate RoB | 12 | 0.17 (-0.38, 0.71) | | | | | | |
| | | High RoB | 16 | 0.07 (-0.45, 0.60) | | | | | | |
| | ROBINS-I | Moderate RoB (intercept) | 4 | 0.31 (-0.12, 0.75) | 0.08 | .778 | 87.80 | < .001 | 67% | 0% |
| | | Serious RoB | 27 | 0.06 (-0.40, 0.53) | | | | | | |
| Confounding | NOS | No RoB (intercept) | 29 | 0.41 (0.25, 0.58) | 1.10 | .303 | 86.31 | < .001 | 69% | 0% |
| | | Yes RoB | 2 | -0.34 (-1.00, 0.32) | | | | | | |
| | Q-Coh | No RoB (intercept) | 9 | 0.26 (0.02, 0.51) | 1.25 | .273 | 83.17 | < .001 | 65% | 0% |
| | | Yes RoB | 22 | 0.17 (-0.14, 0.47) | | | | | | |
| | ROBINS-I | Moderate RoB (intercept) | 8 | 0.27 (-0.00, 0.55) | 0.81 | .375 | 87.36 | < .001 | 67% | 0% |
| | | Serious RoB | 23 | 0.14 (-0.18, 0.47) | | | | | | |
| Selection | NOS | No RoB (intercept) | 20 | 0.35 (0.16, 0.55) | 0.59 | .450 | 84.71 | < .001 | 68% | 0% |
| | | Yes RoB | 10 | 0.13 (-0.22, 0.48) | | | | | | |
| | Q-Coh | No RoB (intercept) | 19 | 0.39 (0.21, 0.58) | 0.13 | .721 | 87.54 | < .001 | 67% | 0% |
| | | Yes RoB | 12 | -0.05 (-0.36, 0.25) | | | | | | |
| | ROBINS-I | Low RoB (intercept) | 3 | 0.85 (0.42, 1.29) | 5.80 | .023 | 69.53 | < .001 | 58% | 29% |
| | | Moderate RoB | 28 | -0.54 (-1.00, -0.08) | | | | | | |
| Exposure | NOS | No RoB (intercept) | 15 | 0.47 (0.24, 0.69) | 0.89 | .355 | 81.97 | < .001 | 67% | 0% |
| | | Yes RoB | 16 | -0.15 (-0.47, 0.17) | | | | | | |
| | Q-Coh | No RoB (intercept) | 25 | 0.40 (0.23, 0.57) | 0.43 | .519 | 86.85 | < .001 | 67% | 0% |
| | | Yes RoB | 6 | -0.12 (-0.48, 0.25) | | | | | | |
| | ROBINS-I | Low RoB (intercept) | 29 | 0.38 (0.22, 0.54) | 0.07 | .788 | 87.86 | < .001 | 67% | 0% |
| | | Serious RoB | 2 | -0.07 (-0.64, 0.49) | | | | | | |

| | Tool | Category | n | β (95% CI) | Test of the model | | Goodness of fit | | | $R^2$ analog |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | F | P-value | Q | P-value | $I^2$ value | |
| Outcome | NOS | No RoB (intercept)<br>Yes RoB | 17<br>14 | 0.40 (0.18, 0.61)<br>-0.01 (-0.34, 0.32) | 0.00 | .958 | 85.64 | < .001 | 68% | 0% |
| | Q-Coh | No RoB (intercept)<br>Yes RoB | 21<br>10 | 0.40 (0.21, 0.59)<br>-0.08 (-0.38, 0.23) | 0.28 | .600 | 84.62 | < .001 | 66% | 0% |
| | ROBINS-I | Low RoB (intercept)<br>Serious RoB<br>No information | 29<br>1<br>1 | 0.32 (0.20, 0.43)<br>-0.06 (-0.48, 0.36)<br>1.38 (0.73, 2.03) | 9.56 | .001 | 50.04 | .006 | 44% | 52% |
| Missing data | NOS | No RoB (intercept)<br>Yes RoB | 26<br>4 | 0.53 (0.15, 0.90)<br>-0.10 (-0.89, 0.69) | 0.09 | .776 | 44.75 | < .001 | 78% | 0% |
| | Q-Coh | No RoB (intercept)<br>Yes RoB | 7<br>6 | 0.59 (0.17, 1.00)<br>-0.22 (-0.88, 0.45) | 0.52 | .489 | 43.59 | < .001 | 77% | 0% |
| | ROBINS-I | Low RoB (intercept)<br>Moderate RoB<br>Serious RoB | 13<br>12<br>6 | 0.44 (0.20, 0.68)<br>-0.12 (-0.46, 0.22)<br>-0.08 (-0.51, 0.34) | 0.28 | .761 | 85.07 | < .001 | 67% | 0% |
| Selective reporting | ROBINS-I | Moderate RoB (intercept)<br>Serious RoB | 26<br>5 | 0.38 (0.21, 0.55)<br>-0.04 (-0.42, 0.35) | 0.03 | .854 | 85.98 | < .001 | 66% | 0% |

Note: Risk of bias domains are based on the classification presented in the Table 2 of the main report. Analyses were performed following a random-effects model with the Knapp Hartung adjustment (two-sided p-value). RoB = risk of bias; n = number of effect sizes in each category; CI = confidence interval.

[a]Meta-regression for overall RoB of the NOS was computed for both quantitative score and categories.