Universitat Autònoma de Barcelona

APLICACIÓN DE LAS TÉCNICAS DE REVISIÓN SISTEMÁTICA A CUESTIONES CLÍNICAS DE PREVALENCIA, PRONÓSTICO, DIAGNÓSTICO E INTERVENCIÓN

Marta Roqué i Figuls

Universitat Autònoma de Barcelona

Facultad de Medicina

Departamento de Pediatría, de Obstetricia y Ginecología,
y de Medicina Preventiva y Salud Pública

APLICACIÓN DE LAS TÉCNICAS DE REVISIÓN SISTEMÁTICA A CUESTIONES CLÍNICAS DE PREVALENCIA, PRONÓSTICO, DIAGNÓSTICO E INTERVENCIÓN

TESIS DOCTORAL

Marta Roqué i Figuls

Directores:

Xavier Bonfill

Javier Zamora

Noviembre de 2020

Universitat Autònoma de Barcelona

Facultad de Medicina

Departamento de Pediatría, de Obstetricia y Ginecología,
y de Medicina Preventiva y Salud Pública

Programa de Doctorado en Metodología de la Investigación Biomédica
y Salud Pública

APLICACIÓN DE LAS TÉCNICAS DE REVISIÓN SISTEMÁTICA A CUESTIONES CLÍNICAS DE PREVALENCIA, PRONÓSTICO, DIAGNÓSTICO E INTERVENCIÓN

Marta Roqué i Figuls

Noviembre de 2020

Memoria de tesis como compendio de publicaciones presentada por Marta Roqué i Figuls para optar al grado de doctor en Medicina por la Universitat Autònoma de Barcelona y realizada bajo la dirección del Dr. Xavier Bonfill y el Dr. Javier Zamora.

Universitat Autònoma de Barcelona

Facultad de Medicina

Departamento de Pediatría, de Obstetricia y Ginecología,
y de Medicina Preventiva y Salud Pública

APLICACIÓN DE LAS TÉCNICAS DE REVISIÓN SISTEMÁTICA A CUESTIONES CLÍNICAS DE PREVALENCIA, PRONÓSTICO, DIAGNÓSTICO E INTERVENCIÓN

TESIS DOCTORAL

Marta Roqué i Figuls

Directores:

Xavier Bonfill

Javier Zamora

Noviembre de 2020

Agradecimientos

Aquesta tesi, realitzada des de la maduresa personal i professional, ha estat una experiència llarga i enriquidora. Em sento orgullosa de la feina feta, que no hagués estat possible sense la participació i la col·laboració de moltes persones, a les quals vull expressar el meu agraïment.

Als meus directors de tesi per creure en mi i encoratjar-me sempre en aquesta cursa de fons. Al Xavier, amb qui he desenvolupat tota la carrera professional i que m'ha transmès valors com l'equanimitat i el rigor, i el ferm convenciment que tot és possible. Al Javier pel seu entusiasme contagiós i tots el coneixement que ha compartit generosament amb mi.

A la Laura Martínez per tots els consells i el suport que m'ha donat per a fer la tesi. Gràcies per fer-me donar el millor de mi mateixa.

Als meus companys del Centre Cochrane Iberoamericà i el Servei d'Epidemiologia Clínica i Salut Pública de Sant Pau pel bon ambient de treball i per ser un equip humà tan acollidor. Als meus companys de la Fundació Salut i Envelliment per totes les experiències compartides.

Especialment, a l'Ivan, el Gerard, l'Ignasi i l'Àlex pels cafès i els dinars, les rialles i les converses. A la Maria José, la Marta i la Yasmín, que posen els fonaments per als projectes de tots. A la Ingrid i l'Aureli per tot el que he après amb ells.

A la Montse Rué, una gran investigadora i millor persona, que em va ensenyar què era la bioestadística. Amb ella vaig iniciar la meva trajectòria com a investigadora i des d'aleshores ha sigut un model a seguir.

A les meves amigues, a la colla dels matemàtics i a la dels informàtics, perquè sense amics la vida és menys interessant i tenir-los al meu costat durant tant de temps és un luxe.

Als meus pares, que em van transmetre els seus valors, la humanitat, el sentit de la família i la importància de l'esforç. Sé que us hagués fet il•lusió viure aquesta tesi i us tinc sempre presents. Als meus germans perquè sempre hi són.

A les meves filles, Anna, Laia i Núria, que sou el motor i la raó de tot. Al Pere perquè és un privilegi viure la vida al teu costat. Ara acaba un projecte vital i espero viure'n molts més amb vosaltres.

Have you heard of tiny Melinda Mae
who ate a monstrous whale?
She though she could,
she said she would,
so she started in right at the tail
/... /
... And in eighty-nine years she ate that whale
because she said she would!

Shel Silverstein

Índice

Resumen

1 Resumen

1.1 Resumen

Antecedentes

Las revisiones sistemáticas (RS) utilizan un método sistemático y explícito para sintetizar la evidencia que responde a una pregunta de investigación específica. La diversidad, complejidad y dispersión de recursos y métodos disponibles para su desarrollo hace necesario identificar, organizar y poner en práctica los mejores recursos y los más apropiados para cada tipo de pregunta de investigación.

Objetivos

Los objetivos de este trabajo de tesis son describir y aplicar distintas técnicas de revisión para la evaluación de la atención sanitaria en cuestiones de prevalencia, pronóstico, exactitud diagnóstica e intervención.

Métodos

Se han realizado cuatro trabajos con diferentes metodologías: 1) una compilación de recursos para la realización de revisiones sistemáticas de prevalencia, pronóstico, exactitud diagnóstica y de intervención, 2) una RS sobre el efecto de los programas de ejercicio en la función física de la población mayor frágil; 3) una RS Cochrane sobre el efecto de la fisioterapia respiratoria en la bronquiolitis aguda del lactante, y 4) una RS Cochrane de la exactitud diagnóstica de la prueba PET-CT para identificar la afectación de los ganglios linfáticos torácicos en pacientes con cáncer de pulmón de célula no pequeña (NSCLC) potencialmente resecable.

Resultados

En la compilación de recursos se identificaron manuales metodológicos desarrollados por la Colaboración Cochrane para desarrollar RS de exactitud diagnóstica y RS de efectos de las intervenciones. Asimismo, se identificaron manuales del Joanna Briggs Institute para desarrollar RS pronóstico, y manuales del grupo GRADE para la evaluación de la calidad de la evidencia en la mayoría de tipos de RS. Los manuales identificados se complementaron con estudios primarios y guías de presentación de informes.

En la RS de fragilidad se identificaron 19 ensayos aleatorizados, 12 comparaban el ejercicio con un control inactivo. La mayoría de los programas de ejercicio eran multicomponentes. En comparación con las intervenciones de control, el ejercicio mejora la velocidad normal y rápida de la marcha, y el rendimiento físico. No se obtuvieron resultados concluyentes para los desenlaces de resistencia, equilibrio y movilidad funcional. La evidencia que compara diferentes modalidades de ejercicio es escasa y heterogénea.

En la RS de la fisioterapia respiratoria, se incluyeron 12 ensayos aleatorizados, que compararon fisioterapia con ninguna intervención. Las técnicas de espiración pasiva forzada no mostraron un efecto beneficioso en pacientes con bronquiolitis grave y están relacionadas con un mayor riesgo de desestabilización respiratoria transitoria y vómitos durante el procedimiento (evidencia de alta calidad). Se necesita más evidencia de calidad sobre el efecto de las técnicas de espiración pasiva lenta, y las técnicas de espiración pasiva forzada en pacientes con bronquiolitis leve a moderada

En la RS del estadiaje del NSCLC, se incluyeron 45 estudios que evaluaron la prueba PET-CT en 6095 participantes. Los hallazgos respaldan las recomendaciones actuales, según las cuales, la PET-CT es útil para el estadiaje ganglionar de los pacientes con NSCLC potencialmente resecable, pero por sí sola es insuficiente para tomar decisiones sobre tratamiento quirúrgico, y puede ser necesario complementarla con una biopsia.

Conclusiones

Aunque existen manuales y recursos metodológicos para desarrollar RS de los principales tipos epidemiológicos, es necesario profundizar en algunos aspectos poco desarrollados como la valoración del riesgo de sesgo en RS de prevalencia, la valoración de la calidad de la evidencia en RS de prevalencia y modelos pronóstico, o las guías de reporte para las RS pronósticas.

También es conveniente que las RS publicadas incorporen una política de actualización, y falta más evidencia sobre los criterios que deben regir las políticas de actualización (basadas en procesos de priorización o en criterios temporales).

1.2 Resum

Antecedents

Les revisions sistemàtiques (RS) utilitzen un mètode sistemàtic i explícit per sintetitzar l'evidència que respon a una pregunta d'investigació específica. La diversitat, complexitat i dispersió de recursos i mètodes disponibles per al seu desenvolupament fa necessari identificar, organitzar i posar en pràctica els recursos millors i més apropiats per a cada tipus de pregunta d'investigació.

Objectius

Els objectius d'aquest treball de tesi són descriure i aplicar les tècniques de revisió per a l'avaluació de l'atenció sanitària en qüestions de prevalença, pronòstic, exactitud diagnòstica i intervenció.

Mètodes

S'han realitzat quatre treballs amb diferents metodologies: 1) un recull de recursos per a la realització de revisions sistemàtiques de prevalença, pronòstic, precisió diagnòstica i d'intervenció, 2) una RS sobre l'efecte dels programes d'exercici en la funció física de la població fràgil d'edat avançada, 3) una RS Cochrane sobre l'efecte de la fisioteràpia respiratòria en la bronquiolitis aguda del lactant, i 4) una RS Cochrane de la precisió diagnòstica de la prova PET-CT per identificar l'afectació dels ganglis limfàtics toràcics en pacients amb càncer de pulmó de cèl·lula no petita (NSCLC) potencialment resecable.

Resultats

En la compilació de recursos es van identificar manuals metodològics desenvolupats per la Col·laboració Cochrane per desenvolupar RS de precisió diagnòstica i RS d'efectes de les intervencions. Així mateix, es van identificar manuals del Joanna Briggs Institute per desenvolupar RS pronòstic, i manuals del grup GRADE per a l'avaluació de la qualitat de l'evidència en la majoria de tipus de RS. Els manuals identificats es van complementar amb estudis primaris i guies de presentació d'informes.

A la RS de fragilitat es van identificar 19 assaigs aleatoritzats, 12 comparaven l'exercici amb un control inactiu. La majoria dels programes d'exercici eren multicomponents. En comparació amb les intervencions de control, l'exercici millora la velocitat normal i ràpida de la marxa, i el rendiment físic. No es van obtenir resultats concloents per als desenllaços de resistència, equilibri i mobilitat funcional. L'evidència que compara diferents modalitats d'exercici és escassa i heterogènia.

A la RS de la fisioteràpia respiratòria, es van incloure 12 assaigs aleatoritzats, que van comparar fisioteràpia amb no fer cap intervenció. Les tècniques d'espiració passiva forçada no van mostrar un efecte beneficiós en pacients amb bronquiolitis greu i estan relacionades amb un major risc de desestabilització respiratòria transitòria i vòmits durant el procediment (evidència d'alta qualitat). Es necessita més evidència de qualitat sobre l'efecte de les tècniques d'espiració passiva lenta, i les tècniques d'espiració passiva forçada en pacients amb bronquiolitis lleu a moderada

A la RS de l'estadiatge de l'NSCLC, es van incloure 45 estudis, que van avaluar la prova PET-CT en 6095 participants. Les troballes recolzen les recomanacions actuals, segons les quals la PET-CT és útil en l'estadiatge ganglionar dels pacients amb NSCLC potencialment resecable, però per ella mateixa és insuficient per prendre decisions sobre tractament quirúrgic, i pot ser necessari complementar-la amb una biòpsia.

Conclusions

Encara que existeixen manuals i recursos metodològics per desenvolupar RS dels principals tipus epidemiològics, cal aprofundir en alguns aspectes poc desenvolupats com la valoració del risc de biaix en RS de prevalença, la valoració de la qualitat de l'evidència en RS de prevalença i models pronòstic, o les guies d'informe per a les RS pronòstiques.

També és convenient que les RS publicades incorporin una política d'actualització, i fa falta més evidència sobre els criteris que han de regir les polítiques d'actualització (ja sigui basades en processos de priorització o en criteris temporals).

1.3 Abstract

Background

Systematic reviews (SR) use a systematic and explicit method to synthesize the evidence that answers a specific research question. The diversity, complexity and dispersion of resources and methods available for its development makes it necessary to identify, organize and put into practice the best and most appropriate resources for each type of research question.

Objectives

The objectives of this thesis work are to describe and apply review techniques for the evaluation of health care in matters of prevalence, prognosis, diagnostic accuracy and intervention.

Methods

Four studies have been carried out with different methodologies: 1) a compilation of resources for conducting systematic reviews of prevalence, prognosis, diagnostic accuracy and intervention, 2) a SR on the effect of exercise programs on the physical function of the frail elderly population, 3) a Cochrane SR on the effect of chest physiotherapy in acute bronchiolitis in infants, and 4) a Cochrane SR of the diagnostic accuracy of the PET-CT test to identify thoracic lymph node involvement in patients with potentially resectable non-small cell lung cancer (NSCLC).

Results

In the compilation of resources, methodological manuals developed by the Cochrane Collaboration to develop SRs of diagnostic accuracy and SRs of the effects of interventions were identified. Likewise, Joanna Briggs Institute manuals were identified to develop prognostic SR, and GRADE Group manuals for evaluating the quality of evidence in most types of SR. The identified manuals were supplemented with primary studies and reporting guidelines.

In the frailty SR, 19 randomized trials were identified, of which 12 compared exercise with an inactive control. Most of the exercise programs were multicomponent. Compared with control interventions, exercise improves normal and fast gait speed, and physical performance. No conclusive results were obtained for endurance, balance, and functional mobility outcomes. The evidence comparing different exercise modalities is scarce and heterogeneous.

In the chest physiotherapy SR, 12 randomized trials were included, which compared physiotherapy with no intervention. Forced passive expiration techniques did not show a beneficial effect in patients with severe bronchiolitis and are associated with an increased risk of transient respiratory destabilization and vomiting during the procedure (high-quality evidence). More evidence of good quality is needed on the effect of slow passive expiration techniques, and forced passive expiration techniques in patients with mild to moderate bronchiolitis

In the NSCLC staging SR, 45 studies were included that evaluated the PET-CT test in 6095 participants. The findings support current recommendations that PET-CT is useful for lymph node staging in patients with potentially resectable NSCLC, but is insufficient by itself to make decisions about surgical treatment, and may need to be supplemented with a biopsy.

Conclusions

Although there are manuals and methodological resources to develop SRs of the main epidemiological types, further research is needed into some yet underdeveloped aspects such as the assessment of the risk of bias in prevalence SRs, the assessment of the quality of the evidence in prevalence SR and prognostic models SRs, or reporting guidelines for prognostic SRs.

It is also convenient that the published SRs incorporate an updating policy, and there is a lack of more evidence on the criteria that should govern updating policies (based on prioritization processes or temporal criteria).

Introducción

The saddest aspect of life right now is that
science gathers knowledge faster than society gathers wisdom

Isaac Asimov

2 Introducción

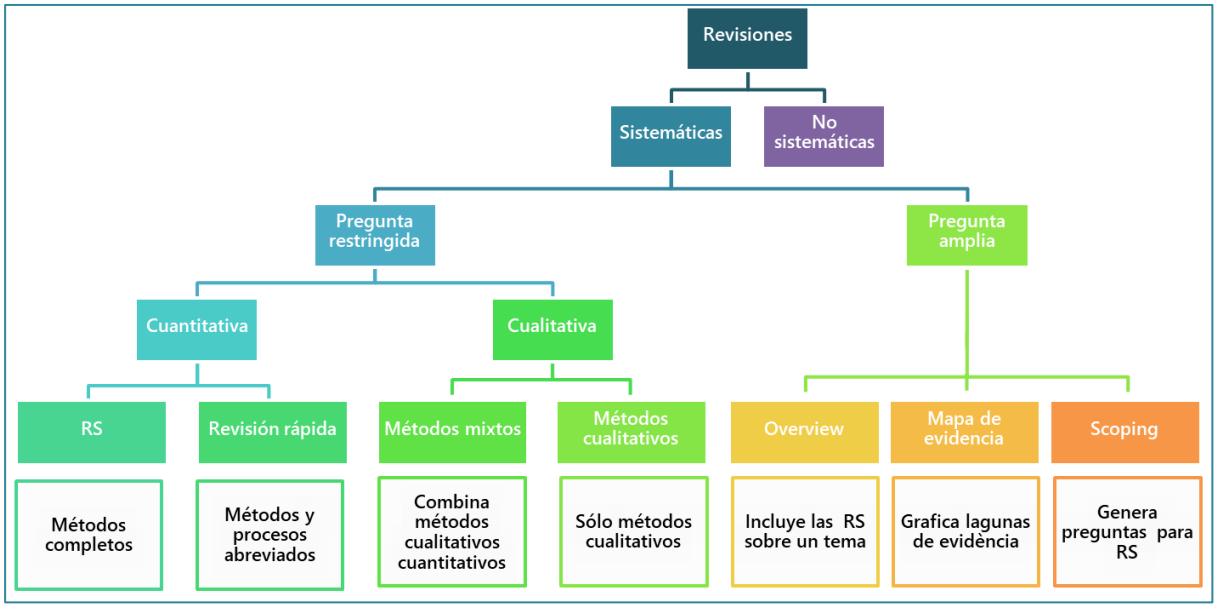
2.1 Revisiones sistemáticas

En muchas áreas del conocimiento la velocidad de generación de nuevas evidencias empíricas sigue un ritmo exponencial con una elevada producción de documentos científicos dispersos en multitud de fuentes. Este fenómeno adquiere especial relevancia en el campo de la investigación biomédica, en el que existe la necesidad de tomar decisiones sobre aspectos que afectan a la salud de la población, fundamentados en la mejor evidencia disponible. En este contexto se hace necesario el desarrollo de metodologías de investigación que permitan tratar la información, dispersa y sin filtrar, para destilar de ella conocimiento relevante, actualizado y de calidad que fundamente la toma de decisiones en salud. Las revisiones sistemáticas se han erigido como una de estas metodologías de investigación.

Una revisión sistemática (RS) se define como un trabajo de investigación que utiliza un método sistemático, explícito y reproducible para identificar, analizar y sintetizar evidencia empírica, y para responder a una pregunta de investigación específica [1]. Las características que hacen que una RS sea sistemática son la aplicación de métodos transparentes y reproducibles, la búsqueda exhaustiva, la evaluación de la calidad de los estudios incluidos, y la síntesis (narrativa o estadística) de la evidencia obtenida.

La creación de la Colaboración Cochrane en 1993 propició el desarrollo de revisiones sistemáticas sobre el efecto de las intervenciones, con el fin de proporcionar evidencia rigurosa que tuviera un impacto sobre la práctica asistencial y la salud de la población. Posteriormente, tanto desde Cochrane como desde otros ámbitos, se realizaron avances metodológicos que permitieron el desarrollo de revisiones sistemáticas con distintos objetivos. Las RS ya no se limitan a analizar el efecto de las intervenciones, sino que también exploran otras cuestiones, como la exactitud diagnóstica de una prueba, las preferencias y valores de los individuos respecto al manejo de su salud, o incluso exploran aspectos metodológicos para la realización de RS. Actualmente, Grant y colaboradores identificaron y describieron hasta 14 tipos de RS, aunque, sin duda, su lista no es exhaustiva [2]. Una clasificación de los principales tipos de RS, organizados según el alcance de la pregunta de investigación de la RS y la metodología aplicada, se muestra en la **figura 1**, y se comentará a continuación.

Figura 1. Clasificación de las revisiones



Inicialmente, las RS respondían a una pregunta restringida y aplicaban metodología exclusivamente cuantitativa, con la que se evaluaba un número limitado de intervenciones, factores de riesgo o pruebas diagnósticas, para los que se determinaba el efecto en forma numérica. Estas RS más tradicionales evolucionaron hacia las revisiones rápidas, que son formas de síntesis del conocimiento en que los componentes del proceso de revisión sistemática se han simplificado u omitido para producir conocimiento en un periodo de tiempo más corto que si se hubiera realizado una RS tradicional, sin disminuir su rigor [3]. Así, por ejemplo, los autores pueden decidir realizar búsquedas bibliográficas menos amplias o sofisticadas, limitar la extracción de datos a una selección de variables clave, o realizar una valoración simplificada de la calidad [2, 4].

Posteriormente, el desarrollo de la investigación cualitativa propició la incorporación de métodos cualitativos en el desarrollo de RS, con el fin de identificar constructos que lleven al desarrollo de una nueva teoría o que amplíen la comprensión de un fenómeno particular, a partir de las opiniones y experiencias de las personas implicadas. Así, por ejemplo, una RS cualitativa permitiría explorar cuestiones sobre aceptabilidad de una intervención, incorporando las preferencias y preocupaciones de los pacientes respecto al tratamiento considerado. Las RS cualitativas pueden tener un enfoque exclusivamente cualitativo (incluyen solo estudios cualitativos), o tener un enfoque mixto, integran tanto evidencia cuantitativa como cualitativa. En las RS de métodos mixtos, la interpretación de datos numéricos puede apoyar las opiniones y perspectivas cualitativas, y viceversa [5].

Aunque al principio las RS tenían un enfoque restringido, posteriormente, y de forma natural, aparecieron otros tipos de RS que tenían enfoques más amplios con el objetivo de dar respuestas a preguntas menos específicas y más complejas, o que ofrecen una visión exploratoria sobre diversos temas. Las revisiones de revisiones (*overview*) son un tipo de RS amplias que cubren todo el espectro de tratamientos disponibles para una condición mediante una síntesis en la que los estudios a incluir son, a su vez, revisiones sistemáticas. Así, por ejemplo, una revisión de revisiones podría integrar la evidencia generada por todas las RS que evalúan las estrategias para el abandono del hábito tabáquico. Otro tipo de RS amplias son los mapeos de evidencia (*evidence mapping*), que son

proyectos exploratorios que muestran gráficamente toda la evidencia disponible para una determinada pregunta de investigación, lo que permite identificar lagunas de evidencia para las que sería necesario desarrollar RS en profundidad. Así, por ejemplo, se podría desarrollar una matriz de evidencia en la plataforma Epistemonikos (https://www.epistemonikos.org) en la que se cruzaran las RS y los ensayos clínicos que evalúan una misma pregunta clínica, lo que permitiría identificar los solapamientos entre RS y, posteriormente, se identificarían estudios no incluidos en las RS. Finalmente, una RS de alcance (*scoping review*) es un proyecto exploratorio para dar una valoración inicial del tamaño y alcance de la investigación existente sobre un tema, mapear la evidencia e informar sobre la investigación futura. Se trata de revisiones especialmente indicadas para examinar la evidencia emergente y poder formular preguntas específicas que puedan ser objeto de síntesis de la evidencia [6].

Son muchas las organizaciones que desarrollan RS en el marco de la salud, como la Colaboración Cochrane, el Joanna Briggs Institute, la Colaboración Campbell, la Organización Mundial de la Salud, o las agencias de evaluación de tecnologías incluidas en la red INAHTA (International Network of Agencies for Health Technology Assessment). Cada organización aplica un enfoque propio a las RS que desarrolla, lo que implica que las temáticas y el enfoque aplicado pueden ser muy distintos, y, en consecuencia, también diferirán las preguntas estudiadas, el tipo de revisión sistemática desarrollada, los métodos aplicados, y el alcance y diseminación de las RS. Para ver un ejemplo de cómo dos organizaciones presentan aproximaciones distintas a un mismo problema de salud, consideremos el problema de la cesación tabáquica en mujeres gestantes. Tanto Cochrane como JBI tienen publicadas RS cualitativas sobre cesación tabáquica en mujeres gestantes, pero la RS Cochrane se dirige a la implementación de una intervención específica (¿qué factores determinan el uso de tratamiento de reemplazo con nicotina en gestantes que fuman?), mientras que la RS de JBI se dirige a los valores y preferencias de las mujeres gestantes para dejar de fumar (¿cuáles son las experiencias y necesidades de cesación tabáquica de las mujeres indígenas gestantes que fuman?) [7, 8]. No es necesario recordar que ambas aproximaciones son válidas, relevantes y necesarias, y pueden ser complementarias en la toma de decisiones.

2.2 Tipos de revisión sistemática en función de su pregunta de investigación

A continuación, se desarrollan las características de las RS cuantitativas. Podemos caracterizar las RS cuantitativas según la pregunta de investigación a la que desean responder, y clasificarlas en cuatro grandes tipos: preguntas de prevalencia, de pronóstico, de exactitud diagnóstica y de intervención.

Las RS de prevalencia tienen como objetivo responder a la pregunta que cuantifica la carga o frecuencia de una determinada condición de salud, por ejemplo, "¿cuál es la prevalencia de fragilidad en las personas mayores que residen en países de bajos ingresos?" [9]

Las RS de pronóstico pueden dividirse en tres tipos de preguntas de investigación [10]: 1) preguntas sobre la incidencia de una condición, medida como los nuevos casos que ocurren dentro de un período de tiempo, como por ejemplo, "¿cuál es la incidencia de demencia en personas mayores durante la última década?" [11]; 2) preguntas explicativas sobre los factores que están asociados o determinan un resultado específico, como por ejemplo, "¿la soledad en las personas mayores es un factor pronóstico de demencia?" [12], y 3) preguntas de predicción de resultados, como por ejemplo,

"¿cuál es el mejor modelo de predicción del riesgo de delirio en personas mayores hospitalizadas?" [13]

Las RS de exactitud diagnóstica (o RS diagnósticas) tienen como objetivo responder a preguntas sobre la exactitud diagnóstica de pruebas para identificar o descartar la presencia de una condición o problema de salud, por ejemplo, "¿son útiles las escalas de fragilidad autorreportada para identificar a las personas mayores residentes en la comunidad que están en situación de fragilidad o prefragilidad?" [14]

Finalmente, las RS de intervención exploran el efecto de las intervenciones en desenlaces de interés de personas con un problema de salud particular, en comparación con una intervención de referencia, por ejemplo, "¿realizar una valoración geriátrica integral a las personas mayores con demencia en el momento del ingreso hospitalario reduce la mortalidad?" [15]

La producción de revisiones sistemáticas se ha acelerado sustancialmente en las últimas cuatro décadas y actualmente hay un gran volumen de publicaciones que se engloban bajo la etiqueta de revisiones sistemáticas. Una búsqueda en Pubmed identifica 192848 referencias bibliográficas de revisiones sistemáticas publicadas entre 2000 y 2020. La distribución aproximada de estas RS (en base a una adaptación de los filtros específicos por diseño metodológico [16]) se muestra en la **tabla 1** y en la **figura 2**. Se puede estimar que un 27.3% de las RS son revisiones de intervención, 21.9% revisiones pronósticas, 5.3% revisiones diagnósticas, y 1.8% revisiones de prevalencia. Estos valores son aproximados, debido a las limitaciones de indexación de Pubmed, pero dan una imagen de la frecuencia relativa de cada tipo de RS. Las 84464 referencias restantes incluyen otros tipos de RS, como *overviews*, *umbrella reviews*, revisiones cualitativas, *scoping reviews*, revisiones sistemáticas metodológicas, mapas de evidencia, metarrevisiones, etc., así como RS de intervención, pronóstico, diagnóstico y prevalencia que no identifican los filtros metodológicos. Como comparación, una búsqueda en las RS publicadas en Cochrane Library identifica 7983 revisiones sistemáticas completas y 1644 protocolos, cuya distribución se muestra en la **tabla 1** y la **figura 3**. Las RS Cochrane se agrupan en 7 grandes ámbitos: RS de intervención, RS diagnósticas, *overviews*, RS metodológicas, RS cualitativas, RS pronósticas y RS rápidas. Del total de RS publicadas, un 96.9% son revisiones de intervención (incluidas las RS rápidas hasta la fecha), un 1.7% son revisiones diagnósticas, y un 0.1% corresponden a revisiones pronósticas. Las 107 referencias restantes corresponden a *overviews* y revisiones cualitativas o metodológicas. Estas búsquedas simples permiten estimar valores aproximados para la frecuencia relativa de cada tipo de RS, y muestran claramente que son las RS de intervención las más frecuentes, seguidas de las revisiones pronósticas en Pubmed, o de las revisiones diagnósticas en Cochrane. Las revisiones pronósticas son un desarrollo reciente en Cochrane, como lo prueba el alto número de protocolos comparado con las revisiones publicadas de este tipo.

Tabla 1. Revisiones en Pubmed y Cochrane Library, clasificadas por tipos

	Revisiones Pubmed	Revisiones Cochrane Library	Protocolos Cochrane Library
Intervención	52601	7733	1497
Pronóstica	42140	5	11
Diagnóstica	10171	140	84
Prevalencia	3472	0	0
Total de RS (incluyendo otros tipos)	192848	7985	1645

Búsquedas en Pubmed.gov (15/10/2020). Términos usados: systematic review[Publication Type]; ((interv*[ti] or treat*[ti] or manag*[ti] or efficacy[ti] or effectiv*[ti] or safety[ti]) or (randomized controlled trial[Publication Type] OR (randomized[Title/Abstract] AND controlled[Title/Abstract] AND trial[Title/Abstract]))); (risk[ti] or Etiolog*[ti] or Prevalen*[ti] or predict*[Title/Abstract] or prognos*[Title/Abstract] OR (first[Title/Abstract] AND episode[Title/Abstract]) OR cohort[Title/Abstract]); (screeni*[ti] or Diagnos*[ti] or specificity[Title/Abstract])]; (Prevalence [ti] or burden[ti]).
Búsquedas ad hoc restringidas al periodo 2000-2020, adaptadas de las búsquedas específicas de Clinical Queries [16]
Búsqueda en Archie/CL por Topic: type of review (9/11/2020)

Figura 2. Distribución de las RS en Pubmed por tipos y año de publicación (2000-2020)

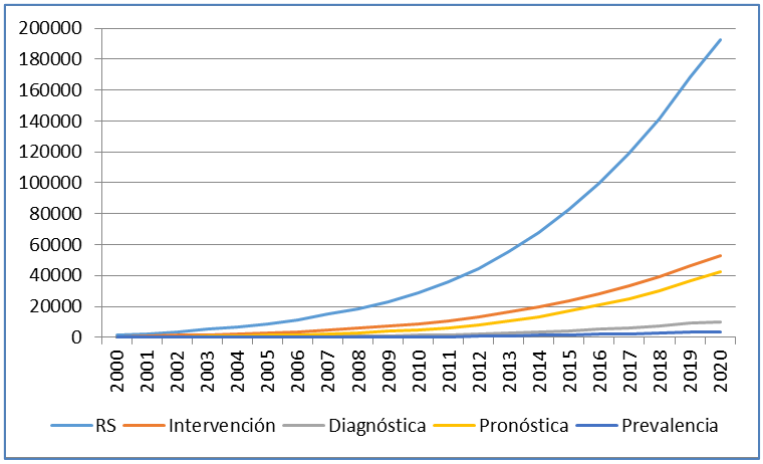
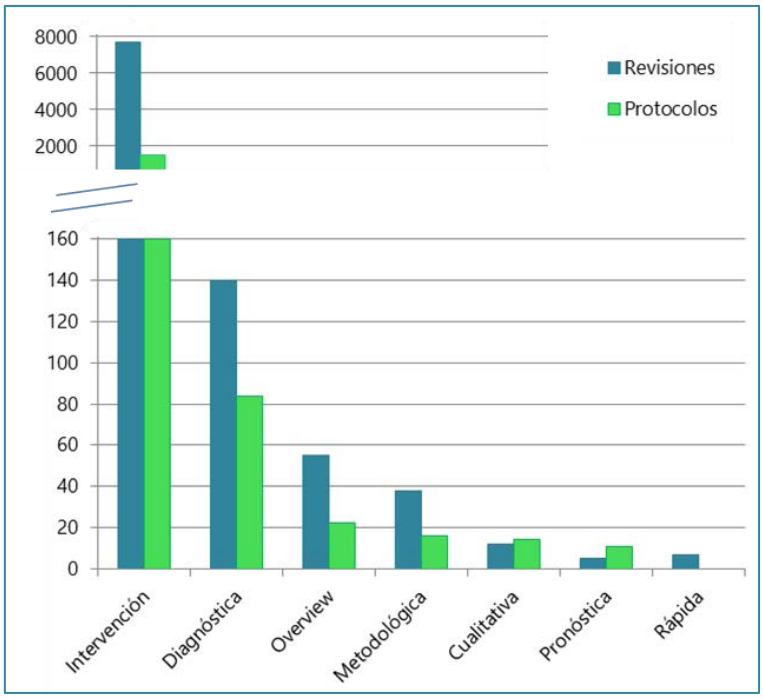


Figura 3. Distribución de las revisiones Cochrane, por ámbitos



2.3 Algunos retos metodológicos en las revisiones sistemáticas

La aplicación de metodologías relevantes y sólidas garantiza la calidad y fiabilidad de los resultados de la síntesis de evidencia. Todos los tipos de RS cuantitativa mencionados comparten una misma estructura, que puede esquematizarse en las siguientes etapas [1]:

1) formulación de la pregunta de investigación,
2) desarrollo del protocolo que describe explícitamente los métodos para realizar la RS,
3) búsqueda exhaustiva de la literatura,
4) evaluación del riesgo de sesgo de los estudios incluidos en la RS,
5) síntesis de los hallazgos de los estudios,
6) evaluación de la calidad de la evidencia obtenida en la RS,
7) informe de resultados y conclusiones de la RS.

A pesar de esta similitud estructural, los diversos tipos de RS requieren la aplicación de métodos diferentes y más o menos complejos según el tipo de RS. La metodología específica para desarrollar cada tipo presenta un nivel de desarrollo y disponibilidad desigual, especialmente en el caso de las RS de prevalencia y pronósticas, para las que faltan herramientas metodológicas específicas para realizar determinadas etapas de la revisión, y las herramientas que sí están disponibles a menudo son poco accesibles o desconocidas para los investigadores que desean realizar RS. Así, por ejemplo, no se dispone de escalas adecuadas para la valoración del riesgo de sesgo de estudios de prevalencia, y las escalas existentes para la valoración del riesgo de sesgo en estudios pronósticos de factores de riesgo son poco conocidas y, en algunos casos, discutidas. Por el contrario, la metodología para desarrollar RS de exactitud diagnóstica y de efecto de las intervenciones está en una etapa de desarrollo más madura y mejor establecida. Son muchas las RS diagnósticas y de intervención publicadas cada año, y se dispone de dos manuales metodológicos desarrollados y mantenidos por la Colaboración Cochrane que guían su desarrollo y publicación, así como un gran número de recursos metodológicos provenientes de otras fuentes. Sin embargo, incluso para estos tipos de RS en que se dispone de un cuerpo metodológico más desarrollado, existen discrepancias y controversias entre determinadas propuestas metodológicas disponibles.

Otro reto metodológico que se plantea para que las síntesis de evidencia sean realmente útiles para la toma de decisiones y la transmisión de conocimiento es que estos resultados se mantengan actualizados, incorporando periódicamente toda nueva evidencia que se genere. La visión de Cochrane es un mundo de mejor salud en el que las decisiones en materia de salud y sanidad están informadas por datos de calidad, relevantes y actualizados procedentes de la investigación. Y en respuesta a esta intrínseca voluntad de generar evidencia válida y relevante, Cochrane ha desarrollado una política de actualización periódica priorizada, por la que de forma periódica se valora la conveniencia de actualizar una revisión en base a la validez de la pregunta clínica que responde, el impacto y usabilidad de la versión actual de la revisión, la necesidad de incorporar mejoras metodológicas, la disponibilidad de estudios o datos adicionales a incorporar en la revisión, y la valoración de cuán estables son las conclusiones actuales en la incorporación de los datos y estudios adicionales en una nueva versión [17]. Además de estas iniciativas regulares de actualización, las revisiones Cochrane también se actualizan cada vez que se reciben comentarios o críticas válidas de parte de los lectores de la revisión. Existe un proceso transparente de recepción de comentarios, que se publican y deben ser atendidos por los autores en un plazo razonable. De ser necesario, se publica una versión modificada de la

revisión con las adiciones o modificaciones sugeridas –por ejemplo, al identificar un estudio adicional o corregir alguna deficiencia metodológica-. Esta política de Cochrane contrasta con la de otras organizaciones, y con la realidad de la publicación de RS en revistas convencionales, que no contemplan la opción de mantener actualizado el cuerpo de evidencia que ponen a disposición del lector, y, por tanto, están generando evidencia de carácter temporal, lo que supone un uso poco óptimo de los recursos humanos y técnicos empleados en el desarrollo de la RS [18].

2.4 Justificación del trabajo de tesis

Como se ha descrito anteriormente, el gran desarrollo de las RS, los retos que enfrentan, la diversidad, los tipos y complejidad de las RS justifican profundizar en diversos aspectos metodológicos relacionados con las mismas, que es el objetivo de este trabajo de tesis.

En primer lugar, la gran variabilidad, complejidad y dispersión de recursos y métodos disponibles para hacer RS hace necesaria una recopilación estructurada y seleccionada de los mejores recursos existentes, que facilite su accesibilidad, comparación y priorización, de modo que los investigadores que deseen desarrollar RS puedan usarlos en los proyectos de investigación.

Adicionalmente, en este trabajo de tesis se ilustra la aplicación de las técnicas de síntesis de la evidencia sobre el efecto de las intervenciones, realizando dos RS sobre intervenciones habituales. Las diferencias que existen entre la síntesis de evidencia del efecto de las intervenciones y la síntesis sobre la precisión de las pruebas diagnósticas se ilustran al comparar los trabajos anteriores con una RS Cochrane diagnóstica. De este modo, se muestran de forma práctica las particularidades de cada uno de los abordajes metodológicos empleados, ilustrando con ejemplos concretos la utilidad de las revisiones sistemáticas, así como las limitaciones en su aplicación en la práctica de investigación.

En tercer lugar, es interesante explorar el impacto derivado de los procesos de desarrollo de revisiones sistemáticas Cochrane en comparación con los de las revistas convencionales. Las revisiones Cochrane se rigen por altos estándares metodológicos, someten su protocolo a un proceso de revisión por pares, se publican en formato electrónico sin límite de extensión, y, además, siguen una política de actualización periódica e incorporación del *feedback*, lo que las distingue de la mayoría de revisiones sistemáticas publicadas en revistas convencionales [19-21].

Asimismo, existe una necesidad especial de disponer de evidencia de calidad sobre la exactitud diagnóstica de las pruebas de imagen, que apoye la toma de decisiones sanitarias. La realización de estas pruebas diagnósticas ha crecido de forma importante en los últimos años y, de todas ellas, la tomografía con emisión de positrones combinada con tomografía computada (PET-CT) es quizás la que ha experimentado un mayor incremento, que se sitúa alrededor del 16.2% en el período 2012-2018 en el Reino Unido [22-23]. El metanálisis en RS diagnósticas muestra un mayor grado de complejidad que las RS de intervención, porque los estudios de exactitud diagnóstica a metanalizar pueden haber utilizado diferentes umbrales, implícitos y explícitos, para definir un resultado positivo en la prueba evaluada [24]. La Colaboración Cochrane es quizás la institución que más ha desarrollado la metodología para la realización de RS de exactitud diagnósticas y la que está promoviendo la aplicación de estándares de alta calidad para las mismas, como ya se ha mencionado anteriormente. Por todo ello, es de interés general realizar RS Cochrane de exactitud diagnóstica de las pruebas de uso habitual en la atención sanitaria, como la PET-CT.

Finalmente, cabe destacar que la aplicación de las técnicas de síntesis se realizará en tres problemas de salud de especial relevancia, porque atañen a poblaciones especialmente vulnerables como son las personas mayores frágiles y los lactantes, o porque se refieren a uno de los cánceres con mayor mortalidad. El primer problema de salud que se considerará es el del mantenimiento de la función física en las personas mayores. A medida que las personas envejecen pueden llegar a una etapa de vulnerabilidad llamada fragilidad que precede y predispone a la discapacidad y dependencia física. La fragilidad es común en los adultos mayores (>65 años) y conlleva un alto riesgo de caídas, empeoramiento de la movilidad, discapacidad, hospitalización y mortalidad [25]. Existe controversia en la conceptualización de la fragilidad, dado que algunos investigadores la entienden como un fenotipo exclusivamente físico, mientras que otros la conceptualizan como una suma de déficits, resultado de la interacción de factores físicos, cognitivos, sociales y psicológicos. A pesar de esta discrepancia, hay unanimidad en considerar que la fragilidad afecta a múltiples dominios de funcionamiento, como puede ser la marcha y movilidad, equilibrio, fuerza muscular, procesamiento motor, cognición, nutrición, fatiga y actividad física [26]. Al buscar intervenciones que permitan abordar la fragilidad, se ha propuesto el ejercicio físico como una herramienta para mejorar las funciones físicas en adultos mayores, como la velocidad de la marcha, el equilibrio, la agilidad y la deambulación. Sin embargo, las revisiones sistemáticas publicadas en el pasado sobre el ejercicio en personas mayores frágiles adolecían de limitaciones metodológicas, estaban desactualizadas o no evaluaban medidas de rendimiento físico, por lo que existe la necesidad de realizar una nueva revisión sistemática que supere estas limitaciones.

El segundo problema de salud que se considera es la bronquiolitis aguda, infección respiratoria viral frecuente en niños menores de dos años, y la infección respiratoria del tracto bajo más frecuente durante el primer año de vida [27]. La mayoría de los niños afectados de bronquiolitis aguda presentan la enfermedad de forma leve y autolimitada en el tiempo, por la que no requieren hospitalización. Sin embargo, un porcentaje de afectados, que presentan la enfermedad de forma más severa y necesitan ser hospitalizados, muestran dificultades para eliminar la flema (secreciones respiratorias mucosas espesas causadas por la infección), lo que conlleva baja oxigenación en sangre y malestar. La fisioterapia torácica se ha propuesto como una intervención que puede ayudar a eliminar las secreciones respiratorias y mejorar la oxigenación del lactante, lo que reduciría la duración del episodio de bronquiolitis. En diversos países se realiza esta intervención en los hospitales de forma habitual, aunque no están claros los beneficios que comporta realmente, ni su seguridad [28]. Por ello, es necesario integrar y organizar la evidencia disponible sobre el conjunto de técnicas de fisioterapia respiratoria y valorar la calidad de la evidencia sobre su eficacia y seguridad.

Finalmente, se considera el cáncer de pulmón no microcítico (NSCLC, por sus siglas en inglés). En ausencia de metástasis a distancia, las opciones de tratamiento del NSCLC dependen de cuánto se haya propagado la enfermedad a los diferentes ganglios linfáticos dentro del tórax, es decir, la etapa de la enfermedad. Si el cáncer no se ha diseminado o no lo ha hecho más allá de los ganglios linfáticos más cercanos (situación que se define como estadios N0 y N1), la cirugía suele ser el tratamiento de elección. Por el contrario, si el cáncer se ha diseminado a los ganglios linfáticos mediastínicos ipsilaterales o a los ganglios linfáticos subcarinales (estadio N2), se descarta el tratamiento quirúrgico con intención curativa y se consideran otras opciones de tratamiento como la radioterapia o la quimioterapia [29]. La prueba PET-CT es un método no invasivo para establecer la propagación del

NSCLC dentro del tórax y en otras partes del cuerpo, que está cada vez más disponible y es utilizada por equipos multidisciplinarios de cáncer de pulmón. Aunque la naturaleza no invasiva de la PET-CT constituye una de las principales ventajas de la prueba, la PET-CT puede ser subóptima en la detección de malignidad en los ganglios linfáticos de tamaño normal y en descartar malignidad en pacientes con enfermedades inflamatorias o infecciosas coexistentes. Por ello, es importante evaluar e integrar la evidencia existente sobre la exactitud diagnóstica de la prueba PET-CT en este contexto mediante una revisión sistemática.

Objetivos

No hi ha cap vent favorable
per aquell que no sap a quin port es dirigeix

Luci Anneu Seneca

3 Objetivos

3.1 Objetivos generales

Describir y aplicar distintas técnicas de revisión para la evaluación sanitaria, poniendo énfasis en técnicas de integración estadística (metanálisis de intervención y metanálisis de exactitud diagnóstica).

3.2 Objetivos específicos

1. Identificar y describir de forma estructurada los mejores y más actuales recursos metodológicos disponibles para desarrollar RS cuantitativas según el tipo de pregunta de investigación que se aborde: revisiones de prevalencia, pronóstico, exactitud diagnóstica y efecto de las intervenciones.
2. Aplicar las técnicas de síntesis de la evidencia a uno o más problemas de salud para proporcionar información de calidad que facilite la toma de decisiones clínicas.
3. Explorar el impacto de la política de actualización de revisiones de Cochrane sobre la incorporación de comentarios y actualización de la evidencia.
4. Determinar la eficacia de un rango de intervenciones en una condición de salud.
5. Determinar la exactitud diagnóstica de una determinada técnica diagnóstica en un problema.

Métodos

Le hasard n'est que la mesure de notre ignorance.
Mais le hasard a des lois.

Henri Poincaré

4 Métodos

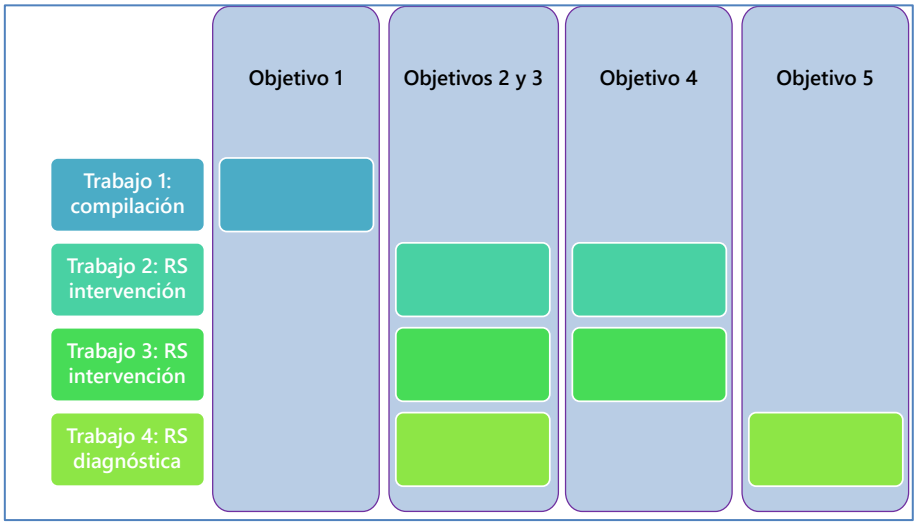
4.1 Métodos del trabajo de tesis

Este trabajo de tesis se presenta en la modalidad de compendio de publicaciones científicas, e incluye cuatro artículos publicados en revistas internacionales indexadas. Los objetivos específicos de los trabajos que conforman el proyecto de tesis son los siguientes:

1. Integrar la evidencia más actual sobre el efecto de las intervenciones de ejercicio en la mejora de las medidas de rendimiento de la función física y los marcadores de fragilidad física en personas mayores que viven en la comunidad, y que están definidas como frágiles según su nivel de función física y las dificultades físicas para realizar actividades de la vida diaria.
2. Integrar la evidencia más actual sobre el efecto y seguridad de las intervenciones de fisioterapia torácica en la mejora de las medidas clínicas en niños y niñas menores de dos años que presentan bronquiolitis aguda, evaluando por separado cada tipo de técnica.
3. Integrar la evidencia más actual sobre la exactitud diagnóstica de la prueba PET-CT para establecer el grado de propagación del cáncer a nivel de ganglios linfáticos en pacientes con NSCLC que son potencialmente candidatos a tratamiento quirúrgico con intención curativa.

En la **Figura 4** se ilustra la relación de los trabajos con los objetivos específicos del trabajo de tesis.

Figura 4. Relación entre los objetivos del proyecto de tesis y los trabajos realizados



A continuación, se describen los métodos resumidos de cada uno de los trabajos que conforman el proyecto de tesis.

4.2 Trabajo 1: Compilación de recursos metodológicos para la elaboración de revisiones sistemáticas

Diseño

Revisión no sistemática de la literatura.

Criterios de elegibilidad

Se incluyeron las mejores y más actuales recomendaciones, guías y herramientas para realizar las etapas de una RS, para cada tipo de RS considerada (RS de prevalencia, RS de pronóstico, RS de diagnóstico y RS de intervención). No se incluyeron los recursos metodológicos para desarrollar otros tipos de RS.

Fuentes de información y estrategia de búsqueda

Se consultaron las pautas de las principales organizaciones que establecen métodos para realizar RS (Cochrane, Joanna Briggs Institute, Red Europea para la Evaluación de la Tecnología de la Salud [EUNETHTA]), Red para la Mejora de la Calidad y Transparencia de la Investigación en Salud [EQUATOR], Grupo de trabajo en la Calificación de la Valoración, Desarrollo y Evaluación de Recomendaciones [GRADE]) para identificar sus recursos propuestos.

Además, se realizó una búsqueda de literatura en MEDLINE (acceso a través de PubMed) en noviembre de 2019. También se realizaron búsquedas de literatura científica *ad hoc* para encontrar otros recursos para cada tipo de RS en relación con la estructura de preguntas de investigación, la estrategia de búsqueda de literatura, la evaluación del riesgo de sesgos y el análisis estadístico.

Selección y extracción de datos

Los autores evaluaron los resultados de búsqueda, seleccionaron los recursos más relevantes y precisos, y resumieron la información más relevante por etapa de desarrollo y tipo de RS.

Los recursos se organizaron en 7 secciones, siguiendo las etapas de desarrollo de una RS: 1) formulación de la pregunta de investigación, 2) desarrollo del protocolo y registro de revisión, 3) estrategia de búsqueda, 4) evaluación de riesgo de sesgo, 5) síntesis estadística de hallazgos, 6) evaluación de la calidad de la evidencia, y 7) informe de resultados y presentación. En cada sección se presentan los recursos organizados por tipo de RS, a fin de ilustrar las diferencias clave entre cada tipo de RS. Adicionalmente, los métodos se ilustran mediante ejemplos de uso en RS reales ya publicadas.

4.3 Trabajo 2: Eficacia del ejercicio físico para mejorar la función física en personas mayores frágiles

Diseño

Revisión sistemática de eficacia de las intervenciones, desarrollada siguiendo la metodología Cochrane.

Criterios de elegibilidad

Se incluyeron ensayos controlados aleatorizados (ECA) que evaluaban el efecto de los programas de ejercicio físico, solo o combinado con otros componentes.

Los participantes de los estudios debían tener 65 años o más, vivir en la comunidad, y ser definidos como frágiles. La fragilidad podía determinarse por criterios estandarizados (p. ej., criterios de Fried), o por presentar la función física reducida al medirla con escalas de rendimiento físico (p. ej., Short Physical Performance Battery) u otras medidas basadas en el rendimiento como la marcha y movilidad, fuerza muscular, ingesta nutricional, cambio de peso, equilibrio, resistencia, fatiga y actividad física. Los participantes podían presentar limitaciones en dos o más medidas de fragilidad basadas en el rendimiento, o podían presentar limitaciones clínicamente significativas en una sola medida.

Los resultados primarios fueron medidas de la función física basadas en el rendimiento, como la movilidad, la marcha, la fuerza muscular, el equilibrio, la resistencia y la discapacidad en las actividades de la vida diaria. Los resultados secundarios fueron número de caídas; institucionalización; efectos adversos del programa de ejercicios, como caídas, fracturas, tendinitis o dolor muscular; calidad de vida relacionada con la salud; síntomas de depresión; hospitalización, y muerte.

Fuentes de información y estrategia de búsqueda

Se realizaron búsquedas en las siguientes bases de datos electrónicas hasta abril de 2013: MEDLINE, The Cochrane Library, PEDro y CINAHL. Se realizaron búsquedas utilizando texto libre y descriptores. La estrategia de búsqueda se adaptó a cada base de datos, se incluyeron los términos de *fragilidad*, *personas mayores*, y múltiples expresiones de ejercicio, y se aplicaron filtros para identificar ensayos controlados aleatorizados.

Extracción de datos

Dos revisores cribaron de forma independiente los resultados de búsqueda, y, posteriormente, realizaron la extracción de datos y la evaluación del riesgo de sesgo de los estudios incluidos. Cualquier discrepancia se resolvió por consenso o consultando con un tercer autor.

Evaluación del riesgo de sesgo en los estudios incluidos

El riesgo de sesgo se evaluó para cada estudio utilizando los criterios descritos en el Manual Cochrane para Revisiones Sistemáticas de Intervenciones [1]. Se evaluaron seis dominios, correspondientes a posibles fuentes de sesgo: (1) generación de secuencia de aleatorización (sesgo de selección); (2) ocultación de la asignación (sesgo de selección); (3) cegamiento de la evaluación de resultados (sesgo

de detección); (4) datos de resultado incompletos (sesgo de atrición); (5) sesgo de informe selectivo, y (6) otras fuentes de sesgo. Para cada ensayo, a partir de las valoraciones de los dominios individuales se derivó una valoración global del riesgo de sesgo del estudio, que podía ser bajo, alto o poco claro.

Síntesis de datos

El efecto del tratamiento se estimó mediante diferencias de medias (DM) y DM estandarizadas en los desenlaces continuos, y razones de riesgo en los desenlaces dicotómicos, con los correspondientes intervalos de confianza (IC) del 95%. Se calcularon medidas de efectos combinadas aplicando el método de metanálisis del inverso de la varianza bajo un modelo de efectos aleatorios.

Exploración de la heterogeneidad: análisis de subgrupos y de sensibilidad

La heterogeneidad se evaluó con el estadístico I^2, considerando valores superiores al 50% como signo de heterogeneidad relevante. Los análisis de subgrupos previstos por grupos de edad y por niveles de función física inicial no se pudieron realizar debido al bajo número de ensayos que proporcionaban esta información, y a la falta de datos detallados en algunos estudios. Igualmente, no se pudo realizar el análisis de sensibilidad, restringido a los ensayos en los que se definía la fragilidad mediante criterios estandarizados.

4.4 Trabajo 3: Eficacia de la fisioterapia torácica en lactantes menores de dos años que presentan bronquiolitis aguda

Diseño

Revisión sistemática de eficacia de las intervenciones, desarrollada como una revisión Cochrane.

Criterios de elegibilidad

Se incluyeron ECA que evaluaban la fisioterapia torácica en lactantes menores de 24 meses que presentaban bronquiolitis aguda (tal como la hubieran definido los autores del ensayo), en cualquier entorno (hospitalario, ambulatorio o domiciliario).

Se incluyeron ensayos que administraban cualquier tipo de fisioterapia torácica (drenaje postural, percusión torácica, vibración, sacudidas torácicas, tos provocada, técnicas de espiración lenta o forzada), comparada con la atención habitual u otras técnicas de fisioterapia, drenaje o respiración.

Las intervenciones se clasificaron en dos categorías principales: vibración y percusión, y técnicas espiratorias pasivas. Estas últimas se subclasificaron en dos categorías: técnicas espiratorias pasivas lentas y técnicas espiratorias pasivas forzadas. Las técnicas de vibración y percusión producen una oscilación del tórax mediante compresión rápida o percusión con las manos del fisioterapeuta. Las técnicas de espiración forzada consisten en aumentar repentinamente el flujo espiratorio al comprimir el tórax o el abdomen. Las técnicas de flujo lento consisten en comprimir la caja torácica y la cavidad abdominal de forma gradual y suave desde la fase espiratoria media hasta el final de la exhalación.

Los desenlaces primarios para la revisión fueron: (1) cambio en el estado de gravedad de la bronquiolitis y (2) tiempo de recuperación. Los resultados secundarios fueron: (1) parámetros respiratorios (niveles de saturación de oxígeno, presión parcial de dióxido de carbono transcutáneo [PaCO2]); (2) duración de la suplementación con oxígeno; (3) duración de la estancia hospitalaria; (4) uso de broncodilatadores y esteroides; (5) valoración de los padres del beneficio de fisioterapia, y (6) eventos adversos. Se consideró evento adverso cualquier resultado no deseado causado por la intervención (por ejemplo, fracturas de costillas, bradicardia, inestabilidad respiratoria, vómitos o discapacidades neurológicas a largo plazo).

Fuentes de información y estrategia de búsqueda

Se realizaron búsquedas en las siguientes bases de datos electrónicas: Registro Cochrane Central de Ensayos Controlados; Registro Especializado del Grupo Cochrane de Infecciones Respiratorias Agudas; MEDLINE y MEDLINE in-process; EMBASE; CINAHL; LILAS Web of Science, y PEDro. El período de búsqueda fue desde octubre de 2011 hasta julio de 2015.

Extracción de datos

Dos revisores cribaron de forma independiente los resultados de búsqueda, y, posteriormente, realizaron la extracción de datos y la evaluación del riesgo de sesgo de los estudios incluidos. Cualquier discrepancia se resolvió por consenso o consultando con un tercer autor.

Evaluación del riesgo de sesgo en los estudios incluidos

El riesgo de sesgo se evaluó para cada estudio utilizando los criterios descritos en el Manual Cochrane para Revisiones Sistemáticas de Intervenciones [1]. Se evaluaron seis dominios, correspondientes a posibles fuentes de sesgo: (1) generación de secuencia (sesgo de selección); (2) ocultación de la asignación (sesgo de selección); (3) cegamiento de la evaluación de resultados (sesgo de detección); (4) datos de resultado incompletos (sesgo de deserción por retiradas, abandonos, desviaciones de protocolo); (5) sesgo de informe selectivo, y (6) otras fuentes de sesgo, en las que se consideró el posible sesgo derivado de la contaminación de los grupos de intervención. Para cada ensayo, a partir de las valoraciones de los dominios individuales, se derivó una valoración global del riesgo de sesgo del estudio, que podía ser bajo, alto o poco claro.

Síntesis de datos

El efecto del tratamiento se estimó mediante DM en los desenlaces continuos, y razones de riesgo en los desenlaces dicotómicos, con sus correspondientes IC del 95%. No se llevó a cabo una combinación estadística de datos mediante metanálisis debido a la heterogeneidad clínica observada entre los estudios, y a la falta de datos apropiados para el metanálisis en los estudios incluidos.

Exploración de la heterogeneidad: análisis de subgrupos y de sensibilidad

La heterogeneidad estadística no se evaluó dada la heterogeneidad clínica mostrada por los estudios incluidos. Se planificó y realizó un análisis de subgrupos según la gravedad de la bronquiolitis, basado en la hipótesis de que el rendimiento de las técnicas de fisioterapia torácica de flujo lento podría depender de la gravedad del paciente. Se clasificaron los ensayos en las categorías de gravedad severa, moderada, o desconocida según los criterios de inclusión del ensayo o las características de los participantes incluidos.

4.5 Trabajo 4: Exactitud diagnóstica de la PET-CT para evaluar la afectación linfática mediastínica en pacientes con cáncer de pulmón potencialmente resecable

Diseño

Revisión sistemática de exactitud diagnóstica, desarrollada como una revisión Cochrane.

Criterios de elegibilidad

Se incluyeron estudios transversales prospectivos o retrospectivos, que evaluaron la exactitud diagnóstica de la PET-CT integrada para diagnosticar la enfermedad N2 en pacientes con sospecha de NSCLC resecable que se consideraron potencialmente adecuados para la resección primaria. Esta revisión no consideró a los pacientes cuyo estadio estaba siendo reevaluado después de la inducción o quimioterapia neoadyuvante.

La prueba PET-CT podía llevarse a cabo en cualquier tipo de equipo de PET-CT integrada disponible (fuera cual fuera el fabricante o el modelo), utilizando cualquier valor de umbral para determinar la prueba como positiva. Sin embargo, no se consideraron los estudios que realizaron la PET-CT con trazadores radioactivos que no fueran el habitual FDG, o que usaran otras técnicas de imagen nuclear, como puede ser la realización de PET independiente sin CT.

La condición de interés de esta revisión fue el NSCLC resecable, definido como NSCLC que no se ha diseminado a los ganglios linfáticos mediastínicos ipsilaterales o a los ganglios linfáticos subcarinales (N2).

El estándar de referencia en los estudios incluidos debía ser la confirmación patológica de los resultados de la PET-CT a partir de muestras obtenidas mediante resección quirúrgica con muestreo mediastínico, mediastinoscopia, cirugía torácica asistida por video, aspiración con aguja transbronquial guiada por ultrasonido endobronquial, aspiración con aguja fina guiada por ultrasonido endobronquial, aspiración con aguja transbronquial, aspiración con aguja transtorácica, biopsias de sitios extratorácicos, o una combinación de cualquiera de los métodos mencionados anteriormente.

Fuentes de información y estrategia de búsqueda

Se realizaron búsquedas en las siguientes bases de datos: The Cochrane Library; MEDLINE a través de OvidSP (desde 1946); Embase a través de OvidSP (desde 1974); PreMEDLINE a través de OvidSP; OpenGrey; y disertaciones y tesis de ProQuest. El período de búsqueda fue hasta el 30 de abril de 2013.

Extracción de datos

Dos revisores cribaron de forma independiente los resultados de búsqueda, y, posteriormente, realizaron la extracción de datos y la evaluación del riesgo de sesgo de los estudios incluidos. Cualquier discrepancia se resolvió por consenso o consultando con un tercer autor.

Evaluación del riesgo de sesgo en los estudios incluidos

La calidad de cada estudio se evaluó utilizando una versión modificada de la herramienta QUADAS-2, a la que se añadieron dos preguntas adicionales: (1) ¿se realizó una definición clara de qué era un resultado positivo de la prueba?, y (2) ¿el estudio estaba libre de financiación comercial?. Se incorporó el ítem correspondiente a la definición de resultados positivos para tener en cuenta la naturaleza subjetiva de la interpretación de imágenes de la prueba PET-CT, que puede basarse en una variedad de criterios diferentes, como puede ser la experiencia clínica, diferentes valores de captación estándar (SUV por sus siglas en inglés), diferentes características morfológicas, o una combinación de las mencionadas anteriormente. Se incluyó el segundo elemento adicional para registrar cualquier sesgo potencial resultante del interés comercial en los resultados. Se resolvió cualquier desacuerdo en las valoraciones independientes de riesgo de sesgo y aplicabilidad mediante discusión.

Síntesis de datos

Se calculó la sensibilidad y la especificidad con IC del 95% para cada estudio. Se representaron gráficamente las estimaciones de las sensibilidades y especificidades observadas junto con su IC del 95% en diagramas de bosque y en un gráfico ROC de sensibilidad versus especificidad para evaluar visualmente la variabilidad entre estudios.

Se ajustó una curva resumen ROC utilizando modelos jerárquicos para el subconjunto de estudios que comparten el mismo umbral de positividad. Se identificó el punto de la curva resumen ROC que maximiza la sensibilidad y especificidad, y se calcularon las sensibilidades y especificidades promedio correspondientes a ese punto. Se representaron gráficamente las estimaciones de precisión promedio con su elipse de confianza del 95% y la región de predicción en el espacio ROC.

Exploración de la heterogeneidad: análisis de subgrupos y de sensibilidad

Se realizó un análisis de subgrupos por factores preespecificados en el protocolo que podían ser fuentes de heterogeneidad, incluyendo cada factor como una covariable en el modelo bivariado. Las fuentes anticipadas de heterogeneidad incluyeron el diseño del estudio, dosis de trazador, poblaciones de pacientes, y diferencias en la adquisición de imágenes por la prueba PET-CT o equipos de escaneo. Se realizó una comparación de la exactitud diagnóstica entre los subgrupos al evaluar si la sensibilidad o la especificidad, o ambas, diferían en los subgrupos de estudios definidos según la covariable. Se utilizó el procedimiento de modelos mixtos no lineales (NLMIXED) en SAS versión 9.1 para Windows (SAS Institute Inc, Cary, NC, EE. UU.) para ajustar los modelos HSROC y bivariados.

Se examinó la robustez de los metanálisis mediante análisis de sensibilidad restringidos a estudios que tenían un bajo riesgo de sesgo y cuya aplicabilidad a los objetivos de la revisión no suscitaba ninguna duda.

Resultados

Women need to shift from thinking "I'm not ready to do that" to thinking "I want to do that- and I'll learn by doing it.''

Sheryl Sandberg

5 Resultados

Este trabajo de tesis es un compendio de cuatro publicaciones originales. En esta sección, para cada una de ellas, se proporcionará la referencia bibliográfica, el impacto, un resumen de los principales resultados, y, finalmente, se adjuntará la publicación.

El impacto de cada publicación se ha evaluado con dos medidas: el factor de impacto (FI) de la revista para el año de publicación, y la puntuación de atención de Altmetric. El FI es un indicador ampliamente utilizado para medir la repercusión de una revista en la comunidad científica. El FI de una revista se calcula mediante el número de veces que se cita por término medio un artículo publicado en dicha revista. Se trata de un indicador con limitaciones importantes, entre ellas, que mide la repercusión de la revista, pero no la de un trabajo concreto publicado en la revista. Por este motivo, también se presenta la puntuación de atención de Altmetric para cada publicación, que sí es una medida específica del propio trabajo. Altmetric es un sistema que rastrea la atención que reciben en línea los resultados de las investigaciones, como, por ejemplo, los artículos académicos y los conjuntos de datos. La puntuación de atención de Altmetric (Altmetric Attention Score) se calcula mediante un recuento ponderado de las menciones que ha recibido el trabajo en las fuentes consideradas por Altmetric; estas fuentes incluyen redes sociales como Twitter y Facebook, medios tradicionales (generalistas o especializados) en diversos idiomas, blogs de organizaciones importantes (como Cancer Research UK) e investigadores individuales, y gestores en línea de referencias como Mendeley y CiteULike. Aunque no existe un baremo que permita valorar la magnitud de una puntuación de atención Altmetric, los desarrolladores dan una regla aproximada por la que una puntuación de 20 o más indica que la publicación está teniendo mayor impacto que sus contemporáneas. Sin duda, las dos medidas propuestas dan solo valoraciones parciales del impacto de cada publicación. En la discusión se resumirá el impacto de cada publicación sobre las guías de práctica clínica de la especialidad, lo que describirá el impacto más específico sobre los profesionales que realizan atención o políticas sanitarias.

5.1 Publicación 1: Toolkit of methodological resources to conduct systematic reviews

Roqué M, Martínez-García L, Solà I, *et al.* Toolkit of methodological resources to conduct systematic reviews [version 3; peer review: 2 approved]. F1000Research 2020, 9:82 (https://doi.org/10.12688/f1000research.22032.3)

FI: la plataforma de publicación no está indexada en Web of Science (aunque sí en Pubmed, Embase y Scopus) y no tiene FI. Puntuación de atención Altmetric: 15

Esta publicación puede consultarse al completo y de forma libre en https://doi.org/10.12688/f1000research.22032.3

En este trabajo se identificaron 69 publicaciones, 21 eran manuales o capítulos de manuales, 8 eran guías de reporte, y 40 correspondían a trabajos primarios de métodos [30]. Los recursos se organizaron por tipo de RS y etapa de la RS, y se ilustraron con 9 ejemplos específicos. En la **tabla 2** puede verse la clasificación de recursos en manuales, guías o trabajos primarios, para cada tipo de RS. Un número importante de trabajos primarios tenían un carácter transversal, ya que presentaban métodos que no eran específicos de un tipo de RS.

En la **tabla 3** se muestran los tipos de pregunta clínica por tipo de RS, vinculados a los ejemplos ilustrativos que se utilizaron a lo largo del trabajo.

Tabla 2. Recursos identificados (tabla adaptada de la publicación 1)

Tipo de RS	Manuales de buenas prácticas	Guías de informe en la publicación	Trabajos primarios de métodos
Prevalencia	1	0	1
Pronóstico	1	0	7
Exactitud diagnóstica	2	2	6
Efecto de las intervenciones	3	4	3
Recursos transversales aplicables a diversos tipos de RS	---	2	23

Tabla 3. Ejemplos de pregunta clínica por tipo de RS (tabla adaptada de la publicación 1)

Tipo de revisión sistemática	Acrónimo para la pregunta de investigación	Ejemplo de pregunta de investigación
Revisión de prevalencia	**CoCoPop-S** (condición, contexto, población y diseño de estudio)	¿Cuál es la prevalencia de fragilidad y pre-fragilidad (condición) en adultos mayores (población) que viven en la comunidad que viven en países de ingresos bajos y medios (contexto)? ¿Cuál es la prevalencia mundial (población) de insuficiente actividad física (condición)?
Revisión pronóstica - pronóstico global	**CoCoPop-S** (condición, contexto, población y diseño de estudio)	¿Cuál es la incidencia de la demencia (condición) en individuos de 60 años o más de edad (población) que viven en países de ingresos altos (contexto)?
Revisión pronóstica - factores pronóstico	**PICOT-S** (población, intervención o factor, comparador, desenlace, marco temporal y diseño de estudio) **PFO-S** (población, factor o modelo, desenlace y diseño de estudio)	¿Es la actividad de la proteasa (factor pronóstico) un factor pronóstico independiente para la cicatrización de heridas (resultado) a las 24 semanas (período de tiempo) en personas con úlceras venosas en las piernas (población)?
Revisión pronóstica - modelos pronóstico	**PICOT-S** (población, intervención o factor, comparador, desenlace, marco temporal y diseño de estudio)	¿Cuál es el mejor modelo de pronóstico para predecir la supervivencia (resultado) general o libre de progresión en pacientes con leucemia linfocítica crónica (población)?
Revisión de exactitud diagnóstica	**PIRD-S** (población, prueba índice, prueba de referencia, condición de interés y diseño de estudio)	¿Los instrumentos de detección de la fragilidad autoinformados (prueba índice) identifican con precisión a las personas mayores (población) en riesgo de fragilidad y prefragilidad (condición de interés)? ¿Es útil el PET 18F florbetapen (prueba índice) en el diagnóstico temprano de demencia (condición) en pacientes con deterioro cognitivo leve (población)?
Revisión de efecto de las intervenciones	**PICO-S** (población, intervención, comparador, desenlace y diseño de estudio)	¿Cuál es el efecto de la ribavirina (intervención) en pacientes con fiebre hemorrágica de Crimea-Congo para prevenir la muerte (resultado)? ¿La evaluación geriátrica integral (intervención) en adultos mayores (población) reduce la mortalidad (resultado)?

RESEARCH ARTICLE

REVISED Toolkit of methodological resources to conduct systematic reviews [version 3; peer review: 2 approved]

Marta Roqué[1,2], Laura Martínez-García[1,2], Ivan Solà[1,2], Pablo Alonso-Coello[1,2], Xavier Bonfill[1-3], Javier Zamora[2,4]

[1]Iberoamerican Cochrane Centre - Sant Pau Biomedical Research Institute (IIB-Sant Pau), Barcelona, Spain
[2]CIBER of Epidemiology and Public Health (CIBERESP), Madrid, Spain
[3]Autonomous University of Barcelona, Bellaterra, Spain
[4]Clinical Biostatistics Unit, Ramón y Cajal Health Research Institute, Madrid, Spain

v3 **First published:** 04 Feb 2020, **9**:82
https://doi.org/10.12688/f1000research.22032.1
Second version: 11 Aug 2020, **9**:82
https://doi.org/10.12688/f1000research.22032.2
Latest published: 14 Oct 2020, **9**:82
https://doi.org/10.12688/f1000research.22032.3

Abstract

Background: Systematic reviews (SR) can be classified by type depending on the research question they are based on. This work identifies and describes the most relevant methodological resources to conduct high-quality reviews that answer health care questions regarding prevalence, prognosis, diagnostic accuracy and effects of interventions.
Methods: Methodological resources have been identified from literature searches and consulting guidelines from institutions that develop SRs. The selected resources are organized by type of SR, and stage of development of the review (formulation of the research question, development of the protocol, literature search, risk of bias assessment, synthesis of findings, assessment of the quality of evidence, and report of SR results and conclusions).
Results: Although the different types of SRs are developed following the same steps, each SR type requires specific methods, differing in characteristics and complexity. The extent of methodological development varies by type of SR, with more solid guidelines available for diagnostic accuracy and effects of interventions SRs.
This methodological toolkit describes the most up-to-date risk of bias instruments: Quality in Prognostic Studies (QUIPS) tool and Prediction model study Risk Of Bias Assessment Tool (PROBAST) for prognostic SRs, Quality assessment of diagnostic accuracy studies tool (QUADAS-2) for diagnostic accuracy SRs, Cochrane risk of bias tool (ROB-2) and Risk of bias in non-randomised studies of interventions studies tool (ROBINS-I) for effects of interventions SRs, as well as the latest developments on the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system.
Conclusions: This structured compilation of the best methodological resources for each type of SR may prove to be a very useful tool for those researchers that wish to develop SRs or conduct methodological

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 3 (revision) 14 Oct 2020	✓ report	
version 2 (revision) 11 Aug 2020	✓ report	✓ report
version 1 04 Feb 2020	? report	? report

1. **Miranda Cumpston**, Monash University, Melbourne, Australia
University of Newcastle, Newcastle, Australia

2. **Edward Purssell**, City, University of London, London, UK

Any reports and responses or comments on the article can be found at the end of the article.

research works on SRs

Keywords
Systematic reviews, prevalence, prognostic, diagnostic accuracy, efficacy of interventions

Corresponding author: Marta Roqué (mroque@santpau.cat)

Author roles: Roqué M: Conceptualization, Methodology, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Martínez-García L**: Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Solà I**: Writing – Original Draft Preparation; **Alonso-Coello P**: Writing – Original Draft Preparation; **Bonfill X**: Writing – Original Draft Preparation; **Zamora J**: Conceptualization, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: The authors are members of CIBERESP (Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública - Biomedical Research Center Network of Epidemiology and Public Health), and hold active roles within Cochrane and the GRADE Working Group.

Grant information: Roqué M is supported by the Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP) as part of a Training Programme call for "Internal mobility: Internships in CIBERESP groups", within the framework of the subprogramme 7.4 "Methodology, clinical records and scientific dissemination." Martínez-García L has a Miguel Servet contract from the Institute of Health Carlos III [CP18/00007].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Roqué M, Martínez-García L, Solà I *et al.* **Toolkit of methodological resources to conduct systematic reviews [version 3; peer review: 2 approved]** F1000Research 2020, **9**:82 https://doi.org/10.12688/f1000research.22032.3

First published: 04 Feb 2020, **9**:82 https://doi.org/10.12688/f1000research.22032.1

REVISED **Amendments from Version 2**

This version incorporates a minor change in response to the peer reviewer comments, and corrected a mistake in the Acknowledgements section.

Any further responses from the reviewers can be found at the end of the article

Introduction

Systematic reviews (SR) are studies that use a systematic and explicit method to identify, analyse and synthesize empirical evidence, and to answer a specific research question[1]. Therefore, SRs are key tools to make informed health choices[2,3].

All SRs are based on a specific research question. Classic epidemiological research questions relate to the prevalence of a medical condition, the associated prognosis of the medical condition (including incidence or global prognosis, prognostic factors associated to the condition's incidence or outcome, and risk profiles defined by prognostic models[4]), diagnostic accuracy of tests that allow us to diagnose the medical condition, and effects of interventions to treat the medical condition. SRs can be classified by the type of research question they answer, as shown in Table 1.

The stages to develop an SR are common to all the types of SRs: 1) Formulating the research question, 2) development of the protocol that explicitly describes the methods to carry out each step of the SR, 3) literature search, 4) risk of bias assessment, 5) synthesis of findings, 6) assessment of the quality of evidence, and 7) report of SR results and conclusions[1]. Although the different types of SRs share the same structure and follow a similar development process, their methods can be different and more or less complex depending on the type of SR.

Nowadays there are numerous methodological resources to conduct reviews, especially for intervention SRs and diagnostic SRs. However, the scattering of these resources and the lack of widely established manuals or recommendations are, in many situations, an obstacle to access them, especially for prevalence SRs and prognostic SRs. Therefore, the objective of this review is to identify and describe the methodological resources available to develop prevalence SRs, prognostic SRs, diagnostic accuracy SRs and effects of interventions SRs.

Table 1. Research question by type of systematic review.

Type of systematic review	Acronym for the research question	Example of research question
Prevalence review	**CoCoPop-S** (*condition, context, population and study design*)	What is the prevalence of frailty and prefrailty (condition) in community-dwelling older adults (population) living in low- and middle-income countries (context)?[5] What is the worldwide (population) prevalence of insufficient physical activity (condition)?[6]
Prognostic review - global prognosis	**CoCoPop-S** (*condition, context, population and study design*)	What is the incidence of dementia (condition) in individuals of at least 60 years of age (population) living in high-income countries (context)?[7]
Prognostic review- prognostic factors	**PICOT-S** (*population, intervention or factor, comparison, outcome, time and study design*) **PFO-S** (*population, factor or model, outcome and study design*)	Is protease activity (prognostic factor) an independent prognostic factor for wound healing (outcome) at 24 weeks (timeframe) in people with venous leg ulcers (population)?[8]
Prognostic review- prognostic models	**PICOT-S** (*population, intervention or factor, comparison, outcome, time and study design*)	What is best prognostic model to predict overall or progression- free survival (outcome) in patients with chronic lymphocytic leukaemia (condition)?[9]
Diagnostic accuracy review	**PIRD-S** (*population, index test, reference test, diagnosis of interest and study design*)	Do self-reported frailty to predict survival in adults with bacterial meningitis screening instruments (index test) accurately identify older people (population) at risk of frailty and prefrailty (condition of interest)?[10] Is PET 18F florbetapen (index test) useful in early diagnosing dementia (condition) in patients with mild cognitive impairment (population)?[11]
Effects of intervention review	**PICO-S** (*population, intervention, comparison, outcome of interest and study design*)	What is the effect of ribavirin (intervention) in patients with Crimean Congo haemorrhagic fever to prevent death (outcome)?[12] Does comprehensive geriatric assessment (intervention) in older adults (population) reduce mortality (outcome)?[13]

Methods

Information sources and search strategy

We consulted the guidelines from the main organizations that establish methods to conduct SRs (Cochrane, Joanna Briggs Institute, European Network for Health Technology Assessment (EUNETHTA), Enhancing the Quality and Transparency of Health Research (EQUATOR) network, Grading of Recommendations Assessment, Development and Evaluation (GRADE)) in order to identify their proposed resources.

Additionally, we performed a literature search in MEDLINE (accessed through PubMed) in November 2019 using the following search syntax: ((“Review Literature as Topic”[Mesh] OR systematic review*[tiab]) AND (handbook*[ti] OR methodolog*[ti] OR manual[ti] OR guide[ti]).

We also performed ad hoc scientific literature searches to find other resources for each type of SR in relation to the research question structure, the literature search strategy, the risk of bias assessment and the statistical analysis.

Eligibility criteria

We included the resources available to design prevalence SRs, prognostic SRs, diagnostic SRs and intervention SRs.

We excluded the methodological resources to develop other types of SRs (methodological, economic evaluation and qualitative research SRs, or overviews).

Data selection and extraction

The authors are members of CIBERESP (Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública - Biomedical Research Center Network of Epidemiology and Public Health), hold active roles within Cochrane and the GRADE Working Group, and are experts in different fields of knowledge (statistics, development of Cochrane reviews, research methodology, information retrieval, development of clinical guidelines). They evaluated the search results, selected the most relevant and accurate resources, and summarized the most relevant information by development stage and type of SR.

The resources were selected based on the authors expert judgement, prioritising those resources which were endorsed or part of a guideline from the organisations cited above, and those which were more recent. The resources were organised in 7 sections, following the development stages of an SR: 1) Formulating the research question, 2) development of the protocol and review registration, 3) search strategy, 4) risk of bias assessment, 5) statistical synthesis of findings, 6) quality of evidence assessment, and 7) results report and presentation. The resources are presented by type of SR in each section, and an example of their use is included[5–13].

For each pre-defined section, the authors selected and summarized the methods that were considered to be more rigorous and widely accepted, prioritizing major methods applicable to all reviews over more controversial methods, or methods which required highly specialized knowledge. The text organises the results pedagogically with the aim to highlight key differences between review types, present the key characteristics of each method, and be a comprehensive tool that contains the most relevant advice based on the authors judgement.

Results

We identified guidance handbooks, primary studies and reporting guidelines as a result of the bibliographic searches. The resources selected are presented in Table 2.

We have identified methodological guidelines dedicated to the development of prevalence SRs[14], global prognosis[15], and prognostic factor SRs[16–18].

During the performed search, we identified methodological manuals to develop prognostic model SRs in the series of publications from the PROGRESS project[19], and in the resource compilation from Cochrane’s Prognosis Methods Group.

For diagnostic accuracy SRs and effects of interventions SRs, we have identified the methodological manuals developed by Cochrane Collaboration are available[1,20]. The recommendations drawn from the guidance handbooks identified are complemented whenever necessary with specific primary method studies identified in our search.

Formulating the research question

The type of SR is determined by the **research question**, which must be formulated in a structured manner as shown in Table 1. Careful development of the research question is vital, since the SR inclusion criteria will stem from it.

Prevalence review. Prevalence SRs aim to answer the question “How common is a health problem in a specific population?” Prevalence SRs focus on existing cases at a given time, measure the global burden of a health problem, and describe the characteristics of the affected population, the geographical distribution of that problem and its variation among subgroups. The structure of the research question must include the elements of condition, context, population and study design (CoCoPop-S)[21], as shown in Table 1. The most adequate study designs to estimate the prevalence would be population registers or cross-sectional studies that include population-representative samples. For instance, Guthold *et al.* (2018) considers studies based on population surveys as a reliable source of information to obtain global prevalence estimators of insufficient physical activity[6].

Prognostic review. SRs of prognosis are mainly based on three types of research questions: 1) “What is the risk of an specific population to have a health problem?”, descriptive question (review of global prognosis) that focuses in new cases occurring within a period of time (incidence), 2) “what factors are associated with or determine a specific outcome?”, an explanatory question (review of prognostic factors), and 3) “are there risk profiles that have higher probability of presenting specific outcomes?”, a result prediction question (review of prognostic models or risk prediction). We have excluded from the

Table 2. Organisation of resources by type.

Best practice manuals and chapters of manuals	Primary methods	Reporting guidelines
JBI manual	Hemingway 2013[4]	Moher 2009[72]
Aromataris 2017[14]	Munn 2018[21]	Moher 2015[73]
(Munn 2017[15]	Iorio 2015[22]	Beller 2013[74]
Moola 2017[16]	Bossuyt 2006[23]	Zorzela 2016[75]
Campbell 2017[31])	Lijmer 1999[25]	McInnes 2018[77]
Cochrane DTA manual	Strauss 2010[26]	Moher 2007[78]
Deeks 2010[20]	Ge 2018[27]	Page 2016[79]
(Bossuyt 2008[24]	Page 2018[28]	Salameh 2019[80]
deVet 2008[32]	Atkinson 2015[30]	
Macaskill 2010[60])	Lefebvre 2013[33]	
Cochrane intervention manual	Glanville 2006[34]	
Higgins 2019[1]	Wilczynski 2004[35]	
(Lefebvre 2019[29]	Beynon. 2013[36]	
Higgins 2019[55]	Sampson 2011[37]	
Deeks 2019[58]	Bramer 2017[38]	
Chaimani 2019[63]	Hartling 2016[39]	
Schünemann 2019[68])	Glanville 2014[40]	
GRADE Working group manual	Isojarvi 2018[41]	
Schünemann 2013[64]	Horsley 2011[42]	
(Schünemann 2020[65]	Gentles 2016[43]	
Schünemann 2020[66]	Hartling 2017[44]	
Santesso 2019[67])	Booth 2016[45]	
PROGRESS project	Rethlefsen 2014[46]	
Riley 2019[17]	Rethlefsen 2015[47]	
Debray 2017[19]	Spencer 2018[48]	
Dekkers 2019[18] manual	Hoy 2012[49]	
	Hayden 2013[50]	
	Morgan 2018[51]	
	Morgan 2019[52]	
	Wolff 2019[53]	
	Whiting 2011[54]	
	Sterne 2016[56]	
	Lau 1997[57]	
	Popay 2006[59]	
	Rutter 2001[61]	
	Rücker 2008[62]	
	Murad 2017[69]	
	Harder 2017[70]	
	Huguet 2013[71]	
	Campbell 2020[76]	

aim of this project a 4th type of prognostic question, known as stratified medicine, and that alludes to the use of prognostic information to individualise therapeutic choices in a group of people with similar characteristics[4].

Structured questions about global prognosis must specify population, outcome, condition to be predicted, context and time frame to determine the incidence (CoCoPop-S). The study designs that provide more reliable incidence estimates are prospective cohort studies with representative samples[15,22]. Structured questions regarding either prognostic factors or models must include population; exposure in terms of the prognostic factor or model of interest, including how it is measured, the intensity and the exposure time; outcome, condition to be

predicted; follow-up time; and context (PICOT-S or PFO-S)[19,21]. The best study designs to evaluate prognostic factors or models are also prospective cohort studies. For instance, Westby *et al.* (2018) published a prognostic factor SR that gives priority to the inclusion of cohort studies and, if none is found, it resorts to including case-control studies, which also explore the association of prognostic factors with the outcome of interest, although with less reliability[8].

Diagnostic accuracy review. Diagnostic SRs aim to answer the question "How good is a test to identify or dismiss the presence of a condition or health problem in a particular population, in comparison with a reference test?" The research question can be posed with the elements of population, index test, reference test, diagnosis of interest and study design (PIRD-S)[21]. The SR approach will depend on the role of the index test in the clinical diagnostic pathway: if it replaces another test, if it will be used in addition to another test to refine the diagnosis, or if it is a triage test previous to other tests[23,24].

Diagnostic SRs preferentially include cross-sectional studies, where the participants are evaluated using the index test and/or the reference test to determine if they have the condition of interest. Case-control designs are subject to risk of bias and their inclusion in diagnostic SRs is not recommended[25]. For instance, Ambagtsheer *et al.* (2017) include in their SR cross-sectional studies where one or more self-reported frailty screening scales have been compared with one of three reference standards: frailty phenotype, frailty index or comprehensive geriatric assessment[10].

Effects of interventions review Interventions SRs aim to answer the question "What effect does a specific intervention have on the relevant outcomes in people with a particular health problem, in comparison with a reference intervention?" The research question is posed with the elements of population, intervention, comparator, outcomes of interest and study design (PICO-S)[1].

The randomised clinical trial (RCT) is the most appropriate study design to evaluate the effects of an intervention, as it is the design with less risk of bias and that best helps to establish causality. In cases where it is not possible to conduct randomised trials for ethical or organizational reasons, non-randomised trials, before-after studies, time series, cohort studies or case-control studies can be considered for their inclusion in the SR[1]. For instance, the SR by Johnson *et al.* (2018) regarding ribavirin for treating Crimean Congo haemorrhagic fever included both RCTs and non-randomised trials to use the available data, given the previous lack of preparedness for experimental research therapeutics in outbreak situations, but concludes that estimates of effect based on the existing literature are highly uncertain due to confounding in non-randomised studies[12].

Development of the protocol and review registration

Writing the SR protocol is a fundamental step that must be done before designing an SR. Herein, the stages and methods to be applied during the development of the SR can be pre-specified. The identified guidelines can be used to identify the methods that need to be stated in the protocol, and some have specific chapters on protocol development[1,14,20].

Similarly to the requirement of clinical trial registration, the SR should also be registered in order to avoid redundancies and, more importantly, to avoid reporting bias, therefore guaranteeing transparency and rigor during the development of the SR[26]. Prospective registration of an SR protocol is recommended by the PRISMA guidelines and is associated with higher SR methodological quality[27]. The largest and most well-known SR register is PROSPERO, produced by the Centre for Reviews and Dissemination in York. With PROSPERO, it is possible to prospectively register any type of review, provided that its aim is a health-related outcome. It contains more than 30,000 entries[28]. All Cochrane SR protocols are published in Cochrane Library and automatically registered in PROSPERO.

Search strategy

Designing a comprehensive research study for an SR is vital in order to reduce bias when identifying studies, and it is important to describe it in the relevant section within the protocol in a transparent and thorough manner to facilitate its evaluation by third parties and its reproducibility.

Methodological reference standards to design comprehensive searches have been published[29,30]. In addition, methodological manuals to develop SRs provide guidelines for diagnostic and effects of interventions SRs[31–33].

The design of the search strategies does not differ by type of SR, but rather their differences are due to the elements of the research question and the design of studies to be identified. In general terms, electronic searches are designed to identify bibliographic references that use a language similar to the elements of the review's clinical question. To this effect, the strategies are built based on the elements of the structured clinical question. Search algorithms use a combination of natural language and the appropriate controlled vocabulary for each bibliographic database. Validated filters can be applied to these strategies to determine specific study designs that can be useful to identify, among others, clinical trials[32–34], or prognostic studies[35]. However, the use of filters is controversial in diagnostic accuracy studies[32,36].

Search performance will vary depending on the type of studies that are included in the SR. Thus, in intervention SRs, the search results for RCTs are more precise (they have a higher proportion of relevant references among all the references that the search has identified), due to better indexation of this type of studies in bibliographic databases. On the contrary, in SRs that include observational studies, like prognostic SRs, identifying studies is more complex given the variability of designs to be included and its poorer indexation in databases, which results in less specific literature searches that lead to a longer and more complex study selection process[17].

Searches must be designed to optimise their sensitivity (the ability to retrieve as many relevant study references as possible), which is a feature that tends to be a detriment to precision, which in

SRs ranges on an average of 3%[37]. To obtain an efficient search with adequate sensitivity, performing searches in MEDLINE and EMBASE may be sufficient, particularly in intervention reviews, as they are the two most frequently used bibliographic databases[38], and they are enough to identify most relevant studies for a specific SR[39]. These searches can be complemented with additional searches in other databases such as PEDro, which provide specific information for certain topics.

Searching in bibliographic databases can be completed with additional strategies, such as checking public trial registers[40,41], searching in the reference list of relevant studies[42], or cross-searching citations[43]. Searching grey literature, understood as any document that is not published in biomedical or scientific journals, has a limited impact in effects of interventions SRs[44], but offers good results in other types of SRs, such as qualitative evaluation SRs[45].

If we take into consideration the methodological and technical challenges that the design and implementation of search strategies pose, involving a medical librarian can be desirable to improve the search quality[46–48].

Risk of bias assessment

Assessing the risk of bias of the included studies is a key element in any SR. It helps evaluate and interpret the included studies results, and it is a determinant of the evidence quality of the SR results. The current tools to assess risk of bias are organised by domains, which roughly correspond to the classic epidemiological biases related to each type of research question. The identified tools to assess risk of bias are presented in Table 3, organised by type of SR and by domain of epidemiological bias assessed.

Each of the domains of these tools includes a number of index questions related to specific aspects of study design or development that can lead to a bias in that domain. The tools can be adapted *a priori* to each review, modifying or deleting questions, or adding new questions specific to the considered research question. The process to assess risk of bias is similar in all the current scales. Firstly, they identify the risk of bias in each domain based on the answers to the questions, and secondly, they integrate these risks in a risk of bias assessment for each health problem, prognostic factor, diagnosed condition or outcome of interest assessed, depending on the type of SR.

Prevalence review. The tool to assess risk of bias by Hoy *et al.* (2012) is available for prevalence SRs. It assesses internal and external validity aspects in the prevalence study[49]. The tool comprises 10 questions where a judgement of high or low risk of bias is made. Based on the answers, the researcher makes a subjective assessment of the study's overall risk of bias as low, moderate or high[49].

Prognostic review. There is no scale available to assess the risk of bias in **global prognostic studies**, although a series of criteria has been proposed to assess risk of bias. These are classified in 1) definition and representativeness of the population, 2) completeness of follow-up, and 3) objective and unbiased measurement of outcome of interest[22]. However, some authors like Roerh *et al.* (2018) use a version of the scale to assess risk of bias designed by Hoy *et al.* (2012), adapted to the assessment of incidence studies considering the duration of the incidence period[7].

For the **prognostic factor** studies, the tools QUIPS and "RoB instrument for NRS of exposures" were identified[50–52]. The QUIPS tool helps assess the risk of bias using 31 questions divided in 6 domains. For each domain, a judgement of high, low or unclear risk of bias is made. Before using the tool, one must carefully consider the potential confounders that can lead to bias. Clinical experts in the specific topic of the SR should participate. The tool "RoB instrument for NRS of exposures" evaluates the risk of bias using 32 questions divided in 7 domains, including a key domain regarding confounders and a domain regarding departures from intended exposures. For each domain, a judgement of critical, serious, moderate or low risk of bias is made. An example of the use of the QUIPS scale can be seen in the review by Westby *et al.* (2018). The authors defined *a priori* two key confounders (age and infection), which the experts and the literature described as prognostic factors for their condition of interest (venous leg ulcers), and which were simultaneously associated with the prognostic factor of interest in the SR (protease activity biomarker). These two confounders were included in the QUIPS scale in the section of control by confounders[8].

We identified the Prediction model Risk Of Bias ASsessment Tool (PROBAST) for the **prognostic** model SRs[53]. This tool assesses the risk of bias using 20 questions divided in 4 domains (participants, predictors, outcome and analysis). For each domain, a judgement of high, low or unclear risk of bias is made. The questions vary according to the aim of the study (development, validation, or development and validation of the prognostic model).

Diagnostic accuracy review The tool QUADAS-2, which evaluates 11 questions divided in 4 domains, is available to assess the risk of bias in diagnostic accuracy studies[54]. For each domain, a judgement of high, low or unclear risk of bias is made. In addition, the external validity or study applicability in relation to the SR is assessed in each domain.

Diagnostic SRs mainly include observational studies, which are more subject to risk of bias, and therefore adapting the QUADAS-2 tool, modifying or adding specific questions to the SR topic, is virtually a requirement during the protocol stage. For instance, the SR by Martínez *et al.* (2017) studied the diagnostic accuracy of an imaging test (amyloid PET) that requires complex visual interpretation. For this reason, a question was included in the QUADAS scale to assess whether the test interpretation was performed by trained readers[11].

Effects of interventions review. For intervention SRs, the Risk of Bias (RoB) 2.0 tool is available to assess the potential bias in randomised clinical trials, and the Risk Of Bias In

Table 3. Tools to assess risk of bias by type of systematic review.

	Scale (n items)	Selection bias (number of items)	Exposure and performance bias (number of items)	Outcome detection bias (number of items)	Attrition bias(number of items)	Confounder bias (number of items)	Selective outcome reporting bias (number of items)	Other biases (number of items)
Prevalence review	**Hoy 2012 (10)**[49]	- Representativeness of population sample (1) - Sample and recruitment (2)	(0)	- Data collection (2) - Case definition and timeframe for prevalence (2) - Reliability of measuring instrument (1)	- Impact of missing data (1)	(0)	(0)	- Appropriate computation of prevalence estimator (1)
Prognostic review-prognostic factors	**QUIPS (31)**[50]	- Study participation (3) - Sample and recruitment (3)	- Prognostic factors definition and measurement (6) - Confounders definition and measurement (4)	- Outcome definition and measurement (3)	- Description and impact of attrition (6)	- Statistical analysis of confounding factors (2)	- Selective reporting of results (1)	- Statistical analysis (3)
	RoB for NRS - exposures (32)[52]	- Selection of participants (5)	- Exposure definition and measurement (5) - Deviations from intended exposure (4)	- Outcome definition and measurement (5)	- Description and impact of attrition (5)	- Statistical analysis of confounding factors (6)	- Selective reporting of results (3)	(0)
Prognostic review-prognostic models	**PROBAST (20)**[53]	- Design of study and selection of participants (2)	- Prognostic factors definition and measurement (3)	- Outcome definition and measurement (6)	- Inclusion of participants in the analysis (2)	(0)	- Selective reporting of results (1)	- Statistical analysis (6)
Diagnostic accuracy review	**QUADAS-2 (11)**[54]	- Selection of participants (3)	- Index test interpretation (1) - Threshold specification for index test (1)	- Adequacy and interpretation of reference test (2) - Time interval between tests, and coverage of reference test (3)	- Inclusion of participants in the analysis (1)	(0)	(0)	(0)
Effects of intervention review	**ROB-2 (16)**[55]	- Selection of participants (randomisation, concealment, and basal imbalances) (3)	- Blinding of participants and personnel (2) - Deviations from intended intervention (2)	- Blinding of outcome detection (2)	- Impact of attrition (3)	(0)	- Selective reporting of results (2)	- Analysis of participants in the allocated intervention arm (2)
	ROBINS-I (35)[56]	- Selection of participants (6)	- Classification of intervention (3) - Deviations from intended intervention (6)	- Outcome measurement (4)	- Description and impact of attrition (5)	- Confounders (8)	- Selective reporting of results (3)	(0)

No risk of bias tool has been identified for global prognosis systematic reviews. The number of items in the risk of bias tools may vary depending on the effect of interest and the included study designs, as well as the addition or suppression of index questions by the researchers to tailor the tool to the SR.

Non-randomised Studies - of Interventions (RoBiNS-I) tool in non-randomised clinical trials[55,56]. The RoB 2.0 tool includes 16 questions divided in 5 domains, including a specific domain for randomisation and a domain for deviations from intended interventions[55]. The number of questions may vary, depending on the effect of interest and the design of the study assessed. For each domain, a judgement is made: high or low risk of bias, or some concerns. For instance, in their SR, Ellis *et al.* (2017) assessed the risk of bias in the evaluation of results separately for the objective outcomes (such as living at home or death) and for the subjective outcomes, showing a lower risk of bias in the evaluation of the objective outcomes[13].

The RoBiNS-I tool assesses the biases that the non-randomised study has when compared with an ideal, pragmatic, unbiased randomised trial, which answers the clinical question of interest (even if this ideal trial may not be feasible or ethical)[56]. RoBiNS-I has 34 questions divided in 7 domains, including a key domain regarding confounders and a domain for deviations from intended interventions. As in the case of prognostic SRs, there should be an *a priori* careful consideration of the potential confounders that must be included in the tool to assess individual studies. A judgement of critical, serious, moderate or low risk of bias is made for each domain. A low risk of bias implies that the non-randomised study is comparable to a well-performed randomised trial. For instance, Johnson *et al.* (2018) excluded from their analyses the non-randomised studies that showed a critical risk of bias according to RoBiNS-I, rejecting 18 out of the 22 included studies[12].

Statistical synthesis of findings

SRs may include a section with a quantitative statistical synthesis or meta-analysis, where a combined estimator of the parameter of interest is obtained from the estimators of the individual studies. Table 4 shows a non-exhaustive compilation of the main characteristics of the meta-analysis methods and the main software commands for each type of SR.

A necessary previous step to any meta-analysis is the evaluation of the existing clinical and statistical heterogeneity in the set of studies, which will inform us 1) if it is reasonable to perform a quantitative synthesis of findings, 2) what meta-analysis model we should apply, and 3) if additional investigation of the causes of heterogeneity is required, for example, subgroup and sensitivity analyses, or meta-regressions[57,58]. In those cases when a quantitative synthesis is precluded, the SR will be restricted to a narrative synthesis. A narrative synthesis should not simply summarize the findings from the included studies in order to draw conclusions about the body of evidence, but instead should be a more formal process which includes a formulation of the theory of how the intervention works, why and for whom, the exploration of the relationships in the data, and the assessment of the robustness of the synthesis[59].

When it is reasonable to perform a statistical synthesis, there are two main models to conduct a meta-analysis: fixed effects model and random effects model. For practical purposes, the chosen model determines how the studies included in the meta-analysis will be numerically weighed. Both models are

Table 4. Methodological characteristics of meta-analysis by type of systematic review.

	Measures to combine	Assessment of heterogeneity	Model	Method	Command (package)
Prevalence review	- Proportion (prevalence)	- Qualitative	- Fixed/Random effects	- Inverse-variance method[a]	- Metaprop (Stata)
Prognostic review - global prognosis	- Cumulative incidence - Incidence rate	- Meta-regression	- Fixed/Random effects	- Inverse-variance method[b]	- Metan (Stata) - Metaprop (Stata) - Review Manager
Prognostic review- prognostic factors	- Hazard Ratio - Odds Ratio	- Meta-regression	- Random effects	- Inverse-variance method	- Metafor (R)
Prognostic review- prognostic models	- Calibration - Discrimination	- Meta-regression	- Random effects	- Multivariate methods	- Metamisc (R)
Diagnostic accuracy review	- Sensitivity - Specificity	- Meta-regression	- Random effects	- HSROC method[c] - Bivariate model	- Metadas (SAS) - Metandi (Stata)
Effects of intervention review	- Mean difference - Risk difference - Standardised mean difference - Hazard Ratio - Incidence rate ratio - Odds Ratio - Risk ratio	- I^2 - Meta-regression	- Fixed/Random effects	- Mantel-Haenzsel method - Multivariate methods	- Metafor (R) - Metan (Stata) - Review Manager

[a] Tukey-Freeman or logit transformation. [b]Transformation for the cumulative incidence. [c] Hierarchical summary receiver operating characteristic (HSROC) method allows estimation of a Receiver operating characteristic (ROC) curve or sensitivity and specificity indexes.

based on different assumptions regarding distribution of effects and heterogeneity in the set of studies, and they differ in their application and interpretation[58].

Finally, there is a variety of resources to conduct meta-analyses, from specific programs to perform meta-analyses (free or paid) to user-defined routines using general statistics packages (SAS, Stata, SPSS), as well as Excel utilities or R libraries. An archive with software and utilities is available from SR Tool Box.

Due to the complexity of the statistical techniques to synthesise results, and the difficulty to standardise methods and decisions to be made during the analysis, it is vital to involve a statistician in the planning and conduct stages of the meta-analysis, especially for prognostic and diagnostic SRs.

Prevalence review. In prevalence SRs, the meta-analysis combines ratios, which are transformed to be meta-analysed using the inverse-variance method[58]. Siriwardhana *et al.* (2018) calculated combined frailty prevalence estimates using a random effects model. The authors assessed that there was high clinical heterogeneity between the studies in terms of actual frailty prevalence, geographic setting, frailty assessment method, cut-off points applied and sample age, although this heterogeneity did not rule out performing a meta-analysis[5].

Prognostic review. In global prognostic SRs, the meta-analysis combines cumulative incidence ratios or incidence rates, while in prognostic factor SRs, the meta-analysis combines odds ratios or hazard ratios, which can be presented in individual studies as raw estimates or as covariate-adjusted estimations derived from logistic or Cox regression models. If combining adjusted estimates, all of them should be adjusted by a minimum set of common factors[17]. In prognostic model SRs, the meta-analysis combines estimates of model discrimination and calibration. These indicators can be synthesised separately or jointly using multivariate models[19].

Prognostic studies usually show significant variability in terms of design, sample case-mix, measurement instruments, analysis methods and presentation of results[17]. Therefore, in prognostic factor and model SRs, it is recommended to perform the meta-analysis using the random effects model, and even to use multivariate meta-analysis methods adjusting for relevant factors[17]. For instance, the SR by Westby *et al.* (2018) describes how the authors dismissed performing a meta-analysis due to the high risk of bias and the extreme heterogeneity across the included studies in terms of population, measurement of the prognostic factor (cut-off points and analytical methods) and outcome measurement[8].

Diagnostic accuracy review. In diagnostic SRs, the meta-analysis combines estimates of sensitivity and specificity of the index test. The meta-analysis in diagnostic SRs shows a higher degree of complexity because the studies may have used different thresholds, both implicit and explicit, to define a positive result in the evaluated test. This leads to a correlation between the sensitivity and specificity indexes, which must be modelled jointly using multivariate methods[60]. The most common available statistical methods are the bivariate hierarchical model and the HSROC model (Hierarchical summary receiver-operating characteristic)[61]. Diagnostic SRs tend to combine studies with very heterogeneous results, and it is recommended to use the random effects model by default and perform a comprehensive examination of the sources of heterogeneity using meta-regression[60]. For instance, the protocol of the SR by Ambagtsheer *et al.* (2017) expects to estimate an average sensitivity and specificity for the frailty scales, when the included studies have applied the same explicit cut-off points to the considered scales. However, given that they are subjective, self-reported scales, the studies could share the same explicit cut-off point, and yet that cut-off point could correspond to different levels of frailty in the studies (implicit thresholds), which will advise against calculating pooled estimates of diagnostic accuracy[10].

Effects of interventions review. In intervention SRs, the meta-analysis combines different measures, depending on the type of outcome: odds ratio or risk ratio for binary outcomes, mean difference or standardised mean difference for continuous outcomes, hazard ratio for time-to-event outcomes, and incidence rate ratios for outcomes that count number of events.

In intervention SRs, the I^2 estimator has been proposed to assess statistical heterogeneity as a supplement to the assessment of clinical and methodological heterogeneity. This indicator is defined as the percentage of the overall variability that cannot be explained by chance, and has values ranging from 0% to 100%; with higher values indicating higher statistical heterogeneity[58]. For instance, I^2 was one of the aspects considered in the SR by Ellis *et al.* (2017) to assess the inconsistency in results, and to decide if a meta-analysis combining the results would be performed[13]. Despite its popularity and ease of interpretation, the use of this indicator is not exempt of controversy due to its dependence on the number of studies and sample size; thus, a small statistical heterogeneity could seem substantial only by the effect of a arge sample size of the included studies[62].

In intervention SRs, pairwise metanalysis has been extended to network metanalysis, which allows the simultaneous comparison of three or more interventions, combining direct and indirect evidence from a network of studies[63].

Quality of evidence

The quality (also confidence or certainty) of evidence in an SR is the degree of confidence that is held against the fact that an estimate of effect or association is close to the actual value of interest[1]. Certainty of evidence is best evaluated with the GRADE system. Certainty in the obtained estimates for each one of the key SR outcomes or factors is classified as high, moderate, low or very low. A level of certainty of evidence is first established from the design of the studies that form the evidence body, which might or might not have an optimal design for the type of considered question. This initial confidence in the evidence body can then decrease in one or two levels if the following is detected: 1) design or execution limitations, 2) inconsistency between estimates, 3) indirect evidence, 4) imprecision in estimates, or 5) publication bias[64].

The certainty of evidence is a key element to interpret and communicate results, and as such, it should be included in the sections of results, discussion, conclusion and abstract, using semi-standardised statements[65]. Additionally, it can be included in a Summary of Findings table, where for each comparison, the key information regarding relative effect and absolute effect magnitude, quantity of available evidence and its certainty is presented[66]. Certainty of evidence can be assessed too when no quantitative synthesis is possible[67].

We will now highlight the specific aspects in which the GRADE system adapts to each type of SR.

Prevalence review. There are no formal adaptations of the GRADE system for prevalence SRs, but there is a proposal to assess the quality of the evidence based on this system[68]. High initial certainty is awarded to survey or cross-sectional study designs with population representativeness that have been properly designed and conducted, while studies with no population representativeness will have lower initial quality.

Prognostic review. There is a GRADE proposal for global prognostic SRs[22] and an adaptation for prognostic factor SR[69]. Guidelines for prognostic model SR are still under development.

In **global prognostic** SRs, the study designs that have high initial certainty are longitudinal cohort studies and pragmatic randomised controlled trials with representative samples[22]. Other observational designs would offer low initial certainty. In **prognostic factor** SRs, explanatory and confirmatory longitudinal designs offer high initial certainty, while exploratory studies are considered to be of moderate quality[69].

In prognostic SRs, the assessment of the limitations follows the general procedure already described, with two particularities: 1) qualitative assessment of inconsistency, because of low reliability of I^2 estimator in the prognostic field[22,69], and 2) possibility of increased certainty in the studies that do not show limitations in the quality of evidence, if (i) the estimated effect magnitude is substantial, or (ii) there is an exposure-response gradient[67]. For instance, the prognostic factor SR by Westby *et al.* (2018) considered the possibility of increasing the certainty of evidence in the studies presenting no limitations. Due to the exploratory nature of the included studies and their high risk of bias, the certainty was not increased in any case and the evidence obtained in the review was of very low quality[8].

Diagnostic accuracy review. There is a GRADE proposal for assessing the certainty of evidence for test accuracy[70,71]. The study designs that start with the highest degree of evidence are cohort or cross-sectional studies where the index test and an appropriate reference standard have been directly compared in patients with diagnostic uncertainty[70]. If the SR included case-control studies, these would offer low-quality initial evidence[25].

Indirectness of evidence would be assessed through any applicability concerns of the patient sample, the intervention and the comparator with respect to the clinical pathway where the test is to be applied. There is uncertainty regarding how to assess inconsistency, because heterogeneity is common and hard to quantify in diagnostic SRs, and it often cannot be explained even if multivariate models are adjusted. Judgments on extent of heterogeneity should be based on similarity of the point estimates, overlap of confidence intervals, and the exploration of possible explanations for the inconsistency from subgroup or sensitivity analyses[70].

Imprecision judgments should be based on both the width of the confidence or credible intervals for sensitivity and specificity, but also on the implications for patient management in terms of true and false positives, and true and false negatives. When the estimated intervals include values that may lead to different conclusions of the test's value, the certainty of the evidence may be lowered[71].

With regard to the criteria to increase the level of evidence, it is unclear whether they should be applied at all and how to do it in diagnostic SRs[71]. The uncertainty surrounding the process of assessing the quality of evidence in diagnostic SRs explains why it is not a requirement in Cochrane SRs at the moment. For instance, the SR by Martínez *et al.* (2017) only included a Summary of Findings table with numerical results and an estimation of the absolute effect that the test would have on a hypothetical cohort of individuals[11].

Effects of interventions review. The GRADE system for assessing the quality of evidence was initially developed for intervention SRs, and it is the indication for which clearer and widely agreed guidelines are available[64]. In terms of study design, RCTs are initially classified as having high certainty, while all non-randomised or observational studies are classified as having low certainty. This proposal for pairwise metanalysis can be extended to network metanalysis[63].

The assessment of the certainty limitations is well-defined in intervention SRs. Inconsistency can be assessed using the I^2 estimator[64]. Imprecision is assessed taking into account whether the review meets the optimal information size, and whether the confidence interval of the effect estimate allows reaching a conclusion, because either it only includes values consistent with a relevant intervention effect, or it completely dismisses it[63]. In observational studies that do not have limitations in the quality of evidence, three criteria are considered to increase certainty: 1) the estimated effect magnitude is important or very important, 2) there is an exposure-response gradient, and 3) all possible biases that could reduce the observed effect confirm the obtained conclusions.

For instance, the SR by Ellis *et al.* (2017) applied the GRADE system to the included randomised trials, and it concluded that there was high certainty of the effect of the comprehensive geriatric assessment on the effects outcomes based on a high number of studies and participants, with a globally low risk of bias, and results consistent among studies. However, the certainty of evidence obtained in cost-effectiveness was low, due to imprecision and inconsistency of results[13].

Results report

It is vital to inform about the methods, results and conclusions of the SRs in a transparent and thorough manner so that their users can interpret, evaluate and apply them. The EQUATOR initiative has developed, and keeps up-to-date, a library with guidelines to communicate the different types of research studies. The PRISMA statement (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) has been proposed in the SR field[72]. This statement consists of a checklist comprised of 27 items and a flow diagram to present the number of studies considered in the SR. In addition, several extensions focusing on reporting specific aspects of SRs have been developed, such as PRISMA-P for reporting SR protocols[73], PRISMA- *Abstracts* for reporting abstracts[74], and PRISMA- *Harms* for reporting harms outcomes in SRs[75]. Additionally, the SWiM guideline is available for reporting intervention SRs where the effects of interventions are synthethised narratively without metanalysis, focusing on the key features of narrative information synthesis (grouping of studies, presentation of data and summary text, and appropriate discussion of limitations of this type of synthesis)[76].

Although the PRISMA statement and the cited extensions are focused on intervention SRs, a specific PRISMA extension has also been developed for diagnostic SRs[77]. On the contrary, no tools have been identified to communicate prevalence or prognostic SRs. In recent years, clarity and transparency in study communications has improved thanks to the development of checklists for scientific paper publication, although there is still room for improvement[78–80].

Discussion

Key results

This review identifies and describes the most relevant methodological resources to conduct prevalence, prognostic, diagnostic accuracy and effects of interventions SRs. This review offers a general and comparative perspective of the methodological resources by SR stage, highlighting the differential elements of each type of SR. This project does not aim to be a standalone tool for a researcher to find complete guidance on how to conduct and report a review, but rather it aims to be a signpost pointing out to the resources where researchers may find in depth guidance to develop their reviews.

Current context

This paper corroborates that developing a rigorous SR is a complex and resource-intensive task[81,82]. In order to tackle the increasing complexity of SRs and ensure the adoption of rigorous methodology, it is necessary that the reviews are made by **multidisciplinary work groups** with knowledge and experience in methodology (such as statistical analysis and information retrieval)[83,84]. In addition, it is important to consider the increasing availability of artificial-intelligence-based **technological tools**, which make it possible to semi-automate the different steps of the SR development, and thus reduce the time and human resources required to conduct the review[85].

Once the rigorous SR has been developed, ensuring the conveyance of the generated knowledge is essential. In this sense, new **formats for synthesis** and presentation of SR results are being explored nowadays to help their dissemination and the adoption of their conclusions in clinical practice and healthcare decision-making. For instance, new formats for result presentation and Summary of Findings tables are being proposed, adapted to the profile of their potential users[86,87].

Limitations and strengths

An inherent limitation of this project is its methodology based on a selection of resources and summary of guidance informed by expert opinion, which may be susceptible to implicit selection biases or lack of comprehensiveness.

The four types of SRs considered in this paper are fundamental to define preventive activities and public health policies, as well as to make health decisions. The selection of resources done is not dependent on whether the reviewer explores questions on efficacy or effectiveness, often described as explanatory or pragmatic questions, and will be useful to the researchers regardless of their intended purpose. However, we have not considered the resources to conduct in-depth exploration of effectiveness issues such as reviews of complex interventions or implementation reviews. Additionally, this research has not considered other types of SRs, such as methodological, economic evaluation and qualitative research SRs, for which it would be convenient to perform similar methodological compilations. Reviews of reviews (or overviews) were also not considered, and as such, there are a number of review-level resources which have not been discussed, for example the risk of bias assessment tool ROBIS or the methodological assessment tool AMSTAR[88,89]. Another limitation of this research is the need to keep it up to date, given the speed at which the methods and methodological resources to develop SRs are updated.

On the other hand, the main strengths of this paper are its transversal approach for the different types of reviews, and the identification of resources for all the stages in the development of an SR. There are few previous publications that offer a transversal perspective of the different types of systematic reviews, and these are focused on a specific stage of the review or on a particular topic. For instance, the work carried out by Munn *et al.* (2018) defined a typology for SRs, characterised from 10 different types of research questions, and delving into the format of each type of question[21]. Pollock *et al.* (2017) review the steps of an SR for 5 types of question, specifically focusing on the particularities of the reviews on stroke rehabilitation[90]. Muka *et al.* (2019) offer a structured compilation of resources for each SR stage, but without delving into the specificities of the different types of SRs[91]. Finally, organising the resources to assess the risk of bias by type of review is a strength and a novelty compared with previous works, which compile the quality assessing tools by type of study design but without linking them to the aim of the study nor the type of systematic review[92,93].

Conclusions

SRs are a key research tool to make decisions in healthcare, public health and medical research. There are methods and resources to develop high-quality reviews to answer most types of clinical questions. This review offers a complete resource guide for prevalence, prognostic, diagnostic and intervention reviews,

and is a very useful tool for those researchers that wish to develop SRs or conduct methodological research works in that field.

Data availability

Underlying data

All data underlying the results are available as part of the article and no additional source data are required.

Acknowledgements

Marta Roqué i Figuls is presently working on her PhD with the PhD Programme in Biomedical Research Methodology and Public Health from the Autonomous University of Barcelona.

The authors are indebted to the helpful peer reviewer comments of Ms Cumpston and Dr Purssell, which have helped improve the manuscript.

References

1. Higgins JPT, Thomas J, Chandler J, *et al.*: (editors). **Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019)**. Cochrane. 2019. [Accessed on 29/11/2019].
Reference Source
2. Urrútia G, Bonfill X: **Revisiones sistemáticas, una herramienta clave para la toma de decisiones clínicas y sanitarias**. *Rev Esp Salud Pública.* 2014; **88**(1): 1–3.
Publisher Full Text
3. Ferreira González I, Urrútia G, Alonso-Coello P: **Systematic reviews and meta-analysis: scientific rationale and interpretation**. *Rev Esp Cardiol.* 2011; **64**(8): 688–96.
PubMed Abstract | Publisher Full Text
4. Hemingway H, Croft P, Perel P, *et al.*: **Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes**. *BMJ.* 2013; **346**: e5595.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Siriwardhana DD, Hardoon S, Rait G, *et al.*: **Prevalence of Frailty and Prefrailty Among Community-Dwelling Older Adults in Low-Income and Middle-Income Countries: A Systematic Review and Meta-Analysis**. *BMJ Open.* 2018; **8**(3): e018195.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Guthold R, Stevens GA, Riley LM, *et al.*: **Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1·9 million participants**. *Lancet Glob Health.* 2018; **6**(10): e1077–86.
PubMed Abstract | Publisher Full Text
7. Roehr S, Pabst A, Luck T, *et al.*: **Is dementia incidence declining in high-income countries? A systematic review and meta-analysis**. *Clin Epidemiol.* 2018; **10**: 1233–1247.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Westby MJ, Dumville JC, Stubbs N, *et al.*: **Protease activity as a prognostic factor for wound healing in venous leg ulcers**. *Cochrane Database Syst Rev.* 2018; **9**(9): CD012841.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Skoetz N, Trivella M, Kreuzer KA, *et al.*: **Prognostic models for chronic lymphocytic leukaemia: an exemplar systematic review and meta-analysis**. *Cochrane Database of Syst Rev.* 2016; **1**: CD012022.
Publisher Full Text
10. Ambagtsheer RC, Thompson MQ, Archibald MM, *et al.*: **Diagnostic test accuracy of self-reported frailty screening instruments in identifying community-dwelling older people at risk of frailty and pre-frailty: a systematic review protocol**. *JBI Database System Rev Implement Rep.* THE JOANNA BRIGGS INSTITUTE, 2017; **15**(10): 2464–2468.
PubMed Abstract | Publisher Full Text
11. Martínez G, Vernooij RW, Fuentes Padilla P, *et al.*: **18F PET with florbetaben for the early diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI)**. *Cochrane Database Syst Rev.* 2017; **11**(11): CD012883.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Johnson S, Henschke N, Maayan N, *et al.*: **Ribavirin for treating Crimean Congo haemorrhagic fever**. *Cochrane Database Syst Rev.* 2018; **6**(6): CD012713.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Ellis G, Gardner M, Tsiachristas A, *et al.*: **Comprehensive geriatric assessment for older adults admitted to hospital**. *Cochrane Database Syst Rev.* 2017; **9**(1): CD006211.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Aromataris E, Munn Z, (Editors): **Joanna Briggs Institute Reviewer's Manual**. The Joanna Briggs Institute, 2017. [Accessed on 10/12/2018].
15. Munn Z, Moola S, Lisy K, *et al.*: **Chapter 5: Systematic reviews of prevalence and incidence**. In: Aromataris E, Munn Z, (Editors). *Joanna Briggs Institute Reviewer's Manual.* The Joanna Briggs Institute. 2017. [Accessed on 10/12/2018].
Publisher Full Text
16. Moola S, Munn Z, Tufanaru C, *et al.*: **Chapter 7: Systematic reviews of etiology and risk**. In: Aromataris E, Munn Z, (Editors). *Joanna Briggs Institute Reviewer's Manual.* The Joanna Briggs Institute. 2017. [Accessed on 10/12/2018].
Publisher Full Text
17. Riley RD, Moons KGM, Snell KIE, *et al.*: **A guide to systematic review and meta-analysis of prognostic factor studies**. *BMJ.* 2019; **364**: k4597.
PubMed Abstract | Publisher Full Text
18. Dekkers OM, Vandenbroucke JP, Cevallos M, *et al.*: **COSMOS-E: Guidance on conducting systematic reviews and meta-analyses of observational studies of etiology**. *PLoS Med.* 2019; **16**(2): e1002742.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Debray TP, Damen JA, Snell KI, *et al.*: **A guide to systematic review and meta-analysis of prediction model performance**. *BMJ.* 2017; **356**: i6460.
PubMed Abstract | Publisher Full Text
20. Deeks JJ, Bossuyt PM, Gatsonis C, (editors): **Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0**. The Cochrane Collaboration. 2010; [Accessed on 21/12/2018].
Reference Source
21. Munn Z, Stern C, Aromataris E, *et al.*: **What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences**. *BMC Med Res Methodol.* 2018; **18**(1): 5.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Iorio A, Spencer FA, Falavigna M, *et al.*: **Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients**. *BMJ.* 2015; **350**: h870.
PubMed Abstract | Publisher Full Text
23. Bossuyt PM, Irwig L, Craig J, *et al.*: **Comparative accuracy: assessing new tests against existing diagnostic pathways**. *BMJ.* 2006; **332**(7549): 1089–92. Erratum in:BMJ.2006 Jun 10;332(7554):1368.
PubMed Abstract | Publisher Full Text | Free Full Text
24. Bossuyt PM, Leeflang MM: **Chapter 6: Developing Criteria for Including Studies**. In: *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.4* [updated September 2008]. The Cochrane Collaboration. 2008.
Reference Source
25. Lijmer JG, Mol BW, Heisterkamp S, *et al.*: **Empirical evidence of design-related bias in studies of diagnostic tests**. *JAMA.* 1999; **282**(11): 1061–6.
PubMed Abstract | Publisher Full Text
26. Straus S, Moher D: **Registering systematic reviews**. *CMAJ.* 2010; **182**(1): 13–14.
PubMed Abstract | Publisher Full Text | Free Full Text
27. Ge L, Tian JH, Li YN, *et al.*: **Association between prospective registration and overall reporting and methodological quality of systematic reviews: a meta-epidemiological study**. *J Clin Epidemiol.* 2018; **93**: 45–55.
PubMed Abstract | Publisher Full Text
28. Page MJ, Shamseer L, Tricco AC: **Registration of systematic reviews in PROSPERO: 30,000 records and counting**. *Syst Rev.* 2018; **7**(1): 32.
PubMed Abstract | Publisher Full Text | Free Full Text
29. Lefebvre C, Glanville J, Briscoe S, *et al.*: **Chapter 4: Searching for and selecting studies**. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, (editors). *Cochrane Handbook for Systematic Reviews of Interventions version 6.0.* (updated July 2019). Cochrane, 2019; [Accessed on 29/11/2018].
Publisher Full Text
30. Atkinson KM, Koenka AC, Sanchez CE, *et al.*: **Reporting standards for literature searches and report inclusion criteria: making research syntheses more transparent and easy to replicate**. *Res Synth Methods.* 2015; **6**(1): 87–95.
PubMed Abstract | Publisher Full Text
31. Campbell JM, Kulgar M, Ding S, *et al.*: **Chapter 9: Diagnostic test accuracy systematic reviews**. In: Aromataris E, Munn Z, (Editors). *Joanna Briggs Institute Reviewer's Manual.* The Joanna Briggs Institute, 2017; [Accessed on 10/12/2018].
Publisher Full Text

32. de Vet HCW, Eisinga A, Riphagen II, *et al.*: **Chapter 7: Searching for Studies**. In: *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.4* [updated September 2008]. The Cochrane Collaboration, 2008.
Reference Source

33. Lefebvre C, Glanville J, Wieland LS, *et al.*: **Methodological developments in searching for studies for systematic reviews: past, present and future?** *Syst Rev.* 2013; **2**: 78.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Glanville JM, Lefebvre C, Miles JN, *et al.*: **How to identify randomized controlled trials in MEDLINE: ten years on.** *J Med Libr Assoc.* 2006; **94**(2): 130–136.
PubMed Abstract | Free Full Text

35. Wilczynski NL, Haynes RB; Hedges Team: **Developing optimal search strategies for detecting clinically sound prognostic studies in MEDLINE: an analytic survey.** *BMC Med.* 2004; **2**(1): 23.
PubMed Abstract | Publisher Full Text | Free Full Text

36. Beynon R, Leeflang MM, McDonald S, *et al.*: **Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE**. *Cochrane Database Syst Rev.* 2013; **2013**(9): MR000022.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Sampson M, Tetzlaff J, Urquhart C: **Precision of healthcare systematic review searches in a cross-sectional sample.** *Res Synth Methods.* 2011; **2**(2): 119–25.
PubMed Abstract | Publisher Full Text

38. Bramer WM, Rethlefsen ML, Kleijnen J, *et al.*: **Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study.** *Syst Rev.* 2017; **6**(1): 245.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Hartling L, Featherstone R, Nuspl M, *et al.*: **The contribution of databases to the results of systematic reviews: a cross-sectional study.** *BMC Med Res Methodol.* 2016; **16**(1): 127.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Glanville JM, Duffy S, McCool R, *et al.*: **Searching ClinicalTrials.gov and the International Clinical Trials Registry Platform to inform systematic reviews: what are the optimal search approaches?** *J Med Libr Assoc.* 2014; **102**(3): 177–83. MANUAL GRADE. Spanish version. [Accessed May 2019].
PubMed Abstract | Publisher Full Text | Free Full Text

41. Isojarvi J, Wood H, Lefebvre C, *et al.*: **Challenges of identifying unpublished data from clinical trials: Getting the best out of clinical trials registers and other novel sources.** *Res Synth Methods.* 2018; **9**(4): 561–578.
PubMed Abstract | Publisher Full Text

42. Horsley T, Dingwall O, Sampson M: **Checking reference lists to find additional studies for systematic reviews.** *Cochrane Database Syst Rev.* 2011; **2011**(8): MR000026.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Gentles SJ, Charles C, Nicholas DB, *et al.*: **Reviewing the research methods literature: principles and strategies illustrated by a systematic overview of sampling in qualitative research.** *Syst Rev.* 2016; **5**(1): 172.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Hartling L, Featherstone R, Nuspl M, *et al.*: **Grey literature in systematic reviews: a cross-sectional study of the contribution of non-English reports, unpublished studies and dissertations to the results of meta-analyses in child-relevant reviews.** *BMC Med Res Methodol.* 2017; **17**(1): 64.
PubMed Abstract | Publisher Full Text | Free Full Text

45. Booth A: **Searching for qualitative research for inclusion in systematic reviews: a structured methodological review.** *Syst Rev.* 2016; **5**: 74.
PubMed Abstract | Publisher Full Text | Free Full Text

46. Rethlefsen ML, Murad MH, Livingston EH: **Engaging medical librarians to improve the quality of review articles.** *JAMA.* 2014; **312**(10): 999–1000.
PubMed Abstract | Publisher Full Text

47. Rethlefsen ML, Farrell AM, Osterhaus Trzasko LC, *et al.*: **Librarian co-authors correlated with higher quality reported search strategies in general internal medicine systematic reviews.** *J Clin Epidemiol.* 2015; **68**(6): 617–26.
PubMed Abstract | Publisher Full Text

48. Spencer AJ, Eldredge JD: **Roles for librarians in systematic reviews: a scoping review.** *J Med Libr Assoc.* 2018; **106**(1): 46–56.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Hoy D, Brooks P, Woolf A, *et al.*: **Assessing risk of bias in prevalence studies: modification of an existing tool and evidence of interrater agreement.** *J Clin Epidemiol.* 2012; **65**(9): 934–9.
PubMed Abstract | Publisher Full Text

50. Hayden JA, van der Windt DA, Cartwright JL, *et al.*: **Assessing bias in studies of prognostic factors.** *Ann Intern Med.* 2013; **158**(4): 280–6.
PubMed Abstract | Publisher Full Text

51. Morgan RL, Thayer KA, Santesso N, *et al.*: **Evaluation of the risk of bias in non-randomized studies of interventions (ROBINS-I) and the 'target experiment' concept in studies of exposures: Rationale and preliminary instrument development.** *Environ Int.* 2018; **120**: 382–387.
PubMed Abstract | Publisher Full Text

52. Morgan RL, Thayer KA, Santesso N, *et al.*: **A risk of bias instrument for non-randomized studies of exposures: A users' guide to its application in the context of GRADE.** *Environ Int.* 2019; **122**: 168–184.
PubMed Abstract | Publisher Full Text

53. Wolff RF, Moons KGM, Riley RD, *et al.*: **PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies.** *Ann Intern Med.* 2019; **170**(1): 51–58.
PubMed Abstract | Publisher Full Text

54. Whiting PF, Rutjes AW, Westwood ME, *et al.*: **QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies.** *Ann Intern Med.* 2011; **155**(8): 529–36.
PubMed Abstract | Publisher Full Text

55. Higgins JPT, Savović J, Page MJ, *et al.*: **Chapter 8: Assessing risk of bias in a randomized trial**. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions*. version 6.0 (updated July 2019). Cochrane, 2019.
Publisher Full Text

56. Sterne JA, Hernán MA, Reeves BC, *et al.*: **ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions.** *BMJ.* 2016; **355**: i4919.
PubMed Abstract | Publisher Full Text | Free Full Text

57. Lau J, Ioannidis JP, Schmid CH: **Quantitative synthesis in systematic reviews.** *Ann Intern Med.* 1997; **127**(9): 820–826.
PubMed Abstract | Publisher Full Text

58. Deeks JJ, Higgins JPT, Altman DG: (editors). **Chapter 10: Analysing data and undertaking meta-analyses**. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, (editors). *Cochrane Handbook for Systematic Reviews of Interventions version 6.0* (updated July 2019). Cochrane. 2019.
Reference Source

59. Popay J, Roberts H, Sowden A, *et al.*: **Guidance on the conduct of narrative synthesis in systematic reviews: A product from the ESRC Methods Programme.** 2006.
Publisher Full Text

60. Macaskill P, Gatsonis C, Deeks JJ, *et al.*: **Chapter 10: Analysing and Presenting Results**. In: Deeks JJ, Bossuyt PM, Gatsonis C (editors), *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0*. The Cochrane Collaboration. 2010.
Reference Source

61. Rutter CM, Gatsonis CA: **A Hierarchical Regression Approach to Meta-Analysis of Diagnostic Test Accuracy Evaluations.** *Stat Med.* 2001; **20**(19): 2865–84.
PubMed Abstract | Publisher Full Text

62. Rücker G, Schwarzer G, Carpenter JR, *et al.*: **Undue Reliance on I^2 in Assessing Heterogeneity May Mislead.** *BMC Med Res Methodol.* 2008; **8**(1): 79.
PubMed Abstract | Publisher Full Text | Free Full Text

63. Chaimani A, Caldwell DM, Higgins T, *et al.*: **Chapter 11: Undertaking network meta-analyses**. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions*. version 6.0 (updated July 2019). Cochrane, 2019.
Publisher Full Text

64. Schünemann H, Brożek J, Guyatt G, *et al.*: (editors). **Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach**. (updated October 2013). [Accessed June 2020].
Reference Source

65. Santesso N, Glenton C, Dahm P, *et al.*: **GRADE Guidelines 26: Informative Statements to Communicate the Findings of Systematic Reviews of Interventions.** *J Clin Epidemiol.* 2020; **119**: 126–135.
PubMed Abstract | Publisher Full Text

66. Schünemann HJ, Higgins JPT, Vist GE, *et al.*: **Chapter 14: Completing 'Summary of findings' tables and grading the certainty of the evidence**. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions version 6.0* (updated July 2019). Cochrane. 2019.
Reference Source

67. Murad MH, Mustafa RA, Schünemann HJ, *et al.*: **Rating the certainty in evidence in the absence of a single estimate of effect.** *Evid Based Med.* 2017; **22**(3): 85–87.
PubMed Abstract | Publisher Full Text | Free Full Text

68. Harder T, Takla A, Eckmanns T, *et al.*: **PRECEPT: An Evidence Assessment Framework for Infectious Disease Epidemiology, Prevention and Control.** *Euro Surveill.* 2017; **22**(40): 16-00620.
PubMed Abstract | Publisher Full Text | Free Full Text

69. Huguet A, Hayden JA, Stinson J, *et al.*: **Judging the Quality of Evidence in Reviews of Prognostic Factor Research: Adapting the GRADE Framework.** *Syst Rev.* 2013; **2**: 71.
PubMed Abstract | Publisher Full Text | Free Full Text

70. Schünemann HJ, Mustafa RA, Brozek J, *et al.*: **GRADE guidelines: 21 part 1. Study design, risk of bias, and indirectness in rating the certainty across a body of evidence for test accuracy.** *J Clin Epidemiol.* 2020; **122**: 129–141.
PubMed Abstract | Publisher Full Text

71. Schünemann HJ, Mustafa RA, Brozek J, *et al.*: **GRADE guidelines: 21 part 2. Test accuracy: inconsistency, imprecision, publication bias, and other domains for rating the certainty of evidence and presenting it in evidence profiles and summary of findings tables.** *J Clin Epidemiol.* 2020; **122**: 142–152.
PubMed Abstract | Publisher Full Text

72. Moher D, Liberati A, Tetzlaff J, *et al.*: **Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement.** *BMJ.* 2009; **339**: b2535.
PubMed Abstract | Publisher Full Text | Free Full Text

73. Moher D, Shamseer L, Clarke M, *et al.*: **Preferred Reporting Items for**

Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 Statement. *Syst Rev.* 2015; **4**(1): 1.
PubMed Abstract | Publisher Full Text | Free Full Text

74. Beller EM, Glasziou PP, Altman DG, *et al.*: **PRISMA for Abstracts: Reporting Systematic Reviews in Journal and Conference Abstracts**. *PLoS Med.* 2013; **10**(4): e1001419.
PubMed Abstract | Publisher Full Text | Free Full Text

75. Zorzela L, Loke YK, Ioannidis JP, *et al.*: **PRISMA Harms Checklist: Improving Harms Reporting in Systematic Reviews**. *BMJ.* 2016; **352**: i157.
PubMed Abstract | Publisher Full Text

76. Campbell M, McKenzie JE, Sowden A, *et al.*: **Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline**. *BMJ.* 2020; **368**: l6890.
PubMed Abstract | Publisher Full Text | Free Full Text

77. McInnes MDF, Moher D, Thombs BD, *et al.*: **Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement**. *JAMA.* 2018; **319**(4): 388–396.
PubMed Abstract | Publisher Full Text

78. Moher D, Tetzlaff J, Tricco AC, *et al.*: **Epidemiology and Reporting Characteristics of Systematic Reviews**. *PLoS Med.* 2007; **4**(3): e78.
PubMed Abstract | Publisher Full Text | Free Full Text

79. Page MJ, Shamseer L, Altman DG, *et al.*: **Epidemiology and Reporting Characteristics of Systematic Reviews of Biomedical Research: A Cross-Sectional Study**. *PLoS Med.* 2016; **13**(5): e1002028.
PubMed Abstract | Publisher Full Text | Free Full Text

80. Salameh JP, McInnes MDF, Moher D, *et al.*: **Completeness of Reporting of Systematic Reviews of Diagnostic Test Accuracy Based on the PRISMA-DTA Reporting Guideline**. *Clin Chem.* 2019; **65**(2): 291–301.
PubMed Abstract | Publisher Full Text

81. Turner T, Green S, Tovey D, *et al.*: **Producing Cochrane systematic reviews-a qualitative study of current approaches and opportunities for innovation and improvement**. *Syst Rev.* 2017; **6**(1): 147.
PubMed Abstract | Publisher Full Text | Free Full Text

82. Borah R, Brown AW, Capers PL, *et al.*: **Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry**. *BMJ Open.* 2017; **7**(2): e012545.
PubMed Abstract | Publisher Full Text | Free Full Text

83. Ioannidis JP, Greenland S, Hlatky MA, *et al.*: **Increasing Value and Reducing Waste in Research Design, Conduct, and Analysis**. *Lancet.* 2014; **383**(9912): 166–75.
PubMed Abstract | Publisher Full Text | Free Full Text

84. Institute of Medicine (US) Committee on Standards for Systematic Reviews of Comparative Effectiveness Research, Eden J, Levit L, *et al.*: **Finding What Works in Health Care: Standards for Systematic Reviews**. Washington (DC): National Academies Press (US); 2011.2, Standards for Initiating a Systematic Review. 2011.
PubMed Abstract | Publisher Full Text

85. Marshall IJ, Wallace BC: **Toward systematic review automation: a practical guide to using machine learning tools in research synthesis**. *Syst Rev.* 2019; **8**(1): 163.
PubMed Abstract | Publisher Full Text | Free Full Text

86. Carrasco-Labra A, Brignardello-Petersen R, Santesso N, *et al.*: **Improving GRADE evidence tables part 1: a randomized trial shows improved understanding of content in summary of findings tables with a new format**. *J Clin Epidemiol.* 2016; **74**: 7–18.
PubMed Abstract | Publisher Full Text

87. Marquez C, Johnson AM, Jassemi S, *et al.*: **Enhancing the uptake of systematic reviews of effects: what is the best format for health care managers and policy-makers? A mixed-methods study**. *Implement Sci.* 2018; **13**(1): 84.
PubMed Abstract | Publisher Full Text | Free Full Text

88. Whiting P, Savović J, Higgins JP, *et al.*: **ROBIS: A new tool to assess risk of bias in systematic reviews was developed**. *J Clin Epidemiol.* 2016; **69**: 225–234.
PubMed Abstract | Publisher Full Text | Free Full Text

89. Shea BJ, Reeves BC, Wells G, *et al.*: **AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both**. *BMJ.* 2017; **358**: j4008.
PubMed Abstract | Publisher Full Text | Free Full Text

90. Pollock A, Berge E: **How to do a systematic review**. *Int J Stroke.* 2018; **13**(2): 138–56.
PubMed Abstract | Publisher Full Text

91. Muka T, Glisic M, Milic J, *et al.*: **A 24-step guide on how to design, conduct, and successfully publish a systematic review and meta-analysis in medical research**. *Eur J Epidemiol.* 2020; **35**(1): 49–60.
PubMed Abstract | Publisher Full Text

92. Harrison JK, Reid J, Quinn TJ, *et al.*: **Using Quality Assessment Tools to Critically Appraise Ageing Research: A Guide for Clinicians**. *Age Ageing.* 2017; **46**(3): 359–65.
PubMed Abstract | Publisher Full Text | Free Full Text

93. Zeng X, Zhang Y, Kwong JS, *et al.*: **The Methodological Quality Assessment Tools for Preclinical and Clinical Studies, Systematic Review and Meta-Analysis, and Clinical Practice Guideline: A Systematic Review**. *J Evid Based Med.* 2015; **8**(1): 2–10.
PubMed Abstract | Publisher Full Text

5.2 Publicación 2: Physical exercise interventions for improving performance-based measures of physical function in community-dwelling, frail older adults: a systematic review and meta-analysis

Giné-Garriga M, Roqué-Fíguls M, Coll-Planas L, Sitjà-Rabert M, Salvà A. Physical exercise interventions for improving performance-based measures of physical function in community-dwelling, frail older adults: a systematic review and meta-analysis. Arch Phys Med Rehabil. 2014 Apr; 95(4):753-769.e3.

FI: 3.395 (2014). Puntuación de atención Altmetric: 24

La RS que constituye el trabajo 2 incluyó 19 estudios (2215 participantes) [31]. Doce de los estudios compararon programas de ejercicio con un control inactivo (1588 participantes), y 7 con un control activo (627 participantes). La mayoría de programas de ejercicio evaluado eran programas multicomponente.

En comparación con las intervenciones de control, se demostró que el ejercicio mejora la velocidad de la marcha normal (DM = 0.07m/s; IC 95%: 0.04 a 0.09), la velocidad de la marcha rápida (DM = 0.08m/s; IC 95%: 0.02 a 0.14), y el SPPB (DM = 2.18; IC 95%: 1.56 a 2.80). Los resultados no son concluyentes para los desenlaces de resistencia, y no se observó un efecto consistente sobre el equilibrio y la movilidad funcional en las actividades de la vida diaria.

La evidencia que compara diferentes modalidades de ejercicio es escasa y heterogénea, tanto en los programas de ejercicio comparados como en los desenlaces considerados.

En conclusión, el ejercicio conlleva algunos beneficios en las personas mayores frágiles, aunque todavía existe incertidumbre con respecto a qué características del ejercicio (tipo, frecuencia, duración) son más efectivas.

En esta RS no se realizó una evaluación de la calidad de la evidencia, y no se construyó una tabla de resumen de hallazgos. Esta revisión no se ha actualizado y no está previsto hacerlo.

Archives of Physical Medicine and Rehabilitation
journal homepage: www.archives-pmr.org
Archives of Physical Medicine and Rehabilitation 2014;95:753-69

REVIEW ARTICLE (META-ANALYSIS)

Physical Exercise Interventions for Improving Performance-Based Measures of Physical Function in Community-Dwelling, Frail Older Adults: A Systematic Review and Meta-Analysis

Maria Giné-Garriga, PhD, PT,[a,b] Marta Roqué-Fíguls, MD,[c,d] Laura Coll-Planas, MD,[c] Mercè Sitjà-Rabert, PhD, PT,[b] Antoni Salvà, MD[c]

From the [a]Department of Physical Activity and Sport Sciences, FPCEE Blanquerna, Universitat Ramon Llull, Barcelona; [b]Department of Physical Therapy, FCS Blanquerna, Universitat Ramon Llull, Barcelona; [c]Institute on Aging, Universitat Autònoma de Barcelona, Barcelona; and [d]Institute of Biomedical Research Sant Pau (IIB-Sant Pau), Barcelona, Spain.

Abstract

Objective: To conduct a systematic review to determine the efficacy of exercise-based interventions on improving performance-based measures of physical function and markers of physical frailty in community-dwelling, frail older people.

Data Sources: Comprehensive bibliographic searches in MEDLINE, the Cochrane Library, PEDro, and CINAHL databases were conducted (April 2013).

Study Selection: Randomized controlled trials of community-dwelling older adults, defined as frail according to physical function and physical difficulties in activities of daily living (ADL). Included trials had to compare an exercise intervention with a control or another exercise intervention, and assess performance-based measures of physical function such as mobility and gait, or disability in ADL.

Data Extraction: Two review authors independently screened the search results and performed data extraction and risk of bias assessment. Nineteen trials were included, 12 of them comparing exercise with an inactive control. Most exercise programs were multicomponent.

Data Synthesis: Meta-analysis was performed for the comparison of exercise versus control with the inverse variance method under the random-effects models. When compared with control interventions, exercise was shown to improve normal gait speed (mean difference [MD] = .07m/s; 95% confidence interval [CI], .04−.09), fast gait speed (MD = .08m/s; 95% CI, .02−.14), and the Short Physical Performance Battery (MD = 2.18; 95% CI, 1.56−2.80). Results are inconclusive for endurance outcomes, and no consistent effect was observed on balance and the ADL functional mobility. The evidence comparing different modalities of exercise is scarce and heterogeneous.

Conclusions: Exercise has some benefits in frail older people, although uncertainty still exists with regard to which exercise characteristics (type, frequency, duration) are most effective.

Archives of Physical Medicine and Rehabilitation 2014;95:753-69

As individuals get older, they may reach a stage of vulnerability called frailty that precedes and predisposes to disability and physical dependence. The terms *frail* and *frailty* are often used in the literature without clear definition or criteria,[1] and there is not yet a consensus on a standardized and valid method of clinically screening for frailty.[2,3] Frailty is considered highly prevalent in old age and to confer a high risk for falls, worsening mobility, disability, hospitalization, and mortality.[4]

Two main definitions of frailty exist. The first one relates frailty to a physical phenotype consisting of solely physical components and has attracted the most attention of researchers.[4] The most well known of these is the frailty phenotype described by Fried et al,[5] which identifies someone as frail when 3 or more of the following criteria are present: unintentional weight loss,

No commercial party having a direct financial interest in the results of the research supporting this article has conferred or will confer a benefit on the authors or on any organization with which the authors are associated.

0003-9993/14/$36 - see front matter © 2014 by the American Congress of Rehabilitation Medicine
http://dx.doi.org/10.1016/j.apmr.2013.11.007

self-reported exhaustion, weakness, slow walking speed, and low levels of physical activity. The second definition has a broader scope and conceptualizes frailty as the result of multiple interacting factors such as having difficulties in activities of daily living (ADL), and social and psychological aspects.[6] This definition was operationalized into the Frailty Index,[7] built as a sum of deficits and able to capture gradations in health status ranging from mild to severe stages, and the risk of adverse outcomes.[8]

A review of the literature by Gobbens et al[9] showed that frailty affects multiple domains of functioning. These include gait and mobility, balance, muscle strength, motor processing, cognition, nutrition (often operationalized as nutritional status or weight change), endurance (including feelings of fatigue and exhaustion), and physical activity.

Frailty is common in older adults (>65y), but different operationalization of frailty status results in widely differing prevalences between studies. In a recent systematic review,[10] the weighted prevalence was 9.9% for physical frailty and 13.6% for the broad definition of frailty. The design of effective interventions to prevent or delay functional decline and disability in older persons is a public health priority. Most likely to benefit from such interventions are community-dwelling frail individuals, without disability or with only early disability, and who are at high risk of becoming functionally dependent.[11] Frail individuals who are institutionalized or hospitalized present a more deteriorated health status and functioning[12] and may need different types of interventions to prevent or minimize complications.

The benefits of exercise in delaying physical dependence in an elderly population have long been recognized,[13,14] and randomized controlled trials[15,16] have shown promising early results of physical exercise. Exercise seems to be beneficial in improving physical functions, such as sit-to-stand performance, balance, agility, and ambulation, in older adults.[17-19] Although there are 6 systematic reviews[2,20-24] exploring the benefits of exercise in frail older adults, a definite conclusion has not yet been reached. Four of the reviews[20,22-24] applied a very broad definition of frailty that included both nonfrail and prefrail participants. The other 2 reviews[2,21] applied consistent definitions of frailty but need to be updated with studies published recently in community-dwelling populations. The most recent reviews[23,24] did not identify some of the studies included in the present review, and both also included non–performance-based measures as main outcomes.

This systematic review aims to integrate the most current evidence on the effect of exercise interventions on improving performance-based measures of physical function and markers of physical frailty in community-dwelling older people defined as frail according to physical function and physical difficulties in ADL. Specifically, we aimed to (1) examine the effectiveness of exercise compared with control interventions; (2) determine which exercise modalities are most effective; and (3) determine whether there are adverse effects within the exercise interventions.

List of abbreviations:

ADL	activities of daily living
BBS	Berg Balance Scale
CI	confidence interval
MD	mean difference
RCT	randomized controlled trial
SPPB	Short Physical Performance Battery
TUG	Timed Up and Go

Methods

We included randomized controlled trials (RCTs) evaluating the effect of physical exercise programs with or without other components on functional performance-based measures of physical function among community-dwelling, frail older adults. Inclusion criteria were as follows: participants should be (1) 65 years and older; (2) living in the community; and (3) defined as frail according to standardized criteria (eg, Fried's), or considered frail according to reduced physical function measured with physical performance scales (eg, Short Physical Performance Battery [SPPB]) or performance-based measures such as gait and mobility, muscle strength, nutritional intake, weight change, balance, endurance, fatigue, and physical activity. Participants either had to have limitations in 2 or more performance-based frailty measures or had to have clinically significant limitations in a single measure. Exclusion criteria were as follows: (1) inclusion of participants with disability (eg, advanced disability in performing ADL, dementia, or end-stage disease); (2) inclusion of prefrail participants (eg, those with nonsignificant impairment in frailty indicators); (3) inclusion of institutionalized participants; and (4) crossover design studies.

Primary outcomes were performance-based measures of physical function such as mobility, gait, muscular strength, balance, endurance, and disability in ADL. Secondary outcomes were number of falls; institutionalization; adverse effects of the exercise program such as falls, fractures, tendinitis, or muscular soreness; health-related quality of life; symptoms of depression; hospitalization; and death.

Searches were conducted in MEDLINE, The Cochrane Library, PEDro, and CINAHL databases (April 2013). All databases were searched using free text and descriptors. The search strategy was adapted for each database, including terms for frailty, older people, multiple expressions of exercise, and limiting for randomized controlled trial; the full search strategy is included in supplemental appendix S1 (available online only at http://www.archives-pmr.org/). The search results were treated using bibliographic management software (Biblioscape 7.41[a]), allowing for duplicate consolidation and further refining of the article list. In addition, reference lists from previous systematic reviews[19,21,22,25] on exercise for the elderly were hand searched to identify trials on frail community-dwelling individuals. Two review authors (M.R., M.S., L.C., or M.G.) independently screened the search results and performed data extraction and risk of bias assessment. Any discrepancies were resolved by consensus or consulting with a third author.

We used the tool for assessing risk of bias proposed by the Cochrane Collaboration.[26] For each trial, we assessed the risk of bias of the following domains: random sequence generation, allocation concealment, blinding of assessments, incomplete outcome data, and selective outcome reporting. For each trial, an overall assessment of risk of bias was derived as low, high, or unclear based on the previous assessments. If any domain was at high risk of bias, the trial was considered to be at high risk of bias. Trials with 4 or 5 domains at low risk of bias were considered to be at low risk of bias. Otherwise, risk of bias of the trial was considered to be unclear.

We pooled data as presented in the original trials, either as intention to treat or not. Heterogeneity was assessed with the I^2 statistic, considering values greater than 50% as a sign of relevant heterogeneity. The effect of treatment was estimated by mean

differences (MDs) and standardized MDs in continuous outcomes and risk ratios in dichotomous outcomes. Pooled effect measures were computed applying the inverse-variance method in a random-effects model. Planned subgroup analyses on age and baseline performance, as well as sensitivity analyses with trials where frailty had been defined following Fried's criteria, could not be conducted because of the lack of detailed data and a low number of trials.

Results

A total of 38 citations providing data from 19 trials were included in this systematic review.[27-63] The flow chart of references and the causes of exclusion are presented in figure 1. We faced some challenges in assessing inclusion criteria: we included trials whose participants had moderate dependence in mobility, but we excluded trials on participants with dependence in basic ADL.[64,65] In 3 included trials,[34,36,62] it cannot be ruled out that a small percentage of participants had mild to moderate cognitive impairment. In the study by Boshuizen et al,[34] participants were excluded if they had a self-reported disease or condition that would be adversely affected by the exercises involved in the program, and in the study by de Jong et al,[36] reasons for dropout included (terminal) disease, but participants needed to have the ability to understand study procedures. In the aforementioned 2 studies, no clear cognitive condition is stated, so we assumed participants were not cognitively impaired. The mean baseline Mini-Mental State Examination scores in the study by Worm et al[62] were 23.9±4.1 and 23.5±5.2 in the exercise group and in the control group, respectively, so some participants included in the study could have had mild to moderate cognitive impairment.

Twelve trials compared an exercise intervention with an inactive control, presented in table 1. The exercise interventions studied differed in content (resistance, stretching, strength, flexibility, balance), setting (facility/home), delivery (individual/group), duration, and frequency. The exercise interventions tested were combinations of aerobic, balance, flexibility, endurance, and strength exercises,[44,52,58,59,62] combinations of balance and strength exercises,[47,49,53] strength exercise programs,[34,61] a stretching intervention,[60] and finally activities related to maintain and improve performance in ADL.[36]

Seven trials compared different modalities of exercise, described in table 2. Of these, 2 trials[30,35] compared facility-based with home-based exercise and could be assessed through meta-analysis. The other trials explored specific exercise modalities such as a progressive resistance-training program using weighted vests,[27,28] the addition of visual computer feedback to balance training,[50] combining whole-body vibration with exercise,[51] or performing Tai Chi.[57] The results of these trials are described in the text.

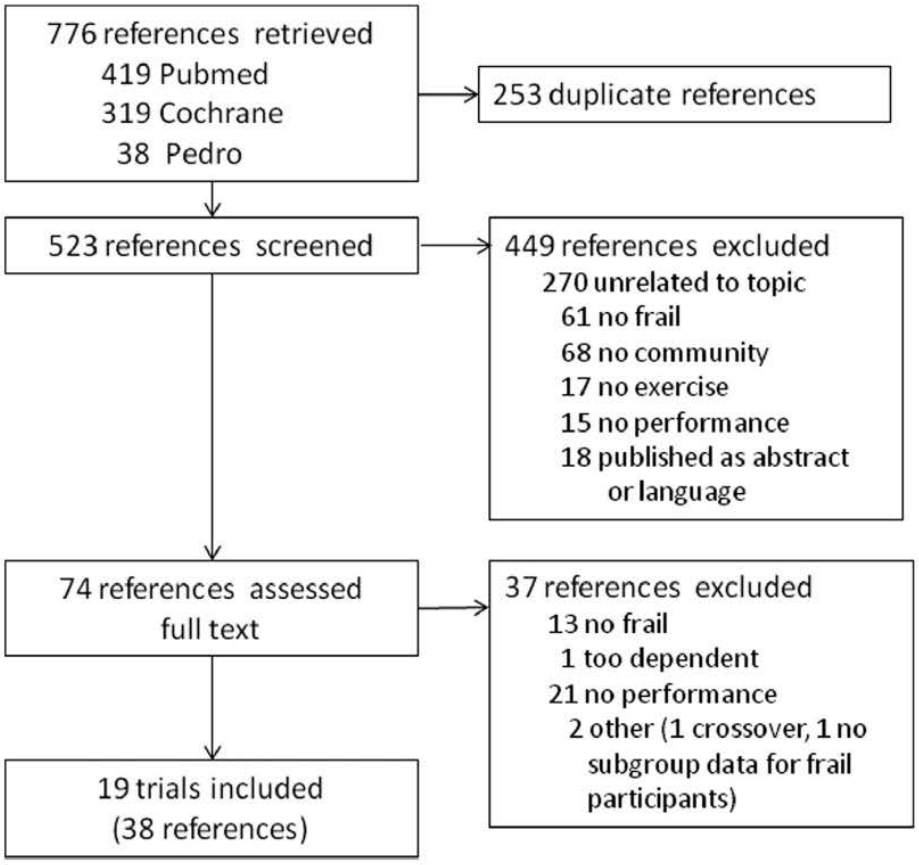


Fig 1 Flow chart.

Table 1 Description of studies comparing exercise versus control

Study	N*	Frailty Criteria	Control	Intervention	Frequency and Duration	Follow-up	Performance and Safety Outcomes	RoB†
Boshuizen,[34] 2005‡	72	(1) Elders experiencing difficulty in getting up from a chair, and (2) maximum knee-extensor torque in both legs not superior to 87.5Nm (25-kg force).	Control group received no training and was asked to remain habitually active.	Strength exercises for the thigh muscles, with use of elastic bands. Individual SSs at a facility and home USs. Arm 1: 2 SSs + 1 US Arm 2: 1 SS + 2 USs	60min 3 times a week for 10wk	EOI, 6mo FUP	Gait speed, tandem test, step test, TUG, GARS scale	U
de Jong,[36-43] 1999§	217	(1) Individuals who require health care, such as home care or Meals-on-Wheels service; (2) age ≥70y; (3) no regular exercise; (4) BMI<25kg/m^2 on the basis of self-reported weight and height or recent weight loss; (4) no use of multivitamin supplements; and (5) ability to understand the study procedures.	Supervised social program for 90min biweekly. In the off week, they were visited at their homes. Subjects were asked not to engage in other exercise programs during the study period.	Supervised group training to maintain or improve mobility and performance of daily activities. Exercises of moderate, gradually increasing intensity comprised walking, exercise-to-music routines, skills training, stretching, and relaxation activities.	45min, twice a week, for 17wk	EOI	Mobility score, gait speed, SPPB, fitness score, ADL score, self-care score, well-being	U
Fairhall,[44,45] 2012	241	Adults ≥70y, with 3 or more of the CHS Frailty criteria: (1) slow gait speed, (2) weak grip strength, (3) exhaustion, (4) low energy expenditure, (5) weight loss.	Usual care available to older residents of the Hornsby Ku-ring-gai area from their general practitioner and community services, which may include medical management of health conditions, allied health input, assessment of care needs, and provision of care.	Multifactorial interdisciplinary intervention targeting the CHS frailty phenotype. Participants classified as having grip weakness, slow 4-m walk time, or low physical activity level received up to 10 home-based physiotherapy sessions and performed a targeted, goal-focused, home-based strength, balance, and endurance training regimen. Mobility aids and other equipment were recommended, if needed. For participants meeting the weight loss	45–60min, 3–5 weekly sessions. 10 physiotherapy sessions over 12mo	EOI	Gait speed, SPPB, mobility score, Barthel Index, Geriatric Depression Score, EQ-5D quality of life	L

(*continued on next page*)

Table 1 (*continued*)

Study	N*	Frailty Criteria	Control	Intervention	Frequency and Duration	Follow-up	Performance and Safety Outcomes	RoB†
				criterion, a clinical evaluation of nutritional intake at home was made, and home-delivered meals and nutritional supplementation were offered, if appropriate. Participants reporting exhaustion and a high score in the Geriatric Depression Scale score were referred to a psychiatrist or psychologist. If applicable, chronic disease management programs were put in place or reinforced, and medication adequacy and compliance were reviewed.				
Gill,[15,46-48] 2002	188	(1) Individuals who required >10s to perform a rapid gait test or (2) could not stand up from a seated position in a hardback chair with their arms folded.	Educational program designed to provide attention and health education. Monthly visits for 6mo, followed by monthly follow-up telephone calls for 6 additional months.	Home training program. Periodic assessments of mobility, balance, and environmental hazards by a physical therapist, followed by individually tailored interventions that target physical impairments. Progressive exercises could be proposed for ROM, balance (10min daily), and muscle conditioning with resistant elastic bands (30min, 3 times/wk).	45—60min assessment visits, 16 visits in 6mo	EOI, 6 mo FUP	Falls, death, fractures, adverse effects	U
Giné-Garriga,[49] 2010	51	Participants verifying 1 of the following: (1) required more than 10s to perform a rapid-gait test, or (2) could not stand up 5 times from a seated position in a hardback chair with their arms folded, or	Control group subjects continued their routine daily activities and had weekly social meetings at the training facility, including four 60-min health education sessions.	Functional circuit training program focused on functional balance and lower body strength-based. Supervised group sessions held at a facility.	45min, twice a week for 12wk	EOI, 6mo FUP	Gait speed, gait test, kick test, semitandem, 1-limb stand, standup test, modified TUG, Barthel Index, fractures, adverse effects	L

(*continued on next page*)

Table 1 (*continued*)

Study	N*	Frailty Criteria	Control	Intervention	Frequency and Duration	Follow-up	Performance and Safety Outcomes	RoB†
		(3) categorized as frail by the exhaustion criterion.						
Rejeski,[52] 2008	412	(1) Summary score <10 on the SPPB, (2) ability to complete the 400-m walk test within 15min without sitting and without the use of an assistive device (including a cane) or the help of another person, and (3) sedentary lifestyle.	Active control (successful aging) with group workshops on health topics and an instructor-led program (5–10min) of upper extremity stretching exercises. Weekly for 24wk, and monthly thereafter.	Physical activity program combining aerobic activities, strength, balance, and flexibility exercises. Exercise training initially center-based transitioning to home-based exercise. Ten weekly closed-group counseling sessions that focus on physical activity and prevention of physical disability.	60min, 1–3 times a week (facility, by study phase) + 1–5 times a week (home), for 12mo	2y EOI	400-m walk efficacy, death, hospitalization, adverse effects	L
Rydwik,[53-56] 2008‖	96	(1) Unintentional weight loss >5% during the last 12mo and/or BMI <20kg/m^2, (2) low PA level graded with a PA scale.	General physical training advice to take walks 3 times per week for at least 20min, to use staircases instead of an elevator from time to time, and to do 30min of physical activity each day. General diet advice to eat 3 main courses and 2–3 between-meal snacks combined with fluid.	Group training led by an instructor comprising aerobic, muscle strength, and balance (Qigong) for 12wk in a facility. Afterwards, home-based exercises for 6mo. Subjects were encouraged to perform Qigong, functional muscle strength training and to take regular walks several times per week.	60min, twice a week for 12wk (facility)	EOI, 6 and 21mo FUP	Leg muscle strength, tandem, 1-limb stand, step test, TUG, death, hospitalization	H
Vestergaard,[58] 2008	63	(1) Unable to get outdoors without a walking aid or help, and/or (2) score ≤3 mobility-tiredness scale; and (3) able to get out of bed/chair.	Control group subjects were asked not to change their usual daily habits.	Home-based training with exercises for flexibility, balance, strength using elastic bands (upper and lower extremities), aerobic.	26min, 3 times a week for 5mo	EOI	Gait speed, SPPB, handgrip, biceps strength, semitandem, chair rise, mobility-tiredness score, EQ-5D, EQ-VAS	H
Villareal,[59] 2011¶	107	Obese elders had to meet 2 of the following operational criteria: (1) modified PPT of 18–32;	Control group subjects did not receive advice to change their diet or activity habits and	Physical therapist led group training combining aerobic, resistance, flexibility, and balance exercises.	90min, 3 times a week for 12mo	EOI, 12mo FUP	SF-36, gait speed, SPPB, total 1 RM, 1-limb stand, obstacle course, FSQ score, adverse effects	L

(*continued on next page*)

Table 1 (*continued*)

Study	N*	Frailty Criteria	Control	Intervention	Frequency and Duration	Follow-up	Performance and Safety Outcomes	RoB†
		(2) Vo_2peak of 11 −18mL/kg/min; or (3) difficulty in performing 2 IADL or 1 BADL.	were prohibited from participating in any weight loss or exercise program. Received general information about a healthy diet in monthly visits with the staff.					
Watt,[60] 2011	74	(1) Individuals with a low IADL score (<3/5); (2) a major orthopedic diagnosis in the lower back, pelvis, or lower extremities since the age of 50y.	Shoulder abductor stretching exercise performed at home. Participants were supervised twice each week by a rehabilitation clinician.	Hip extension stretching exercise performed at home. Participants were supervised twice each week by a rehabilitation clinician.	8min, daily for 10wk	EOI	Gait speed	H
Westhoff,[61] 2000	21	(1) Elders experiencing difficulty in getting up from a chair, and (2) maximum knee-extensor torque in both legs not >87.5Nm (25-kg force).	Control group received no training and was asked to remain habitually active.	Strength exercises for the thigh muscles, with use of elastic bands. Two individual supervised sessions at a facility and 1 home session unsupervised.	60min, 3 times a week for 10wk	EOI, 6mo FUP	Gait speed, tandem, balance test, box stepping, GARS ADL	U
Worm,[62,63] 2001	46	(1) Elders aged >70y; and (2) living in their home; and (3) not able to leave their home unaided or without mobility aids.	Control group was not involved in any intervention.	Group-based training at a facility consisting of flexibility, aerobic, rhythm, balance, strength, and endurance. Home-based muscle and flexibility training.	60min, twice a week (facility) + 5—8min daily (home), for 12wk	EOI	SF-36, gait speed, shoulder abductors, BBS, step maximum speed	U

Abbreviations: BADL, basic activities of daily living; BMI, body mass index; CHS, Cardiovascular Health Study; EOI, end of intervention; EQ-5D, EuroQol quality-of-life scale; EQ-VAS, EuroQol visual analog scale; FSQ, Functional Status Questionnaire; FUP, follow-up after end of intervention; GARS, Global Assessment of Recent Stress; H, high risk; IADL, instrumental activities of daily living; L, low risk; PA, physical activity; PPT, Physical Performance Test; RM, repetition maximum; RoB, risk of bias; ROM, range of motion; SF-36, Medical Outcomes Study 36-Item Short-Form Health Survey; SS, supervised session; U, unknown risk; US, unsupervised session; Vo_2peak, peak oxygen consumption.

* Number of randomized patients, which may differ from number of analyzed patients.

† Risk of bias categories: low risk (L), unknown risk (U), high risk (H).

‡ Two exercise arms grouped into a single intervention arm.

§ Four arms grouped in 1 comparison: exercise ± diet vs control ± diet.

‖ Four arms grouped in 1 comparison: exercise ± vitamin D vs control ± vitamin D.

¶ Four arms analyzed in 2 comparisons: exercise vs control; exercise + diet vs control + diet.

Definitions of frailty used in the trials were often not explicit and when so, they were quite diverse. All trials assessed outcomes at the end of the intervention, and only 5 trials reported longer follow-up data at 6 to 12 months. Overall risk of bias was low in 5 trials.[30,44,49,52,59] The rest of the trials had an unknown or a high risk of bias.

Trials with a 4-arm design[54,59] had their data analyzed by pairing treatment arms (comparing exercise plus diet, with diet and exercise with control). A 2×2 factorial trial[36] testing exercise and a diet intervention presented its data combined into a comparison of exercise versus control. In a trial[34] of 2 exercise arms and 1 control arm, the exercise arms were combined.

According to the methodological quality of the included trials, 5 studies showed a low risk of bias, 4 showed a high risk of bias, and 10 reported an unknown risk of bias (see tables 1 and 2).

Effects of interventions

Results for exercise compared with a control intervention at the end of treatment (12 trials) are presented in figures 2 through 4, and in supplemental appendix S2 (available online only at http://www.archives-pmr.org/). Exercise showed a significant and homogeneous effect on gait speed (see fig 2). Exercising participants walked faster than control participants, with a gait speed that on average was .06m/s higher for normal gait (95% confidence interval [CI], .04−.08) and .08m/s higher for fast gait speed (95% CI, .02−.14). Exercise also had a significant benefit on gait test results, decreasing in 1.73 seconds the time needed to walk 10m (95% CI, .26−3.20; I^2=48%) (see supplemental fig S1, located in supplemental appendix S2). Gait speed refers to a test that requires the individual to walk a certain distance (eg, 8m) in a comfortable fast pace, to derive the gait speed in meters per second. The gait test includes, for example, the rapid gait test, which requires the subject to walk a shorter distance (eg, 3m), turn, and return to the initial position; the latest test is more related to general mobility and is usually assessed with seconds needed to perform the test.

Exercise significantly increased the performance measure SPPB by 1.87 units (95% CI, 1.17−2.57), although no differences were observed in general physical function scales or the Timed Up and Go (TUG) test, both measures showing heterogeneity (I^2=61% and 72%, respectively) (see fig 3).

Exercise did not prove to have a consistent effect on balance measures (see fig 4). Results for tandem and 1-limb tests were not significant albeit highly heterogeneous, while the semitandem test showed a significant increase of 2.93 seconds in the exercise group (95% CI, 1.24−4.62). The Berg Balance Scale (BBS) showed a significant increase of 17.40 points on a single trial (95% CI, 7.76−27.04).

A significant effect of exercise in endurance was observed on the chair rise test, reducing the time needed to stand up 5 times by 2.35 seconds (95% CI, .35−4.35) (see supplemental fig S2, located in supplemental appendix S2).

Functional mobility was assessed through different scales of dependence on ADL activities, and no significant effect of exercise was observed (standardized MD=.39; 95% CI, .07−.71; I^2=67%) (see supplemental fig S3, located in supplemental appendix S2). In particular, the only trial[41] focused on maintaining and improving ADL failed to show a significant effect of exercise on a disability score that included self-rated disabilities in 16 daily activities, a mobility score as the sum of 4 items (such as move outdoors, use stairs, and walk at least 400m), and a self-care ability score as the sum of 7 items (such as walk between rooms, use the toilet, and get dressed).

Only 7 trials presented data on adverse effects related to exercise (see supplemental fig S4, located in supplemental appendix S2). The frequency and characteristics of the adverse effects depended on the type of exercise tested and the setting, either home-based or facility-based. In a facility-based trial[49] aiming to improve balance and strength, no cases of fractures or muscular soreness were observed, and only 1 case of tendinitis occurred. In a second facility-based trial[56] combining aerobic, resistance, flexibility, and balance exercises, only 1 case of fractures and tendinitis was observed, and 4% of participants fell. The 3 home-based trials[42,44,51] presented higher overall incidences of adverse effects. In a trial[52] testing an exercise intervention to improve endurance, strength, and flexibility that was initially center-based and later transitioned to home-based, the overall incidence of muscular soreness was 81.60%. In a trial[15] testing a multimodal intervention with competency-based exercises, overall incidences of fractures, musculoskeletal problems leading to restriction in usual activities, and falls were 3.26%, 31.5%, and 56.5%, respectively. In the last trial[44] testing a multifactorial intervention targeting the Cardiovascular Health Study frailty phenotype, the incidence of back pain was 1%. None of the included trials reported adverse effects such as institutionalization or hospitalization. There were no significant differences between intervention and control in any of the trials. Graphic results for endurance, functional mobility, and adverse effects are shown in supplemental appendix S2.

Seven studies compared different types of exercise and did not show a consistent effect of any of them on performance measures. Supervised facility-based programs of variable composition and intensity were compared with home-based, unsupervised flexibility exercises.[30,35] Benefits were observed for some measures (eg, SPPB, Physical Performance Test, ADL disability) in the more intense programs compared with flexibility exercises in 175 participants.[30,35]

A progressive resistance-training program using weighted vests for resistance (InVest) failed to show a significant effect in performance measures (eg, SPPB, chair stand), either a slow-velocity low-resistance program[27] or the National Institute on Aging's strength-training program.[28]

Balance training using visual computer feedback did not have an effect on performance measures (eg, TUG test, BBS, 6-min walk test) compared with conventional balance training.[50]

The addition of whole-body vibration to strength and balance exercises failed to show an effect on performance measures (eg, TUG test, BBS).[51]

Performing Tai Chi significantly reduced the risk of falling compared with conventional physiotherapy (risk ratio, .74; 95% CI, .56−.98), although the mean number of falls was not different in both groups.[57]

There is scarce evidence available on the effect of exercise past the end of the intervention, and suggests that its benefits are short-lived when exercise is discontinued.[49,51,53,59,61] Nevertheless, there is a significant effect on fast gait speed between 6 and 12 months after discontinuing exercise compared with control (pooled results from 3 trials[49,53,59] not shown), as well as a significant effect on SPPB.[59] The effect on balance, endurance, and functional status dissipates. From the limited evidence comparing different types of exercises at follow-up, similar conclusions can be derived on the dissipation of effect once the exercise is discontinued.[51]

Table 2 Description of studies comparing 2 types of exercise

Study	N*	Frailty Criteria	Intervention 1	Intervention 2	Frequency and Duration	Follow-up	Performance and Safety Outcomes	RoB†
Bean,[27] 2004	21	Individuals with SPPB scores between 4 and 10.	Small-group supervised exercise sessions in a facility. Exercises addressed major muscle groups of the trunk and limbs, emphasizing task-specific movement patterns (InVEST). Progressive resistance program using a weighted vest.	Supervised exercises at a facility consisting of slow-velocity, low-resistance exercises using body or limb weight for resistance.	30min, 3 times a week for 12wk	EOI	Gait test, SPPB, leg press power, unilateral stance, chair rise	U
Bean,[28,29] 2009	138	Individuals with SPPB scores between 4 and 10 who were able to climb a flight of stairs independently or using a device (eg, cane).	Small-group supervised exercise sessions in a facility. Exercises addressed major muscle groups of the trunk and limbs, emphasizing task-specific movement patterns (InVEST). Progressive resistance program using a weighted vest.	Small-group supervised exercise sessions in a facility. Exercises followed the National Institute on Aging training program. Resistance program using free weights.	45—60min, 3 times a week for 16wk	EOI	SPPB, Late Life Function Disability Instrument	U
Binder,[30-33] 2002	119	Individuals had to meet at least 2 of the following 3 criteria: (1) score between 18 and 32 on the modified PPT, (2) report of difficulty or need for assistance with up to 2 IADL or 1 ADL, or (3) achievement of a Vo_2peak between 10 and 18mL/(kg*min).	Facility-supervised exercise program 3 times a week. 1st phase: in group format, focused on flexibility, balance, coordination, speed of reaction, and strength. 2nd phase: progressive resistance training. 3rd phase: endurance training.	Home exercise program comprising 9 activities that challenge flexibility, 2—3 times a week plus a monthly exercise class at a facility.	9mo	EOI	SPPB, 1-limb stand, BBS, Functional Status Scale	L
Brown,[35] 2000	84	Participants had to score <32 points on the PPT.	Supervised exercise program designed to challenge all major muscle groups and to enhance flexibility, balance,	Home exercise performing 9 activities that challenge range of motion. Participants were invited to exercise onsite under	About 3mo	EOI	Gait speed, SPPB, tandem, semitandem test, 1-limb stand, BBS, reach test	U

(*continued on next page*)

Table 2 (*continued*)

Study	N*	Frailty Criteria	Intervention 1	Intervention 2	Frequency and Duration	Follow-up	Performance and Safety Outcomes	RoB†
			strength. 22 exercises with 3 levels of difficulty, 3 times a week for 36 sessions (about 3mo).					
Hagedorn,[50] 2010	35	Patients referred to a falls and balance clinic. (1) Dynamic Gait Index score <19, (2) able to see visual feedback pictures, and (3) able to follow instructions for testing and training.	Computer feedback balance training plus progressive resistance muscle strength, and physical fitness training. Balance training where participant controlled a computer game on the screen through weight shifts. Individual training held at a facility.	Traditional balance training plus progressive resistance muscle strength and physical fitness training. Balance training on different surfaces with open and closed eyes. One legged balance training, walking on a line, passing an obstacle course. Individual training held at a facility.	90min, twice a week for 12wk	EOI	Gait test, tandem, 1-limb stand, BBS, balance test, raise test, 6-min walk test, Fall-Efficacy Scale	H
Pollock,[51] 2012	78	(1) Participants had 2 or more falls in the last 12mo or (2) 1 fall plus TUG test of >15s.	Vibration group: whole-body vibration therapy session after each exercise session, consisting in 5×1-min bouts, separated by 30s of rest, on an asynchronous whole-body vibration platform. Frequency and amplitude up to 30Hz and 8-mm peak to peak.	Exercise group: supervised exercise program at a facility, focused on progressive strength, balance, and functional mobility training.	60min, 3 times a week for 8wk	EOI, 4mo FUP	SF-12, gait speed, BBS, TUG, fear of falling score	U
Tousignant,[57] 2012		People requiring a minimum of 3 services (ie, occupational therapy, physiotherapy, neuropsychology, nursing, or physician)	Supervised Tai Chi: movements that included a combination of body alignment and specific orientations, weight transfer, and changes of direction.	Conventional physiotherapy: balance program that consisted of weight transfer, strengthening, and walking exercises.	60min, twice a week for 15wk	EOI	Falls, number of falls	U

(*continued on next page*)

Table 2 (*continued*)

Study	N*	Frailty Criteria	Intervention 1	Intervention 2	Frequency and Duration	Follow-up	Performance and Safety Outcomes	RoB†
		and (1) referred for a recent fall problem, and (2) identified as being at high risk of fall (BBS≤49/56 and at least 1 accidental fall in the previous 6mo).						

Abbreviations: EOI, end of intervention; FUP, follow-up after end of intervention; H, high risk; IADL, instrumental activities of daily living; InVEST, Increased Velocity Exercise Specific to Task; L, low risk; PPT, Physical Performance Test; RoB, risk of bias; SF-12, 12-Item Short-Form Health Survey; U, unknown risk; V_{O_2}peak, peak oxygen consumption.

* Number of randomized patients, which may differ from number of analyzed patients.

† Risk of bias categories: low risk (L), unknown risk (U), high risk (H).

Discussion

This systematic review has identified the available evidence on the effect of exercise in frail elderly people. When compared with control interventions, exercise has shown to improve gait speed and the SPPB in the frail elderly.

Results are inconclusive for endurance outcomes, and no consistent effect was observed on balance and functional status. The evidence comparing different modalities of exercise is scarce, and it is not possible to pinpoint which exercise characteristics (type, frequency, intensity, duration, setting, combinations) are most effective. Most of the trials included in the review have an unclear or a high risk of bias in their results.

The strong points of this project are as follows: (1) its specific focus on a well-defined population (community-dwelling frail elderly excluding prefrail individuals); (2) its restrictive inclusion of RCTs; (3) the inclusiveness of all types of physical activity interventions and comparisons; and (4) the robust outcomes assessed (performance outcomes), which are relevant indicators of disability for rehabilitation and geriatric specialists. We have focused on frail older adults without dementia and dependency, because this is a population in whom prevention of disability through physical activity is likely. For this reason we have excluded hospitalized and institutionalized individuals, more likely to be dependent or in an unstable clinical condition, and in whom prevention of disability requires further attention. Prefrail individuals were also excluded because different types of exercise programs should be applied.

There are several systematic reviews published on the benefits of physical activity in older adults; however, to our knowledge, there are only 6 systematic reviews[2,20-24] published specifically on the benefits of exercise in frail older adults.

Our review provides an up-to-date search and quantifies the effect of exercise on different performance parameters through meta-analysis. Without regular updates, systematic reviews become outdated quickly, especially in areas of science with many active researchers.[66] Of the 6 previous systematic reviews, only 2[22,24] performed a meta-analysis. De Vries et al[24] could not use weighted MDs in their analysis because of the large variation and the large number of studies that did not report sufficient data; therefore, some of the analysis was based on only a few studies and small samples, resulting in inconclusive CIs. Chou et al[22] also performed a meta-analysis but used a broad definition of frailty that could have included nonfrail and prefrail participants.

Chin A Paw et al[2] examined the effect of exercise on the functional ability of frail older adults. They included all studies that were published between 1995 and 2007, considering any setting and using at least 1 performance-based measure of physical function. No standardized definition of frail was considered, and the included trials presented a variable range of functional abilities. From a qualitative assessment of the trials, the authors concluded that regular exercise training (resistance and multicomponent training) could improve functional outcomes in this population, although more high-quality studies were needed.

Daniels et al[20] examined the effect of any type of intervention on disability in community-dwelling, physically frail older adults. The review included studies verifying at least 1 of the frailty indicators described by Ferrucci et al[11] to identify their participants as frail but focused solely on the outcome disability. Since frailty is thought to be caused by multisystem reduction, the presence of only 1 frailty indicator does not necessarily warrant that participants were frail. With our more strict frailty criteria,

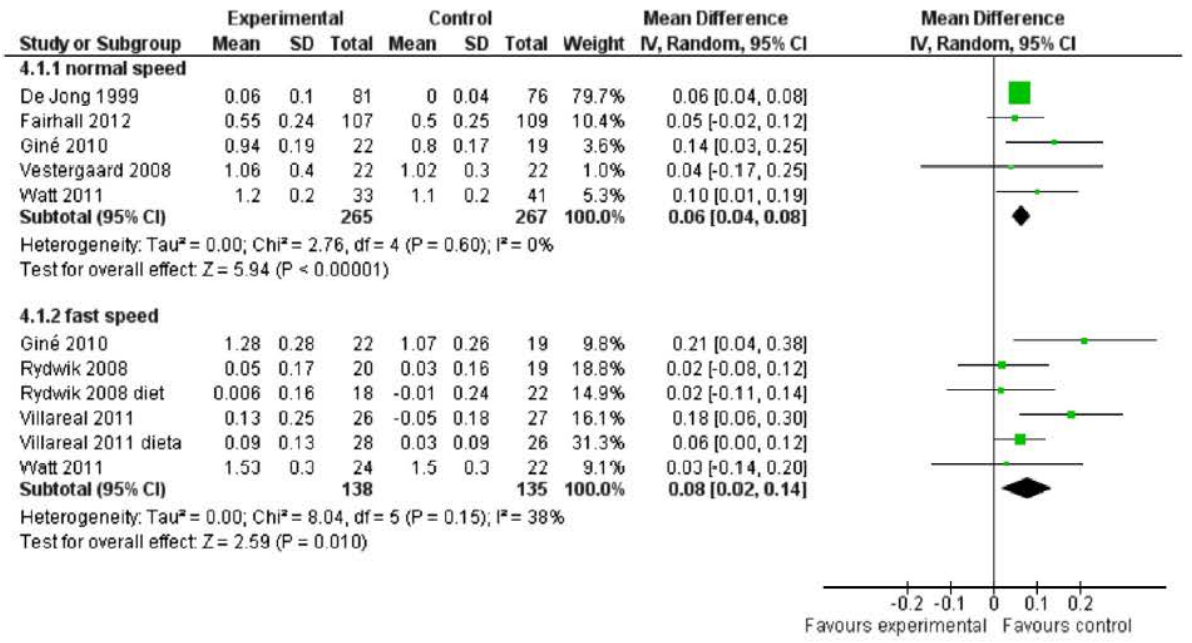


Fig 2 Gait results of exercise compared with control. Abbreviations: df, degrees of freedom; IV, Inverse Variance method.

only 5[15,30,34,36,62] of the 10 studies in Daniels were included in our review. The authors suggested that multicomponent exercise training reduced disability impact, especially in moderately frail people. Nevertheless, the subset of trials verifying our more strict frailty criteria[15,30,34,36,62] showed conflicting results with regard to prevention of disability, and this result is in agreement with the uncertainty identified in our review and in a more general overview.[21] Particularly relevant is that the only included trial[36] that focused on maintaining and improving ADL in community-dwelling frail individuals failed to show a significant effect of exercise on a disability score.

In a qualitative overview, Theou et al[21] examined the effectiveness of current exercise interventions for the management of frailty. The authors included frail subjects who were community dwelling, in retirement homes and mixed settings, in the hospital, and in long-term care. The authors found that only 3 trials used a validated definition of frailty to categorize participants, while the rest of the trials either used a nonvalidated definition or did not include an operational definition of frailty. This key finding that limits the applicability of its results shows the urgent need for a clear and widely accepted definition of frailty. Despite these limitations, the authors pointed out some characteristics of exercise programs that seemed to show superior outcomes: multicomponent training with a duration of $\geq$5 months and performed 3 times per week for 30 to 45 minutes per session. Nevertheless, the applicability of these conclusions is limited given the broad spectrum of participants' settings and interventions considered, the limitations in frailty definition observed, and the qualitative nature of the comparisons performed. Further evidence from specific randomized trials or providing a meta-analysis is necessary to confirm these conclusions.

Chou[22] performed a meta-analysis that aimed to determine the effect of exercise on the physical function, ADL, and quality of life of frail older adults living in the community or institutionalized. Their inclusion criterion for frailty was based on the Fried Frailty Index, Speechley and Tinetti's criteria, and the Falls Efficacy Scale, with a very broad perspective that could have included nonfrail or prefrail participants as well as dependent participants who are past the frailty predisability stage. Regardless of including studies published between 2001 and June 2010, they did not include most trials in Theou's review.[21] The results of their meta-analysis on community and noncommunity trials agree with our findings, showing a significant benefit of exercise in gait speed (their results show an improvement of .07m/s, and our results show an improvement of .06m/s) and BBS, but also great heterogeneity in results for the TUG test and performance in ADL.

De Vries et al[24] also performed a meta-analysis that aimed to assess the effects of physical exercise therapy on physical functioning, mobility, physical activity, and quality of life. Meta-analysis limitations of this trial have been previously discussed. Their inclusion criterion for frailty was based on the presence of mobility problems, physical disability, multimorbidity, or a combination of these, so that nonfrail or prefrail participants could have been included. They found that physical exercise therapy had a positive effect on mobility and physical functioning. High-intensity exercise interventions seem to be more effective in improving physical functioning than low-intensity exercise interventions.

Cadore et al[23] aimed to recommend training strategies that improve the functional capacity in physically frail older adults, focusing specially on supervised exercise programs that improve muscle strength, fall risk, balance, and gait ability. They showed that multicomponent exercise intervention seemed to be the best strategy to improve the rate of falls, gait ability, balance, and strength performance in physically frail older individuals. They included studies that defined subjects as prefrail and mild-to-moderate frail, and there were no restriction to RCTs.

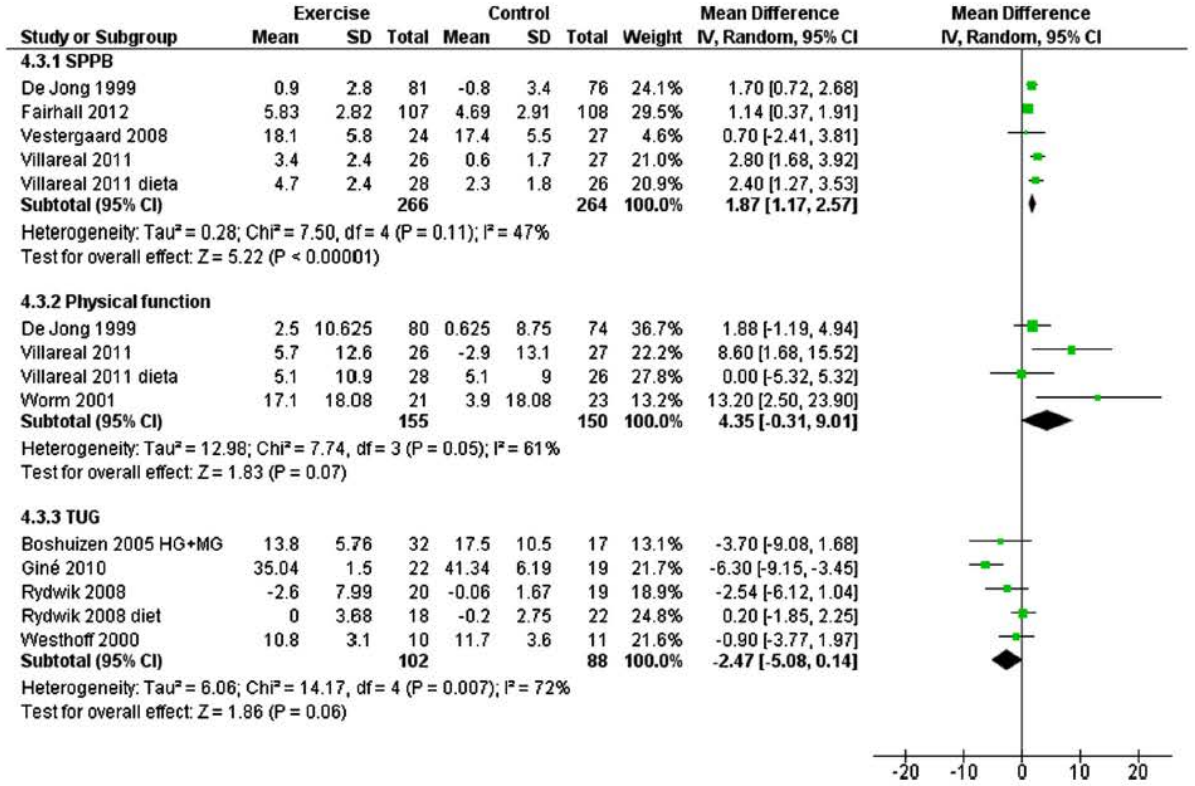


Fig 3 Combined performance results of exercise compared with control. Note that De Jong results have been rescaled from 0 to 16 (16 worse result) to 0 to 100. Abbreviations: df, degrees of freedom; IV, Inverse Variance method.

Our systematic review is in agreement with the systematic reviews cited in that the most studied exercise protocol for frail older adults is a multicomponent training. We have found moderate evidence to support exercise training for improving gait speed and combined performance measures such as SPPB, in line with other authors,[22] but we have found the evidence to be inconclusive regarding the effect of exercise training for improving functional mobility or balance, in contrast to other reviews.[20,22]

In our systematic review, exercise has shown to improve gait speed and performance in the frail elderly, which is similar to Chou's findings.[22] Gait speed slower than .60m/s was a common feature in the frail older adults.[67] Additionally, slowed gait speed in the older adult population has been related to an increased risk for falls,[68] which, in turn, often leads to a loss of independent living and to institutionalization. As an outcome measure, gait speed has been shown to be a predictor of functional decline, nursing home placement, and mortality.[69] Specifically, a decrease in gait speed of 0.1m/s has been associated with a 10% decrease in the ability to perform instrumental ADL.[70] Reduced muscle strength or poor balance results in a decrease in gait speed. Exercise training has shown to increase gait speed; thus, frail older adults might improve in ambulation and require less dependence and assistance in performing ADL. Clinical practice guidelines explicitly recommend lower limb strength exercises and balance training to prevent falls. Gait speed and performance should also be considered.[71]

Improvements in balance and functional mobility might be linked to the exercise program characteristics. Despite the lack of clear evidence of the effect of exercise on ADL, there is an argument for task-oriented or functional practice. Previous studies[49,72] have shown the importance of the exercise being task specific if functional ability is to be improved. The duration of training has also been suggested to be an important contributing factor to the retention of neuromuscular adaptations once training has ended,[73] so longer-duration programs might be recommended.

Some authors argue that the number of adverse events is minimal and rarely life threatening, while the gains of regular exercise clearly outweigh the risks.[15,42,44,49,51,52,56] However, depending on the exercise type, we have found that some important adverse events have been detected, such as fractures or falls. Soreness had been reported as an adverse effect in different trials[52]; however, soreness is a normal consequence of the training process in this population. Therefore, exercise programs should be well designed, and conducted and monitored by well-trained physiotherapists and physical activity specialists. Moreover, trials should systematically report adverse effects (eg, type, when does it appear and disappear, its severity, and whether it causes a hospitalization). This register could allow future assessments on the risk-benefit of the intervention.

While more research is still needed, most evidence shows that regular physical activity or exercise is beneficial for older adults who are frail or at high risk of frailty. Rehabilitation and physical activity specialists should recommend regular physical activity or exercise training to frail older adults as a means to modify frailty and its adverse outcomes.[74] However, the exercise recommendations for a healthy older adult will likely be different from those targeting frail older adults. Specifically, frail older adults may

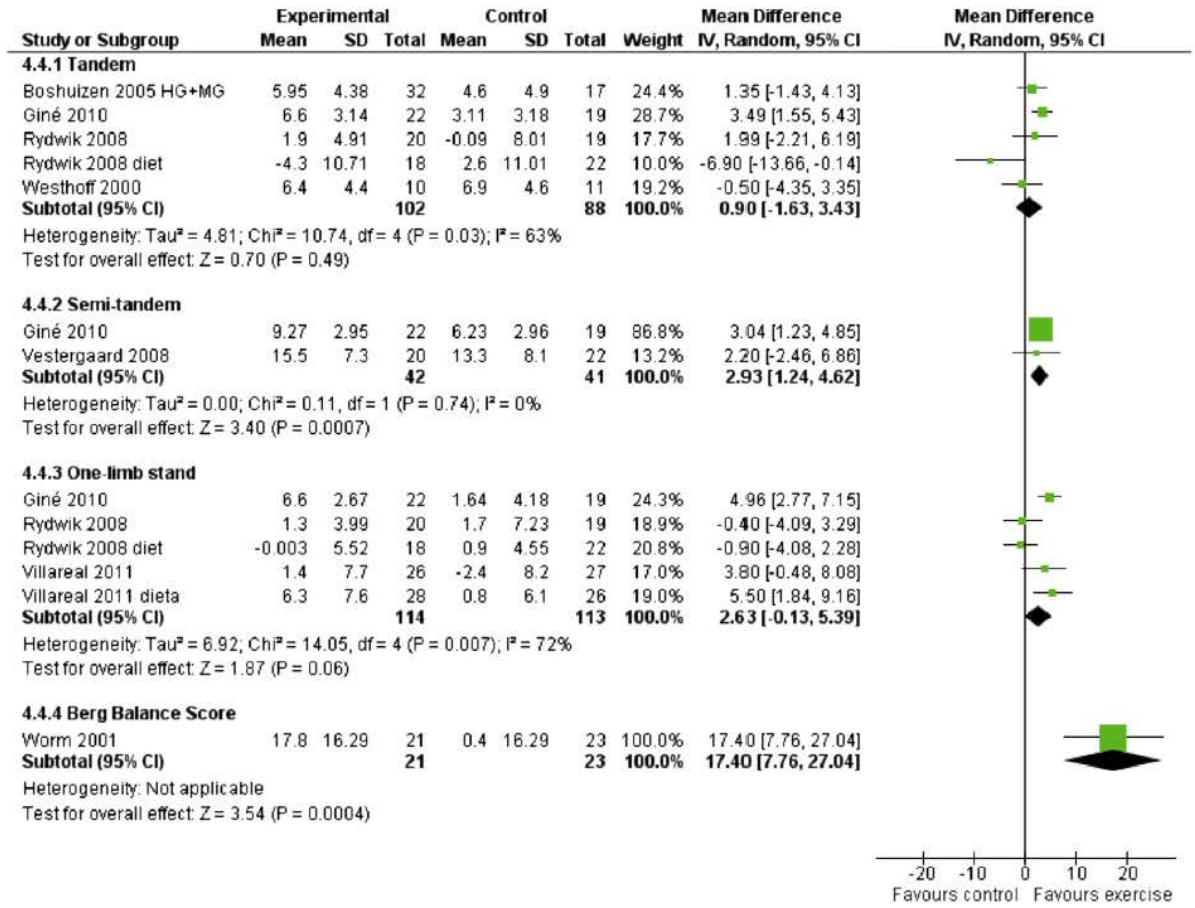


Fig 4 Balance results of exercise compared with control. Abbreviations: df, degrees of freedom; IV, Inverse Variance method.

need functional-based exercise programs with shorter-duration sessions compared with healthy older adults. Physical activity programs linked to local community facilities offering exercise programs for older adults could offer some advantages over home-based programs, facilitating the continuity of a functional-based exercise program linked or not to social activities, but they have other disadvantages in terms of costs, difficulties in transport and comfort, and preferences of users.

With increasing age, there is a well-described decline in voluntary physical activity leading to an increase risk of frailty.[74] In the present systematic review, we have restricted the inclusion criteria to individuals older than 65 years. Liu and Fielding[75] reviewed the literature investigating the utility of aerobic and resistance exercise training as an intervention for frailty in older adults. The authors concluded that gains of regular exercise clearly outweigh its risks (mainly musculoskeletal complaints, rare cases of falls and cardiovascular risks) if the exercise is appropriately designed. According to our results, there is little evidence to guide interventions to prevent or reduce functional mobility and mobility-related disability in frail older people. The optimal intervention to improve these parameters in daily situations remains unclear. Studies should also follow Consolidated Standards of Reporting Trials (CONSORT) recommendations for nonpharmacologic trials[76] to report risk of bias with a total transparency, and make effective interventions reproducible in the clinical practice.

Moreover, several related areas need further investigation. Adherence to an exercise regimen is necessary to observe beneficial effects, and strategies to increase adherence need to be developed in order to effectively implement exercise as a treatment modality on a wide scale. Also, more studies should assess the sustainability of the effects of exercise. In future studies, researchers should also assess whether significant results translate into significant benefits in clinical practice.

Study limitations

Regarding the project's limitations, one common finding in the present review is the variability in participant and intervention characteristics, and outcome measures used across studies, similar to previous reviews. Given the multisystem nature of frailty, this variability is to be expected, since multicomponent interventions need to be proposed to affect different indicators of frailty, which will need to be assessed with different outcome measures. Nevertheless, this great heterogeneity hinders the ability to draw conclusions about the appropriate design of the exercise program and, to some extent, the ability to quantify the effect of exercise interventions. Additional limitations of the project are the sample

sizes and the risk of bias of the trials included in the review, which limit the strength of the conclusions drawn. In the future, it would be desirable to have larger trials with more rigorous methodology conducted to provide more robust evidence on this topic.

Conclusions

Exercise has some benefits in frail older people, although uncertainty still exists with regard to which exercise characteristics (type, frequency, intensity, duration, setting, combinations) are most effective. When compared with control interventions, exercise has shown to improve gait speed and the SPPB in the frail elderly. However, results are inconclusive for endurance outcomes, and no consistent effect was observed on balance and functional mobility.

Some aspects to be taken into account for future research are the need for larger trials with more rigorous methodology, focusing on a well-defined population of community-dwelling frail elderly. Such trials should test the sustainability over time of the effects of physical activity interventions, particularly task-oriented or functional practice programs, incorporating strategies to increase adherence and assessing performance outcomes in the medium- and long-term. Finally, despite significant work over the past decade, a clear consensus definition of frailty does not emerge from the literature.[3] Important areas for further research include whether disability should be considered a component or an outcome of frailty. A consensus on what is frailty and the criteria to be applied in clinical practice will guide the research and the practice recommendations to clearly defined, homogeneous populations.

Supplier

a. CG Information, 740 Granbury Way, Alpharetta, GA 30022.

Keywords

Exercise; Frail elderly; Meta-analysis; Rehabilitation; Review, systematic

Corresponding author

Maria Giné-Garriga, PhD, PT, Department of Physical Activity and Sport Sciences, FPCEE Blanquerna, Universitat Ramon Llull, C/ Císter 34, 08022 Barcelona, Spain. *E-mail address:* mariagg@blanquerna.url.edu.

Acknowledgments

We thank Àlex Domingo (Institute on Aging, Barcelona) for developing and conducting the bibliographic search strategy, and Dr. Rydwik, MD, PhD, for providing additional information on her trial. Marta Roqué-Fíguls is a doctorate candidate at the Pediatrics, Obstetrics and Gynecology and Preventive Medicine Department, Universitat Autònoma de Barcelona, Spain.

References

1. Markle-Reid M, Browne G. Conceptualizations of frailty in relation to older adults. J Adv Nurs 2003;44:58-68.
2. Chin A Paw MJ, van Uffelen JGZ, Riphagen I, van Mechelen W. The functional effects of physical exercise training in frail older people: a systematic review. Sports Med 2008;38:781-93.
3. Rodríguez-Mañas L on behalf of the FOD-CC group. Searching for an operational definition of frailty: a Delphi method based consensus statement. The Frailty Operative Definition-Consensus Conference Project. J Gerontol A Biol Sci Med Sci 2012;68:62-7.
4. Sternberg SA, Wershof Schwartz A, Karunananthan S, Bergman H, Mark Clarfield A. The identification of frailty: a systematic literature review. J Am Geriatr Soc 2011;59:2129-38.
5. Fried LP, Tangen CM, Walston J, et al. Frailty in older adults: evidence for a phenotype. J Gerontol A Biol Sci Med Sci 2001;56:M146-56.
6. Rockwood K, Stadnyk K, MacKnight C, McDowell I, Hébert R, Hogan DB. A brief clinical instrument to classify frailty in elderly people. Lancet 1999;353:205-6.
7. Rockwood K, Song X, MacKnight C, et al. A global clinical measure of fitness and frailty in elderly people. CMAJ 2005;173:489-95.
8. Rockwood K, Mitnitski A. Frailty in relation to the accumulation of deficits. J Gerontol A Biol Sci Med Sci 2007;62:722-7.
9. Gobbens RJ, Luijkx KG, Wijnen-Sponselee MT, Schols JM. Toward a conceptual definition of frail community dwelling older people. Nurs Outlook 2010;58:76-86.
10. Collard RM, Boter H, Schoevers RA, Oude Voshaar RC. Prevalence of frailty in community-dwelling older persons: a systematic review. J Am Geriatr Soc 2012;60:1487-92.
11. Ferrucci L, Guralnik JM, Studenski S, et al. Designing randomized controlled trials aimed at preventing or delaying functional decline and disability in frail older persons: a consensus report. J Am Geriatr Soc 2004;52:625-34.
12. Bergman H, Ferrucci L, Guralnik J, et al. Frailty: an emerging research and clinical paradigm—-issues and controversies. J Gerontol A Biol Sci Med Sci 2007;62:731-7.
13. American College of Sports Medicine, Chodzko-Zajko WJ, Proctor DN, et al. Exercise and physical activity for older adults. Position stand. Med Sci Sports Exerc 2009;41:1510-30.
14. Gates S, Fisher JD, Cooke MW, Carter YH, Lamb SE. Multifactorial assessment and targeted intervention for preventing falls and injuries among older people in community and emergency care settings: systematic review and meta-analysis. BMJ 2008;336:130-3.
15. Gill TM, Baker DI, Gottschalk M, Peduzzi PN, Allore H, Byers A. A program to prevent functional decline in physically frail, elderly persons who live at home. N Engl J Med 2002;347:1068-74.
16. Littbrand H, Lundin-Olsson L, Gustafson Y, Rosendahl E. The effect of a high-intensity functional exercise program on activities of daily living: a randomized controlled trial in residential care facilities. J Am Geriatr Soc 2009;57:1741-9.
17. Peri K, Kerse N, Robinson E, Parsons M, Parsons J, Latham N. Does functionally based activity make a difference to health status and mobility? A randomised controlled trial in residential care facilities (The Promoting Independent Living Study; PILS). Age Ageing 2008;37:57-63.
18. Hiroyuki S, Uchiyama Y, Kakurai S. Specific effects of balance and gait exercises on physical function among the frail elderly. Clin Rehabil 2003;17:472-9.
19. Liu CJ, Latham NK. Progressive resistance strength training for improving physical function in older adults. Cochrane Database Syst Rev 2009;(3):CD002759.
20. Daniels R, van Rossum E, de Witte L, Kempen GI, van den Heuvel W. Interventions to prevent disability in frail community-dwelling elderly: a systematic review. BMC Health Serv Res 2008;8:278.
21. Theou O, Stathokostas L, Roland KP, et al. The effectiveness of exercise interventions for the management of frailty: a systematic review. J Aging Res 2011;2011:569194.
22. Chou CH, Hwang CL, Wu YT. Effect of exercise on physical function, daily living activities, and quality of life in the frail older adults: a meta-analysis. Arch Phys Med Rehabil 2012;93:237-44.
23. Cadore EL, Rodríguez-Mañas L, Sinclair A, Izquierdo M. Effects of different exercise interventions on risk of falls, gait ability, and

balance in physically frail older adults: a systematic review. Rejuvenation Res 2013;16:105-14.
24. de Vries NM, van Ravensberg CD, Hobbelen JS, Olde Rikkert MG, Staal JB, Nijhuis-van der Sanden MW. Effects of physical exercise therapy on mobility, physical functioning, physical activity and quality of life in community-dwelling older adults with impaired mobility, physical disability and/or multi-morbidity: a meta-analysis. Ageing Res Rev 2012;11:136-49.
25. Howe TE, Rochester L, Neil F, Skelton DA, Ballinger C. Exercise for improving balance in older people. Cochrane Database Syst Rev 2011;(11):CD004963.
26. Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated March 2011]. The Cochrane Collaboration: 2011. Available at: http://www.cochrane-handbook.org. Accessed March 21, 2013.
27. Bean JF, Herman S, Kiely DK, et al. Increased Velocity Exercise Specific to Task (InVEST) training: a pilot study exploring effects on leg power, balance, and mobility in community-dwelling older women. J Am Geriatr Soc 2004;52:799-804.
28. Bean JF, Kiely DK, LaRose S, O'Neill E, Goldstein R, Frontera WR. Increased Velocity Exercise Specific to Task (InVEST) training vs. the National Institute on Aging's (NIA) strength training program: changes in limb power and mobility. J Gerontol A Biol Sci Med Sci 2009;64:983-91.
29. Bean JF, Kiely DK, LaRose S, Goldstein R, Frontera WR, Leveille SG. Are changes in leg power responsible for clinically meaningful improvements in mobility in older adults? J Am Geriatr Soc 2010;58:2363-8.
30. Binder EF, Schechtman KB, Ehsani AA, et al. Effects of exercise training on frailty in community-dwelling older adults: results of a randomized, controlled trial. J Am Geriatr Soc 2002;50:1921-8.
31. Binder EF, Yarasheski KE, Steger-May K, et al. Effects of progressive resistance training on body composition in frail older adults: results of a randomized, controlled trial. J Gerontol A Biol Sci Med Sci 2005;60:1425-31.
32. Ehsani AA, Spina RJ, Peterson LR, et al. Attenuation of cardiovascular adaptations to exercise in frail octogenarians. J Appl Physiol 2003;95:1781-8.
33. Villareal DT, Steger-May K, Schechtman KB, et al. Effects of exercise training on bone mineral density in frail older women and men: a radomised controlled trial. Age Ageing 2004;33:309-12.
34. Boshuizen HC, Stemmerik L, Westhoff MH, Hopman-Rock M. The effects of physical therapists' guidance on improvement in a strength-training program for the frail elderly. J Aging Phys Act 2005;13:5-22.
35. Brown M, Sinacore DR, Ehsani AA, Binder EF, Holloszy JO, Kohrt WM. Low-intensity exercise as a modifier of physical frailty in older adults. Arch Phys Med Rehabil 2000;81:960-5.
36. de Jong N, Chin A Paw MJ, de Groot LC, de Graaf C, Kok FJ, van Staveren WA. Functional biochemical and nutrient indices in frail elderly people are partly affected by dietary supplements but not by exercise. J Nutr 1999;129:2028-36.
37. de Jong N, Chin A Paw MJ, de Graaf C, van Staveren WA. Effect of dietary supplements and physical exercise on sensory perception, appetite, dietary intake and body weight in frail elderly subjects. Br J Nutr 2000;83:605-13.
38. de Jong N, Chin A Paw MJ, de Groot LC, Hiddink GJ, van Staveren WA. Dietary supplements and physical exercise affecting bone and body composition in frail elderly persons. Am J Public Health 2000;90:947-54.
39. de Jong N. Sensible aging: using nutrient-dense foods and physical exercise with the frail elderly. Nutr Today 2001;36:202-7.
40. de Jong N, Chin A Paw MJ, de Groot LC, et al. Nutrient-dense foods and exercise in frail elderly: effects on B vitamins, homocysteine, methylmalonic acid, and neuropsychological functioning. Am J Clin Nutr 2001;73:338-46.
41. Chin A Paw MJ, de Jong N, Schouten EG, Hiddink GJ, Kok FJ. Physical exercise and/or enriched foods for functional improvement in frail, independently living elderly: a randomized controlled trial. Arch Phys Med Rehabil 2001;82:811-7.
42. Chin A Paw MJ, de Jong N, Schouten EG, van Staveren WA, Kok FJ. Physical exercise or micronutrient supplementation for the wellbeing of the frail elderly? A randomised controlled trial. Br J Sports Med 2002;36:126-31.
43. Chin A Paw MJ, de Jong N, Pallast EG, Kloek GC, Schouten EG, Kok FJ. Immunity in frail elderly: a randomized controlled trial of exercise and enriched foods. Med Sci Sports Exerc 2000;32:2005-11.
44. Fairhall N, Sherrington C, Kurrle SE, Lord SR, Lockwood K, Cameron ID. Effect of a multifactorial interdisciplinary intervention on mobility-related disability in frail older people: randomised controlled trial. BMC Med 2012;10:120.
45. Cameron ID, Fairhall N, Langron C, et al. A multifactorial interdisciplinary intervention reduces frailty in older people: randomized trial. BMC Med 2013;11:65.
46. Gill TM, Baker DI, Gottschalk M, et al. A prehabilitation program for physically frail community-living older persons. Arch Phys Med Rehabil 2003;84:394-404.
47. Gill TM, Baker DI, Gottschalk M, Peduzzi PN, Allore H, Van Ness PH. A prehabilitation program for the prevention of functional decline: effect on higher level physical function. Arch Phys Med Rehabil 2004;85:1043-9.
48. Peduzzi P, Guo Z, Marottoli RA, Gill TM, Araujo K, Allore HG. Improved self-confidence was a mechanism of action in two geriatric trials evaluating physical interventions. J Clin Epidemiol 2007;60:94-102.
49. Giné-Garriga M, Guerra M, Pagès E, Manini TM, Jiménez R, Unnithan VB. The effect of functional circuit training on physical frailty in frail older adults: a randomized controlled trial. J Aging Phys Act 2010;18:401-24.
50. Hagedorn DK, Holm E. Effects of traditional physical training and visual computer feedback training in frail elderly patients. A randomized intervention study. Eur J Phys Rehabil Med 2010;46:159-68.
51. Pollock RD, Martin FC, Newham DJ. Whole-body vibration in addition to strength and balance exercise for falls-related functional mobility of frail older adults: a single-blind randomized controlled trial. Clin Rehabil 2012;26:915-23.
52. Rejeski WJ, King AC, Katula JA, et al. Physical activity in prefrail older adults: confidence and satisfaction related to physical function. J Gerontol B Psychol Sci Soc Sci 2008;63:P19-26.
53. Rydwik E, Lammes E, Frändin K, Akner G. Effects of a physical and nutritional intervention program for frail elderly people over age 75. A randomized controlled pilot treatment trial. Aging Clin Exp Res 2008;20:159-70.
54. Rydwik E, Frandin K, Akner G. Effects of a physical training and nutritional intervention program in frail elderly people regarding habitual physical activity level and activities of daily living—a randomized controlled pilot study. Arch Gerontol Geriatr 2010;51:283-9.
55. Rydwik E, Gustafsson T, Frandin K, Akner G. Effects of physical training on aerobic capacity in frail elderly people (75+ years). Influence of lung capacity, cardiovascular disease and medical drug treatment: a randomized controlled pilot trial. Aging Clin Exp Res 2010;22:85-94.
56. Lammes E, Rydwik E, Akner G. Effects of nutritional intervention and physical training on energy intake, resting metabolic rate and body composition in frail elderly. A randomised, controlled pilot study. J Nutr Health Aging 2012;16:162-7.
57. Tousignant M, Corriveau H, Roy PM, Desrosiers J, Dubuc N, Hébert R. Efficacy of supervised Tai Chi exercises versus conventional physical therapy exercises in fall prevention for frail older adults: a randomized controlled trial. Disabil Rehabil 2013;35:1429-35.
58. Vestergaard S, Kronborg C, Puggaard L. Home-based video exercise intervention for community-dwelling frail older women: a randomized controlled trial. Aging Clin Exp Res 2008;20:479-86.
59. Villareal DT, Chode S, Parimi N, et al. Weight loss, exercise, or both and physical function in obese older adults. N Engl J Med 2011;364:1218-29.

60. Watt JR, Jackson K, Franz JR, Dicharry J, Evans J, Kerrigan DC. Effect of a supervised hip flexor stretching program on gait in frail elderly patients. PM R 2011;3:330-5.
61. Westhoff MH, Stemmerik L, Boshuizen HC. Effects of a low-intensity strength-training program on knee-extensor strength and functional ability of frail older people. J Aging Phys Act 2000;8:214-27.
62. Worm CH, Vad E, Puggaard L, Stivring H, Lauritsen J, Kragstrup J. Effects of a multicomponent exercise program on functional ability in community-dwelling, frail older adults. J Aging Phys Act 2001;9:414-24.
63. Frederiksen H, Bathum L, Worm C, Christensen K, Puggaard L. ACE genotype and physical training effects: a randomized study among elderly Danes. Aging Clin Exp Res 2003;15:284-91.
64. Sato D, Kaneda K, Wakabayashi H, Shimoyama Y, Baba Y, Nomura T. Comparison of once and twice weekly water exercise on various bodily functions in community-dwelling frail elderly requiring nursing care. Arch Gerontol Geriatr 2011;52:331-5.
65. Szturm T, Betker AL, Moussavi Z, Desai A, Goodman V. Effects of an interactive computer game exercise regimen on balance impairment in frail community-dwelling older adults: a randomized controlled trial. Phys Ther 2011;91:1449-62.
66. Dijkers MP, Bushnik T, Heinemann AW, et al. Systematic reviews for informing rehabilitation practice: an introduction. Arch Phys Med Rehabil 2012;93:912-8.
67. Peterson MJ, Giuliani C, Morey MC, et al. Physical activity as a preventative factor for frailty: the Health, Aging, and Body Composition Study. J Gerontol A Biol Sci Med Sci 2009;64:61-8.
68. Bootsma-van der Wiel A, Gussekloo J, De Craen AJ, Van Exel E, Bloem BR, Westendorp RG. Common chronic diseases and general impairments as determinants of walking disability in the oldest-old population. J Am Geriatr Soc 2002;50:1405-10.
69. Brach JS, Van Swearingen JM, Newman AB, Kriska AM. Identifying early decline in physical function in community-dwelling older women: performance-based and self-report measures. Phys Ther 2002;82:320-8.
70. Judge JO, Schechtman K, Cress E. The relationship between physical performance measures and independence in instrumental activities of daily living. J Am Geriatr Soc 1996;44:1332-41.
71. Gillespie L, Robertson M, Gillespie W, et al. Interventions for preventing falls in older people living in the community. Cochrane Database Syst Rev 2012;(9):CD007146.
72. Skelton DA, Young A, Greig CA, Malbut KE. Effects of resistance training on strength, power, and selected functional abilities of women aged 75 and older. J Am Geriatr Soc 1995;43:1081-7.
73. Smith K, Winegard K, Hicks AL, McCartney N. Two years of resistance training in older men and women: the effects of three years of detraining on the retention of dynamic strength. Can J Appl Physiol 2003;28:462-74.
74. Walston J, Hadley EC, Ferrucci L, et al. Research agenda for frailty in older adults: toward a better understanding of physiology and etiology: summary from the American Geriatrics Society/National Institute on Aging Research Conference on Frailty in Older Adults. J Am Geriatr Soc 2006;54:991-1001.
75. Liu CK, Fielding RA. Exercise as an intervention for frailty. Clin Geriatr Med 2011;27:101-10.
76. Boutron I, Moher D, Altman DG, Schulz KF, Ravaud P. CONSORT Group. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. Ann Intern Med 2008;148:295-309.

5.3 Publicación 3: Chest physiotherapy for acute bronchiolitis in paediatric patients between 0 and 24 months old

Roqué i Figuls M, Giné-Garriga M, Granados Rugeles C, Perrotta C, Vilaró J. Chest physiotherapy for acute bronchiolitis in paediatric patients between 0 and 24 months old. Cochrane Database Syst Rev. 2016 Feb 1;2: CD004873.

FI: 6.124 (2016). Puntuación de atención Altmetric: 107

Esta publicación puede consultarse al completo y de forma libre en https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD004873.pub5/full

La RS que constituye el trabajo 3 incluyó 12 ECA (1249 participantes) [32]. Cinco ensayos (246 participantes) evaluaron técnicas convencionales de fisioterapia (vibración y percusión más drenaje postural), y siete ensayos (1003 participantes) evaluaron técnicas espiratorias pasivas orientadas al flujo: técnicas espiratorias pasivas lentas, en cuatro ensayos, y técnicas espiratorias pasivas forzadas, en tres ensayos.

Las técnicas convencionales no lograron mostrar un beneficio en el resultado primario del cambio en el estado de gravedad de la bronquiolitis, medida por medio de puntuaciones clínicas (cinco ensayos, 241 participantes analizados). La seguridad de las técnicas convencionales se ha estudiado solo anecdóticamente, con un caso de atelectasia (colapso o cierre del pulmón que resulta en un intercambio de gases reducido o nulo) reportado en el brazo de control de un ensayo.

Las técnicas espiratorias pasivas lentas no mostraron un beneficio en los resultados primarios del estado de gravedad de la bronquiolitis y en el tiempo de recuperación (evidencia de baja calidad). Tres ensayos midieron la gravedad de la bronquiolitis a través de puntuaciones clínicas, sin diferencias significativas entre los grupos en ninguno de estos ensayos, realizados en pacientes con enfermedad moderada y grave. Solo un ensayo observó una pequeña mejora significativa transitoria en la puntuación clínica de la escala de Wang, inmediatamente después de la intervención en pacientes con gravedad moderada de la enfermedad. Hay evidencia de muy baja calidad de que las técnicas de espiración pasiva lenta parecen ser seguras, ya que dos estudios informaron de que no se observaron efectos adversos.

Las técnicas de espiración pasiva forzada no tuvieron ningún efecto sobre la gravedad de la bronquiolitis en términos de tiempo de recuperación (dos ensayos, 509 participantes) y tiempo hasta la estabilidad clínica (un ensayo, 99 participantes analizados). Esta evidencia es de alta calidad y corresponde a pacientes con bronquiolitis severa. Además, también hay evidencia de alta calidad de que estas técnicas están relacionadas con un mayor riesgo de desestabilización respiratoria transitoria (razón de riesgo [RR] = 5.4, IC 95%: 1.6 a 18.4, un ensayo) y vómitos durante el procedimiento (RR = 10.2, IC 95%: 1.3 a 78.8, un ensayo). Se obtuvieron resultados no concluyentes para bradicardia con desaturación (RR = 1.0, IC 95%: 0.2 a 5.0, un ensayo) y bradicardia sin desaturación (RR = 3.6, IC: 95%

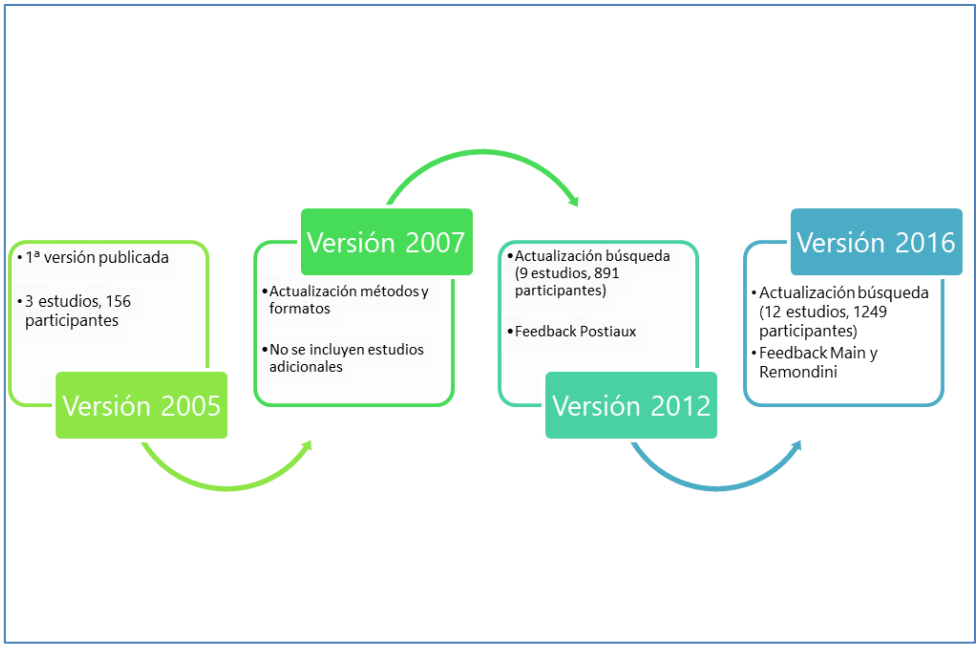
0.7 a 16.9, un ensayo), debido a la limitada precisión de los estimadores. Hay evidencia muy débil del efecto de la espiración forzada combinada con técnicas convencionales en pacientes con bronquiolitis leve a moderada, en los que se observó un alivio inmediato y transitorio de la gravedad de la enfermedad (un ensayo, 13 participantes).

En resumen, ninguna de las técnicas de fisioterapia torácica analizadas en esta revisión (técnicas espiratorias pasivas convencionales o técnicas espiratorias forzadas) ha demostrado una reducción en la gravedad de la enfermedad. Las técnicas de vibración y percusión no se recomiendan en la práctica habitual en entornos hospitalarios debido a la falta de beneficios y al riesgo de posibles eventos adversos. Existe evidencia de alta calidad de que las técnicas de espiración forzada en la bronquiolitis severa no presentan ningún beneficio clínico, y, sin embargo, están relacionadas con efectos adversos como vómitos, bradicardia con desaturación o desestabilización respiratoria transitoria. Existe evidencia de baja calidad que sugiere que las técnicas de flujo lento no proporcionan un beneficio general claro, pero podrían proporcionar algunos beneficios transitorios en algunos niños con bronquiolitis. Los ensayos incluidos tienen un riesgo de sesgo incierto o alto, a excepción de un único ensayo, que administró espiración forzada. El riesgo de sesgo de los ensayos y la imprecisión de las estimaciones llevaron a la baja calidad. Los estudios futuros deberían evaluar el efecto potencial de las técnicas espiratorias pasivas lentas en pacientes no hospitalizados con enfermedad leve a moderada y en pacientes con virus sincitial respiratorio positivo. Además, podrían explorar la combinación de fisioterapia torácica con salbutamol o solución salina hipertónica.

La valoración de calidad de la evidencia de esta RS se presenta en dos tablas de resumen de hallazgos. En la primera, se resumen los resultados para la comparación de las técnicas de espiración pasiva lenta con la no realización de fisioterapia, y, en la segunda, se resumen para la comparación de las técnicas de espiración forzada con la no realización de fisioterapia. Los resultados incluyen: tiempo de recuperación-estabilidad clínica, puntuación clínica y efectos adversos. Como no se realizó un metanálisis en la revisión, las tablas no presentan riesgos comparativos ilustrativos. Se evaluó la calidad de la evidencia utilizando el sistema GRADE, aplicando las pautas del grupo GRADE y desarrollando tablas de resúmenes de hallazgos con el software GRADE profiler.

La publicación 3 corresponde a la versión actual, publicada en The Cochrane Library, de una revisión Cochrane que se publicó originalmente en 2005 y se actualizó en 2007, 2012 y 2016 [32-35]. Está previsto que esta revisión Cochrane vuelva a actualizarse en 2021. La política Cochrane establece que las revisiones deben ser actualizadas con la realización de una nueva búsqueda bibliográfica, aproximadamente cada 2 años, a menos que exista una justificación que indique lo contrario [17]. Las revisiones publicadas se clasifican en 3 categorías (revisión actualizada, pendiente de actualización y revisión estable que no se actualizará) mediante un sistema basado en la propuesta de Garner y colegas [36]. La actualización de una revisión es una oportunidad para considerar cualquier comentario de los lectores, y responder adecuadamente a los mismos. Las actualizaciones de esta RS han consistido en: 1) la actualización de la estrategia de búsqueda e incorporación de nuevos estudios a la revisión, 2) la incorporación de cambios y correcciones sugeridos por el *feedback* de lectores [37-39], y 3) la incorporación de avances metodológicos. En la **figura 5** se resume el historial de actualización de la publicación 3.

Figura 5. Historial de actualizaciones de la publicación 3



Cochrane Database of Systematic Reviews

Chest physiotherapy for acute bronchiolitis in paediatric patients between 0 and 24 months old (Review)

Roqué i Figuls M, Giné-Garriga M, Granados Rugeles C, Perrotta C, Vilaró J

Roqué i Figuls M, Giné-Garriga M, Granados Rugeles C, Perrotta C, Vilaró J.
Chest physiotherapy for acute bronchiolitis in paediatric patients between 0 and 24 months old.
Cochrane Database of Systematic Reviews 2016, Issue 2. Art. No.: CD004873.
DOI: 10.1002/14651858.CD004873.pub5.

www.cochranelibrary.com

[Intervention Review]

Chest physiotherapy for acute bronchiolitis in paediatric patients between 0 and 24 months old

Marta Roqué i Figuls[1], Maria Giné-Garriga[2], Claudia Granados Rugeles[3], Carla Perrotta[4], Jordi Vilaró[5]

[1]Iberoamerican Cochrane Centre - Biomedical Research Institute Sant Pau (IIB Sant Pau), CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain. [2]FPCEE Blanquerna. Department of Physical Activity and Sport Sciences, Universitat Ramon Llull, Barcelona, Spain. [3]Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine, Pontificia Universidad Javeriana, Bogotá, Colombia. [4]Family Medicine, Hospital Italiano de Buenos Aires, Ciudad Autonoma de Buenos Aires, Argentina. [5]Department of Health Sciences, Ramon Lull University, Barcelona, Spain

Contact address: Jordi Vilaró, Department of Health Sciences, Ramon Lull University, Padilla, 326-332, Barcelona, 08025, Spain. jordivc@blanquerna.url.edu, jordi.gestos@gmail.com.

Editorial group: Cochrane Acute Respiratory Infections Group.
Publication status and date: Edited (no change to conclusions), comment added to review, published in Issue 7, 2017.

Citation: Roqué i Figuls M, Giné-Garriga M, Granados Rugeles C, Perrotta C, Vilaró J. Chest physiotherapy for acute bronchiolitis in paediatric patients between 0 and 24 months old. *Cochrane Database of Systematic Reviews* 2016, Issue 2. Art. No.: CD004873. DOI: 10.1002/14651858.CD004873.pub5.

ABSTRACT

Background

This Cochrane review was first published in 2005 and updated in 2007, 2012 and now 2015. Acute bronchiolitis is the leading cause of medical emergencies during winter in children younger than two years of age. Chest physiotherapy is sometimes used to assist infants in the clearance of secretions in order to decrease ventilatory effort.

Objectives

To determine the efficacy of chest physiotherapy in infants aged less than 24 months old with acute bronchiolitis. A secondary objective was to determine the efficacy of different techniques of chest physiotherapy (for example, vibration and percussion and passive forced exhalation).

Search methods

We searched CENTRAL (2015, Issue 9) (accessed 8 July 2015), MEDLINE (1966 to July 2015), MEDLINE in-process and other non-indexed citations (July 2015), EMBASE (1990 to July 2015), CINAHL (1982 to July 2015), LILACS (1985 to July 2015), Web of Science (1985 to July 2015) and Pedro (1929 to July 2015).

Selection criteria

Randomised controlled trials (RCTs) in which chest physiotherapy was compared against no intervention or against another type of physiotherapy in bronchiolitis patients younger than 24 months of age.

Data collection and analysis

Two review authors independently extracted data. Primary outcomes were change in the severity status of bronchiolitis and time to recovery. Secondary outcomes were respiratory parameters, duration of oxygen supplementation, length of hospital stay, use of bronchodilators and steroids, adverse events and parents' impression of physiotherapy benefit. No pooling of data was possible.

Main results

We included 12 RCTs (1249 participants), three more than the previous Cochrane review, comparing physiotherapy with no intervention. Five trials (246 participants) evaluated conventional techniques (vibration and percussion plus postural drainage), and seven trials (1003 participants) evaluated passive flow-oriented expiratory techniques: slow passive expiratory techniques in four trials, and forced passive expiratory techniques in three trials.

Conventional techniques failed to show a benefit in the primary outcome of change in severity status of bronchiolitis measured by means of clinical scores (five trials, 241 participants analysed). Safety of conventional techniques has been studied only anecdotally, with one case of atelectasis, the collapse or closure of the lung resulting in reduced or absent gas exchange, reported in the control arm of one trial.

Slow passive expiratory techniques failed to show a benefit in the primary outcomes of severity status of bronchiolitis and in time to recovery (low quality of evidence). Three trials analysing 286 participants measured severity of bronchiolitis through clinical scores, with no significant differences between groups in any of these trials, conducted in patients with moderate and severe disease. Only one trial observed a transient significant small improvement in the Wang clinical score immediately after the intervention in patients with moderate severity of disease. There is very low quality evidence that slow passive expiratory techniques seem to be safe, as two studies (256 participants) reported that no adverse effects were observed.

Forced passive expiratory techniques failed to show an effect on severity of bronchiolitis in terms of time to recovery (two trials, 509 participants) and time to clinical stability (one trial, 99 participants analysed). This evidence is of high quality and corresponds to patients with severe bronchiolitis. Furthermore, there is also high quality evidence that these techniques are related to an increased risk of transient respiratory destabilisation (risk ratio (RR) 5.4, 95% confidence interval (CI) 1.6 to 18.4, one trial) and vomiting during the procedure (RR 10.2, 95% CI 1.3 to 78.8, one trial). Results are inconclusive for bradycardia with desaturation (RR 1.0, 95% CI 0.2 to 5.0, one trial) and bradycardia without desaturation (RR 3.6, 95% CI 0.7 to 16.9, one trial), due to the limited precision of estimators. However, in mild to moderate bronchiolitis patients, forced expiration combined with conventional techniques produced an immediate relief of disease severity (one trial, 13 participants).

Authors' conclusions

None of the chest physiotherapy techniques analysed in this review (conventional, slow passive expiratory techniques or forced expiratory techniques) have demonstrated a reduction in the severity of disease. For these reasons, these techniques cannot be used as standard clinical practice for hospitalised patients with severe bronchiolitis. There is high quality evidence that forced expiratory techniques in severe patients do not improve their health status and can lead to severe adverse events. Slow passive expiratory techniques provide an immediate and transient relief in moderate patients without impact on duration. Future studies should test the potential effect of slow passive expiratory techniques in mild to moderate non-hospitalised patients and patients who are respiratory syncytial virus (RSV) positive. Also, they could explore the combination of chest physiotherapy with salbutamol or hypertonic saline.

PLAIN LANGUAGE SUMMARY

Chest physiotherapy for acute bronchiolitis in children younger than two years of age

Review question

We reviewed the evidence about the effect of chest physiotherapy in infants younger than two years of age with acute bronchiolitis.

Background

Acute bronchiolitis is a frequent viral respiratory infection in children younger than two years of age. Most children have a mild disease and do not require hospitalisation. Those who do need to be hospitalised sometimes have difficulty clearing phlegm (thick mucous respiratory secretions caused by the infection). It has been proposed that chest physiotherapy may assist in the clearance of respiratory secretions and improve breathing. There are three different types of chest physiotherapy available: vibration and percussion, forced expiratory techniques and slow flow techniques that avoid blockage of the airway.

Study characteristics

The evidence is current to July 2015. This review has included 12 trials with a total of 1249 participants. By type of chest physiotherapy, five trials tested vibration and percussion techniques in 246 participants, three trials tested forced expiratory techniques in 624 participants, and four trials tested slow flow techniques in 375 participants.

Key results

Vibration and percussion techniques produce a thorax (chest) oscillation by fast compression or percussion with the physiotherapist's hands. Neither manoeuvre was shown to improve the clinical scores of patients with acute bronchiolitis in the trials. These techniques did not show improvements in respiratory measurements, time on oxygen therapy or length of hospital stay. There were no data on time to recovery from acute bronchiolitis, use of bronchodilators or steroids, or parents' assessment of physiotherapy benefit. The trials included in this review did not present data on adverse effects related to the intervention, but the literature cites cases of relevant adverse effects such as rib fractures related to these techniques.

Forced expiratory techniques consist of suddenly increasing the expiratory flow by compressing the thorax or the abdomen. In participants with severe bronchiolitis, such techniques failed to reduce time to recovery or time to clinical stability when compared to no physiotherapy. They also failed to improve clinical scores, oxygen saturation or respiratory rates except in mild to moderate bronchiolitis patients. There were no data on secondary outcomes such as duration of oxygen supplementation, length of hospital stay, or use of bronchodilators and steroids. Two studies reported no significant differences in parents' impression of the benefit of physiotherapy compared to controls. One of the trials reported a higher number of transient episodes of vomiting and respiratory instability after forced expiratory physiotherapy. This trial found no differences for bradycardias (decreases in heart rate), with and without desaturation (reduced oxygen levels in blood).

Slow flow techniques consist of compressing the rib cage and the abdominal cavity gradually and gently from the mid-expiratory phase up to the end of exhalation. Slow flow techniques showed an overall lack of benefit on clinical scores of severity of the disease. However, in two trials they provided either a short-lived relief in terms of clinical scores or a decrease in the need for oxygen support in children with moderate bronchiolitis. There were no changes in length of hospital stay, use of bronchodilators or steroids. There were no data on changes in time to recovery, change in respiratory measurements, or parents' impression of physiotherapy benefit. No severe adverse events were reported in the trials.

Quality of the evidence

Vibration and percussion techniques are not recommended in routine practice in hospital settings due to a lack of benefit and risk of potential adverse events. There is high quality evidence that forced expiratory techniques in severe bronchiolitis present no clinical benefit, while being related to adverse effects such as vomiting, bradycardia with desaturation, or transient respiratory destabilisation. There is low quality evidence that suggests that slow flow techniques do not provide a clear overall benefit, but could provide some transient benefits in some children with bronchiolitis. Except for one trial, related to forced expiration, the included trials are at unclear or high risk of bias. The risk of bias of the trials and the imprecision of the estimates led to the low quality of evidence for the effect of slow flow techniques on clinical scores. Further trials are needed before reaching firm conclusions.

SUMMARY OF FINDINGS FOR THE MAIN COMPARISON *[Explanation]*

Forced expiration compared with no physiotherapy for acute bronchiolitis

Patient or population: paediatric patients between 0 and 24 months old with acute bronchiolitis
Settings: hospital
Intervention: forced expiration
Comparison: standard care

Outcomes	Relative effect (95% CI)	No. of participants (studies)	Quality of the evidence (GRADE)	Comments
Time to recovery/time to clinical stability (follow-up until hospital discharge)	Studies reported that no differences in time to recovery/clinical stability were observed	624 (3 trials)	⊕⊕⊕⊕ **high**	Participants with severe bronchiolitis (Gajdos 2010; Rochat 2010) Participants with mild-moderate bronchiolitis (Remondini 2014)
Adverse events (follow-up until hospital discharge)	Bradycardia with desaturation (RR 1.0, 95% CI 0.2 to 5.0) Bradycardia without desaturation (RR 3.6, 95% CI 0.7 to 16.9) Transient respiratory destabilisation (RR 5.4, 95% CI 1.6 to 18.4) Vomiting during procedure (RR 10.2, 95% CI 1.3 to 78.8)	496 (2 trials)	⊕⊕⊕⊕ **high**	Participants with severe bronchiolitis (Gajdos 2010; Rochat 2010)

*The basis for the **assumed risk** (e.g. the median control group risk across studies) is provided in footnotes. The **corresponding risk** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).
CI: confidence interval; **RR:** risk ratio

GRADE Working Group grades of evidence
High quality: Further research is very unlikely to change our confidence in the estimate of effect.
Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
Very low quality: We are very uncertain about the estimate.

BACKGROUND

Description of the condition

Acute bronchiolitis is the leading cause of Emergency Department visits during winter in children younger than two years of age. It results in high utilisation of healthcare resources and is an increasing burden on outpatient practices, Emergency Departments and hospitals (Carroll 2008). It also results in significant morbidity for infants. Infant mortality rates vary depending upon the population. In high-income countries, incidence of bronchiolitis-associated deaths is low, and due mainly to patients with severe comorbidities (e.g. congenital heart disease, etc.). For example, it was reported to be 2 per 10,000 live births in the USA in the 1990s (Holman 2003) and 1.82 per 100,000 in the UK in 2000 (Panickar 2005). Furthermore, there is strong evidence of irreversible airway damage and reduced lung function in adults who were hospitalised for bronchiolitis in infancy (Sigurs 2010). Children who have had respiratory syncytial virus (RSV) disease in early life have been shown to have a higher incidence of asthma/wheezing in later life (odds ratio 3.84: Régnier 2013).

Some years ago, the American Academy of Pediatrics published a statement on the diagnosis and treatment of bronchiolitis (AAPs 2006). However, criteria for diagnosing acute bronchiolitis vary greatly. Most doctors agree that the case definition for an episode of acute bronchiolitis should include children aged 24 months or younger who have a first episode of acute wheezing accompanied by physical findings of viral infection (for example, coryza, cough and fever) (González 2001; Videla 1998; Wainwright 2003). The most prevalent virus identified with the disease is RSV.

Most cases of acute bronchiolitis are mild and can be treated on an outpatient basis; 1% to 3% (depending on the severity of the disease) will require hospitalisation (Ralston 2014). Risk factors associated with the need for hospitalisation are young age, premature birth, chronic lung disease, congenital heart disease and a deficient immune system (AAPs 2006). In low-income countries the most frequent risk factors associated with hospitalisation and severe disease include living in a low-income family, malnourishment, low birthweight, age of the mother, mother's education level, being bottle-fed and premature birth (Smyth 2006; Spencer 1996).

Description of the intervention

The standard treatment of acute bronchiolitis is to ensure adequate oxygenation, fluid intake and feeding of the infant (AAP 2006; SIGN 2006). Pharmacological strategies considered in acute bronchiolitis include bronchodilators, antibiotics and steroids but their effectiveness remains quite uncertain and current guidelines do not recommend their use (AAPs 2006; SIGN 2006). There is no evidence to support the use of glucocorticoids or antibiotics (Farley 2014; Fernandes 2013), and although there is some evidence that bronchodilators, nebulised hypertonic saline, epinephrine and heliox therapy may have some benefit in terms of improving clinical scores (Gadomski 2014; Hartling 2011; Liet 2010; Umoren 2011; Zhang 2011), this benefit must be weighed against the lack of benefit in reducing the duration or severity of illness, costs and adverse effects.

Chest physiotherapy has been proposed to assist in the clearance of tracheo-bronchial secretions. The main goal is to clear the airway obstruction, reduce airway resistance, enhance gas exchange and reduce the work of breathing. Different techniques are used in paediatric patients: 1) the conventional chest physical therapy (cCPT) such as chest percussion and vibration in combination with postural drainage positions, chest shaking and directed coughing and 2) the flow-based techniques: slow or forced passive expiration may help to mobilise secretions towards the trachea and trigger coughing that helps to remove secretions. Specific measures are recommended to prevent spreading of the disease during the procedure, such as cohort segregation, hand washing and wearing gowns, masks, gloves and goggles (Hall 1981). However, conventional chest physiotherapy techniques may have drawbacks: it has been claimed that they might cause distress to the infant and concerns have arisen about the safety of the procedure, especially in relation to rib fractures in patients at risk (Beeby 1998; Chalumeau 2002; Chanelière 2006).

Why it is important to do this review

At the time of the first publication of this review, there was uncertainty about the efficacy of conventional physiotherapy techniques (vibration and percussion). The review challenged their application in daily practice, prompting the recommendation that chest physiotherapy based on vibration and percussion not be applied routinely in hospital settings (AAP 2006; BGT 2005; SIGN 2006). However, chest physiotherapy is still being applied in outpatient and inpatient settings (Barben 2008; González 2010a). Parents' expectation and demand for chest physiotherapy in clinical daily practice may explain its widespread use (Sanchez 2007). New and gentler passive expiratory physiotherapy techniques have become mainstream in several countries. In France, passive forced exhalation techniques are recommended by a consensus panel both for inpatient and outpatient cases (Beauvois 2001; Consensus 2001), with extremely high implementation in outpatient settings (David 2010; Halna 2005; Touzet 2007). However, lately there seems to be contrary practice to the routine use of respiratory physiotherapy in bronchiolitis. Other countries such as Chile also report using chest physiotherapy in outpatient and inpatient settings, although it is not clear which techniques are applied (Girardi 2001). These changes motivated a shift in the focus of the review, in order to assess the efficacy and safety of passive expiratory techniques, and to explore the differential effect of chest physiother-

apy depending on the technique used, severity of the patients and setting of implementation.

OBJECTIVES

To determine the efficacy of chest physiotherapy in infants aged less than 24 months old with acute bronchiolitis. A secondary objective was to determine the efficacy of different techniques of chest physiotherapy (for example, vibration and percussion and passive forced exhalation).

METHODS

Criteria for considering studies for this review

Types of studies

We included randomised controlled trials (RCTs) evaluating chest physiotherapy in acute bronchiolitis.

Types of participants

Infants younger than 24 months of age with acute bronchiolitis as defined by the trial authors, in all settings.

Types of interventions

We included trials that compared any type of chest physiotherapy (postural drainage, chest percussion, vibration, chest shaking, directed coughing, slow or forced expiration techniques) versus standard care or other physiotherapy, drainage or breathing techniques.

The interventions are classified into two main categories: vibration and percussion, and passive expiratory techniques. Passive expiratory techniques are further subdivided into slow passive expiratory techniques and forced passive expiratory techniques.

Types of outcome measures

Primary outcomes

1. Change in the severity status of bronchiolitis.
2. Time to recovery.

Secondary outcomes

1. Respiratory parameters (oxygen saturation levels, transcutaneous carbon dioxide partial pressure ($PaCO_2$)).
2. Duration of oxygen supplementation.
3. Length of hospital stay.
4. Use of bronchodilators and steroids.
5. Parents' impression of physiotherapy benefit.
6. Adverse events. We defined adverse events as any undesired outcome due to the intervention. For example, rib fractures, bradycardia, respiratory instability, vomiting or long-term neurological disabilities. We took all outcomes into consideration. We described the method used to measure any adverse events.

Search methods for identification of studies

Electronic searches

In this update we searched the Cochrane Central Register of Controlled Trials (CENTRAL 2015, Issue 9) (accessed 8 July 2015), the Cochrane Acute Respiratory Infections Group's Specialised Register (October 2011 to July 2015), MEDLINE and MEDLINE in-process and other non-indexed citations (October 2011 to July 2015), EMBASE (October 2011 to July 2015), CINAHL (October 2011 to July 2015), LILACS (October 2011 to July 2015), Web of Science (October 2011 to July 2015) and Pedro (October 2011 to July 2015). See Appendix 1 for details of previous searches.

We used the search strategy described in Appendix 2 to search CENTRAL and MEDLINE. We did not combine the search strategy with a filter for identifying randomised trials as there were too few results. We adapted the search strategy to search MEDLINE in-process (Appendix 3); EMBASE (Appendix 4); CINAHL (Appendix 5); LILACS (Appendix 6) and Web of Science (Appendix 7).

Searching other resources

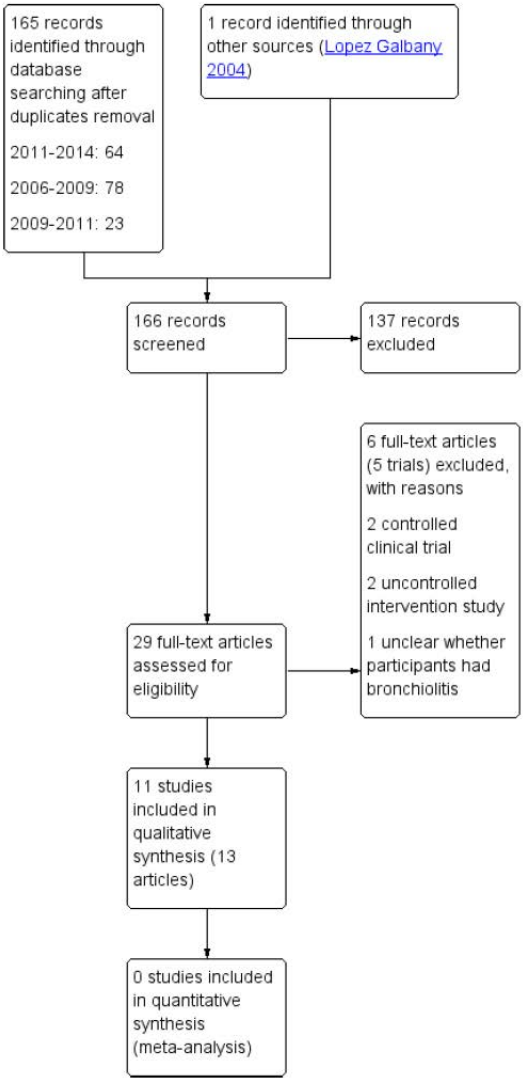
In the first publication of this review, we examined the reference lists of general paediatric, infectious diseases, pneumatology and physiotherapy textbooks. We reviewed reference lists of all selected articles and recent review articles and also examined published abstracts from the Pediatric Academic Societies' Annual Meetings (US) (1999 to 2003). We handsearched the French journals *Journal Pédiatrie Puériculture* (1999 to May 2004) and *Archives de Pédiatrie* (1994 to 1997; 2000 to May 2004). We also searched the World Health Organization (WHO) International Clinical Trials Registry Platform (ICTRP) and www.ClinicalTrials.gov trials registers with the search terms bronchiolitis AND "chest physiotherapy" for completed and ongoing studies (latest search 8 July 2015).

Data collection and analysis

Selection of studies

Three review authors (CG, MG, MR) independently screened the results from the initial search of all the databases and reference lists to identify citations that seemed relevant to this review. We obtained the full-text articles once pertinent abstracts or titles were identified. Four review authors (CG, MG, MR, JV) independently decided on which trials to include using a standard form. There were no disagreements in relation to the included trials. We recorded the selection process in sufficient detail to complete a PRISMA flow diagram (see Figure 1) (Moher 2009) and Characteristics of excluded studies table.

Figure 1. Study flow diagram



Data extraction and management

Two review authors (MR, MG) independently extracted the data. We used a standard form to extract the following data.

1. Characteristics of the study (design, method of randomisation, withdrawals, drop-outs).
2. Participants (age, gender, low birth weight or normal weight, ambulatory or hospital patients, disease severity, nutritional status).
3. Intervention (type of chest physiotherapy, administration, co-interventions) and its comparator.
4. Outcomes (types of outcome measures, timing of outcomes, adverse effects).
5. Results.

Assessment of risk of bias in included studies

Two review authors (MG, MR) independently assessed risk of bias for each study using the criteria outlined in the *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins 2011). We resolved any disagreement by discussion.

1. Sequence generation (selection bias)

We described for each included study the methods used to generate the allocation sequence in sufficient detail to allow an assessment of whether it should produce comparable groups. We assessed the methods as:

- low risk of bias (any truly random process, e.g. random number table; computer random number generator);
- high risk of bias (any non random process, e.g. odd or even date of birth; hospital or clinic record number); or
- unclear risk of bias.

2. Allocation concealment (selection bias)

We described for each included study the method used to conceal the allocation sequence in sufficient detail to determine whether intervention allocation could have been foreseen in advance of, or during recruitment, or changed after assignment. We assessed the methods as:

- low risk of bias (e.g. telephone or central randomisation; consecutively numbered, sealed, opaque envelopes);
- high risk of bias (open random allocation; unsealed or non-opaque envelopes, alternation; date of birth); or
- unclear risk of bias.

3. Blinding of outcome assessment (detection bias)

Blinding of study participants and personnel was not possible due to the characteristics of the interventions studied. We described for each included study all the methods used, if any, to blind outcome assessors from knowledge of which intervention a participant received. We also provided information on whether the intended blinding was effective. Where blinding was not possible, we assessed whether the lack of blinding was likely to have introduced bias. We assessed the methods as:

- adequate;
- high risk of bias; or
- unclear risk of bias.

4. Incomplete outcome data (attrition bias through withdrawals, drop-outs, protocol deviations)

We described for each included study and for each outcome or class of outcomes the completeness of data including attrition and exclusions from the analysis. We stated whether attrition and exclusions were reported, the numbers included in the analysis at each stage (compared with the total randomised participants), reasons for attrition or exclusion where reported and whether missing data were balanced across groups or were related to outcomes. We assessed whether each study was at risk of attrition bias:

- low risk of bias;
- high risk of bias; or
- unclear risk of bias.

5. Selective reporting bias

We described for each included study how the possibility of selective outcome reporting bias was examined by us and what we found. We assessed the methods as:

- low risk of bias (where it is clear that all of the study's pre-specified outcomes and all expected outcomes of interest to the review have been reported);
- high risk of bias (where not all of the study's pre-specified outcomes have been reported; one or more reported primary outcomes were not pre-specified; outcomes of interest are reported incompletely and so cannot be used; study fails to include results of a key outcome that would have been expected to have been reported); or
- unclear risk of bias.

6. Other sources of bias

We described for each included study any important concerns we have about other possible sources of bias, in particular about

contamination. We assessed whether each study was free of other problems that could put it at risk of bias:

- low risk of bias;
- high risk of bias; or
- unclear risk of bias.

Measures of treatment effect

We estimated the effect of treatment by mean differences (MDs) in continuous outcomes and risks ratios (RRs) in dichotomous outcomes, with their corresponding confidence intervals (CIs).

Unit of analysis issues

We would have assessed their data analysis in search of possible unit of analysis errors if any cluster-randomised trials had been included in the review. We would have combined them with individually randomised trials if no errors were observed. We did not expect to identify any cross-over randomised trial on this topic given the short course of bronchiolitis.

Dealing with missing data

We assessed the impact of missing data on the results from the 'Risk of bias' assessment, considering for each trial the magnitude of missing data and how it was dealt with. We tried to assess how many patients were excluded from the trials analysis, which treatment group they belonged to, what were the causes for excluding them and whether their exclusion was biased the trials results. If a quantitative analyses had been performed, the main analysis would be based on available data and a secondary intention-to-treat (ITT) sensitivity analysis would have been performed for dichotomous outcomes. The ITT sub-analysis would have used imputation, assuming that all missing data corresponded to a negative outcome.

Assessment of heterogeneity

We would have assessed statistical heterogeneity with the I^2 statistic, considering values $I^2 \geq 50\%$ as a sign of moderate to high heterogeneity if the trials included had been similar enough to perform a quantitative analysis (Higgins 2003).

Assessment of reporting biases

We did not explore publication bias and other reporting biases statistically or graphically due to the lack of statistical data in the included studies.

Data synthesis

We did not perform a meta-analysis due to clinical heterogeneity and statistical considerations. We described the individual results with the effect measures described in the original trials. If the included trials had been similar enough to combine them, a statistical pooling of effect measures would have been performed with a random-effects model, applying the inverse-variance method. We wrote the review using Review Manager 5.3 (RevMan 2014).

GRADE and 'Summary of findings' table

We added 'Summary of findings' tables to this 2014 update, comparing slow passive expiration techniques with no physiotherapy and forced expiration techniques with no physiotherapy. The outcomes included: time to recovery/clinical stability, clinical score and adverse effects. Since we did not perform a meta-analysis in the review, we did not present illustrative comparative risks in the tables. We assessed the quality of evidence using the GRADE system. We used the guidelines of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group to assess the quality of the evidence related to selected outcomes (Guyatt 2008). The GRADE system assesses the quality of evidence based on the extent to which users can be confident that an association reflects the item being evaluated (Guyatt 2008). Assessment of the quality of evidence included risk of bias, heterogeneity, directness of the evidence, risk of publication bias and precision of effect estimates, among others (Guyatt 2011; Guyatt 2011a; Guyatt 2011b; Guyatt 2011c; Guyatt 2011d; Guyatt 2011e; Guyatt 2011f; Guyatt 2011g). We developed 'Summaries of findings' tables with the GRADE profiler software (GRADEpro GDT 2015).

Subgroup analysis and investigation of heterogeneity

In the 2014 update, we proposed two subgroup analyses based on the hypothesis that performance of slow flow chest physiotherapy techniques could depend on the patient's severity and, consequently, on setting (inpatient versus outpatient). We introduced a subgroup analysis by patient severity, classifying trials into severe/moderate/unknown categories depending on the inclusion criteria of the trial, or on the characteristics of the included participants. We proposed a subgroup analysis by setting, classifying trials into inpatient/outpatient categories, under the hypothesis that patients with more severe bronchiolitis would be seen in inpatient settings, while outpatient settings would attend a variable pool of patients, but mostly with moderate or low levels of bronchiolitis severity.

Sensitivity analysis

If a quantitative analyses had been performed, we would have carried out an ITT sensitivity analysis for dichotomous outcomes, imputing all missing data as a negative outcome.

RESULTS

Description of studies

Results of the search

In the search update up to July 2015, we retrieved 129 unique records from the databases searched, and we included three new trials (Gomes 2012; Remondini 2014; Sanchez Bayle 2012). The Gomes paper corresponds to an ongoing trial included in previous review versions (Clinicaltrials.gov identifier NCT00884429). We identified one ongoing trial (Bella Lisbôa 2008).

Included studies

See Characteristics of included studies table.

We included 12 RCTs in this review totaling 1249 participants (Aviram 1992; Bohe 2004; De Córdoba 2008; Gajdos 2010; Gomes 2012; Lopez Galbany 2004; Nicholas 1999; Postiaux 2011; Remondini 2014; Rochat 2010; Sanchez Bayle 2012; Webb 1985).

A description of included trials by type of intervention is shown in Table 1. Five trials assessed percussion and vibration techniques in 246 randomised participants (Aviram 1992; Bohe 2004; De Córdoba 2008; Nicholas 1999; Webb 1985), while six trials assessed different passive flow-oriented expiratory techniques in 974 randomised participants. Three of these trials assessed forced expiration techniques (Gajdos 2010; Remondini 2014; Rochat 2010), and four trials assessed slow flow techniques (Gomes 2012; Lopez Galbany 2004; Postiaux 2011; Sanchez Bayle 2012). The Gomes 2012 trial assessed the effect of slow flow passive expiratory techniques (slow flow) against vibration and percussion techniques. All 12 trials evaluated the efficacy of chest physiotherapy in hospitalised infants with a clinical diagnosis of acute bronchiolitis.

The trials were classified by the clinical severity of the included infants, as reported in the papers or as estimated by the review authors. Clinical severity of participants was mild in one trial (De Córdoba 2008 1.9 mean Silverman-Anderson score at baseline, out of 10 maximum score), moderate in six trials (Bohe 2004 5.7 mean Wang score at baseline; Gomes 2012 75% of participants with a four to eight points in Wang score; Postiaux 2011 5.75 mean Wang score at baseline; Webb 1985 11 mean clinical score at admission over 30 maximum score; Lopez Galbany 2004 5.6 mean Wang score at baseline; Remondini 2014 5.8 mean respiratory distress assessment instrument (RDAI) score at baseline) and severe in four trials (Gajdos 2010; Nicholas 1999; Rochat 2010; Sanchez Bayle 2012). Also, in these studies with severe bronchiolitis patients, they included infants who required nasogastric feeding or intravenous fluid. The severity of bronchiolitis in one trial was unknown (Aviram 1992).

The studies were carried out in the UK (Nicholas 1999; Webb 1985), Spain (Lopez Galbany 2004; Sanchez Bayle 2012), Brazil (De Córdoba 2008; Gomes 2012; Remondini 2014), France (Gajdos 2010), Belgium (Postiaux 2011), Israel (Aviram 1992), Argentina (Bohe 2004), and Switzerland (Rochat 2010).

Two of the included trials are unpublished and we contacted the trial authors for further clarification and data gathering (Aviram 1992; Lopez Galbany 2004). We contacted the authors of several trials asking for clarification and additional information, with positive responses (Aviram 1992; Gomes 2012; Lopez Galbany 2004; Postiaux 2011; Rochat 2010; Sanchez Bayle 2012).

Finally, only two studies reported specific funding from governmental organisations (Gajdos 2010; Rochat 2010). Two declared no conflicts of interest (Postiaux 2011; Sanchez Bayle 2012), and the other studies did not specify any conflicts of interest.

Published trials

A recent trial was conducted in Brazil included 29 infants younger than one year admitted to hospital with a diagnosis of acute bronchiolitis (Remondini 2014). Patients that presented with congenital heart diease, neuropathy, underlying lung disease, indication for ventilatory support, RDAI score $\leq$ four associated to $SpO_2 \geq$ 92% were excluded. Patients were randomly allocated in two intervention groups. One (n = 16) underwent postural drainage associated to percussion and tracheal aspiration and the other group (n = 13), underwent postural drainage associated with forced passive expiratory technique and tracheal aspiration. Patients were assessed three times a day (before, 10 and 60 minutes after the physiotherapy intervention) by the same therapist. The endpoint was to compare the efficacy of both techniques in improving RDAI and SpO_2. Trial authors considered discharging patients from the study when the RDAI score was $\leq$ four, which was associated with adequate oxygenation ($SpO_2 \geq$ 92%). The total number of sessions was 83; 48 in conventional group and 35 in force expiratory group. The physiotherapist in charge of the infant determined the number of sessions according to the disease severity. The session numbers ranged from one to four a day.

A trial conducted in Spain and recruited 293 infants less than seven months old admitted to hospital with a diagnosis of first episode of acute bronchiolitis by the McConnochie 1993 criteria and at least one of the following signs: toxic aspect; history of apnoea or cyanosis; respiratory rate > 60; or pulse oxymetry < 94%. Inclusion criteria and signed informed consents were conducted after randomisation, leading to the exclusion of 40 randomised participants not meeting the criteria, and 16 participants whose parents refused consent because of the blinded design of the study that prevented knowing the intervention received (Sanchez Bayle 2012). Participants were allocated to receive either prolonged slow expiratory technique with manual vibration and assisted cough (n = 136) or postural changes plus oxygen therapy until pulse oximetry oxygen saturation (SpO_2) >= 94% (n = 100). All interventions were administered twice a day and only the physiotherapists were aware of the allocation group of the infants. Parents, doctors and

nurses were unaware of the treatment allocation during the study. The two groups were similar with regard to age, sex, duration of symptoms prior to hospital admission, fever, respiratory distress, clinical and respiratory severity score on admission, respiratory syncytial virus (RSV) positive, oxygen saturation and biochemical results. Two-thirds of the participants were RSV-positive. The primary outcomes were duration of oxygen supplementation and length of hospital stay. Secondary outcomes were salbutamol use, ipratropium bromide use, antibiotics use, adrenaline use and incidence of pneumonia.

Another recent trial was conducted in Brazil and included 30 infants up to two years of age, previously healthy, with a clinical diagnosis of acute viral bronchiolitis and positive outcome of RSV in nasopharyngeal aspirate detected by immunofluorescence technique (Gomes 2012). Participants were allocated to receive either prolonged slow expiration (slow passive and progressive expiration from the functional residual capacity into the expiratory reserve volume) and rhinopharyngeal retrograde clearance (forced inspiratory manoeuvre through the nose) (n = 10) or vibrations, expiratory compression, modified postural drainage only in the lateral decubitus position and clapping (n = 10) or suction of the upper airways (n = 10). The third group was only assessed at admission, and afterwards followed the standard chest physiotherapy regimen in the hospital; this group was not considered in this review. The two groups were similar with regard to age, sex, weight and clinical score. The primary outcomes were Wang's clinical score. Secondary outcomes were retractions and SpO_2. Assessors were blinded to the treatment groups.

A trial conducted in Belgium recruited 20 infants with acute RSV bronchiolitis, with a mean age of 4.19 months (Postiaux 2011). Infants were randomised to inhalation of a 3% hypertonic saline solution and salbutamol (n = 8) or to a physiotherapy protocol combining prolonged slow expiration technique and coughing provoked after the same inhalation of saline solution and salbutamol (n = 12). The two groups were similar with regards to age, sex and Wang clinical severity score on admission (Wang 1992). The trial main outcome is Wang's clinical score, which assigns a value between zero and three to each of the four variables: respiratory rate, wheezing, retractions and general condition. The maximum Wang score is 12 and a higher Wang score indicates a worse condition. Secondary outcomes were SpO_2 and heart rate (HR). All outcomes were assessed before the session, at the end of the session and two hours afterwards. Both of the paediatric evaluators were blinded to the applied treatment and goals. Physiotherapists in charge of administering the treatments were instructed to ignore the results of each evaluation until the end of the study. The participants' parents were unaware of the group in which their child was included. In both groups the periods of time spent in the room were identical, so outside observers were blinded to the applied treatment.

The largest trial was conducted in France, randomising 496 hospitalised infants with a first acute bronchiolitis episode between the ages of 15 days and 24 months (mean age two months, range 1.3 to 3.9 months) (Gajdos 2010). Infants had to present with at least one of the following on admission: toxic aspect; history of apnoea or cyanosis; respiratory rate > 60/minute, pulse oxymetry < 95%, alimentary intake < two-thirds of the daily food requirements. The control group presented with a higher proportion of RSV-positive patients than the intervention group (76.4% versus 73.3%), as well as the proportion of cases of lung atelectasis diagnosis on chest X-ray (12.9% versus 7.6%). Patients were allocated to receive either the passive forced exhalation technique with assisted cough (n = 246) or nasal suction (n = 250). All interventions were administered three times a day, with the physiotherapist staying alone with the infant in a room with a covered window pane. The primary outcome was time to recovery, defined as eight hours without oxygen supplementation associated with minimal or no chest recession and ingesting more than two-thirds of the daily food requirements. Survival analyses of time to recovery were adjusted for prognostic baseline covariates (personal eczema or history of atopy, age in months, hypoxaemia at randomisation, need for intravenous (IV) fluids at randomisation, atelectasis at randomisation, duration of symptoms, use of mucolytic before randomisation or RSV infection). The therapists were not involved in the evaluation of time to recovery. Secondary outcomes were intensive care unit admissions, artificial ventilation, antibiotic treatment, description of side effects during procedures and parental perception of comfort.

Rochat 2010 analysed 99 infants admitted to a Swiss hospital with bronchiolitis during two consecutive RSV seasons (2005 to 2006 and 2006 to 2007). Participants had a mean age of 3.9 months. All infants received standard care including oxygen therapy and rhinopharyngeal suctioning. Infants were either randomised to additionally receive a physiotherapy protocol combining prolonged slow expiratory technique, slow accelerated expiratory technique and coughing provoked (n = 51), or randomised to no physiotherapy (n = 53). The two groups were similar with regard to age, sex, clinical and respiratory severity score on admission, proportion who were RSV Enzyme-Linked ImmunoSorbent Assay (ELISA) positive (overall proportion 75%) and history of eczema (overall proportion 7%). The trial assessed time to clinical stability, clinical and respiratory scores, respiratory rate, pulse oximetry oxygen saturation (SpO_2) and complications such as transfer to the intensive care unit.

De Córdoba 2008 randomised 24 hospitalised infants below two years of age, in Brazil. Nineteen of those infants were analysed, of whom five were allocated to vibration and postural drainage, eight to percussion and postural drainage and six to the control group (bronchial aspiration). Infants had to present clinical and laboratory signs of acute viral bronchiolitis and bronchial hypersecretion (pulmonary auscultation). There was no information on percentage of RSV patients or patients with lung collapse/consolidation at baseline or during the trial. The three groups were similar with regard to age, sex, oxygen saturation and cardiac and respiratory

frequency on admission. Mean age was 93 days, 131 days and 125 days in each intervention group. The main outcomes were: saturation of oxygen pulse, cardiac frequency, respiratory frequency, Silverman-Anderson Score of respiratory discomfort (Silverman 1956), and amount of inhaled secretions. Outcomes were assessed immediately after treatment and 15 minutes later. Results were expressed as means and standard deviations (SDs).

In the Bohe 2004 study conducted in Argentina, 16 infants were randomly allocated to the physiotherapy group and 16 to the control group. Patients were included if they had a clinical diagnosis of acute bronchiolitis defined by an acute upper respiratory infection plus fever, tachypnoea or increase of respiratory effort. The mean age of the participants was 2.8 months and 78.1% of participants were positive for RSV. There was no information on the percentage of patients with atelectasis/consolidation at baseline or during the trial. The intervention was percussion, postural drainage, vibration and nasopharyngeal aspiration twice a day. The control group received only nasopharyngeal aspiration. The endpoints were length of hospital stay and a severity score constructed out of five clinical variables: respiratory rate, heart rate, lung auscultation and accessory muscle use.

A trial conducted in the UK randomly allocated 50 infants to control (n = 24) or treatment (n = 26) groups; their mean age was 2.8 months (range 0.4 to 7.6 months). Infants had to present clinical diagnoses of acute bronchiolitis and severe respiratory distress requiring nasogastric tube feeding or intravenous fluids (Nicholas 1999). The intervention and control groups presented similar proportions of RSV-positive patients (79% versus 85%). There was no information on atelectasis/consolidation at study entry or afterwards. The physiotherapy protocol established manual techniques of percussion and vibrations performed in postural drainage positions with possible modifications as required in relation to infant tolerance. The main outcomes were clinical status and length of hospital stay. Secondary endpoints were oxygen requirements and change in oxygen saturation levels after physiotherapy; these outcomes were measured only in the intervention arm. Results were expressed using means but standard deviations (SDs) were not reported. The trial author could not provide clarification as she was no longer in possession of the complete database.

The oldest trial was conducted in the UK and analysed 90 infants with a mean age of 4.6 months (range 0 to 15 months) presenting a clinical diagnosis of acute viral bronchiolitis (Webb 1985). Forty-four infants were allocated to physiotherapy and 46 infants to the control group. The two groups were similar with regards to age, sex, severity score on admission, proportion who were RSV-positive (overall proportion 69%), proportion with a first-degree family history of atopy (overall proportion 36%), those participants with smokers in their household (overall proportion 66%) and participants with some degree of atelectasis/consolidation on chest X-rays (overall proportion 24.5%). The intervention tested consisted of "chest percussion with a cupped hand for three minutes in each of five postural drainage positions followed by assisted coughing" or "gentle oropharyngeal suction performed twice each day while in the hospital". Three medical doctors made clinical assessments of the severity of the illness at a fixed time every day. A score of zero to three was allocated for each of 10 clinical signs: heart rate, respiratory rate, hyperinflation, use of accessory muscles, recession, rhinitis, wheeze, cough, crepitations and rhonchi, to give a total severity clinical score of a maximum of 30 points. At hospital discharge, parents were asked to maintain a symptom record diary and children were reviewed in outpatient clinics after two weeks. The main outcomes were: clinical score on admission, every day and after five days, length of hospital stay and total length of illness. Results were expressed as medians and ranges. The trial author was unable to provide the mean and SD of each parameter because the raw data were no longer available.

Unpublished trials

In the Lopez Galbany 2004 pilot study conducted in Spain, 30 infants with RSV-positive bronchiolitis were randomly allocated to receive physiotherapy with slow expiratory technique (n = 15) or no intervention (n = 15). Outcomes assessed were the Bierman Pierson modified severity clinical score and length of hospital stay. The Aviram 1992 study was a randomised controlled intervention study conducted in Israel, which included 50 infants aged one to five months, paired by age and clinical severity score. Participants were allocated to receive chest physiotherapy or not, in addition to salbutamol inhalations every six hours. Although there is no information on the physiotherapy technique applied, it is assumed to be based on vibration and percussion. Outcomes assessed were length of stay in hospital, improvement in clinical score and changes in SaO_2. Clinical scoring was performed in a blinded manner.

Excluded studies

See Characteristics of excluded studies table.

We excluded six studies. One study was a single-blind randomised clinical trial including infants under two years of age with moderate acute wheezing episodes attending an outpatient clinic (Castro 2014). The study randomised 48 participants to receive salbutamol with or without chest physiotherapy using slow and long expiratory flow and assisted cough techniques. After inclusion of the participant by a family physician, those infants in the chest physiotherapy group received physiotherapy for one hour. Afterwards the patient was assessed by the including family physician, blinded to intervention status, for re-evaluation of his or her clinical status, clinical score and SpO_2 level. If the patient met the criteria of improvement, he or she was discharged. Otherwise, the participant received a second hour of treatment, according to his or her original randomised group. After the second hour, the participant was assessed again by the original family physician and referred to the hospital for admission if the criteria of improvement based on the clinical score was still not achieved. The study endpoints were

clinical score, SpO_2, number of hospital admissions and parents satisfaction.

Three other excluded studies were uncontrolled intervention studies (Bernard-Narbone 2003; Postiaux 2004; Quitell 1988), and the last two were non-randomised comparative trials (Belcastro 1984; Pupin 2009).

The two comparative trials' details are as follows:

Belcastro 1984 was a pilot study with 12 patients that compared:

1. osteopathic manipulative treatment to postural drainage in a non-randomised fashion (first three patients received osteopathy and the rest postural drainage); and
2. bronchodilators to placebo in a randomised, double-blind fashion.

The endpoints were number of hospital days and mean daily respiratory rates.

Pupin 2009 was a comparative controlled intervention study which included 81 infants with clinically and radiologically diagnosed acute viral bronchiolitis. Participants were non-randomly allocated to receive expiratory flow increase technique (EFIT), vibration plus postural drainage or a control procedure (no respiratory therapy, only manual contact of the physical therapist on the thorax). Each procedure consisted of a single therapeutic session performed in the morning for 10 minutes. Heart rate, respiratory rate and SpO_2 were assessed before the procedure and at 10, 30 and 60 minutes after it. The authors conclude that "In terms of overall improvement of cardiorespiratory parameters, neither the EFIT nor vibration/PD provided any benefit to infants with acute viral bronchiolitis. However, over time, respiratory physical therapy seems to contribute to decreasing the respiratory rate in these patients".

Risk of bias in included studies

The overall risk of bias for the comparison of vibration and percussion techniques is moderate to high, because of the uncertainties and limitations associated with the assessment of risk of bias in the five trials in this comparison (Figure 2; Figure 3).

Figure 2. 'Risk of bias' graph: review authors' judgements about each methodological quality item presented as percentages across all included studies

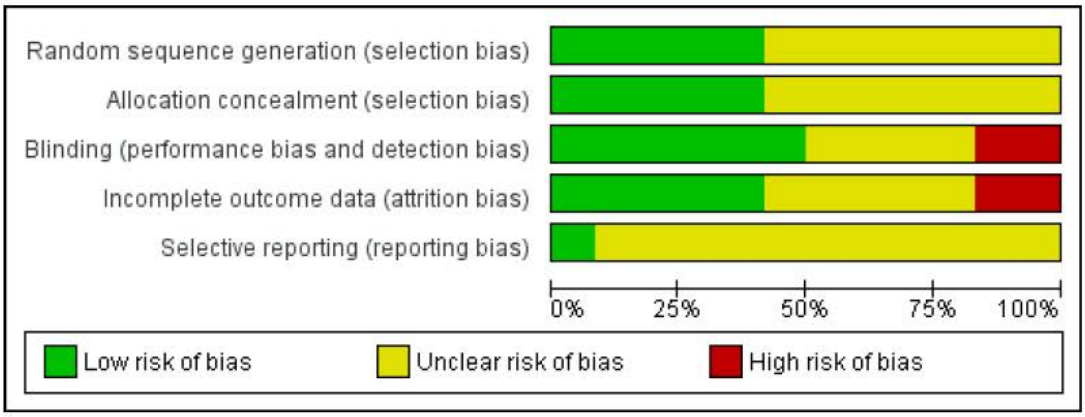
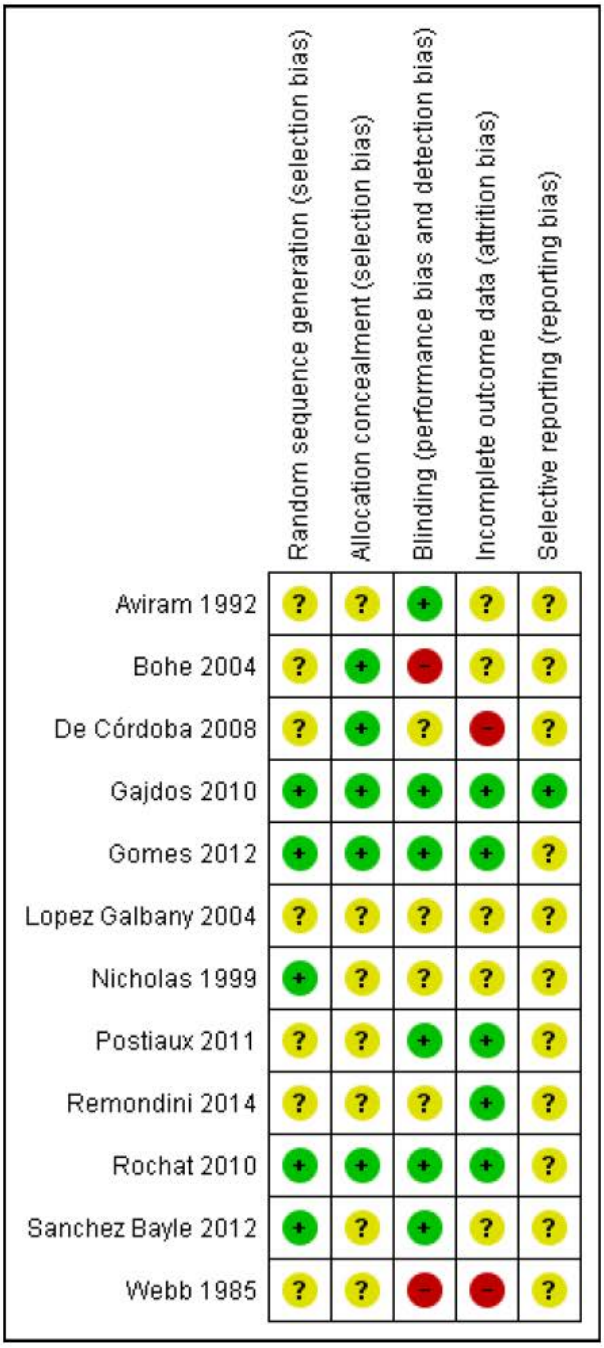


Figure 3. 'Risk of bias' summary: review authors' judgements about each methodological quality item for each included study



The overall risk of bias for the comparison of passive expiratory techniques is uncertain. However, the two trials comparing forced expiration techniques are at low risk of bias (Gajdos 2010; Rochat 2010). The comparison of slow flow techniques has one low risk of bias trial (Gomes 2012), and four trials of uncertain risk of bias (Lopez Galbany 2004; Postiaux 2011; Remondini 2014; Sanchez Bayle 2012).

Allocation

Scant information was provided regarding randomisation methods and allocation concealment. Five trials described adequate sequence generation procedures (Gajdos 2010; Gomes 2012; Nicholas 1999; Rochat 2010; Sanchez Bayle 2012). Five trials either described procedures to conceal allocation (De Córdoba 2008; Gajdos 2010; Gomes 2012; Rochat 2010), or claimed to have concealed allocation (Bohe 2004).

Blinding

Masking of outcome assessment was most likely absent in all but two of the included trials. Five trials implemented rigorous procedures to mask outcome assessments (Gajdos 2010; Gomes 2012; Postiaux 2011; Rochat 2010; Sanchez Bayle 2012), but the other trials were admittedly open (Bohe 2004; Rochat 2010; Webb 1985), or most likely so (Aviram 1992; De Córdoba 2008; Lopez Galbany 2004; Nicholas 1999). Even though some outcomes were objective and not subject to bias (oxygen saturation, heart rate), other outcomes depended on observation and could be more vulnerable (clinical scores and respiratory discomfort questionnaire).

Incomplete outcome data

A single trial had a large sample size and had an adequate description of attrition of participants, as well as a description of how they were handled (ITT analysis) (Gajdos 2010). Another trial had a large sample and an adequate description of attrition of participants (Rochat 2010). The rest of the included trials were small and the attrition of participants was either null (Gomes 2012; Postiaux 2011), or low and unclearly dealt with (Bohe 2004; De Córdoba 2008; Nicholas 1999; Sanchez Bayle 2012; Webb 1985).

Selective reporting

A single trial had a low risk of selective reporting bias, as shown by comparing the trial protocol with the published paper (Gajdos 2010). Assessment of selective reporting bias is not possible for the rest of the trials due to the scarcity of available data.

Effects of interventions

See: **Summary of findings for the main comparison** Forced expiration compared with standard care for acute bronchiolitis; **Summary of findings 2** Slow passive expiration compared with standard care for acute bronchiolitis

Although the included trials provided some data on change in the severity status of bronchiolitis (using clinical scores) and length of hospital stay, due to clinical and statistical considerations we were unable to pool the data. First of all, the clinical scores assessed in the included trials were heterogeneous:

1. the studies used different scores, although admittedly based on similar recordings;
2. the timing of the assessments was quite variable (15 minutes after the intervention (De Córdoba 2008), two hours after the intervention (Postiaux 2011), at hospital discharge (Bohe 2004), on the fifth day (Lopez Galbany 2004); and
3. not all trials provided data for this outcome, in particular the largest, most valid trial (Gajdos 2010).

It seems unreliable to present a statistical analysis that only partially incorporates the available evidence, lacking the most influential trial with a sample size that doubles that of the rest of the trials. Finally, length of hospital stay is quite an asymmetric variable, often presented as medians, and the usual meta-analysis methods, based on symmetry, are not the right tools to analyse it.

Postural drainage plus percussion and vibration techniques

Primary outcomes

1. Change in the severity status of bronchiolitis

Five trials (241 analysed participants) in this comparison assessed the severity of bronchiolitis by means of clinical scores and none of them showed statistical differences between groups at day five (Aviram 1992; Bohe 2004; De Córdoba 2008; Nicholas 1999; Webb 1985).

Nicholas 1999 and Webb 1985 assessed this outcome using a common clinical score. In the Webb 1985 study there were no statistically significant differences between groups in relation to the clinical score or to the proportion who remained in hospital at day five. The clinical score was similar in both groups at baseline and on each of the first five days of assessment at the hospital. In the control group the median score on admission was 12 (range 4 to 24) in 46 participants and in the physiotherapy group the median score was 10 (range 4 to 22) in 44 participants. On the fifth day, 18 participants who remained in hospital had a median score

of five (range 1 to 11) in the control group; 11 participants in the physiotherapy group had a median score of six (range not presented in the original article). The study also assessed the length of illness, which was not significantly different between the groups (Mann-Whitney test (Mann 1947)). In the control group the median length of illness was 14 (range 4 to 27) and in the physiotherapy group the median was 13 (range 7 to 26). Nicholas 1999 expressed clinical scores using means but did not report standard deviations (SDs). There were no differences in the admission mean clinical scores (intervention group 9.1 versus control group 10.9) between groups. The trial authors reported that clinical scores did not show any statistically significant differences between groups during the five day trial. Data were provided on a graph but could not be extracted. Bohe 2004 used a different clinical severity score to the one used in the other two trials. The score at day five or the day of discharge was 3.25 (SD 1.27) in the physiotherapy group and 3.12 (SD 1.15) in the control group (mean difference (MD) 0.13, 95% confidence interval (CI) -0.71 to 0.97). The unpublished trial did not describe the clinical score used and it also failed to show differences between treatment groups (Aviram 1992).

2. Time to recovery

No trial presented data on time to recovery.

Secondary outcomes

1. Respiratory parameters

Data for respiratory parameters are available in only one of the included trials, assessed immediately after treatment and at 15 minutes (De Córdoba 2008). No significant differences were observed in oxygen saturation levels nor in respiratory frequency between the treatment groups in their 15-minute results (Kruskal Wallis test (Kruskal 1952)). The amount of aspired secretions was significantly smaller in the control group than in the intervention groups ($P = 0.02$, Kruskal Wallis test). Respiratory discomfort was assessed by means of the Silverman-Andersen Questionnaire (Silverman 1956), which significantly improved ($P < 0.05$, Friedman analysis of variance) post 15 minutes with respect to baseline in the two treatment groups but not in the control group. It is not clear from the paper whether differences across the groups were tested but it can be assumed that the lack of data means that there were not significant differences across the groups.

2. Duration of oxygen supplementation

Nicholas 1999 found that the mean number of hours with supplemental oxygen in the control group was 63 (range 2.3 hours to 128 hours) compared with 86 (range 36 hours to 148 hours) in the physiotherapy group. Differences were reported as not significant using a non-parametric test.

3. Length of hospital stay

In Bohe 2004, mean length of hospital stay was four days (SD 2) in the treatment group and 3.9 days (SD 1.3) in the control group. There were no statistically significant differences between them (MD 0.13, 95% CI -1 to 1.26). In the Nicholas 1999 study, mean length of hospital stay was 6.6 days (range 2.3 days to 11.5 days) in the control group and 6.7 days (range 3 days to 9.5 days) in the physiotherapy arm. Webb 1985 showed a median length of hospital stay of four days (range one day to 15 days) in the control group and a median of four days (range two days to 11 days) in the physiotherapy group.

4. Use of bronchodilators and steroids

No trial presented data on use of bronchodilators and steroids.

5. Parents' impression of physiotherapy benefit

No trial presented data on parents' impression of the benefit of physiotherapy in this comparison.

6. Adverse events

In the Bohe 2004 study one case of atelectasis was reported in the control arm. The participant was withdrawn from the trial and assigned to receive chest physiotherapy.

Passive expiratory techniques - forced passive expiratory techniques

Primary outcomes

A summary of results is presented in Summary of findings for the main comparison.

1. Change in the severity status of bronchiolitis

One trial (103 participants) assessed severity of bronchiolitis through a clinical score assessing feeding, vomiting and sleep (Rochat 2010). No differences were observed in changes in the clinical score (mixed linear models $P = 0.37$).

One trial (29 participants) compared the addition of forced passive expiratory techniques to postural drainage. The trial assessed severity of bronchiolitis using respiratory distress assessment instrument (RDAI) (Remondini 2014). They observed significant differences immediately after forced passive expiratory physiotherapy + postural drainage (10 and 60 minutes post intervention; $P < 0.001$). However, when compared to conventional physiotherapy (postural drainage + manual percussion or tapping), no differences were found.

2. Time to recovery

Three trials (628 participants) in this comparison assessed resolution of bronchiolitis in terms of time to recovery (Gajdos 2010; Remondini 2014), and time to clinical stability (Rochat 2010). Overall, there were no significant differences between groups in any of these trials.

In Gajdos 2010, the physiotherapy intervention (forced expiratory technique with assisted cough) had no significant effect on time to recovery as assessed by the logrank test and a Cox regression. The median time to recovery was 2.31 days (95% CI 1.97 to 2.73) for the control group and 2.02 days (95% CI 1.96 to 2.34) for the physiotherapy group (hazard ratio (HR) 1.09, 95% CI 0.91 to 1.31, P = 0.33). In Rochat 2010, time to clinical stability, assessed as a primary outcome, was similar for increased exhalation technique (IET) and placebo (2.9 ± 2.1 versus 3.2 ± 2.8 days, logrank test P =0.45).

For both primary outcomes, the quality of the evidence using GRADE was high.

Secondary outcomes

One trial comparing the addition of forced passive expiratory physiotherapy to postural drainage, Remondini 2014, did not observed differences in SpO_2 during and after the intervention. There were no data on secondary outcomes such as duration of oxygen supplementation, length of hospital stay and use of bronchodilators and steroids.

1. Respiratory parameters

In Rochat 2010, the rate of improvement of a respiratory score, defined as secondary outcome, only showed a slightly faster improvement of the respiratory score in the prolonged slow expiration (PSE) technique group when including stethacoustic properties (mixed linear model P =0.044). No differences were observed in oxygen saturation (SpO_2) (mixed linear models P = 0.85) or respiratory rates (mixed linear models P = 0.24).

2. Duration of oxygen supplementation

No trial presented data on duration of oxygen supplementation.

3. Length of hospital stay

No trial presented data on length of hospital stay.

4. Use of bronchodilators and steroids

No trial presented data on use of bronchodilators and steroids.

5. Parents' impression of physiotherapy benefit

Two trials provided data on the parents' impression on the benefit of chest physiotherapy.

Remondini 2014 presented data on the parents' impression on the benefit of physiotherapy compared to conventional physiotherapy. Parents in both groups reported satisfaction related to improvements of breathing, feeding and nasal congestion, but no difference was observed between the intervention groups. Gajdos 2010 reported they did not observe any significant difference in the way the parents rated the influence of physiotherapy on respiratory status (risk ratio (RR) 0.99, 95% CI 0.90 to 1.08, P = 0.89) or comfort (RR 0.99, 95% CI 0.94 to 1.05, P = 0.84).

6. Adverse events

In the only trial in the review that specifically monitored adverse events, there were no significant differences between groups in the proportion of children who experienced one episode of bradycardia with desaturation (risk ratio (RR) 1.0, 95% CI 0.2 to 5.0, P = 1.00) or without desaturation (RR 3.6, 95% CI 0.7 to 16.9, P = 0.10) (Gajdos 2010). Conversely, in the IET physiotherapy group there were a higher proportion of children who had transient respiratory destabilisation (RR 5.4, 95% CI 1.6 to 18.4, P = 0.002) or vomited during the procedure (RR 10.2, 95% CI 1.3 to 78.8, P = 0.005).

Regarding the physiotherapy technique, in Rochat's study, complications were defined as concomitant bacterial infection or transfer to the intensive care unit due to respiratory fatigue (Rochat 2010). The trial authors state that complications related to bronchiolitis severity were rare and occurred more frequently in the control group (n = 19; 12 in the control group, seven in the intervention group), albeit not significantly (P = 0.21). Also, they state that no direct complications of physiotherapy, such as respiratory deterioration, occurred.

Remondini 2014 did not report any adverse events.

For adverse events, the quality of the evidence using GRADE was high.

Passive expiratory techniques - slow passive expiratory techniques

Primary outcomes

A summary or results is presented in the Summary of findings 2.

1. Change in the severity status of bronchiolitis

Three trials analysing 286 participants assessed severity of bronchiolitis through clinical scores (Gomes 2012; Lopez Galbany 2004; Postiaux 2011). Overall, there were no significant differences between groups in any of these trials. Furthermore, the quality of the evidence for this outcome using GRADE was low.

In Lopez Galbany 2004 no significant differences were observed between groups in change from baseline values (P = 0.175). Mean values for a modified version of the Bierman Pierson score (Bierman 1974; Tal 1983) at five days were 2.46 for the physiotherapy group and 2.79 for the control group.

In Postiaux 2011, a significant small improvement in the Wang clinical score was observed immediately after the intervention in the group receiving slow flow physiotherapy and salbutamol (3.6 versus 5.1, ANOVA P = 0.02), which disappeared two hours later (4.6 versus 3.7, ANOVA P = 0.21). The authors report a "day-to-day baseline improvement in Wang score significantly better [in the CPT group] than that in the control group" but this conclusion is based on within-group tests on a diminishing sample due to discharge of patients ("After 5 days, 6 of the 8 control group patients had been discharged, whereas all 12 of the new-method-CPT group had been discharged").

One trial (30 participants) compared severity of clinical scores between both physiotherapy techniques (Gomes 2012). The authors only applied statistical tests to within-groups comparisons pre versus post. They found significant within-group differences in clinical score values and retractions assessed at 48 hours for both physiotherapy regimens, and significant differences in clinical score and oxygen saturation assessed at 72 hours for the slow flow physiotherapy. Although not statistically tested, endpoint values at 48 and 72 hours for the clinical score and all its sub-scales appear to be equal between both physiotherapy groups.

2. Time to recovery

No trial presented data on time to recovery.

Secondary outcomes

1. Respiratory parameters

No data were presented for this outcome.

2. Duration of oxygen supplementation

One trial (236 participants) compared the average hours with oxygen supplementation in the physiotherapy and control groups, which showed no statistically significant differences (Sanchez Bayle 2012). Mean hours of oxygen therapy needed were 49.98 ± 37.10 in the physiotherapy group and 53.53 ± 38.87 in the control group.

3. Length of hospital stay

This outcome was assessed in three trials (286 participants), and none of them detected statistically significant differences between the length of hospital stay of the physiotherapy and control groups. Mean length of stay in Sanchez Bayle 2012 was 4.56 ± 2.07 days in the physiotherapy group and 4.54 ± 1.72 days in the control group. Mean length of stay in Lopez Galbany 2004 was 6.18 days in the physiotherapy group and 5.88 in the control group. Average hospital stay in Postiaux 2011 was 5.3 ± 1.8 days in the physiotherapy group and 6.3 ± 2 days in the control group (Mann-Whitney U test P = 0.25).

4. Use of bronchodilators and steroids

One trial including 236 participants recorded the percentages of participants that received salbutamol, ipratropium bromide or antibiotics, which showed no statistical differences between the intervention and control groups (Sanchez Bayle 2012).

5. Parents' impression of physiotherapy benefit

No trial presented data on parents' impression of physiotherapy benefit.

6. Adverse events

Two studies explicitly stated that no adverse events were observed but there is no definition on the events considered (Postiaux 2011; Sanchez Bayle 2012).

The quality of the evidence for adverse events using GRADE was very low.

Subgroup analyses

The subgroup analysis by participant severity was confused by interaction with techniques. Four trials included participants with severe bronchiolitis, corresponding to the comparison of vibration and percussion (Nicholas 1999), slow passive expiration (Sanchez Bayle 2012), and forced expiration (Gajdos 2010; Rochat 2010). Five trials included moderate cases of bronchiolitis, corresponding to the comparison of slow passive expiration (Gomes 2012; Lopez Galbany 2004; Postiaux 2011), and vibration and percussion (Bohe 2004; Webb 1985). One trial of vibration and percussion techniques included mild cases of bronchiolitis (De Córdoba 2008). While no formal meta-analysis or test of subgroups could be conducted due to lack of data, it became clear that the evidence for the slow flow chest physiotherapy techniques was unevenly distributed, with slow flow techniques studied in less severe participants than forced expiratory techniques.

It was not possible to conduct the subgroup analysis by setting, since all the trials included hospitalised participants.

Subgroup analysis performed on the included trials

Sanchez Bayle 2012 conducted subgroup analyses of the effect of physiotherapy on length of hospital stay and duration of oxygen supplementation by subgroups of respiratory syncytial virus (RSV) status. They found statistical differences in the number of hours with oxygen supplementation in the subgroup of RSV-positive

participants that received physiotherapy compared to those RSV-positive participants in the control group (mean hours 48.80 ± 37.70 versus 58.68 ± 36.78; P = 0.042, Mann-Whitney test). There were no other statistical differences.

Gajdos 2010 performed subgroup analyses by personal eczema or history of atopy, RSV-positive infection and hypoxaemia at randomisation. There was no statistically significant quantitative interaction on time to recovery between any of these subgroups.

Nicholas 1999 performed a subgroup analysis between participants who had more than 10 points on the baseline clinical score and those with a baseline clinical score below 9.5. There were no differences between the physiotherapy and control groups in this subgroup analysis.

Webb 1985 reports that there were no differences between treatments in daily scores or length of illness in the subset of participants with some degree of collapse/consolidation on chest X-rays.

ADDITIONAL SUMMARY OF FINDINGS *[Explanation]*

Slow passive expiration compared with no physiotherapy for acute bronchiolitis

Patient or population: paediatric patients between 0 and 24 months old with acute bronchiolitis
Settings: hospital
Intervention: slow passive expiration
Comparison: standard care

Outcomes	Relative effect (95% CI)	No. of participants (studies)	Quality of the evidence (GRADE)	Comments
Change in the severity status of bronchiolitis Wang score and Bierman Pearson score (follow-up ranging from 2.5 hours to discharge)	2 studies did not find changes. 1 study found a transient small effect	286 (3 trials)	⊕⊕◯◯ **low**[1]	Participants with moderate bronchiolitis (Gomes 2012; Lopez Galbany 2004; Postiaux 2011)
Adverse events (follow-up)	Studies reported that no adverse events were observed	256 (2 trials)	⊕◯◯◯ **very low**[2]	Participants with moderate and severe bronchiolitis (Postiaux 2011; Sanchez Bayle 2012)

*The basis for the **assumed risk** (e.g. the median control group risk across studies) is provided in footnotes. The **corresponding risk** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).
CI: confidence interval; **RR:** risk ratio

GRADE Working Group grades of evidence
High quality: Further research is very unlikely to change our confidence in the estimate of effect.
Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
Very low quality: We are very uncertain about the estimate.

[1]Downgraded quality due to uncertain risk of bias and imprecision of estimates.
[2]Downgraded quality due to uncertain risk of bias, imprecision of estimates and indirectness of assessments because the trials were unclear on the adverse effects assessment.

DISCUSSION

Summary of main results

This review included 12 trials and 1249 participants exploring the efficacy of three physiotherapy modalities (vibration and percussion, slow passive expiratory techniques and forced passive expiratory techniques), compared to no intervention in hospitalised infants with acute bronchiolitis not on mechanical ventilation. None of the included trials showed a significant benefit of either chest physiotherapy techniques in change of disease severity, respiratory parameters, length of hospital stay or oxygen requirements in this population. One trial found transient immediate respiratory score improvements in moderate bronchiolitis patients that received slow expiratory techniques. The included trials did not report severe adverse events. In Gajdos 2010, a significant risk of vomiting (risk ratio (RR) > 10) and respiratory instability (RR > 5) was reported in children receiving physiotherapy with passive increased exhalation technique and assisted cough, while no complications related to physiotherapy and few complications related to bronchiolitis severity were observed in trials applying prolonged slow expiration techniques (Postiaux 2011; Rochat 2010).

Quality of the evidence

The quality of the evidence in the review is variable depending on the comparisons considered. While there is high quality evidence for forced expiration techniques, the quality of evidence is low for the techniques based on slow passive expiration and very low for vibration and percussion. The assessments of quality of evidence have relied heavily on the risk of bias of the trials and the imprecision of their results, mainly due to small sample sizes. For adverse events, there were concerns regarding indirectness of assessments for trials that were not clear enough on the adverse events assessment procedure.

The high quality evidence for forced expiration techniques in severe patients stems from the overall low risk of bias of the trials considered, the large number of patients considered and the consistency of the trials' results. Although the three trials assessed recovery with two different measures (time to recovery and time to clinical stability), the results were homogeneous and led to similar conclusions of no effect of the physiotherapy techniques. One of the trials had a very large sample size and good methodological quality, and was designed to detect a 20% decrease in time to recovery, assessed eight-hourly (Gajdos 2010). Since this adequately powered trial was negative, our confidence in the lack of effect observed with this physiotherapy techniques is high. Also, the negative results are consistent in all the assessed outcomes, including respiratory parameters, which are more sensitive to the treatment and nevertheless do not show a statistical benefit. There are also negative results in length of hospital stay, a less relevant outcome since it is a crude measure of length of illness and it is sensitive to unrelated factors (i.e. hospital discharge practices, day of the week, parental wishes, etc).

The low quality of evidence for the slow flow techniques in moderate/severe patients stems from their uncertain risk of bias, moderate sample sizes and methodological limitations in adverse effects assessment. The included trials used different measures of clinical severity and some of them presented incomplete data. Although most data on clinical efficacy were negative overall, a transient effect was observed in one trial, leading to concerns of potential inconsistency in results and potential lack of power. The largest trial in the comparison and second largest trial in the review did not perform an a priori sample size estimation and thus we cannot assess the power of the trial or the potential lack of power of the conclusions (Sanchez Bayle 2012). The quality of evidence on the safety of the slow passive expiration techniques stems from the doubts regarding how was safety assessed in the trials. The safety issues observed in the forced expiratory techniques are related to the intrinsic characteristics of forcing expiration and it could be argued that these issues would be minor or non-existent in the slow passive expiration procedures due to their gentler nature.

The very low quality of evidence for the vibration and percussion techniques stems from their high risk of bias and small sample sizes. However, the consistency between trials in showing a lack of effect and the external reports on safety of the procedures, give strength to a negative conclusion (Beeby 1998; Chalumeau 2002; Harding 1998; Knight 2001).

A methodological issue in the trials was the implementation of a valid placebo. Since all but one of the trials had a non-intervention group, the researchers would have been expected to establish an outcome assessment procedure that prevented bias. Again, this was effectively and imaginatively established in the Gajdos 2010, Postiaux 2011 and Sanchez Bayle 2012 trials. Gajdos and Sanchez Bayle compared chest physiotherapy with nasal suctioning or postural changes, respectively. Postiaux administered in both groups an aerosol composed of albuterol (3 mL) and hypertonic saline (3% NaCl) and added to the intervention group the slow passive expiration techniques. However, none of these alternatives were shown to have an impact on the overall trial results as this lack of placebo alternative will usually over-estimate the results, favouring the intervention.

Finally, it is important to consider that a limitation of the majority of the studies was that they did not analyse the effectiveness of the techniques in terms of duration of oxygen supplementation, time to recovery or other treatments used, such as bronchodilators and corticosteroids. Due to their importance in terms of disease improvement, it would be important to take these variables into account in future research,

Potential biases in the review process

To avoid biases in the review process, we have applied robust methods for searching, study selection, data collection and 'Risk of bias' assessment. To guarantee the comprehensiveness of the search, we sought both published and unpublished trials and contacted trial authors when possible to gather additional information about unpublished trials. Although pooling of data was not possible, we have considered its potential impact and performed a careful assessment of individual trials. In addition, we have performed a rigorous 'Risk of bias' assessment for the included trials.

Agreements and disagreements with other studies or reviews

The first publication of this review in 2005, Perrotta 2005, prompted the recommendation that chest physiotherapy based on vibration and percussion not be applied routinely in hospital settings (AAP 2006; BGT 2005; SIGN 2006). During recent years, a few systematic reviews have been published on this topic based on the same evidence and reaching similar conclusions to ours (Bourke 2010; González 2010b; Schechter 2007; Wainwright 2010). Also, in France, due to Cochrane evidence, two studies analysed the use of forced expiratory technique (AFE in French). They observed a decrease in chest physiotherapy prescription (Branchereau 2013), and a recommendation to not systematically prescribe chest physiotherapy for ambulatory patients (Verstraete 2014). As a consequence, this updated review includes the most recent randomised controlled trials (RCTs) and remains the main source of evidence on chest physiotherapy for acute bronchiolitis.

AUTHORS' CONCLUSIONS

Implications for practice

Conventional chest physical therapy (postural drainage plus percussion and vibration techniques) has not been shown to improve the severity of bronchiolitis and has been associated with adverse events. For these reasons, conventional techniques cannot be not used in clinical practice for patients with bronchiolitis.

Chest physiotherapy using passive flow-oriented expiratory techniques (which includes both forced expiratory techniques and slow flow techniques) has not been shown to improve the severity of bronchiolitis by means of clinical scores, nor to reduce time to recovery or length of stay in hospitalised patients. There is high quality evidence that forced expiratory techniques in severe patients do not improve their health status and can lead to severe adverse events. For these reasons, there are no argument in favour of routine use of these techniques as standard clinical practice for hospitalised patients with severe bronchiolitis.

However, there is a gap in the knowledge regarding the effects of slow passive expiratory techniques in patients with moderate bronchiolitis or respiratory syncytial virus (RSV)-positive disease. There is low quality evidence from individual trials that slow passive expiratory techniques could have a short-lived effect in reducing respiratory scores in patients presenting with moderate bronchiolitis and in reducing the need for oxygen supplementation in RSV-positive patients with severe bronchiolitis. The findings of the review are that there is low quality evidence that slow flow techniques could induce temporary relief in some children, and for this reason we conclude that, under clinician judgement, these techniques could be considered in specific situations, to improve respiratory performance.

Implications for research

Based on the review results, it seems clear that conventional and forced expiratory techniques will not change the course of the disease in hospitalised patients with severe disease. Therefore, further studies using these techniques in this population should not be a research priority.

However, there is uncertainty about the role of slow passive expiratory physiotherapy during a bronchiolitis episode, and the clinical relevance of transient short-term relief for patients who are RSV-positive should be discussed and studied. Other areas for further research are the effect of slow flow physiotherapy techniques combined or not with salbutamol or hypertonic saline, as well as the effect of chest physiotherapy in moderate bronchiolitis. Any research conducted on this topic should include a specific assessment of adverse events.

Finally, we recommend exploring the effects of slow passive expiratory techniques in mild to moderate non-hospitalised patients. Until now, all reviewed studies were conducted in a hospital setting and the generalisation of these results to non-hospitalised patients may not be straightforward due to differences in the health conditions and severity of disease between these two populations.

ACKNOWLEDGEMENTS

We thank Dr Gadjós, Dr Asher Tal, Ms Núria Lopez, Dr Postiaux, Dr Gomes, Dr Rochat and Dr Sanchez Bayle for their help in providing information regarding their studies. We thank the following people for commenting on this 2014 update: Eman Sobh, Martin Chalumeau, Mark Jones and Hans van der Wouden.

Marta Roqué i Fíguls is a PhD candidate at the Department of Paediatrics, Obstetrics and Gynecology and Preventive Medicine, Universitat Autònoma de Barcelona, Spain.

REFERENCES

References to studies included in this review

Aviram 1992 *{published and unpublished data}*
Aviram M, Damri A, Yekutielli C, Bearman J, Tal A. Chest physiotherapy in acute bronchiolitis [Abstract]. *European Respiratory Journal* 1992;**5**(Suppl 15):229–30. CENTRAL: CN–00492981]

Bohe 2004 *{published data only}*
Bohe L, Ferrero ME, Cuestas E, Polliotto L, Genoff M. Indications of conventional chest physiotherapy in acute bronchiolitis. *Medicina de Buenos Aires* 2004;**64**(3): 198–200.

De Córdoba 2008 *{published data only}*
De Córdoba F, Rodrigues M, Luque A, Cadrobbi C, Faria R, Solé D. Fisioterapia respiratória em lactentes com bronquiolite: realizar ou não?. *Mundo Saúde* 2008;**32**(2): 183–8.

Gajdos 2010 *{published data only}*
* Gajdos V, Katsahian S, Beydon N, Abadie V, de Pontual L, Larrar S, et al. Effectiveness of chest physiotherapy in infants hospitalized with acute bronchiolitis: a multicenter, randomized, controlled trial. *PLoS Medicine* 2010;**7**(9): 1–11.

Gomes 2012 *{published and unpublished data}*
Gomes ELFD, Postiaux G, Medeiros DRL, Monteiro KKDS, Sampaio LMM, Costa D. Chest physical therapy is effective in reducing the clinical score in bronchiolitis: randomized controlled trial [A fisioterapia respiratória é eficaz na redução de escore clínico na bronquiolite:ensaio controlado randomizado]. *Revista Brasileira de Fisioterapia* 2012;**16**:241–7.

Lopez Galbany 2004 *{unpublished data only}*
Lopez Galbany N. Oral presentation in local meeting. Presentation slides on file 2004 (accessed 1 January 2011).

Nicholas 1999 *{published data only}*
Nicholas KJ, Dhouieb MO, Marshal TG, Edmunds AT, Grant MB. An evaluation of chest physiotherapy in the management of acute bronchiolitis. Changing clinical practice. *Physiotherapy* 1999;**85**(12):669–74.

Postiaux 2011 *{published data only}*
Postiaux G, Louis J, Gerroldt J, Kotik A-C, Lemuhot A, Patte C. Effects of a new chest physiotherapy protocol in infant RSV bronchiolitis, a RCT. European Respiratory Society Annual Congress, Berlin, Germany, October 4-8. 2008:E1772. CENTRAL: CN–00679586]
* Postiaux G, Louis J, Labasse HC, Gerroldt J, Kotik AC, Lemuhot A, et al. Effects of an alternative chest physiotherapy regimen protocol in infants with RSV bronchiolitis. *Respiratory Care* 2011;**56**(7):989–94. DOI: 10.4187/respcare.00721; PUBMED: 21352671

Remondini 2014 *{published data only}*
Remondini R, Zamprônio dos Santos A, de Castro G, do Prado C, Ribeiro Ferreira da Silva Filho LV. Comparative analysis of the effects of twochest physical therapy interventions in patients withbronchiolitis during hospitalization period [Análise comparativa dos efeitos de duas intervenções de fisioterapia respiratória empacientes com bronquiolite durante o período de internação hospitalar]. *Einstein (Sao Paulo)* 2014;**12**(4):452–8. Empty name]">DOI: 10.1590/S1679–45082014AO3230; PubMed: 25628196]]

Rochat 2010 *{published data only}*
Rochat I, Leis P, Bouchardy M, Oberli C, Sourial H, Friedli-Burri M, et al. Chest physiotherapy in bronchiolitis: a randomised trial assessing passive expiratory manoeuvres. *Paediatric Respiratory Reviews* 2010;**11**(Suppl 1526):85–6.
* Rochat I, Leis P, Bouchardy M, Oberli C, Sourial H, Friedli-Burri M, et al. Chest physiotherapy using passive expiratory techniques does not reduce bronchiolitis severity: a randomised controlled trial. European Journal of Pediatrics 2011 Sep 17 Epub ahead of print].

Sanchez Bayle 2012 *{published and unpublished data}*
Sanchez Bayle M, Martin Martin R, Cano Fernandez J, Martínez Sánchez G, Gómez Martín J, Yep Chullen G, et al. Chest physiotherapy and bronchiolitis in the hospitalised infant. Double-blind clinical trial [Estudio de la eficacia y utilidad de la fisioterapia respiratoria en labronquiolitis aguda del lactante hospitalizado. Ensayo clínico aleatorizado y doble ciego]. *Anales de Pediatria* 2012;**77**:5–11.

Webb 1985 *{published data only}*
Webb MS, Martin JA, Cartlidge PH, Ng YK, Wright NA. Chest physiotherapy in acute bronchiolitis. *Archives of Disease in Childhood* 1985;**60**:1078–9.

References to studies excluded from this review

Belcastro 1984 *{published data only}*
Belcastro M, Backes C, Chila A. Bronchiolitis: a pilot study of osteopathic manipulative treatment, bronchodilators and other therapy. *Journal of the American Osteopathic Association* 1984;**83**(9):672–5.

Bernard-Narbone 2003 *{published data only}*
Bernard-Narbonne F, Daoud P, Castaing H, Rousset A. Effectiveness of chest physiotherapy in ventilated children with acute bronchiolitis [Efficacité de la kinésithérapie respiratorire chez des enfants intubés ventilés atteinets de bronchiolite aiguë]. *Archives de Pédiatrie* 2003;**10**:1043–7.

Castro 2014 *{published data only}*
Castro-Rodriguez JA, Sanchez I. Chest physiotherapy for acute wheezing episodes: an inappropriate interpretation of the first trial in outpatient infants. *Acta Paediatrica, International Journal of Paediatrics* 2014;**103**(17):e326–e7. DOI: doi.org/10.1111/apa.12670
* Castro-Rodriguez JA, Silva R, Tapia P, Salinas P, Tellez A, Leisewitz T, et al. Chest physiotherapy is not clinically indicated for infants receiving outpatient care for acute wheezing episodes. *Acta Pediatrica* 2014;**103**(23):518–23.

Postiaux 2004 *{unpublished data only}*
Postiaux G, Dubois R, Marchand E, Jacquy J, Mangiaracina M. Chest physiotherapy in infant bronchiolitis: a new approach - nCPT. Proceedings of the 6th International Meeting of Pediatric Neumonology, Lisboa, Portugal. 2004; Vol. Suppl:117–25.

Pupin 2009 *{published data only}*
Pupin M, Riccetto A, Ribeiro J, Baracat E. Comparison of the effects that two different respiratory physical therapy techniques have on cardiorespiratory parameters in infants with acute viral bronchiolitis. *Jornal Brasileiro De Pneumologia: Publicacao Oficial Da Sociedade Brasileira De Pneumologia E Tisilogia* 2009;**35**(9):860–7.

Quitell 1988 *{published data only}*
Quitell LM, Wolfson MR, Schidlow DV. The effectiveness of chest physical therapy in infants with bronchiolitis. *American Review of Respiratory Disease* 1988;**137**:406A.

References to ongoing studies

Bella Lisbôa 2008 *{published data only}*
Bella Lisbôa A. Chest physiotherapy effectiveness in infants with acute bronchiolitis. www.anzctr.org.au/ACTRN12608000601336.aspx 2008.

Additional references

AAP 2006
American Academy of Pediatrics. Diagnosis and management of bronchiolitis. *Pediatrics* 2006;**118**:1774-93.

AAPs 2006
Subcommittee on Diagnosis and Management of Bronchiolitis. Diagnosis and management of bronchiolitis. *Pediatrics* 2006;**118**:1774–93.

Barben 2008
Barben J, Kuehni CE, Trachsel D, Hammer J, on behalf of the Swiss Paediatric Respiratory Research Group. Management of acute bronchiolitis: can evidence based guidelines alter clinical practice?. *Thorax* 2008;**63**:1103–9. DOI: 10.1136/thx.2007.094706

Beauvois 2001
Beauvois E. Role of physiotherapy in the treatment of acute bronchiolitis in the infant [Place de la kinésithérapie dans le traitement des bronchiolites aiguës du norrisson. Conférence de consensus]. *Archives de Pédiatrie* 2001;**8** (Suppl 1):128–31.

Beeby 1998
Beeby PJ, Henderson-Smart DJ, Lacey JL, Rieger I. Short and long term neurological outcomes following neonatal chest physiotherapy. *Journal of Pediatrics and Child Health* 1998;**34**:60–2.

BGT 2005
Bronchiolitis Guideline Team. Evidence based clinical practice guideline for medical management of bronchiolitis in infants year of age or less presenting with a first time episode. Cincinnati Children's Hospital Medical Center 2005.

Bierman 1974
Bierman CW, Pierson WE. The pharmacologic management of status asthmaticus in children. *Pediatrics* 1974;**54**:245–7.

Bourke 2010
Bourke T, Shields M. Bronchiolitis. *Clinical Evidence* 2011; **4**(308):1–43.

Branchereau 2013
Branchereau E, Branger B, Launay E, Verstraete M, Vrignaud B, Levieux K, et al. Management of bronchiolitis in general practice and determinants of treatment being discordant with guidelines of the HAS [État des lieux des pratiques médicales en médecine générale en matière de bronchiolite et déterminants deprises en charge thérapeutiques discordantes par rapport aux recommandations de l'HAS]. *Archives de Pédiatrie* 2013;**20**: 1369-75.

Carroll 2008
Carroll KN, Gebretsadik T, Griffin MR, Wu P, Dupont WD, Mitchel EF, et al. The increasing burden and risk factors for bronchiolitis-related medical visits in infants enrolled in a state healthcare insurance plan. *Pediatrics* 2008;**122**(1):58–64. [PUBMED: PMC2655142]

Chalumeau 2002
Chalumeau M, Foix-L'Helias L, Scheinmann P, Zuani P, Gendrel D, Ducou-le-Pointe. Rib fractures after chest physiotherapy for bronchiolitis or pneumonia in infants. *Pediatric Radiology* 2002;**32**(9):644–7.

Chanelière 2006
Chanelière C, Moreux N, Pracros JP, Bellona G, Reixa P. Rib fractures after chest physiotherapy: a report of 2 cases [Fractures costales au cours des bronchiolites aiguës virales: à propos de 2 cas]. *Archives de Pédiatrie* 2006;**13**:1410-2.

Consensus 2001
Agence nationale d'accréditation d'évaluation en santé. Consensus conference on acute bronchiolitis in infants [Conférence de consensus sur la bronchiolite du norrisson]. *Archives de Pédiatrie* 2001;**8**(Suppl 1):11–23.

David 2010
David M, Luc-Vanuxem C, Loundou A, Bosdure E, Auquier P, Dubus JC. Assessment of the French consensus conference for acute viral bronchiolitis on outpatient management: progress between 2003 and 2008. *Archives of Pediatrics* 2010;**17**(2):125–31. [PUBMED: 19959347]

Dick 1991
Dick KJ. Investigation and evaluation of physiotherapy intervention in acute bronchiolitis of infancy. Masters thesis. Department of Physiotherapy, Queen Margaret College, Edinburgh 1991.

Dickersin 1994
Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. *BMJ* 1994;**309**:1286–9.

Farley 2014
Farley R, Spurling GKP, Eriksson L, Del Mar CB. Antibiotics for bronchiolitis in children under two years of

age. *Cochrane Database of Systematic Reviews* 2014, Issue 10. DOI: 10.1002/14651858.CD005189.pub4

Fernandes 2013
Fernandes RM, Bialy LM, Vandermeer B, Tjosvold L, Plint AC, Patel H, et al. Glucocorticoids for acute viral bronchiolitis in infants and young children. *Cochrane Database of Systematic Reviews* 2013, Issue 6. DOI: 10.1002/14651858.CD004878.pub4

Gadomski 2014
Gadomski AM, Scribani MB. Bronchodilators for bronchiolitis. *Cochrane Database of Systematic Reviews* 2014, Issue 6. DOI: 10.1002/14651858.CD001266.pub4

Girardi 2001
Girardi G, Astudillo P, Zuniga F. The IRA program in Chile: milestones and history [El programa IRA en Chile: hitos e historia]. *Revista Chilena de Pediatria* 2001;**72**(4):292–300.

González 2001
Gonzalez Caballero D, González Pérez-Yarza E. Acute bronchiolitis: basics for rational guidelines [Bonquilitis aguda: bases para un protocolo racional]. *Anales Españoles de Pediatria* 2001;**55**:355–64.

González 2010a
González J, Ochoa C, Grupo Investigador del Proyecto aBREVIADo (BRonquiolitis-Estudio de Variabilidad, Idoneidad y ADecuación). Study of variability in the management of acute bronchiolitis in Spain in relation to age of patients. National multicenter study (aBREVIADo project) [Estudio de variabilidad en el abordaje de la bronquiolitis aguda en España en relación con la edad de los pacientes]. *Anales de Pediatria (Barcelona)* 2010;**72**(1): 4–18.

González 2010b
González de Dios J, Ochoa Sangrador C. Consensus conference on acute bronchiolitis (IV): treatment of acute bronchiolitis. Review of scientific evidence [Conferencia de Consenso sobre bronquiolitis aguda (IV): tratamiento de la bronquiolitis aguda. Revisión de la evidencia científica]. *Anales de Pediatria (Barcelona)* 2010;**72**(4):285.e1–285.e42.

GRADEpro GDT 2015 [Computer program]
McMaster University (developed by Evidence Prime, Inc.). GRADEpro GDT: GRADEpro Guideline Development Tool. Available from www.gradepro.org. McMaster University (developed by Evidence Prime, Inc.), 2015.

Guyatt 2008
Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schunemann HJ. What is "quality of evidence" and why is it important to clinicians?. *BMJ* 2008;**336**(7651):995–8. [PUBMED: 18456631]

Guyatt 2011
Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. *Journal of Clinical Epidemiology* 2011;**64**(12):1294–302. [PUBMED: 21803546]

Guyatt 2011a
Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. *Journal of Clinical Epidemiology* 2011;**64**(12):1277–82. [PUBMED: 21802904]

Guyatt 2011b
Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. *Journal of Clinical Epidemiology* 2011;**64**(12):1303–10. [PUBMED: 21802903]

Guyatt 2011c
Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). *Journal of Clinical Epidemiology* 2011;**64**(4):407–15. [PUBMED: 21247734]

Guyatt 2011d
Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. *Journal of Clinical Epidemiology* 2011;**64**(12):1311–6. [PUBMED: 21802902]

Guyatt 2011e
Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. *Journal of Clinical Epidemiology* 2011;**64**(12):1283–93. [PUBMED: 21839614]

Guyatt 2011f
Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology* 2011;**64**(4):383–94. [PUBMED: 21195583]

Guyatt 2011g
Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. *Journal of Clinical Epidemiology* 2011;**64**(4):395–400. [PUBMED: 21194891]

Hall 1981
Hall CB, Douglas RG Jr. Modes of transmission of respiratory syncytial virus. *Journal of Pediatrics* 1981;**99**: 100.

Halna 2005
Halna M, Leblond P, Aissi E, Dumonceaux A, Delepoulle F, El Kohen R, et al. Impact of the consensus conference on outpatient treatment of infant bronchiolitis. Three-year study in the Nord district of France. *Presse Médicale* 2005; **34**:277–81.

Harding 1998
Harding J, Miles F, Becroft D, Allen B, Knight D. Chest physiotherapy may be associated with brain damage in extremely premature infants. *Journal of Pediatrics* 1998;**132**: 440–4.

Hartling 2011
Hartling L, Bialy LM, Vandermeer B, Tjosvold L, Johnson DW, Plint AC, et al. Epinephrine for bronchiolitis. *Cochrane Database of Systematic Reviews* 2011, Issue 6. DOI: 10.1002/14651858.CD003123.pub2

Higgins 2003
Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;**327**: 557–60.

Higgins 2011
Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

Holman 2003
Holman RC, Shay DK, Curns AT, Lingahojr Anderson LJ. Risk factors for bronchiolitis associated deaths among infants in the US. *Pediatric Infectious Diseases Journal* 2003; **22**(6):483–90.

Knight 2001
Knight DB, Bevan CJ, Harding JE, Teele RL, Kuschel CA, Battin MR, et al. Chest physiotherapy and porencephalic brain lesions in very preterm infants. *Journal of Pediatrics and Child Health* 2001;**37**(6):554–8.

Kruskal 1952
Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 1952;**47**(260):583-621.

Liet 2010
Liet JM, Ducruet T, Gupta V, Cambonie G. Heliox inhalation therapy for bronchiolitis in infants. *Cochrane Database of Systematic Reviews* 2010, Issue 4. DOI: 10.1002/14651858.CD006915.pub2

Mann 1947
Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 1947;**18**(1):50–60. DOI: 10.1214/aoms/1177730491

McConnochie 1993
McConnochie KM. Bronchiolitis: what's in the name?. *American Journal of Diseases of Children (1960)* 1993;**137**: 11–3.

Moher 2009
Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *BMJ* 2009;**339**: 2535.

Panickar 2005
Panickar JR, Dodd SR, Smyth RL, Couriel JM. Trends in deaths from respiratory illness in children in England and Wales from 1968 to 2000. *Thorax* 2005;**60**:1035–8.

Ralston 2014
Ralston S, Comick A, Nichols E, Parker D, Lanter P. Effectiveness of quality improvement in hospitalization for bronchiolitis: a systematic review. *Pediatrics* 2014;**134**(3): 571–81.

RevMan 2014 [Computer program]
The Nordic Cochrane Centre, The Cochrane Collaboration. Review Manager (RevMan). Version 5.3. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2014.

Régnier 2013
Régnier SA, Huels J. Association between respiratory syncytial virus hospitalizations in infants and respiratory sequelae: systematic review and meta-analysis. *Pediatric Infectious Disease Journal* 2013;**32**(8):820–6.

Sanchez 2007
Sánchez Etxaniz J, Benito Fernández J, Mintegi Raso S. Bronquiolitis aguda: ¿por qué no se aplica lo que se publica? Barreras en la transmisión del conocimiento. *Evidencias en Pediatria* 2007;**3**:88.

Schechter 2007
Schechter MS. Airway clearance applications in infants and children. *Respiratory Care* 2007;**52**(10):1382-90; discussion 1390-1.

SIGN 2006
SIGN 91. Bronchiolitis in children. A national clinical guideline. http://www.sign.ac.uk (accessed 13 April 2011).

Sigurs 2010
Sigurs N, Aljassim F, Kjellman B, Robinson PD, Sigurbergsson F, Bjarnason R, et al. Asthma and allergy patterns over 18 years after severe RSV bronchiolitis in the first year of life. *Thorax* 2010;**65**:1045–52.

Silverman 1956
Silverman W, Anderson D. A controlled clinical trial of effects of water mist on obstructive respiratory signs, death rate and necropsy findings among premature infants. *Pediatrics* 1956;**17**(1):1–10.

Smyth 2006
Smyth RL, Openshaw PJ. Bronchiolitis. *Lancet* 2006;**368**: 312-22.

Spencer 1996
Spencer N, Logan S, Scholey S, Gentle S. Deprivation and bronchiolitis. *Archives of Disease in Childhood* 1996;**74**: 50–2.

Tal 1983
Tal A, Bavilski C, Yohai D, Bearman JE, Gorodischer R, Moses SW. Dexamethasone and salbutamol in the treatment of acute wheezing in infants. *Pediatrics* 1983;**71**:13–8.

Touzet 2007
Touzet S, Réfabert L, Letrilliart L, Ortolan B, Colin C. Impact of consensus development conference guidelines on primary care of bronchiolitis: are national guidelines being followed?. *Journal of Evaluation in Clinical Practice* 2007;**13** (4):651–6. [PUBMED: 176833310]

Umoren 2011
Umoren R, Odey F, Meremikwu MM. Steam inhalation or humidified oxygen for acute bronchiolitis in children up to three years of age. *Cochrane Database of Systematic Reviews* 2011, Issue 1. DOI: 10.1002/14651858.CD006435.pub2

Verstraete 2014
Verstraete M, Crosb P, Gouina M, Oillica H, Bihouea T, Denoualb H, et al. Update on the management of acute viral bronchiolitis: proposed guidelines of Grand Ouest University Hospitals [Prise en charge de la bronchiolite aiguë du nourrisson de moins de 1 an: actualisation et consensus

médical au sein des hôpitaux universitaires du Grand Ouest (HUGO)]. *Archives de Pédiatrie* 2014;**21**:53–62.

Videla 1998
Videla C, Carballal G, Misirlian A, Aguilar M. Acute lower respiratory infections due to respiratory syncytial virus and adenovirus among hospitalized children from Argentina. *Clinical and Diagnostic Virology* 1998;**10**:17–23.

Wainwright 2003
Wainwright C, Altamirano L, Cheney M, Cheney J, Barber S, Price D, et al. A multicenter, randomized, double-blind, controlled trial of nebulized epinephrine in infants with acute bronchiolitis. *New England Journal of Medicine* 2003; **349**:27–35.

Wainwright 2010
Wainwright C. Acute viral bronchiolitis in children - a very common condition with few therapeutic options. *Paediatric Respiratory Reviews* 2010;**11**(1):39–45.

Wang 1992
Wang EE, Milner RA, Navas L, Maj H. Observer agreement for respiratory signs and oxymetry in infants hospitalized with lower respiratory infections. *American Review of Respiratory Disease* 1992;**145**(1):106–9.

Wood 1972
Wood DW. A clinical score system for the diagnosis of respiratory failure. *American Journal of Diseases of Children* 1972;**123**:227–9.

Zhang 2011
Zhang L, Mendoza-Sassi RA, Wainwright C, Klassen TP. Nebulised hypertonic saline solution for acute bronchiolitis in infants. *Cochrane Database of Systematic Reviews* 2013, Issue 7. DOI: 10.1002/14651858.CD006458.pub3

References to other published versions of this review

Perrotta 2004
Perrotta C, Ortiz Z, Roque M, Gallo M. Chest physiotherapy for acute bronchiolitis in paediatric patients between 0 and 24 months old. *Cochrane Database of Systematic Reviews* 2004, Issue 3. DOI: 10.1002/14651858.CD004873

Perrotta 2005
Perrotta C, Ortiz Z, Roque M. Chest physiotherapy for acute bronchiolitis in paediatric patients between 0 and 24 months old. *Cochrane Database of Systematic Reviews* 2005, Issue 2. DOI: 10.1002/14651858.CD004873.pub3

Perrotta 2007
Perrotta C, Ortiz Z, Roque M. Chest physiotherapy for acute bronchiolitis in paediatric patients between 0 and 24 months old. *Cochrane Database of Systematic Reviews* 2007, Issue 1. DOI: 10.1002/14651858.CD004873.pub3

Roqué 2012
Roqué i Figuls M, Giné-Garriga M, Granados Rugeles C, Perrotta C. Chest physiotherapy for acute bronchiolitis in paediatric patients between 0 and 24 months old. *Cochrane Database of Systematic Reviews* 2012, Issue 2. DOI: 10.1002/14651858.CD004873.pub4

* *Indicates the major publication for the study*

5.4 Publicación 4: PET-CT for assessing mediastinal lymph node involvement in patients with suspected resectable non-small cell lung cancer

Schmidt-Hansen M, Baldwin DR, Hasler E, Zamora J, Abraira V, Roqué i Figuls M. PET-CT for assessing mediastinal lymph node involvement in patients with suspected resectable non-small cell lung cancer. Cochrane Database of Systematic Reviews 2014, Issue 11. Art. No.: CD009519.

FI: 6.035 (2014). Puntuación de atención Altmetric: 9

Esta publicación puede consultarse al completo y de forma libre en https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD009519.pub2/full

La RS que constituye el trabajo 4 incluyó 45 estudios, que se clasificaron en tres grupos según los criterios que se usaron para definir un resultado positivo de la prueba PET-CT: actividad observada en los ganglios mayor que la actividad en el entorno (18 estudios, número de participantes = 2823, prevalencia de ganglios N2 y N3 = 679/2328), valores máximos de captación estándar (SUVmax) >=2.5 (12 estudios, N = 1656 , prevalencia de ganglios N2 y N3 = 465/1656), y otro criterio o criterio mixto (15 estudios, N = 1616, prevalencia de ganglios N2 y N3 = 400/1616). Ninguno de los estudios informó de la ocurrencia de eventos adversos [40].

En el grupo actividad > entorno, la prueba PET-CT identificó con precisión el 77.4% (IC 95%: 65.3 a 86.1) de los participantes con NSCLC diseminado más allá de los nodos N1, y el 90.1% (IC 95%: 85.3 a 93.5) de los participantes con NSCLC no propagado más allá de los nodos N1.

Hubo evidencia de alta heterogeneidad entre estudios y falta de precisión. Los análisis de sensibilidad sugirieron que la estimación general de la sensibilidad era especialmente susceptible a los sesgos de selección; el estándar de referencia considerado; tener una definición clara de positividad de la prueba en el estudio, y, en menor medida, el sesgo asociado a la prueba índice y el sesgo de financiación comercial. El análisis de sensibilidad restringido a los estudios de bajo riesgo de sesgo resultó en estimaciones combinadas de sensibilidad más bajas que las del análisis principal que incluyó todos los estudios.

En el grupo SUVmax >= de 2.5, la prueba PET-CT identificó con precisión el 81.3% (IC 95%: 70.2 a 88.9) de los participantes con diseminación más allá de los nodos N1, y el 79.4% (IC 95%: 70 a 86.5) de los participantes sin propagación más allá de nodos N1. En este grupo, hubo evidencia de una muy alta heterogeneidad entre los estudios y una clara falta de precisión. Los análisis de sensibilidad sugirieron que las estimaciones de sensibilidad y especificidad eran marginalmente susceptibles al sesgo de flujo y temporalidad y al sesgo de financiación comercial, asociados a estimaciones ligeramente más altas de sensibilidad y especificidad.

Los resultados variaron mucho entre los estudios en cada análisis, y estuvieron muy influidos por los factores de calidad y tamaño de los estudios, el país de realización, el porcentaje de participantes con adenocarcinoma, la dosis de trazador FDG y el tipo de escáner PET-CT.

Los resultados de esta revisión muestran que la exactitud diagnóstica de la prueba PET-CT es insuficiente para tomar decisiones sobre tratamiento basadas solo en los resultados de la prueba. Por lo tanto, los hallazgos respaldan las recomendaciones del Instituto Nacional de Salud y Atención de Excelencia (NICE) sobre este tema, por las que la prueba PET-CT es solo una guía para los médicos al decidir el siguiente paso: ya sea la realización de una biopsia o, cuando los ganglios negativos son pequeños, directamente la realización de cirugía.

En esta revisión no se realizó una evaluación de la calidad de la evidencia, y no se construyeron tablas de resumen de los hallazgos.

Esta revisión se publicó originalmente en 2014 y no se ha actualizado desde su primera publicación en The Cochrane Library, debido a que la política de actualización de Cochrane, mencionada anteriormente, no se aplica a las revisiones de exactitud diagnóstica.

Cochrane Library

Cochrane Database of Systematic Reviews

PET-CT for assessing mediastinal lymph node involvement in patients with suspected resectable non-small cell lung cancer (Review)

Schmidt-Hansen M, Baldwin DR, Hasler E, Zamora J, Abraira V, Roqué i Figuls M

Schmidt-Hansen M, Baldwin DR, Hasler E, Zamora J, Abraira V, Roqué i Figuls M.
PET-CT for assessing mediastinal lymph node involvement in patients with suspected resectable non-small cell lung cancer.
Cochrane Database of Systematic Reviews 2014, Issue 11. Art. No.: CD009519.
DOI: 10.1002/14651858.CD009519.pub2.

www.cochranelibrary.com

[Diagnostic Test Accuracy Review]

PET-CT for assessing mediastinal lymph node involvement in patients with suspected resectable non-small cell lung cancer

Mia Schmidt-Hansen[1], David R Baldwin[2], Elise Hasler[1], Javier Zamora[3], Víctor Abraira[4], Marta Roqué i Figuls[5]

[1]National Guideline Alliance, Royal College of Obstetricians and Gynaecologists, London, UK. [2]Department of Respiratory Medicine, Nottingham University Hospitals, NHS Trust, Nottingham City Hospital, Nottingham, UK. [3]Clinical Biostatistics Unit, Ramon y Cajal Institute for Health Research (IRYCIS), CIBER Epidemiology and Public Health (CIBERESP), Madrid (Spain) and Queen Mary University of London, Madrid, Spain. [4]Clinical Biostatistics Unit, Ramon y Cajal Institute for Health Research (IRYCIS), CIBER Epidemiology and Public Health (CIBERESP) and Cochrane Collaborating Centre, Madrid, Spain. [5]Iberoamerican Cochrane Centre - Biomedical Research Institute Sant Pau (IIB Sant Pau), CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain

Contact address: Mia Schmidt-Hansen, National Guideline Alliance, Royal College of Obstetricians and Gynaecologists, 27 Sussex Pl, Regent's Park, London, NW1 4RG, UK. sapms@cf.ac.uk.

Editorial group: Cochrane Lung Cancer Group.
Publication status and date: Edited (no change to conclusions), published in Issue 11, 2016.
Review content assessed as up-to-date: .

Citation: Schmidt-Hansen M, Baldwin DR, Hasler E, Zamora J, Abraira V, Roqué i Figuls M. PET-CT for assessing mediastinal lymph node involvement in patients with suspected resectable non-small cell lung cancer. *Cochrane Database of Systematic Reviews* 2014, Issue 11. Art. No.: CD009519. DOI: 10.1002/14651858.CD009519.pub2.

ABSTRACT

Background

A major determinant of treatment offered to patients with non-small cell lung cancer (NSCLC) is their intrathoracic (mediastinal) nodal status. If the disease has not spread to the ipsilateral mediastinal nodes, subcarinal (N2) nodes, or both, and the patient is otherwise considered fit for surgery, resection is often the treatment of choice. Planning the optimal treatment is therefore critically dependent on accurate staging of the disease. PET-CT (positron emission tomography-computed tomography) is a non-invasive staging method of the mediastinum, which is increasingly available and used by lung cancer multidisciplinary teams. Although the non-invasive nature of PET-CT constitutes one of its major advantages, PET-CT may be suboptimal in detecting malignancy in normal-sized lymph nodes and in ruling out malignancy in patients with coexisting inflammatory or infectious diseases.

Objectives

To determine the diagnostic accuracy of integrated PET-CT for mediastinal staging of patients with suspected or confirmed NSCLC that is potentially suitable for treatment with curative intent.

Search methods

We searched the following databases up to 30 April 2013: *The Cochrane Library*, MEDLINE via OvidSP (from 1946), Embase via OvidSP (from 1974), PreMEDLINE via OvidSP, OpenGrey, ProQuest Dissertations & Theses, and the trials register www.clinicaltrials.gov. There were no language or publication status restrictions on the search. We also contacted researchers in the field, checked reference lists, and conducted citation searches (with an end-date of 9 July 2013) of relevant studies.

Selection criteria

Prospective or retrospective cross-sectional studies that assessed the diagnostic accuracy of integrated PET-CT for diagnosing N2 disease in patients with suspected resectable NSCLC. The studies must have used pathology as the reference standard and reported participants as the unit of analysis.

Data collection and analysis

Two authors independently extracted data pertaining to the study characteristics and the number of true and false positives and true and false negatives for the index test, and they independently assessed the quality of the included studies using QUADAS-2. We calculated sensitivity and specificity with 95% confidence intervals (CI) for each study and performed two main analyses based on the criteria for test positivity employed: *Activity > background* or *SUVmax ≥ 2.5* (SUVmax = maximum standardised uptake value), where we fitted a summary receiver operating characteristic (ROC) curve using a hierarchical summary ROC (HSROC) model for each subset of studies. We identified the average operating point on the SROC curve and computed the average sensitivities and specificities. We checked for heterogeneity and examined the robustness of the meta-analyses through sensitivity analyses.

Main results

We included 45 studies, and based on the criteria for PET-CT positivity, we categorised the included studies into three groups: *Activity > background* (18 studies, N = 2823, prevalence of N2 and N3 nodes = 679/2328), *SUVmax ≥ 2.5* (12 studies, N = 1656, prevalence of N2 and N3 nodes = 465/1656), and *Other/mixed* (15 studies, N = 1616, prevalence of N2 to N3 nodes = 400/1616). None of the studies reported (any) adverse events. Under-reporting generally hampered the quality assessment of the studies, and in 30/45 studies, the applicability of the study populations was of high or unclear concern.

The summary sensitivity and specificity estimates for the '*Activity > background* PET-CT positivity criterion were 77.4% (95% CI 65.3 to 86.1) and 90.1% (95% CI 85.3 to 93.5), respectively, but the accuracy estimates of these studies in ROC space showed a wide prediction region. This indicated high between-study heterogeneity and a relatively large 95% confidence region around the summary value of sensitivity and specificity, denoting a lack of precision. Sensitivity analyses suggested that the overall estimate of sensitivity was especially susceptible to selection bias; reference standard bias; clear definition of test positivity; and to a lesser extent, index test bias and commercial funding bias, with lower combined estimates of sensitivity observed for all the low 'Risk of bias' studies compared with the full analysis.

The summary sensitivity and specificity estimates for the *SUVmax ≥ 2.5* PET-CT positivity criterion were 81.3% (95% CI 70.2 to 88.9) and 79.4% (95% CI 70 to 86.5), respectively.In this group, the accuracy estimates of these studies in ROC space also showed a very wide prediction region. This indicated very high between-study heterogeneity, and there was a relatively large 95% confidence region around the summary value of sensitivity and specificity, denoting a clear lack of precision. Sensitivity analyses suggested that both overall accuracy estimates were marginally sensitive to flow and timing bias and commercial funding bias, which both lead to slightly lower estimates of sensitivity and specificity.

Heterogeneity analyses showed that the accuracy estimates were significantly influenced by country of study origin, percentage of participants with adenocarcinoma, (^{18} F)-2-fluoro-deoxy-D-glucose (FDG) dose, type of PET-CT scanner, and study size, but not by study design, consecutive recruitment, attenuation correction, year of publication, or tuberculosis incidence rate per 100,000 population.

Authors' conclusions

This review has shown that accuracy of PET-CT is insufficient to allow management based on PET-CT alone. The findings therefore support National Institute for Health and Care (formally 'clinical') Excellence (NICE) guidance on this topic, where PET-CT is used to guide clinicians in the next step: either a biopsy or where negative and nodes are small, directly to surgery. The apparent difference between the two main makes of PET-CT scanner is important and may influence the treatment decision in some circumstances. The differences in PET-CT accuracy estimates between scanner makes, NSCLC subtypes, FDG dose, and country of study origin, along with the general variability of results, suggest that all large centres should actively monitor their accuracy. This is so that they can make reliable decisions based on their own results and identify the populations in which PET-CT is of most use or potentially little value.

PLAIN LANGUAGE SUMMARY

PET-CT scanning to assess the spread of non-small cell lung cancer within the chest

In the absence of distant metastasis, treatment options for non-small cell lung cancer depend on how much the disease has spread to the different lymph nodes within the chest, that is, the stage of the disease. If the cancer has not spread beyond the nearest (N1) lymph nodes, surgery is often the treatment of choice. Other treatment options for these patients include treatment with either radiotherapy, chemotherapy, or both. Planning the optimal treatment is therefore critically dependent on accurate staging of the disease. PET-CT

scanning is a non-invasive method of establishing the spread of NSCLC within the chest and elsewhere in the body, which is increasingly available and used by lung cancer multi-disciplinary teams. Although the non-invasive nature of PET-CT constitutes one of the major advantages of the test, PET-CT may be suboptimal in detecting malignancy in normal-sized lymph nodes and in ruling out malignancy in patients with coexisting inflammatory or infectious diseases. We examined the accuracy of PET-CT scanning in establishing the spread of cancer in patients with suspected or confirmed NSCLC that is potentially suitable for surgical treatment with curative intent.

We included 45 studies, and based on the criteria for a positive PET-CT scan, we performed two main analyses. In the 18 studies (2823 participants) in the *Activity > background* group, PET-CT was found to accurately identify 77.4% (95% CI 65.3 to 86.1) of the participants with NSCLC spread beyond the N1 nodes and 90.1% (95% CI 85.3 to 93.5) of the participants without spread beyond the N1 nodes. In the 12 studies (1656 participants) in the *SUVmax of ≥ 2.5* group, PET-CT accurately identified 81.3% (95% CI 70.2 to 88.9) of the participants with spread beyond the N1 nodes and 79.4% (95% CI 70 to 86.5) of the participants without spread beyond the N1 nodes. However, the results varied a lot between the studies in each analysis, and the quality and size of the studies themselves, country of study origin, percentage of participants with adenocarcinoma, FDG dose, and type of PET-CT scanner influenced the results. We believe that the results of this review show that the accuracy of PET-CT is insufficient to allow management based on PET-CT alone.

BACKGROUND

Accurately determining the diagnosis and stage of lung cancer is important to ensure that patients are offered the best possible treatment. However, the process is often complex. The symptoms and signs of lung cancer can be difficult to distinguish from those of other diseases (some of which may coexist in lung cancer patients), and many lung cancers are diagnosed via other routes (e.g., emergency or Accident & Emergency admissions; through other specialities; or as incidental findings on imaging, such as chest radiographs and computed tomography (CT)) (Department of Health 2011). The diagnosis is made by means of a variety of different biopsies and imaging techniques, some of which yield information about both diagnosis and staging (NICE 2011). The need to consider the location of the primary tumour; patient preferences; and the fitness of the patient, which itself may influence both diagnostic and treatment decisions and may require a change to the diagnostic and staging pathway, augments the complexity.

Target condition being diagnosed

A major determinant of treatment offered to patients with non-small cell lung cancer (NSCLC) is the intrathoracic (mediastinal) nodal status (for a glossary, see Appendix 1). If the disease has not spread to either the ipsilateral mediastinal nodes, subcarinal (N2) nodes, or both, and the patient is otherwise considered fit for surgery, resection is often the treatment of choice (Manser 2005). Other treatment options for these patients include combination or single-modality treatment with either radiotherapy, chemotherapy, or both (O'Rourke 2010). Planning the optimal treatment is therefore critically dependent on accurate staging of the disease.

Lung cancer staging is performed using an arsenal of different complementary tests; some of these are non-invasive (e.g., various types of imaging) (NICE 2011; Silvestri 2013), and some are invasive (e.g., surgical staging, mediastinoscopy) or minimally invasive (e.g., endobronchial ultrasound-guided transbronchial needle aspiration (EBUS-TBNA), endoscopic ultrasound-guided fine needle aspiration (EUS-FNA), and various other methods for obtaining biopsies) (Detterbeck 2007; NICE 2011). The most definitive test is surgical staging - by means of resection of the primary tumour and systematic nodal dissection, with prior mediastinoscopy to assess the contralateral nodes. However, surgical staging is highly invasive and thus not appropriate for many patients without first acquiring further information about the likely suitability for resection with curative intent. Imaging tests (including combined positron emission tomography and CT (PET-CT)) - to assess the probability of malignant involvement and detect extrathoracic metastases and mediastinal lymph node metastasis that would preclude treatment with curative intent - will most often determine suitability for resection with curative intent. One or more biopsies may have to follow imaging findings to pathologically confirm the results of these tests. Occasionally, when imaging tests are unequivocally positive for cancer, the findings alone will be enough to exclude patients from radical treatment. This, in effect, means that only the patients who receive resection will receive the ultimate reference standard (i.e., surgical staging). Those patients who are found to have unresectable NSCLC will usually have had their cancer stage pathologically confirmed by a number of other tests that are considered suitable for the location of the affected lymph node(s).

Therefore, the reference standard for this review necessarily had to consist of a number of invasive tests that all yield pathologically

confirmable information and can be collectively considered as tests that provide cytohistological confirmation of tumour extent. The secondary aims of this review reflect our consideration of this issue as we have tried to consider potential differences in the reference standard as a source of heterogeneity between the studies.

Index test(s)

PET-CT is a non-invasive staging method of the mediastinum, which is increasingly available and used by lung cancer multidisciplinary teams. PET-CT is most commonly performed using

(^{18}F)-2-fluoro-deoxy-D-glucose (FDG) as a tracer to provide a measure of glucose uptake, with simultaneous CT to aid localisation. Before receiving a PET-CT scan, most patients will already have received a CT scan, and PET-CT is most commonly used to confirm early-stage disease in patients who have no significant nodes ($\geq$ 1 cm on the short axis) on CT or to clarify nodal status, in which case PET-CT is not always the first test after CT. That is, currently, the role of PET-CT is primarily in triaging patients, by identifying patients with no spread to the mediastinum who may therefore be candidates for resection, and distinguishing from those patients with either distant or mediastinal metastases, or both, that may need to be biopsied before their treatment plan can be developed.

Clinical pathway

NSCLC patients present with a variety of symptoms and signs as described in the NICE guidelines on referral for suspected cancer (NICE 2005). In England, about 38% of patients first present through the emergency route (i.e., Accident & Emergency or medical admissions). General practitioners urgently refer the majority of the remainder. In most cases, with the exception of those who are too ill to be helped by further diagnostic attempts, the first diagnostic step when lung cancer is suspected is imaging that is either chest radiography or multidetector CT. The latter should ideally be done with the administration of intravenous contrast with contiguous slices from the lower neck to upper abdomen. The secondary care pathway begins with this CT and a clinical assessment where the history is taken, a physical examination, and basic blood tests and lung function obtained. From this information, the first estimate of fitness is made. As part of the two-way communication with either the patient or carer (or both) the potential diagnosis is explained, and some idea about the patient's preference is formulated. The next step in the pathway is to choose the test that gives the most diagnostic and staging information with least risk of harm, provided that the patient is agreeable to this and that further information will likely help the patient. This choice is heavily dependent on what is shown by the CT, and NICE clinical guideline 121 gives detailed guidance on the most appropriate choice of test (NICE 2011; see also De Leyn 2014). The most relevant part of this guidance for the purposes of this review concerns the staging of the mediastinum, which has a separate and more detailed algorithm within the NICE guideline. Once diagnosis, stage, and fitness assessment is confirmed, treatment may be offered on the basis of this information, and follow-up is usually supervised in secondary care in liaison with community services. On relapse, patients may be reassessed, which may be as detailed as the initial work-up, but is usually less so, with treatment offered again on the basis of the findings of the reassessment. Patients in the UK have access to the support of lung cancer nurse specialists throughout the pathway, and there should be holistic needs assessment at all stages of diagnosis and treatment.

Role of index test(s)

PET-CT is central to the assessment of patients who might potentially be suitable for treatment with curative intent. This test is able to define more clearly whether lung cancer has spread to lymph nodes or further. It is in routine use in the UK and a standard of care. NICE recommends PET-CT when the CT does not show significantly enlarged lymph nodes or where nodes of intermediate probability of malignancy are seen (NICE 2011). In reality, many PET-CTs are done for larger high-probability (of cancer) nodes (on CT) prior to minimally invasive sampling, although this practice is unlikely to be cost-effective. Thus, PET-CT forms part of a sequence of tests in the work-up of patients potentially suitable for surgery and is increasingly being done early in the pathway after a baseline CT scan. However, PET-CT is not a perfect test, and it is important to quantify its accuracy and be aware of factors that might alter this.

Alternative test(s)

Other imaging modalities can provide similar information to PET-CT, and these include contrast-enhanced magnetic resonance imaging (MRI) and single photon emission-computed tomography (SPECT). Neither of these tests are as widely available as PET-CT, and importantly, unlike PET-CT, they are also not embedded in the lung cancer pathway. Other tests include the minimally invasive lymph node sampling procedures (i.e., EBUS-TBNA, EUS-FNA) and may be used ahead of PET-CT when treatment might be determined by the result from a single nodal station. Where this sample is negative for malignancy, this approach risks the need for further sampling if PET-CT suggests malignancy in other node stations. For the purposes of this review, the minimally invasive sampling techniques are not considered as alternative tests per se, but rather as part of the techniques that all provide pathological information and thereby collectively constitute the reference standard (see also Target condition being diagnosed).

Rationale

Although the non-invasive nature of PET-CT constitutes one of the major advantages of the test, PET-CT may be suboptimal in detecting malignancy in normal-sized lymph nodes and in ruling out malignancy in patients with coexisting inflammatory or infectious diseases (Cerfolio 2005; Kim 2006; Lee 2007a; Shim 2005; Tournoy 2007; Yi 2007). The role of PET-CT in the accurate staging pathway for patients with lung cancer is therefore still debated, and a crucial question is when a biopsy sample is needed to increase the sensitivity and specificity of PET-CT. Multidisciplinary teams must have a clear idea of the likelihood of false positive and negative PET-CT results in a given circumstance (in particular, the size of mediastinal nodes) in order to best manage patients and advise them whether or not a biopsy is necessary. A false negative rate that is consistently above 20% would cause clinicians to question the utility of the test. However, the question is complex. For example, in the case of detection of distant metastases in patients otherwise fit for surgery, a 20% false negative rate might lead to only one patient in 100 having futile surgery (as the baseline rate of distant metastases is around 5%). In the case of assessing mediastinal nodes by PET-CT, the overall impact will again depend on the prevalence. However, it is also noted that resection of these nodes may not necessarily mean that an operation was the wrong thing to do, as we know from the National Lung Cancer Audit that outcomes are better when patients have had surgery, even for N2 disease (NICE 2011). On balance, we have focused on nodal metastases in this review. False negative outcomes of PET-CT should only apply to nodes that are not significantly enlarged on (a prior) CT, as enlarged nodes should be biopsied. False positives are of a lesser concern since they should always be followed by a further test to confirm.

This review represents an extension to a previous review we have undertaken in this area for the 2011 NICE updated guideline on the diagnosis and treatment of lung cancer (NICE 2011); this included fewer studies and no meta-analysis.

OBJECTIVES

To determine the diagnostic accuracy of integrated PET-CT for mediastinal staging of patients with suspected or confirmed NSCLC that is potentially suitable for treatment with curative intent.

Secondary objectives

To assess potential sources of heterogeneity, including study design (e.g., retrospective/prospective, consecutive/random series); patient populations (number and characteristics, e.g., T- and N-stage, significant nodes on prior CT, country); different cut-off values for test positivity (malignancy); differences in either PET-CT image acquisition, scanning equipment, or both; and potential differences in reference standard (mediastinoscopy/pathological or surgical staging).

METHODS

Criteria for considering studies for this review

Types of studies

Prospective or retrospective cross-sectional studies that assessed the diagnostic accuracy of integrated PET-CT for diagnosing N2 disease in patients with suspected resectable NSCLC. The studies must have used pathology as the reference standard and reported participants as the unit of analysis.

Participants

Patients with suspected/confirmed NSCLC who were considered potentially suitable for primary resection. This review did not consider patients who were being restaged after induction or neoadjuvant chemotherapy.

Index tests

PET-CT carried out on the various available integrated PET-CT scanners with cut-off values for test positivity as reported in the included studies. The type of integrated PET-CT scanner, scanner manufacturer, and cut-off values did not influence whether we included a study or not; rather, as part of the secondary objectives, we examined the potential contribution of these factors to systematic between-study variation as potential sources of heterogeneity. However, we did not consider studies that employed tracers other than FDG or other nuclear medicine imaging, such as single photon emission-computed tomography (SPECT) or stand-alone PET.

Target conditions

Resectability of lung cancer depends on the locoregional spread of the disease. NSCLC is generally not considered resectable if it has spread beyond N1 disease. Thus, the target condition of this review was resectable NSCLC, which for the present purposes, was defined as NSCLC that has not spread to either the ipsilateral mediastinal lymph nodes, the subcarinal (N2) lymph nodes, or both.

Reference standards

Pathological confirmation of PET-CT results from samples obtained via either surgical resection with mediastinal sampling, mediastinoscopy, video-assisted thoracic surgery (VATS), endobronchial ultrasound-guided transbronchial needle aspiration (EBUS-TBNA), EUS-FNA, TBNA (transbronchial needle aspiration), TTNA (transthoracic needle aspiration), biopsies of extrathoracic sites, or a combination of the aforementioned.

Search methods for identification of studies

Electronic searches

We searched the following databases up to 30 April 2013, using the search terms and strategies identified in Appendix 2:

- *The Cochrane Library* (specifically, the Cochrane Central Register of Controlled Trials (CENTRAL), the Health Technology Assessment (HTA) database, the Database of Abstracts of Reviews of Effects (DARE), and the Cochrane Methodology Register (CMR));
- MEDLINE via OvidSP (from 1946);
- Embase via OvidSP (from 1974);
- PreMEDLINE via OvidSP;
- OpenGrey; and
- ProQuest Dissertations & Theses.

We also searched the trials register http://clinicaltrials.gov for research projects in process on 30 April 2013. We used Web of Science (or Scopus if the citation was not on Web of Science) to track records citing those studies, which we included in the final review, with an end-date of 9 July 2013. There were no language or publication status restrictions on the search.

Searching other resources

We handsearched the reference lists of the included articles along with the reference lists of any relevant review articles identified through the search. We also contacted the authors of the included studies and other experts in the field of lung cancer staging for information about any ongoing or unpublished studies. We imposed no language or publication status restrictions on the search.

Data collection and analysis

Selection of studies

Firstly, one of the review authors (MSH) assessed for potential inclusion the titles and abstracts of all the studies identified by the search. This first stage of screening excluded all records that were not studies of PET-CT in patients with NSCLC. Secondly, two of the review authors (MSH and DRB) assessed for potential inclusion the titles and abstracts of the remaining records. Thirdly, two of the review authors (MSH and DRB) independently considered the full records of all potentially relevant studies for inclusion by applying the selection criteria outlined in the Types of studies section. We resolved any disagreements by discussion.

Data extraction and management

Using a standardised data extraction form, two authors (MSH and DRB or MRF) extracted data pertaining to study design, participant detail, index and reference tests, and funding (see Table 1). We resolved any disagreements by discussion. With studies where only a subgroup of the participants met the inclusion criteria for the current review, we only extracted data on this subgroup. For the comparison of the index test with the reference standard, we extracted the number of true and false positives and true and false negatives for the index test when these numbers were presented in the studies. Otherwise, we reconstructed the two-by-two table of true and false positives and negatives from the information reported in the studies, and if this was not possible, we contacted the study authors for the data.

Assessment of methodological quality

Two of three of the authors (MSH and DRB or MFR) independently assessed the quality of each study using a modified version of the QUADAS-2 tool (Whiting 2011), as outlined in Table 2. QUADAS-2 consists of four domains that each require a 'Risk of bias' judgement of low, high, or unclear. For three of these domains, a further judgement needs to be made rating concerns of applicability as low, high, or unclear in terms of how applicable the individual study results are to the question posed by the review. Signalling questions that require a yes, no, or unclear response support the 'Risk of bias' judgements. We included two additional signalling questions on our checklist:

1. Was there a clear definition of a positive result? (We included this under the 'Index test' domain.)
2. Was the study free of commercial funding?

We included the item pertaining to the definition of positive results to take into account the subjective nature of PET-CT image interpretation, which may be based on a variety of different criteria, such as extensive clinical experience, different standard uptake values (SUV), different morphological features, or a combination of the aforementioned. We included the second additional item in order to record any potential bias resulting from commercial interest in the results. We resolved any disagreements between the risk of bias and applicability concern ratings through discussion.

Statistical analysis and data synthesis

We extracted the numbers of true positives, false positives, true negatives, and false negatives for each study based only on the

ability of PET-CT to distinguish between N0 and N1 mediastinal disease and N2 and N3 mediastinal disease. Therefore, we considered both N0 and N1 disease as negatives and both N2 and N3 as positives. If PET-CT indicated N1 disease that was shown by the reference standard to be N0 disease (and vice versa), the PET-CT results were still considered a true negative because N0 and N1 disease were both considered resectable disease. The same principle applied to N2 and N3 disease, that is, if PET-CT indicated N2 disease that was shown to be N3 disease by the reference standard (and vice versa), the PET-CT results were still considered a true positive. However, if PET-CT indicated N0 or N1 disease that the reference standard showed to be N2 or N3 disease, the FDG PET result was considered a false negative. Similarly, if PET-CT indicated N2 or N3 disease that was shown by the reference standard to be N0 or N1 disease, the PET-CT result was considered to be a false positive. If data for more than one positivity threshold were reported, we extracted all the data, but only analysed the threshold most commonly used by all the studies. We only extracted data with participant as the unit of analysis, not, for example, lymph node.

We calculated sensitivity and specificity with 95% confidence intervals (CI) for each study. We plotted the estimates of the observed sensitivities and specificities together with their 95% CI in forest plots and in a receiver operating characteristic (ROC) plot of sensitivity versus 1-specificity in order to visually assess the between-study variability. We fitted a summary ROC curve using the HSROC model for the subset of studies sharing the same positivity threshold (Harbord 2007; Rutter 2001). We selected one threshold per study in the special case of a single study reporting data for more than one threshold. If the studies showed sufficient clinical homogeneity (see Investigations of heterogeneity), we derived summary accuracy estimates for the studies using the same criteria for test positivity for all participants (i.e., *SUVmax* $\geq$ *2.5*, *Activity* > *background*). In the case of different thresholds used in the studies for the analyses, we selected the most frequently used, clinically relevant threshold among the included studies. We identified the average operating point on the SROC curve and computed average sensitivities and specificities. We plotted averaged accuracy estimates with their 95% confidence ellipse and prediction region in ROC space. We had planned to compute the positive and negative likelihood ratios from the pooled estimates of sensitivity and specificity, but given the high degree of heterogeneity we found, the accuracy estimates should be interpreted with caution. As a consequence, we did not compute the likelihood ratios in order to separate the results of our review from their use in clinical practice for a specific patient (i.e., updating post-test probability after a test result).

Investigations of heterogeneity

Several factors can contribute to heterogeneity in diagnostic accuracy of a test across studies. We checked for heterogeneity as part of the planned meta-analysis. Anticipated sources of heterogeneity included study design (e.g., retrospective/prospective, consecutive/random series); FDG dose; patient populations (year, country, sample size, percentage of adenocarcinoma, country tuberculosis rate); and differences in PET-CT image acquisition or scanning equipment (or both).

We could not explore potential differences in reference standard (mediastinoscopy/pathological or surgical staging), one of the planned sources of heterogeneity, because of lack of adequate data. We replaced another planned source of heterogeneity (different cut-off values for test positivity) by the type of test positivity (surrounding activity, SUVmax, and other criteria).

We conducted a subgroup analysis for each factor anticipated to be a heterogeneity source by including the factor as a covariate in the bivariate model (Reitsma 2005). We performed comparison of diagnostic accuracy between subgroups by testing whether either sensitivity or specificity, or both, differed in subgroups of studies defined according to the covariate. The analysis aimed to estimate valid measures of diagnostic accuracy taking into account the effect of any confounding variables. We used the non-linear mixed models (NLMIXED) (Macaskill 2004) procedure in SAS version 9.1 for Windows (SAS Institute Inc, Cary, NC, USA) to fit the HSROC and bivariate models.

Sensitivity analyses

We examined the robustness of the meta-analyses by conducting sensitivity analyses using different components of the 'Risk of bias' assessment. We performed these analyses by limiting inclusion in the meta-analysis to those studies in the primary analyses that had low risk of bias and low concerns about potential applicability. We also excluded from the analyses studies according to other characteristics that could potentially introduce bias into the results (i.e., whether a clear definition for test positivity was used and whether commercial funding was provided).

RESULTS

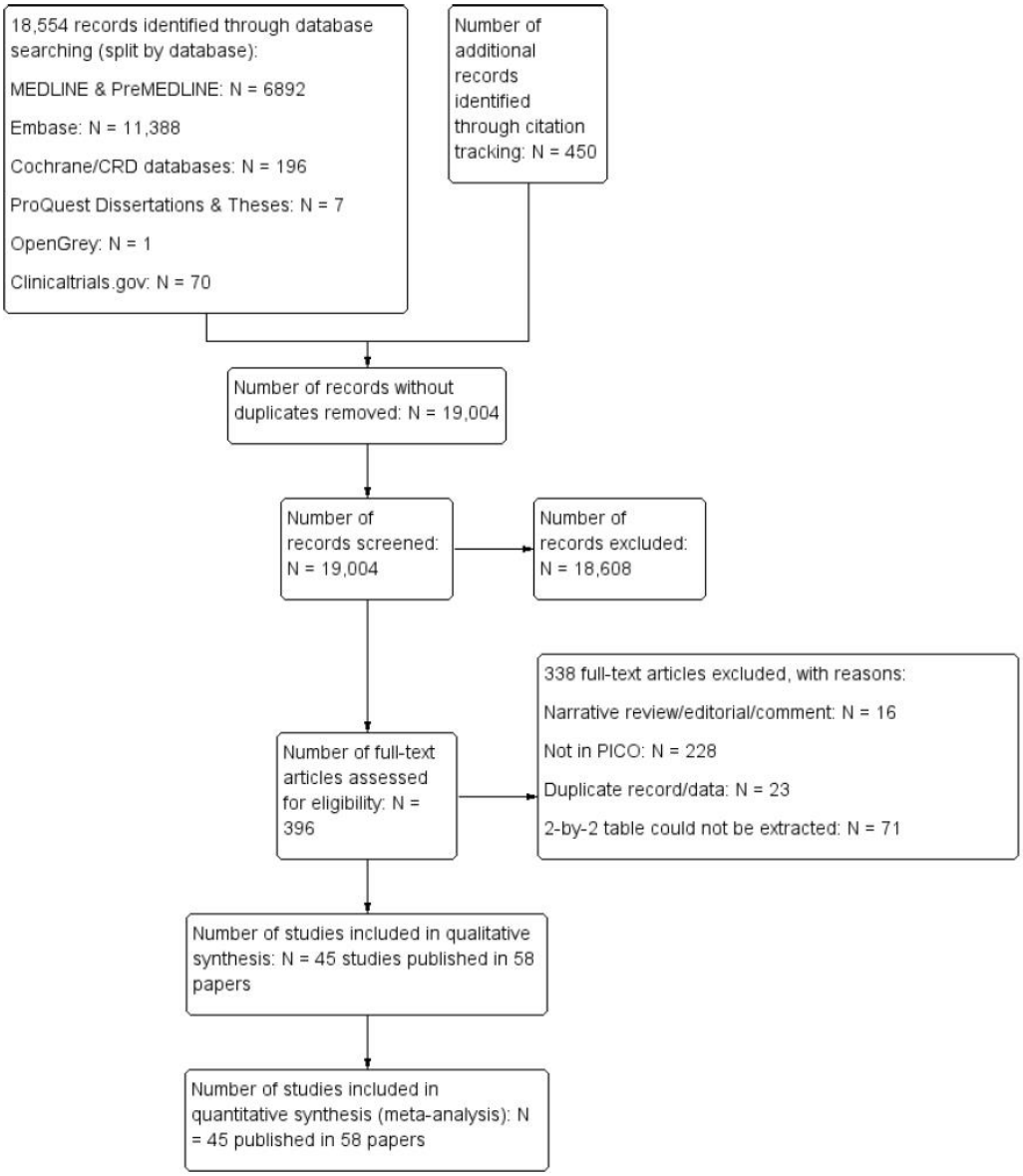
Results of the search

Our search strategy identified 19,004 records, of which we excluded 18,608 as not relevant, based on the title/abstract, and obtained the full publications of 396 records. Of the full publications, 45 studies published in 58 papers met the inclusion criteria while we excluded 338 articles for the following reasons: narrative review/editorial/comment (N = 16); not meeting the PICO (population, intervention, comparison, outcome) criteria (i.e., either the population or tests did not match the current target population or index/reference tests, or the study did not examine the accuracy of PET-CT for mediastinal staging) (N = 228); duplicate

record or data (N = 23); or the two-by-two table could not be extracted for patient level N0 and N1 versus N2 and N3 data (N = 71) (see also Figure 1 and the 'Characteristics of excluded studies' tables). The 45 included studies had a total of 6095 participants available for analysis (median = 112, interquartile range (IQR) = 54 to 169), 4551 of whom were N0 and N1 and 1544 participants of whom were N2 and N3. The prevalence of positive nodes (N2 and N3) varied amongst the studies, ranging from as low as 4% (Lee 2012) to 83% (Uskul 2009), with a median of 22% (IQR = 18 to 30). Thirty-two studies reported the percentage of participants with adenocarcinoma, which ranged from 20.5% to 87.2%.

The studies were categorised according to the incidence rate of tuberculosis (TB, which also included HIV (human immunodeficiency virus)) as reported by the World Health Organization (WHO) (www.who.int/tb/country/data/profiles/en/index.html). Two thirds of the studies (N = 30) had incidence rates lower than 50 per 100,000 population. Half of the studies were performed in Asia (N = 22), while Europe provided 11 studies; North America, a further three studies; and nine studies were from other countries (Turkey, Egypt). All the studies were published after 2005 (2006 to 2009: N = 17; 2010 to 2011: N = 17; and 2012 to 2013: N = 11).

Figure 1. Study flow diagram



The studies also varied in which PET-CT scanner they used, with 19 studies using a Discovery scanner, 14 studies using a Biograph scanner, and the remaining 12 studies employing other/mixed/not reported scanners. We also observed between-study variation in terms of the FDG dose used for the PET-CT scans. Where the total dose was not reported directly, the data were converted to total FDG dose in MBq (megabecquerel) in the following manner: When the dose was reported as MBq/kg, we calculated a total dose for a participant weighing 70 kg. When the FDG dose was reported as a range, we used the mean value. According to these calculations, 12 studies used up to 300 MBq, 25 studies used 301 to 500 MBq, and four studies used > 500 MBq. The remaining four studies did not report FDG dose. There was little difference in injection-to-scan time between the studies (> 45 minutes: N = 1; 30 to 60 minutes: N = 1; 40 to 60 minutes: N = 1; 45 minutes: N = 2; 45 to 60 minutes: N = 1; 50 minutes: N = 3; 60 minutes: N = 26; 55 to 65 minutes: N = 1; 50 to 70 minutes: N = 2; 60 to 120 minutes: N = 1; 75 minutes: N = 1; not reported: N = 5). Twenty-nine studies used attenuation correction; one study did not; and 15 studies did not report whether they undertook attenuation correction. The included studies used different criteria for test positivity. Based on these criteria, we categorised the included studies into three groups of criteria for test positivity: *Activity > background* (18 studies, N = 2823, prevalence of N2 and N3 nodes = 679/2823), *SUVmax ≥ 2.5* (12 studies, N = 1656, prevalence of N2 and N3 nodes = 465/1656), and *Other/mixed* (15 studies, N = 1616, prevalence of N2 and N3 nodes = 400/1616). None of the studies reported (any) adverse events. For full detail study details, see the 'Characteristics of included studies' tables.

Methodological quality of included studies

We have summarised below the methodological quality of the included studies as assessed by QUADAS-2 and per study and per QUADAS-2 item in Figure 2 and Figure 3, respectively. Inspection of Figure 2 and Figure 3 reveals that a substantial amount of under-reporting in the original studies, which led to many judgements of unclear, hampered the quality of the data.

Figure 2. 'Risk of bias' and applicability concerns summary: review authors' judgements about each domain for each included study

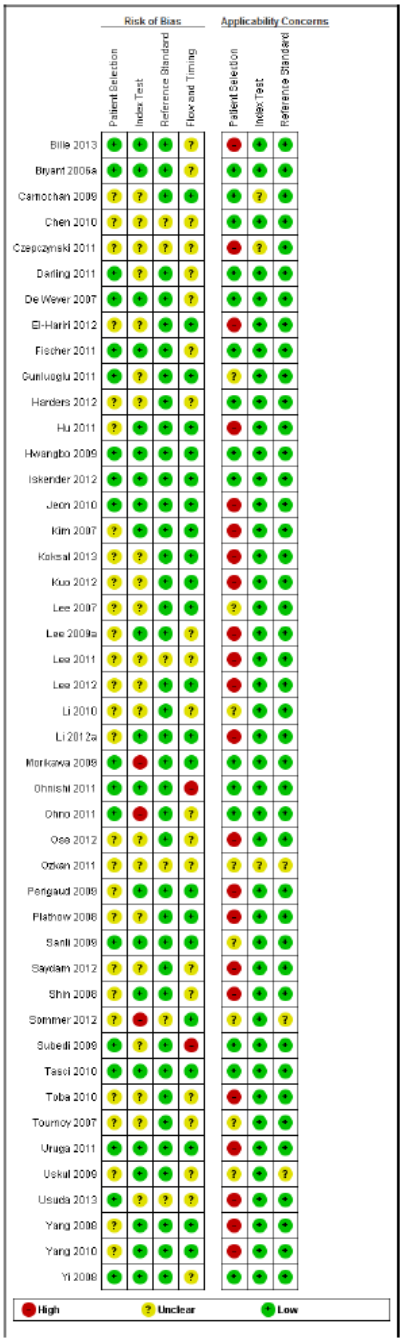
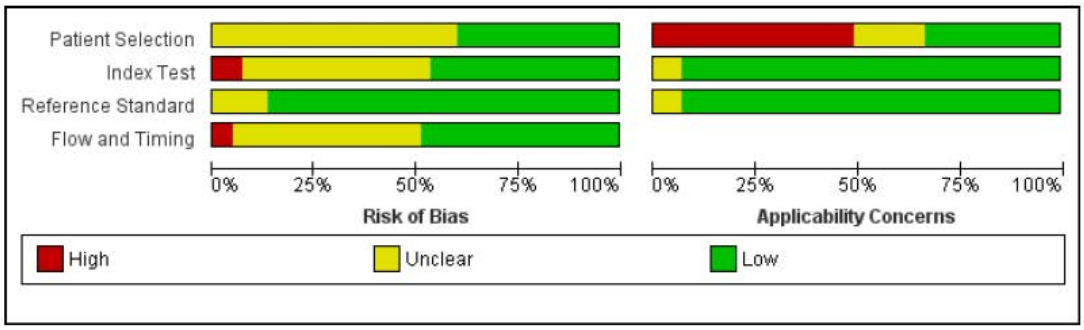


Figure 3. 'Risk of bias' and applicability concerns graph: review authors' judgements about each domain presented as percentages across included studies



Participant selection

Risk of bias

We judged participant selection to be at low risk of bias in 18 of the studies and at unclear risk of bias in the remaining 27 studies.

Applicability concerns

A substantial number of the included studies only included participants who had received resection for NSCLC (Bille 2013; Czepczynski 2011; El-Hariri 2012; Hu 2011; Jeon 2010; Kim 2007; Koksal 2013; Kuo 2012; Lee 2009a; Lee 2011; Li 2012a; Ose 2012; Perigaud 2009; Plathow 2008; Toba 2010; Uruga 2011; Usuda 2013; Yang 2008; Yang 2010), while other studies only included participants with T1 NSCLC (Lee 2012; Shin 2008) or who were retired coal workers (Saydam 2012). All of these inclusion restrictions artificially narrow the range of patients who would receive FDG/PET-CT in standard practice, in particular, the patients with N2 and N3 disease, which in turn gives rise to high concern about the applicability of the populations to the objective of this review. Eight studies did not provide enough information for this item to be rated (i.e., we classified these studies as unclear concerns about applicability), while the populations of the remaining studies were directly applicable to the current question (thus, we classified them as low concern about applicability).

Index test

Risk of bias

The index test was of low or unclear risk in the vast majority of the included studies. However, in three of the included studies, the risk of bias for the index test was high because the results were based on a posthoc specification of the optimal threshold (Morikawa 2009; Ohno 2011) or on more data than just the PET-CT images (Sommer 2012). This was along with a flexible/non-systematic use of SUVs without a general cut-off value (Sommer 2012).

Applicability concerns

We rated three of the included studies as unclear for applicability of the index test because not enough information was reported to assess this question (Carnochan 2009; Czepczynski 2011; Ozkan 2011). We considered the index test as employed by the remaining studies to be applicable to the aims of this review.

Reference standard

Risk of bias

We considered all of the included studies to be at low risk of bias with the exception of Chen 2010; Czepczynski 2011; Lee 2011; Ozkan 2011; Sommer 2012; Usuda 2013, which were all of unclear risk of bias for the reference standard.

Applicability concerns

We considered the reference standard to be applicable to the review in all the included studies apart from three (Ozkan 2011; Sommer

2012; Uskul 2009), which we rated as unclear because of a lack of information reported in the papers, making it impossible to assess the applicability of the reference standard in these cases.

Flow and timing

Risk of bias

Most of the studies were of low or unclear risk of bias, but we considered two of the studies to be at high risk of bias for flow and timing because of missing data (Ohnishi 2011; Subedi 2009).

Other assessed 'Risk of bias' items

Prespecified cut-off values for PET-CT positivity

We selected this item for preplanned sensitivity analyses to assess if the results were sensitive to whether the cut-off values for test positivity were specified a priori or posthoc. However, on appraising the included studies, it became apparent that this item did not apply to at least half of the included studies, that is, the studies that did not use an explicitly quantitative test measure (i.e., SUV). Because when no quantitative criterion has been employed, the answer to this item is 'no' without this in itself giving rise to a problem. We therefore decided to just incorporate this potential source of bias into the 'Risk of bias' assessment for the index test and to limit the assessment of the influence of this item to the sensitivity analysis of the risk of bias for the index test.

PET-CT test positivity clearly defined

Only in nine studies were the criteria for PET-CT positivity either unclearly defined (Ohno 2011; Ozkan 2011; Plathow 2008; Sommer 2012; Tournoy 2007) or not defined (Carnochan 2009; Chen 2010; Czepczynski 2011; Darling 2011); whereas, the remaining 36 studies clearly defined test positivity.

Commercial funding of the studies

Only 19 studies reported any details about funding, and of those studies, 14 had received non-commercial funding (Darling 2011; Fischer 2011; Hu 2011; Hwangbo 2009; Kuo 2012; Lee 2012; Li 2010; Li 2012a; Morikawa 2009; Shin 2008; Usuda 2013; Yang 2008; Yang 2010; Yi 2008); two had received commercial funding (Ohno 2011; Sommer 2012); and three studies reported that they had received no funding (Saydam 2012; Tournoy 2007; Uruga 2011).

Findings

Accuracy of integrated PET-CT for mediastinal staging

Figure 4, Figure 5, and Figure 6 show forest plots of PET-CT sensitivity and specificity for assessing mediastinal lymph node involvement for all the 45 studies included in the review, grouped by the criteria for test positivity employed, i.e., *Activity > background*, *SUVmax ≥ 2.5*, or *Other/mixed/unclear*. Both sensitivity and specificity estimates varied greatly within all three groups. Indeed, sensitivity estimates varied by more than 50% in all three groups, with specificity estimates varying by at least 27% within the groups.

Figure 4. Forest plot of studies with Activity > background as the criterion for test positivity

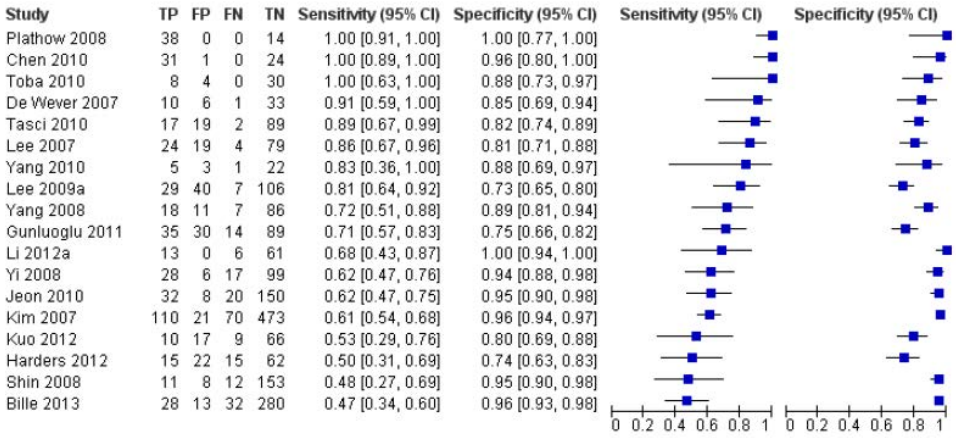


Figure 5. Forest plot of studies with SUVmax ≥ 2.5 as the criterion for test positivity

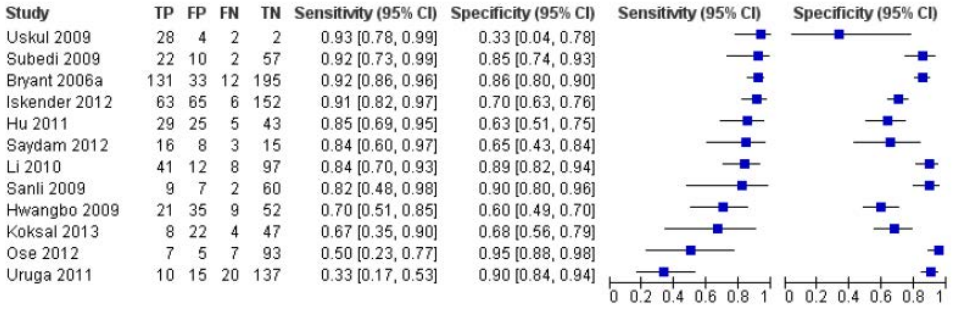
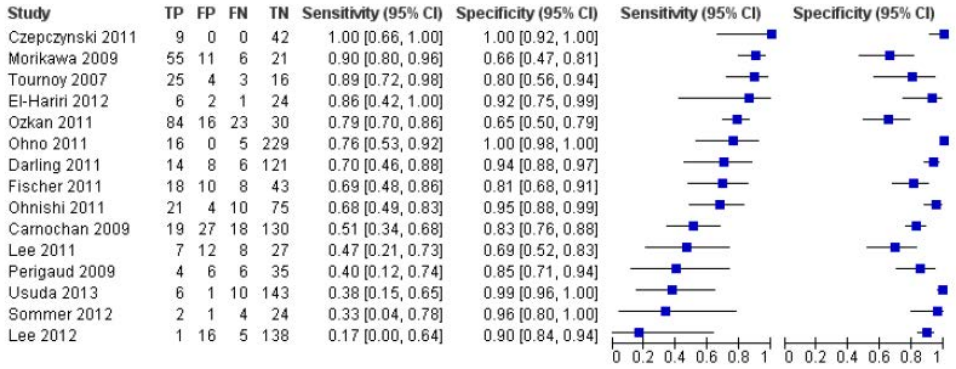
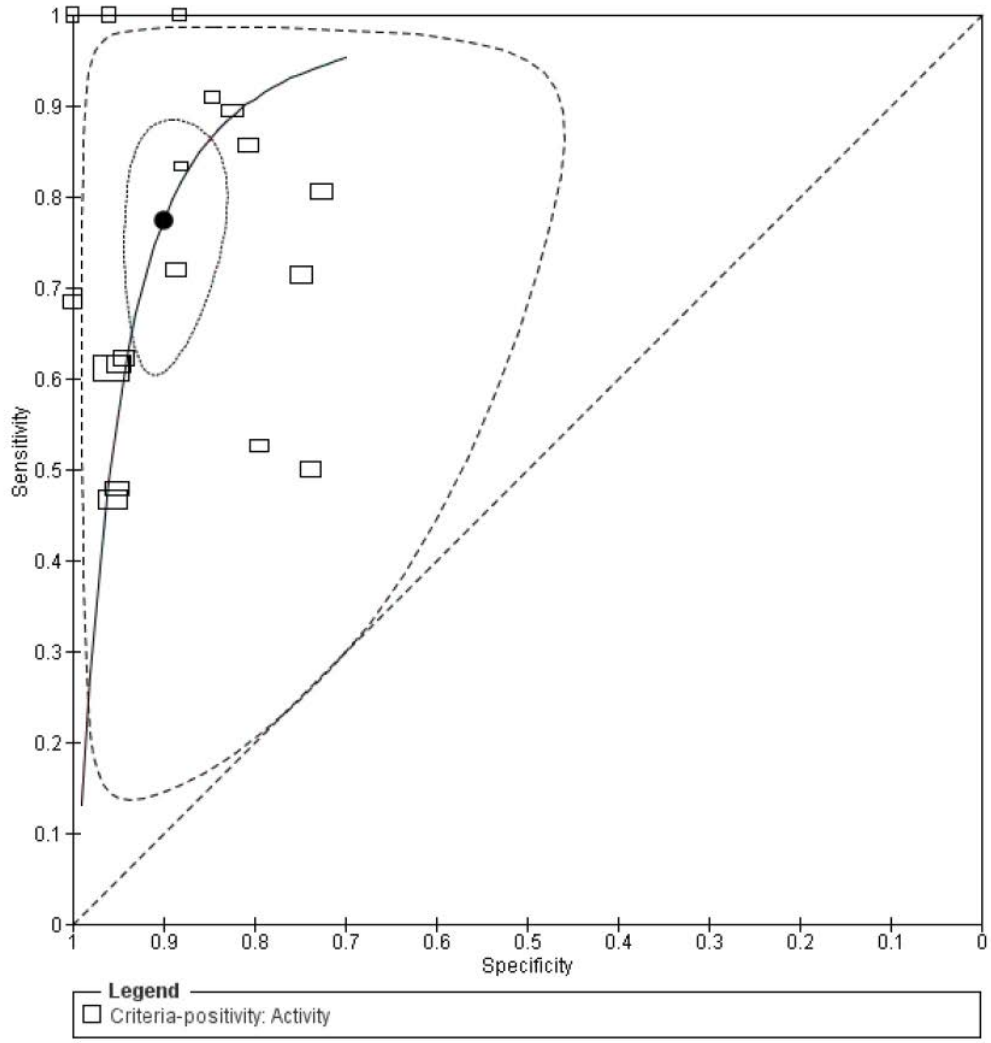


Figure 6. Forest plot of studies with Other/mixed/unclear criteria for test positivity



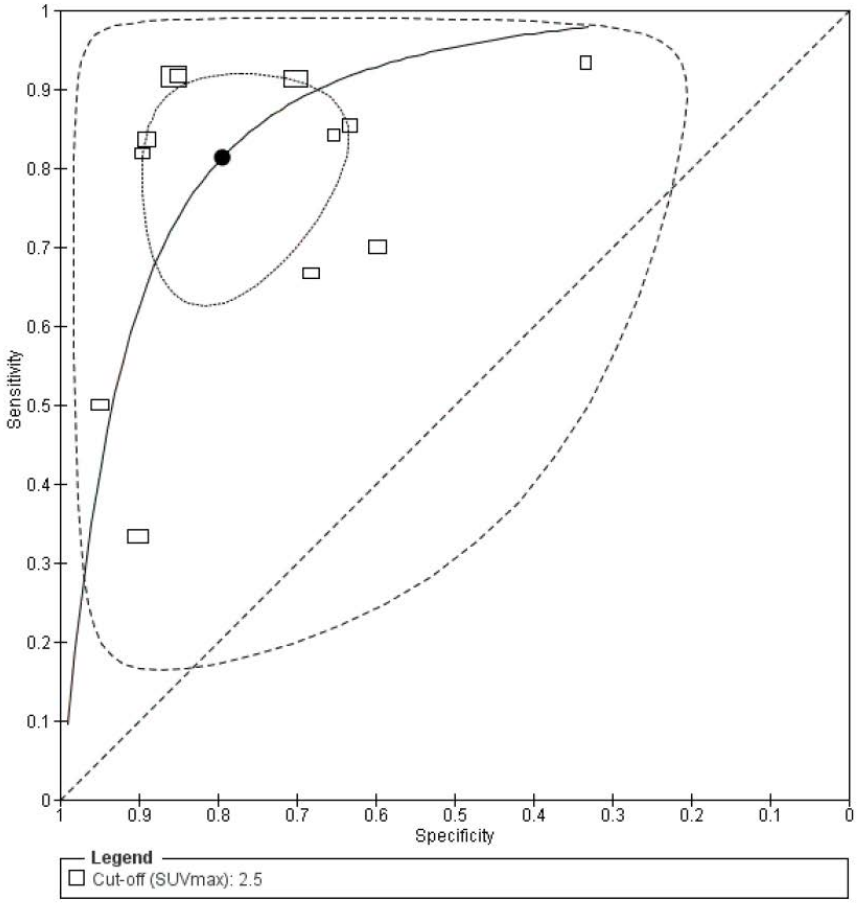
We conducted two primary analyses based on the criteria for test positivity: Regarding the *Activity > background* group, 18 of the included studies employed a qualitative criterion for test positivity based on the relative activation between the lymph nodes and the surrounding tissue (Bille 2013; Chen 2010; De Wever 2007; Gunluoglu 2011; Harders 2012; Jeon 2010; Kim 2007; Kuo 2012; Lee 2007; Lee 2009a; Li 2012a; Plathow 2008; Shin 2008; Tasci 2010; Toba 2010; Yang 2008; Yang 2010; Yi 2008). The summary sensitivity and specificity estimates for this criterion for test positivity were 77.4% (95% CI 65.3 to 86.1) and 90.1% (95% CI 85.3 to 93.5), respectively. Figure 7 shows the accuracy estimates of these studies in ROC space along with a summary ROC curve fitted with the HSROC model. The wide area of the prediction region illustrates that between-study heterogeneity is still high, and the 95% confidence region around the summary value of sensitivity and specificity is also relatively large, denoting lack of precision.

Figure 7. Summary ROC Plot of studies with Activity > background as the criterion for test positivity. Empty squares represent individual study estimates, with the size of the square proportional to the study sample size. The solid line represent the SROC curve. The filled circle is the summary point representing the average sensitivity and specificity estimates. The ellipses around this summary point are the 95% confidence region (dotted line) and the 95% prediction region (dashed line). The dashed upward diagonal represents the completely uninformative test



Regarding the *SUVmax ≥ 2.5* group, 12 studies used a common cut-off value of SUVmax of ≥ 2.5 (Bryant 2006a; Hu 2011; Hwangbo 2009; Iskender 2012; Koksal 2013; Li 2010; Ose 2012; Sanli 2009; Saydam 2012; Subedi 2009; Uruga 2011; Uskul 2009). The summary sensitivity and specificity estimates for this most common threshold were 81.3% (95% CI 70.2 to 88.9) and 79.4% (95% CI 70 to 86.5), respectively. Figure 8 shows the accuracy estimates of these studies in ROC space, along with a summary ROC curve fitted with the HSROC model. The wide area of the prediction region illustrates that between-study heterogeneity is very high. The 95% confidence region around the summary value of sensitivity and specificity is also large, denoting a clear lack of precision.

Figure 8. Summary ROC Plot of studies with SUVmax > 2.5 as the criterion for test positivity. Empty squares represent individual study estimates, with the size of the square proportional to the study sample size. The solid line represent the SROC curve. The filled circle is the summary point representing the average sensitivity and specificity estimates. The ellipses around this summary point are the 95% confidence region (dotted line) and the 95% prediction region (dashed line). The dashed upward diagonal represents the completely uninformative test

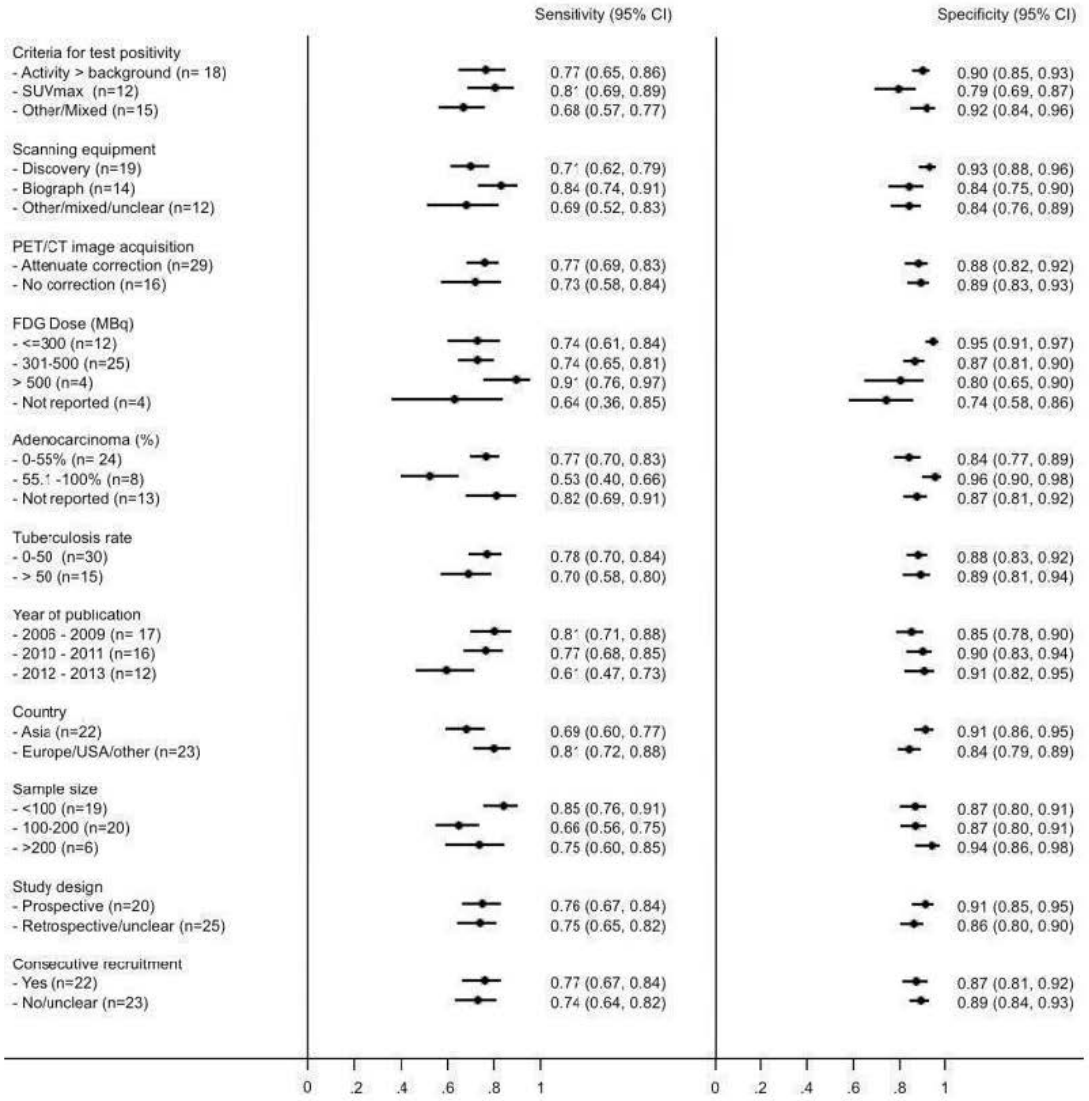


We did not conduct any further analyses for the studies using *Other/mixed/unclear* criteria for test positivity as the large variation in these criteria would have made any further analyses meaningless.

Investigations of heterogeneity

We report on subgroups based on preplanned covariates that were anticipated to contribute to heterogeneity. We did not conduct these analyses separately for the two main analyses as we found no overall effect of test positivity criteria (*Activity > background* (sensitivity = 0.77, 95% CI 0.65 to 0.86; specificity = 0.9, 95% CI 0.85 to 0.93) versus *SUVmax ≥ 2.5* (sensitivity = 0.81, 95% CI 0.69 to 0.89; specificity = 0.79, 95% CI 0.69 to 0.87) versus *Other/mixed* (sensitivity = 0.68, 95% CI 0.57 to 0.77; specificity = 0.92, 95% CI 0.84 to 0.96); P = 0.77). Figure 9 presents the results of the subgroup analysis.

Figure 9. Investigations of possible sources of heterogeneity



In summary, there were differences in accuracy for a number of factors (country, type of PET-CT scanner, percentage of participants with adenocarcinoma, FDG dose, and study size), while we observed no differences for design characteristics, year of publication, and tuberculosis incidence rate. The detailed results for each factor follow.

Country of origin was significantly associated with diagnostic accuracy. Studies performed in western countries (Europe/USA/other: sensitivity = 0.81, 95% CI 0.72 to 0.88; specificity = 0.84, 95% CI 0.79 to 0.89) showed greater sensitivity (p (pair-wise) = 0.045) and lower specificity (P (pair-wise) = 0.035) than studies performed in Asian countries (sensitivity = 0.69, 95% CI 0.6 to 0.77; specificity = 0.91, 95% CI 0.86 to 0.95)); P (overall effect) = 0.04). The type of PET-CT scanner was also associated with different diagnostic accuracy: Compared to Discovery (sensitivity = 0.71, 95% CI 0.62 to 0.79; specificity = 0.93, 95% CI 0.88 to 0.96), Biograph scanning equipment (sensitivity = 0.84, 95% CI 0.74 to 0.91; specificity = 0.84, 95% CI 0.75 to 0.9) showed greater sensitivity (P (pair-wise) = 0.039) and lower specificity (P (pair-wise) = 0.047; P (overall effect) = 0.039). Thirty-two studies reported the percentage of participants with adenocarcinoma and

ranged from 20.5% to 87.2%. There was a clear split in the data between 54.8% and 69.1%, and we therefore employed a cut-off of 55% to analyse this covariate with three levels: 0% to 55% (in effect, this is 20.5% to 54.8%; N = 24), 55.1% to 100% (in effect, 69.1% to 87.2%; N = 8), and not reported (N = 13).

The percentage of participants with adenocarcinoma also influenced the diagnostic accuracy of PET-CT (P (overall effect) = 0.003) with the sensitivity being significantly higher (P (pair-wise) = 0.004) and specificity (P (pair-wise) = 0.001) being significantly lower in studies with ≤ 55% adenocarcinoma participants (sensitivity = 0.77, 95% CI 0.7 to 0.83; specificity = 0.84, 95% CI 0.77 to 0.89) compared with studies with > 55% adenocarcinoma participants (sensitivity = 0.53, 95% CI 0.4 to 0.66; specificity = 0.96, 95% CI 0.9 to 0.98). We also found that FDG dose was associated with different diagnostic accuracy estimates (P (overall effect) = 0.015): Sensitivity was significantly higher (P (pair-wise) = 0.031) and specificity significantly lower (P (pair-wise) = 0.044) in studies using > 500 MBq (sensitivity = 0.91, 95% CI 0.76 to 0.97; specificity = 0.8, 95% CI 0.65 to 0.9) compared with studies using 300 or less MBq (sensitivity = 0.74, 95% CI 0.61 to 0.84; specificity = 0.95, 95% CI 0.91 to 0.97); specificity was also significantly lower (P (pair-wise) = 0.003) in studies using 301 to 500 MBq (sensitivity = 0.74, 95% CI 0.65 to 0.81; specificity = 0.87, 95% CI 0.81 to 0.9) compared with studies using 300 or less MBq, and sensitivity was significantly higher in studies using > 500 MBq compared with those using 301 to 500 MBq (P (pair-wise) = 0.007). The heterogeneity analyses revealed one final covariate that influenced sensitivity and specificity, namely, study size (P (overall effect) = 0.025) with significantly higher sensitivity in studies with < 100 participants (sensitivity = 0.85, 95% CI 0.76 to 0.91; specificity = 0.87, 95% CI 0.8 to 0.91) compared with studies with 100 to 199 participants ((sensitivity = 0.66, 95% CI 0.56 to 0.75; specificity = 0.87, 95% CI 0.8 to 0.91); P (pair-wise) = 0.003) and significantly higher specificity in studies with > 200 participants (sensitivity = 0.75, 95% CI 0.6 to 0.85; specificity = 0.94, 95% CI 0.86 to 0.98) compared with studies with < 100 participants (P (pair-wise) = 0.045) and studies with 100 to 199 participants (P (pair-wise) = 0.0495).

No other analysed covariates were associated with different diagnostic accuracy of the test: Design (prospective (sensitivity = 0.76, 95% CI 0.67 to 0.84; specificity = 0.91, 95% CI 0.85 to 0.95) versus retrospective/unclear (sensitivity = 0.75, 95% CI 0.65 to 0.82; specificity = 0.86, 95% CI 0.8 to 0.9); P = 0.444), consecutive recruitment (yes (sensitivity = 0.77, 95% CI 0.67 to 0.84; specificity = 0.87, 95% CI 0.81 to 0.92) versus no/unclear (sensitivity = 0.74, 95% CI 0.64 to 0.82; specificity = 0.89, 95% CI 0.84 to 0.93); P = 0.933), attenuation correction (yes (sensitivity = 0.77, 95% CI 0.69 to 0.83; specificity = 0.88, 95% CI 0.82 to 0.92) versus no/unclear (sensitivity = 0.73, 95% CI 0.58 to 0.84; specificity = 0.89, 95% CI 0.83 to 0.93); P = 0.55), year of publication (2006 to 2009 (sensitivity = 0.81, 95% CI 0.71 to 0.88; specificity = 0.85, 95% CI 0.78 to 0.9) versus 2010 to 2011 (sensitivity = 0.77, 95% CI 0.68 to 0.85; specificity = 0.9, 95% CI 0.83 to 0.94) versus 2012 to 2013 (sensitivity = 0.61, 95% CI 0.47 to 0.73; specificity = 0.91, 95% CI 0.82 to 0.95); P = 0.139), and tuberculosis incidence rate per 100,000 population (0 to 50 (sensitivity = 0.78, 95% CI 0.7 to 0.84; specificity = 0.88, 95% CI 0.83 to 0.92) versus > 50 (sensitivity = 0.7, 95% CI 0.58 to 0.8; specificity = 0.89, 95% CI 0.81 to 0.94); P = 0.688).

Where the overall effect of the covariate was significant but one of the levels of the covariate was not reported (adenocarcinoma), *Other/mixed/unclear* (scanning equipment), or unclear (FDG dose), we have not reported any pair-wise comparisons involving that level of the covariate because we would not be able to make any useful statements about such analyses.

Sensitivity analysis

In Table 3, we present this restricted analysis including only studies with low risk of bias or low concerns about applicability. Table 3 seems to suggest that in the *Activity > background* group, the overall estimate of sensitivity especially is sensitive to selection bias; reference standard bias; and clear definition of test positivity; and to a lesser extent, index test bias and commercial funding bias, with lower combined estimates of sensitivity observed for all the low 'Risk of bias' studies compared with the full analysis. In the *SUVmax ≥ 2.5* group, the sensitivity analyses suggest that both overall accuracy estimates are much less sensitive to the exclusion of studies according to the covariates analysed. Only flow and timing bias and commercial funding bias led to slightly lower estimates of both sensitivity and specificity. We did not make any formal statistical comparison given the scarce number of studies analysed after the exclusions in the sensitivity analysis.

Summary of findings

PET-CT for assessing mediastinal lymph node involvement in participantswith suspected resectable non-small cell lung cancer					
Population	Participants with suspected/confirmed NSCLC who are considered potentially suitable for primary resection				
Index test	PET-CT carried out on the various available integrated PET-CT scanners with cut-off values for test positivity as reported in the included studies				
Target condition	Resectable NSCLC defined as NSCLC that has not spread to either the ipsilateral mediastinal lymph nodes, subcarinal (N2) lymph nodes, or both				
Reference standard	Pathological confirmation of PET-CT results				
Included studies	45 studies with 6095 participants available for analysis (median = 112, interquartile range (IQR) = 54 to 169), 4551 of whom were N0 or N1 and 1544 participants were N2 or N3 Different criteria for test positivity were used in the included studies: *Activity > background* (18 studies; N = 2823; prevalence of N2 and N3 nodes = 679/2328) *SUVmax* $\geq$ *2.5* (12 studies; N = 1656; prevalence of N2 and N3 nodes = 465/1656) *Other/mixed* criteria for test positivity (15 studies; N = 1616; prevalence of N2 and N3 nodes = 400/1616) None of the studies reported (any) adverse events				
Test subgroup	**Number of participants (studies)**	**Prevalence %**	**Summary accuracy % (95% CI)**	**Implications**	**Quality and comments**
Activity > background	2823 (18)	29.2	Sensitivity: 77.4 (65.3 to 86.1) Specificity: 90.1 (85.3 to 93.5)	With the observed prevalence, there will be 66 missed cases and 70 cases who will receive futile surgery	Participant selection, index test, and flow and timing poorly reported Population spectrum narrower than in standard clinical practice in a substantial number of studies Results sensitive to selection bias, reference standard bias, and clear definition of test positivity Substantial heterogeneity was observed

SUVmax ≥ 2.5	1656 (12)	28.1	Sensitivity: 81.3 (70.2 to 88.9) Specificity: 79.4 (70.0 to 86.5)	With the observed prevalence, there will be 53 missed cases and 148 cases who will receive futile surgery	Participant selection, index test and flow, and timing poorly reported Population spectrum narrower than in standard clinical practice in a substantial number of studies Results sensitive to flow and timing bias and commercial funding bias Substantial heterogeneity was observed
All included studies	6095	25.3	Heterogeneity analyses showed significant contributions to between-study heterogeneity from the following covariates: country of study origin, percentage of participants with adenocarcinoma, FDG dose, type of PET-CT scanner, and study size. Study design, consecutive recruitment, attenuation correction, year of publication, and tuberculosis incidence rate per 100,000 population did not contribute significantly to the observed heterogeneity		

CAUTION: The results in this table should **not** be interpreted in isolation from the results of the individual included studies contributing to each summary test accuracy measure. These are reported in the main body of the review

CI = confidence interval.

FDG = (^{18} F)-2-fluoro-deoxy-D-glucose.

IQR = interquartile range.

NSCLC = non-small cell lung cancer.

PET-CT = positron emission tomography-computed tomography.

SUVmax = maximum standardised uptake value.

DISCUSSION

We have conducted an up-to-date review of studies that have examined the role of PET-CT in determining whether there has been N2 or N3 disease. This is important because it is often crucial in planning treatment. People with N2 or N3 disease do not usually undergo radical treatment with surgery as their primary treatment and instead receive palliative treatment. Where radical treatment is to be offered, there needs to be careful planning, and knowledge of the status of N2 or N3 nodes is essential.

Summary of main results

We found that there was considerable variation in sensitivity and specificity amongst the 45 studies evaluated. Our two main analyses showed that in the studies employing a criterion of *Activity > background*, the summary sensitivity and specificity estimates were 77.4% (95% CI 65.3 to 86.1) and 90.1% (95% CI 85.3 to 93.5), respectively, and for studies employing *SUVmax ≥ 2.5* as the criterion for test positivity, the sensitivity and specificity estimates for this threshold were 81.3% (95% CI 70.2 to 88.9) and 79.4% (95% CI 70 to 86.5), respectively. However, it was the case for both analyses that the prediction and confidence regions were large, and further analyses found that the following covariates partly explained between-study variability: country of origin, with studies performed in western countries showing greater sensitivity and lower specificity than studies performed in Asian countries; type of PET-CT scanner, with Biograph scanning equipment showing greater sensitivity and lower specificity than Discovery; the percentage of participants with adenocarcinoma, with the sensitivity being significantly higher and specificity significantly lower in studies with ≤ 55% adenocarcinoma participants compared with studies with > 55% adenocarcinoma participants; FDG dose, with significantly higher sensitivity and significantly lower specificity in studies using > 500 MBq compared with studies using 300 or less MBq, significantly lower specificity in studies using 301 to 500 MBq compared with studies using 300 or less MBq, and significantly higher sensitivity in studies using > 500 MBq compared with those using 301 to 500 MBq; and study size, with significantly higher sensitivity in studies with < 100 participants compared with studies with 100 to 199 participants and significantly higher specificity in studies with 200+ participants compared with studies with > 100 participants and studies with 100 to 199 participants. Sensitivity analyses also suggested that the summary estimates from the two main analyses were sensitive to a number of biases. Specifically, in the *Activity > background* group, the overall estimate of sensitivity especially is sensitive to selection bias; reference standard bias; clear definition of test positivity; and to a lesser extent, index test bias and commercial funding bias, with lower combined estimates of sensitivity observed for all the low 'Risk of bias' studies compared with the full analysis. In the *SUVmax ≥ 2.5* group, the sensitivity analyses suggested that both overall accuracy estimates were somewhat sensitive to flow and timing bias and commercial funding bias, which led to slightly lower estimates of both sensitivity and specificity.

The observation that studies performed in western countries showed higher sensitivity and lower specificity compared with studies performed in Asian countries may be linked to the observation that the sensitivity was significantly higher and specificity significantly lower in studies with ≤ 55% adenocarcinoma participants compared with studies with > 55% adenocarcinoma participants. This is because we know that there are differences in tumour biology of lung cancer in east Asians, with a greater proportion of cancers being adenocarcinoma and in non-smokers (Maemondo 2010). This may influence the FDG uptake, which is lower in adenocarcinoma than in other common forms of NSCLC (Casali 2010; Davidson 2009; Jeong 2002; Lu 2010), and therefore the sensitivity, which also fits well with the finding that studies using a relatively higher dose of FDG (> 500 MBq) had a higher sensitivity than those using a relatively lower dose (< 500 MBq).

The observation that the type of PET-CT scanner employed is associated with different accuracy estimates suggests that the two main integrated PET-CT scanner manufacturers have produced products with different characteristics, and it is of potential importance to lung cancer clinicians to know that the equipment alone may influence the result obtained.

The finding that study size and various biases influenced the results of this review underscores the need for well-designed and adequately powered (and reported) diagnostic test accuracy studies in NSCLC staging research, specifically, but also in diagnostic medical research in general. It should however also be noted that even within the different test positivity criteria subgroups, the actual criteria/cut-offs used varied between the studies (e.g., in the *Other/mixed/unclear* group, SUVmax > 3.5 versus ≥ 4.1 versus ≥ 4.45), which may well explain a significant amount of the remaining between-study heterogeneity. Unfortunately, we were unable to investigate the contribution of this variable in greater detail because of the low number of studies within each test positivity subcategory. We also note that SUVmax may show some variation in value on repeated measurements, but this is minimal in relation to the spread of values normally obtained in studies (i.e., few are on the cut-off value). There may be some variation in the SUVmax measured in different centres, but this is again likely to be minimal as the majority of the measurement is standardised by the software (Lindholm 2014).

Strengths and weaknesses of the review

We performed an extensive search for relevant studies and were able to obtain data from 45 studies for inclusion. With these studies, we were able to show that a number of conceivably connected factors influence the accuracy of PET-CT for mediastinal staging of NSCLC, namely adenocarcinoma; Asian population; and FDG

dose, as well as by scanner type, a finding which we do not believe is linked to the influence of the other covariates after careful examination of potential overlap between the studies that contributed to the different results. However, despite the relatively large number of relevant studies and a number of prespecified heterogeneity and sensitivity analyses, a substantial amount of unexplained heterogeneity still mark the results, which we hypothesise can, at least in part, be explained by the large variation in the criteria used for test positivity in the different studies. Unfortunately, we were unable to examine in detail this hypothesis because too few studies used the same criteria.

While we are also reasonably confident that the reference standard used in our review is robust and clinically appropriate, it should be noted that some of these tests themselves have limitations in their accuracy. Where EBUS-TBNA was positive, there was often no further sampling. False positives would be very uncommon and unlikely to influence the results unless there was a systematic error within the study by the clinicians involved. Where EBUS-TBNA or mediastinoscopy was negative, reliance was placed either on a period of follow-up, confirming negativity, or systematic nodal dissection and sampling as part of surgery. Where the latter was not clearly specified or the nodal sampling potentially was incomplete, we identified this as a potential source of bias. However, we were clearly unable to assess the quality of adherence to the protocol specified in the studies. It is possible that nodal sampling quality varied amongst surgeons and studies, although most of these studies were conducted at large centres where one would expect high standards. There is also a small risk that N3 nodes might have been missed where mediastinoscopy was not performed prior to surgery. This would mean that contralateral nodes would only have been sampled by EBUS-TBNA, as only ipsilateral nodes are sampled in a systematic nodal dissection. A further limitation is that we were unable to find sufficient studies that looked at the accuracy of PET-CT in lymph nodes that were not significantly enlarged by CT criteria ($\leq$ 10 mm maximum short axis diameter). However, those studies with a low prevalence of malignancy were likely to have included people with smaller nodes as nodal size is strongly correlated with malignancy; in these studies, specificity appeared to be high. Lastly, we would have preferred to be able to include more studies in potentially more difficult populations, such as those with a high prevalence of diseases or conditions known to produce false positive results, such as tuberculosis and industrial exposure to pathogens. Unfortunately, this was not possible as not enough relevant studies appear to have been conducted in such populations.

Applicability of findings to the review question

Broadly speaking, our findings are applicable to the review question in terms of the index test and reference standard where, generally, there was good correspondence between the tests used in the included studies and those specified in our review question. However, as outlined in Methodological quality of included studies, a substantial number of the included studies only included participants who had received resection for NSCLC while other studies only included participants with T1 NSCLC or who were retired coal workers. And all of these inclusion restrictions artificially narrow the range of patients who would receive FDG/PET-CT in standard practice, in particular, the patients with N2 or N3 disease, and this, in turn, gives rise to high concerns about the applicability of the populations to the question of the present review. On the other hand, enough studies were available to enable us to address these applicability concerns through sensitivity analyses, which suggested that sensitivity is increased while specificity is decreased relative to the overall estimates within both of the main analyses when we only analysed the studies with low concerns about applicability. We believe that these results are directly applicable to the typical populations seen in routine clinical practice (accepting that these differ in different countries). We have shown clearly that there is variation in the accuracy of PET-CT in the differentiation of N2 and N3 lymph node metastasis in NSCLC and that this variation is related to, among other factors, NSCLC subtype (adenocarcinoma), country of study origin (Asia), FDG dose, and PET-CT scanner type, all of which should be born in mind by the lung cancer diagnostician.

AUTHORS' CONCLUSIONS

Implications for practice

This review has provided up-to-date data on the accuracy of PET-CT scanning in determining N2 and N3 nodal status in non-small cell lung cancer. It has shown that pooled sensitivity and specificity, whilst reasonable at around 0.8, is insufficient to allow management based on PET-CT alone. In clinical practice, PET-CT is a useful test, and this review supports that. However, the review has also confirmed that PET-CT has to form part of a clinical pathway supported by other investigations and cannot be used as a stand-alone test. The findings therefore support NICE guidance on this topic, where PET-CT is used to guide clinicians in the next step, which is either a biopsy or where negative and nodes are small, directly to surgery (NICE 2011). The apparent difference between the two main makes of PET-CT scanner is important, as this appears independent of the operator or other factors. This a new finding, to our knowledge, and may be important for lung cancer multidisciplinary teams to know. The relatively low sensitivity but high specificity of the Discovery could, in some circumstances, such as where the patient is of very borderline fitness, influence the decision. This would, as is recommended, include the wishes of the patient after a fully informed discussion. The difference between makes and the general variability of results suggests that all large centres should actively monitor their accuracy so that they can make reliable decisions based on their own results.

The pooled results by country identified important differences in the accuracy of PET-CT, showing that it may be less sensitive in Asian countries. Again, this calls for centres to audit their results and identify the populations in which PET-CT is of most use or potentially little value.

Implications for research

In radiology, as in many other areas of medicine, technological advances may lead to rapid changes in clinical practice. Newer PET-CT scanners will be introduced that have higher resolution and lower radiation dose. As they are expensive, it will take some time before they become universally used, but it will be important to measure their accuracy as soon as possible. A key question will be how they perform in different populations and according to the size of lymph nodes. Studies should be designed in populations with a high prevalence of tuberculosis or industrial disease and in participants with interstitial lung disease. These patients are commonly encountered in clinical practice, and we do not know exactly how these conditions alter the accuracy of PET-CT. There should be correspondence between the protocols and make of scanners in studies conducted in Asia and in the western countries so that comparisons can be made and so that populations can be identified where PET-CT is of use or no use. The reasons for our observed difference between the make of PET-CT scanners are not clear, and studies should be undertaken to establish the reason for the difference, which is likely to relate to calibration rather than a difference in the accuracy of detectors. Another key question is whether some N2 nodes that are shown to be positive on PET-CT, with or without pathological confirmation, should be resected as part of a definitive operation. This may depend on nodal size, SUV, and number of nodal stations involved. NICE clinical guideline 121 (NICE 2011) has already made recommendations for research into this area.

ACKNOWLEDGEMENTS

We would like to thank the National Institute for Health and Care Excellence (NICE), the National Collaborating Centre for Cancer, and the Guideline Development Group for the NICE clinical guideline 121 (The diagnosis and treatment of lung cancer (update)) (NICE 2011). We would also like to thank all the peer reviewers who have provided us with helpful comments on earlier versions of this review (and the protocol before that) and the Cochrane Lung Cancer group for all their continued advice and assistance.

Marta Roqué i Fíguls is a PhD candidate at the Department of Paediatrics, Obstetrics and Gynecology and Preventive Medicine, Universitat Autònoma de Barcelona, Spain.

REFERENCES

References to studies included in this review

Bille 2013 *{published data only}*
Billè A, Okiror L, Skanjeti A, Errico L, Arena V, Penna D, et al. Evaluation of integrated positron emission tomography and computed tomography accuracy in detecting lymph node metastasis in patients with adenocarcinoma vs squamous cell carcinoma. *European Journal of Cardio-Thoracic Surgery* 2013;**43**(3):574–9. [3420746; PUBMED: 22689182]
Billè A, Okiror L, Skanjeti A, Errico L, Arena V, Penna D, et al. The prognostic significance of maximum standardized uptake value of primary tumor in surgically treated non-small-cell lung cancer patients: analysis of 413 cases. *Clinical Lung Cancer* 2013;**14**(2):149–56. [3420747; PUBMED: 22682667]
Billé A, Pelosi E, Skanjeti A, Arena V, Errico L, Borasio P, et al. Preoperative intrathoracic lymph node staging in patients with non-small-cell lung cancer: accuracy of integrated positron emission tomography and computed tomography. *European Journal of Cardio-Thoracic Surgery* 2009;**36**(3):440–5. [3420748; PUBMED: 19464906]

Bryant 2006a *{published data only}*
Bryant AS, Cerfolio RJ, Klemm KM, Ojha B. Maximum standard uptake value of mediastinal lymph nodes on integrated FDG-PET-CT predicts pathology in patients with non-small cell lung cancer. *The Annals of Thoracic Surgery* 2006;**82**(2):417–22. [3420750; PUBMED: 16863739]
Cerfolio RJ, Bryant AS, Ojha B, Eloubeidi M. Improving the inaccuracies of clinical staging of patients with NSCLC: a prospective trial. *The Annals of Thoracic Surgery* 2005;**80**(4):1207–13. [3420751; PUBMED: 16181842]

Carnochan 2009 *{published data only}*
Carnochan FM, Walker WS. Positron emission tomography may underestimate the extent of thoracic disease in lung cancer patients. *European Journal of Cardio-Thoracic Surgery* 2009;**35**(5):781–4. [3420753; PUBMED: 19272791]

Chen 2010 *{published data only}*
Chen W, Jian W, Li HT, Li C, Zhang YK, Xie B, et al. Whole-body diffusion-weighted imaging vs. FDG-PET for the detection of non-small-cell lung cancer. How do they measure up?. *Magnetic Resonance Imaging* 2010;**28**(5):613–20. [3420755; PUBMED: 20418042]

Czepczynski 2011 *{published data only}*
Czepczynski R, Zielinski P, Stangierski A, Gabryel P, Kasprzak W, Dyszkiewicz W, et al. The value of 18F-FDG PET/CT scan in the evaluation lymph node status in patients with non-small cell lung carcinoma. *European*

Journal of Nuclear Medicine and Molecular Imaging 2011; **Conference**:S282. [3420757]

Darling 2011 *{published data only}*
Darling GE, Maziak DE, Inculet RI, Gulenchyn KY, Driedger AA, Ung YC, et al. Positron emission tomography-computed tomography compared with invasive mediastinal staging in non-small cell lung cancer: results of mediastinal staging in the early lung positron emission tomography trial. *Journal of Thoracic Oncology* 2011;**6**(8):1367–72. [3420759; PUBMED: 21587082]
Maziak DE, Darling GE, Inculet RI, Gulenchyn KY, Driedger AA, Ung YC, et al. Positron emission tomography in staging early lung cancer: a randomized trial. *Annals of Internal Medicine* 2009;**151**(4):221–8. [3420760; PUBMED: 19581636]

De Wever 2007 *{published data only}*
De Wever W, Ceyssens S, Mortelmans L, Stroobants S, Marchal G, Bogaert J, et al. Additional value of PET-CT in the staging of lung cancer: comparison with CT alone, PET alone and visual correlation of PET and CT. *European Radiology* 2007;**17**(1):23–32. [3420762; PUBMED: 16683115]

El-Hariri 2012 *{published data only}*
El-Hariri MA, Gouhar GK, Refat AM. Integrated PET/CT in the preoperative staging of lung cancer: a prospective comparison of CT, PETand integrated PET/CT. *The Egyptian Journal of Radiology and Nuclear Medicine* 2012;**43**(4):613–21. [3420764; DOI: 10.1016/j.ejrnm.2012.09.007]

Fischer 2011 *{published data only}*
Fischer B, Lassen U, Mortensen J, Larsen S, Loft A, Bertelsen A, et al. Preoperative staging of lung cancer with combined PET-CT. *The New England Journal of Medicine* 2009;**361**(1):32–9. [3420766; PUBMED: 19571281]
Fischer BM, Loft A, Bertelsen AK, Mortensen J, Lassen U. Diagnostic accuracy of PET/CT assigning overall TNM-stage in patients with NSCLC according to the 1997 respectively the 2010 TNM classification system. *European Journal of Nuclear Medicine and Molecular Imaging* 2011; **Conference**:S123. [3420767]
Fischer BM, Mortensen J, Hansen H, Vilmann P, Larsen SS, Loft A, et al. Multimodality approach to mediastinal staging in non-small cell lung cancer. Faults and benefits of PET-CT: a randomised trial. *Thorax.* 2011;**66**(4):294–300. [3420768; PUBMED: 21169287]

Gunluoglu 2011 *{published data only}*
Gunluoglu MZ, Melek H, Medetoglu B, Demir A, Kara HV, Dincer SI. The validity of preoperative lymph node staging guidelines of European Society of Thoracic Surgeons in non-small-cell lung cancer patients. *European Journal of Cardio-thoracic Surgery* 2011;**40**(2):287–90. [3420770; PUBMED: 21185733]

Harders 2012 *{published data only}*
Harders SW. LUCIS: lung cancer imaging studies. *Danish Medical Journal* 2012;**59**(11):B4542. [3420772; PUBMED: 23171752]
Harders, SW, Madsen HH, Hjorthaug K, Arveschoug AK, Rasmussen TR, Meldgaard, P, et al. Mediastinal staging in non-small cell lung carcinoma: CT versus F-18-FDG PET/CT. Cancer Imaging in press. [3420773]

Hu 2011 *{published data only}*
Hu M, Han A, Xing L, Yang W, Fu Z, Huang C, et al. Value of dual-time-point FDG PET/CT for mediastinal nodal staging in non-small-cell lung cancer patients with lung comorbidity. *Clinical Nuclear Medicine* 2011;**36**(6): 429–33. [3420775; PUBMED: 21552018]
Hu M, Yu J, Xing L, Han A, Kong L. Diagnostic ability of dual-time-point fdg PET/CT for mediastinal lymph node metastases in non-small cell lung cancer patients. *International Journal of Radiation Oncology Biology Physics* 2010;**78**(3 Suppl):S500. [3420776; DOI: 10.1016/j.ijrobp.2010.07.1169]

Hwangbo 2009 *{published data only}*
Hwangbo B, Kim SK, Lee HS, Lee HS, Kim MS, Lee JM, et al. Application of endobronchial ultrasound-guided transbronchial needle aspiration following integrated PET/CT in mediastinal staging of potentially operable non-small cell lung cancer. *Chest* 2009;**135**(5):1280–7. [3420778; PUBMED: 19118267]

Iskender 2012 *{published data only}*
Iskender I, Kadioglu SZ, Cosgun T, Kapicibasi HO, Sagiroglu G, Kosar A, et al. False-positivity of mediastinal lymph nodes has negative effect on survival in potentially resectable non-small cell lung cancer. *European Journal of Cardio-Thoracic Surgery* 2012;**41**(4):874–9. [3420780; PUBMED: 22423060]
Iskender I, Kadioglu SZ, Kosar A, Atasalihi A, Kir A. Is there any maximum standardized uptake value variation among positron emission tomography scanners for mediastinal staging in non-small cell lung cancer?. *Interactive Cardiovascular and Thoracic Surgery* 2011;**12**(6): 965-9. [3420781; PUBMED: 21441257]
Iskender I, Kapicibasi HO, Kadioglu SZ, Sevilgen G, Tezel C, Kosar A, et al. Comparison of integrated positron emission tomography/computed tomography and mediastinoscopy in mediastinal staging of non-small cell lung cancer: analysis of 212 patients. *Acta Chirurgica Belgica* 2012;**112**(3):219–25. [3420782; PUBMED: 22808763]

Jeon 2010 *{published data only}*
Jeon TY, Lee KS, Yi CA, Chung MP, Kwon OJ, Kim B-T, et al. Incremental value of PET/CT over CT for mediastinal nodal staging of non-small cell lung cancer: comparison between patients with and without idiopathic pulmonary fibrosis. *American Journal of Roentgenology* 2010;**195**(2): 370–6. [3420784; PUBMED: 20651192]

Kim 2007 *{published data only}*
Kim BT, Lee KS, Shim SS, Choi JY, Kwon OJ, Kim H, et al. Stage T1 non-small cell lung cancer: preoperative mediastinal nodal staging with integrated FDG PET/

CT--a prospective study. *Radiology* 2006;**241**(2):501–9. [3420786; PUBMED: 16966480]
Kim YK, Lee KS, Kim BT, Choi JY, Kim H, Kwon OJ, et al. Mediastinal nodal staging of nonsmall cell lung cancer using Integrated 18F-FDG PET/CT in a tuberculosis-endemic country: diagnostic efficacy in 674 patients. *Cancer* 2007; **109**(6):1068–77. [3420787; PUBMED: 7311309]

Koksal 2013 *{published data only}*
Koksal D, Demirag F, Bayiz H, Ozmen O, Tatci E, Berktas B, et al. The correlation of SUVmax with pathological characteristics of primary tumor and the value of tumor/lymph node SUVmax ratio for predicting metastasis to lymph nodes in resected NSCLC patients. *Journal of Cardiothoracic Surgery* 2013;**8**:63–70. [3420789; PUBMED: 23557204]

Kuo 2012 *{published data only}*
Kuo WH, Wu YC, Wu CY, Ho KC, Chiu PH, Wang CW, et al. Node/aorta and node/liver SUV ratios from (18)F-FDG PET/CT may improve the detection of occult mediastinal lymph node metastases in patients with non-small cell lung carcinoma. *Academic Radiology* 2012;**19**(6): 685–92. [3420791; PUBMED: 22459646]

Lee 2007 *{published data only}*
Lee BE, von Haag D, Lown T, Lau D, Calhoun R, Follette D. Advances in positron emission tomography technology have increased the need for surgical staging in non-small cell lung cancer. *The Journal of Thoracis and Cardiovascular Surgery* 2007;**133**(3):746–52. [3420793; PUBMED: 17320577]

Lee 2009a *{published data only}*
Lee JW, Kim BS, Lee DS, Chung JK, Lee MC, Kim S, et al. 18F-FDG PET/CT in mediastinal lymph node staging of non-small-cell lung cancer in a tuberculosis-endemic country: consideration of lymph node calcification and distribution pattern to improve specificity. *European Journal of Nuclear Medicine and Molecular Imaging* 2009;**36**(11): 1794–802. [3420795; PUBMED: 19430783]

Lee 2011 *{published data only}*
Lee SH, Min JW, Lee CH, Park CM, Goo JM, Chung DH, et al. Impact of parenchymal tuberculosis sequelae on mediastinal lymph node staging in patients with lung cancer. *Journal of Korean Medical Sciences* 2011;**26**(1): 67–70. [3420797; PUBMED: 21218032]

Lee 2012 *{published data only}*
Lee SM, Park CM, Paeng JC, Im HJ, Goo JM, Lee HJ, et al. Accuracy and predictive features of FDG-PET/CT and CT for diagnosis of lymph node metastasis of T1 non-small-cell lung cancer manifesting as a subsolid nodule. *European Radiology* 2012;**22**(7):1556–63. [3420799; PUBMED: 22358427]

Li 2010 *{published data only}*
Li XD, Yin JL, Liu WK, Ouyang X, Zhou Z, Qiao GB, et al. [Value of positron emission tomography-computed tomography in the diagnosis of mediastinal lymph node metastasis of non-small cell lung cancer]. *Journal of Southern Medical University* 2010;**30**(3):506–8. [3420801; PUBMED: 20335121]

Li 2012a *{published data only}*
Li M, Wu N, Liu Y, Zheng R, Liang Y, Zhang W, et al. Regional nodal staging with 18F-FDG PET-CT in non-small cell lung cancer: Additional diagnostic value of CT attenuation anddual-time-point imaging. *European Journal of Radiology* 2012;**81**(8):1886–90. [3420803; PUBMED: 21511421]

Morikawa 2009 *{published data only}*
Morikawa M, Demura Y, Ishizaki T, Ameshima S, Miyamori I, Sasaki M, et al. The effectiveness of 18F-FDG PET/CT combined with STIR MRI for diagnosing nodal involvement in the thorax. *Journal of Nuclear Medicine* 2009;**50**(1):81–7. [3420805; PUBMED: 19091887]

Ohnishi 2011 *{published data only}*
Ohnishi R, Yasuda I, Kato T, Tanaka T, Kaneko Y, Suzuki T, et al. Combined endobronchial and endoscopic ultrasound-guided fine needle aspiration for mediastinal nodal staging of lung cancer. *Endoscopy* 2011;**43**(12): 1082–89. [3420807; PUBMED: 21971924]

Ohno 2011 *{published data only}*
Ohno Y, Koyama H, Yoshikawa T, Nishio M, Aoyama N, Onishi Y, et al. N stage disease in patients with non-small cell lung cancer: efficacy of quantitative and qualitative assessment with STIR turbo spin-echo imaging, diffusion-weighted MR imaging, and fluorodeoxyglucose PET/CT. *Radiology* 2011;**261**(2):605–15. [3420809; PUBMED: 21926377]

Ose 2012 *{published data only}*
Ose N, Sawabata N, Minami M, Inoue M, Shintani Y, Kadota Y, Okumura M. Lymph node metastasis diagnosis using positron emission tomography with 2-[F-18] fluoro-2-deoxy-D-glucose as a tracer and computed tomography in surgical cases of non-small cell lung cancer. *European Journal of Cardio-Thoracic Surgery* 2012;**42**(1):89–92. [3420811; PUBMED: 22290887]

Ozkan 2011 *{published data only}*
Ozkan EA, Araz M, Soydal C, Aras G. A retrospective analysis of 18F-FDG PET/CT in the primary staging of non-small cell lung cancer (NSCLC). *European Journal of Nuclear Medicine and Molecular Imaging* 2011;**Conference**: S125. [3420813]

Perigaud 2009 *{published data only}*
Perigaud C, Bridji B, Roussel JC, Sagan C, Mugniot A, Duveau D, et al. Prospective preoperative mediastinal lymph node staging by integrated positron emission tomography-computerised tomography in patients with non-small-cell lung cancer. *European Journal of Cardio-Thoracic Surgery* 2009;**36**(4):731–6. [3420815; PUBMED: 19632852]

Plathow 2008 *{published data only}*
Plathow C, Aschoff P, Lichy MP, Eschmann S, Hehr T, Brink I, et al. Positron emission tomography/computed tomography and whole-body magnetic resonance imaging in staging of advanced nonsmall cell lung cancer-initial results.

Investigative Radiology 2008;**43**(5):290–7. [3420817; PUBMED: 18424949]

Sanli 2009 *{published data only}*
Sanli M, Isik AF, Zincirkeser S, Elbek O, Mete A, Tuncozgur B, et al. Reliability of positron emission tomography-computed tomography in identification of mediastinal lymph node status in patients with non-small cell lung cancer. *The Journal of Thoracic and Cardiovascular Surgery* 2009;**138**(5):1200–5. [3420819; PUBMED: 19660381]

Saydam 2012 *{published data only}*
Saydam O, Gokce M, Kilicgun A, Tanriverdi O. Accuracy of positron emission tomography in mediastinal node assessment in coal workers with lung cancer. *Medical Oncology* 2012;**29**(2):589–94. [3420821; PUBMED: 21380783]

Shin 2008 *{published data only}*
Shin KM, Lee KS, Shim YM, Kim J, Kim BT, Kwon OJ, et al. FDG PET/CT and mediastinal nodal metastasis detection in stage T1 non-small cell lung cancer: prognostic implications. *Korean Journal of Radiology* 2008;**9**(6):481–9. [3420823; PUBMED: 19039263]

Sommer 2012 *{published data only}*
Sommer G, Wiese M, Winter L, Lenz C, Klarhofer M, Forrer F, et al. Preoperative staging of non-small-cell lung cancer: comparison of whole-body diffusion-weighted magnetic resonance imaging and 18F-fluorodeoxyglucose-positron emission tomography/computed tomography. *European Radiology* 2012;**22**(12):2859–67. [3420825; PUBMED: 22772365]

Subedi 2009 *{published data only}*
Subedi N, Scarsbrook A, Darby M, Korde K, Mc SP, Muers MF. The clinical impact of integrated FDG PET-CT on management decisions in patients with lung cancer. *Lung Cancer* 2009;**64**(3):301–7. [3420827; PUBMED: 19004519]

Tasci 2010 *{published data only}*
Tasci E, Tezel C, Orki A, Akin O, Falay O, Kutlu CA. The role of integrated positron emission tomography and computed tomography in the assessment of nodal spread in cases with non-small cell lung cancer. *Interactive Cardiovascular and Thoracic Surgery* 2010;**10**(2):200–3. [3420829; PUBMED: 19933240]

Toba 2010 *{published data only}*
Toba H, Kondo K, Otsuka H, Takizawa H, Kenzaki K, Sakiyama S, et al. Diagnosis of the presence of lymph node metastasis and decision of operative indication using fluorodeoxyglucose-positron emission tomography and computed tomography in patients with primary lung cancer. *Journal of Medical Investigation* 2010;**57**(3-4):305–13. [3420831; PUBMED: 20847531]

Tournoy 2007 *{published data only}*
Tournoy KG, Maddens S, Gosselin R, Van Maele G, van Meerbeeck JP, Kelles A. Integrated FDG-PET/CT does not make invasive staging of the intrathoracic lymph nodes in non-small cell lung cancer redundant: a prospective study. *Thorax* 2007;**62**(8):696–701. [3420833; PUBMED: 17687098]

Uruga 2011 *{published data only}*
Uruga H. PET/CT for mediastinal lymph node staging in non-small-cell lung cancer with interstitial pneumonia. *American Journal of Respiratory and Critical Care Medicine* 2011;**Conference**:1. [3420835]

Uskul 2009 *{published data only}*
Üskül BT, Baysungur V, Aksoy F, Turan FE, Sevilgen G, Turker H, et al. Combined use of transbronchial needle aspiration and PET/CT in mediastinal nodal staging of non small cell lung cancer. *Multidisciplinary Respiratory Medicine* 2009;**4**(1):8–14. [3420837]

Usuda 2013 *{published data only}*
Usuda K, Sagawa M, Motono N, Ueno M, Tanaka M, Machida Y, et al. Advantages of diffusion-weighted imaging over positron emission tomography-computer tomography in assessment of hilar and mediastinal lymph nodes in lung cancer. *Annals of Surgical Oncology* 2013;**20**(5):1676–83. [3420839; PUBMED: 23242821]
Usuda K, Zhao XT, Sagawa M, Matoba M, Kuginuki Y, Taniguchi M, et al. Diffusion-weighted imaging is superior to positron emission tomography in the detection and nodal assessment of lung cancers. *The Annals of Thoracic Surgery* 2011;**91**(6):1689–95. [3420840; PUBMED: 21619964]

Yang 2008 *{published data only}*
Yang W, Fu Z, Yu J, Yuan S, Zhang B, Li D, et al. Value of PET/CT versus enhanced CT for locoregional lymph nodes in non-small cell lung cancer. *Lung Cancer* 2008;**61**(1):35–43. [3420842; PUBMED: 18177978]
Yang WF, Tan GZ, Fu Z, Yu JM. [Evaluation of the diagnostic value of (18)F-FDG PET-CT and enhanced CT for staging of lymph node metastasis in non small cell lung cancer]. *Zhonghua zhong liu za zhi [Chinese Journal of Oncology]* 2009;**31**(12):925–8. [3420843; PUBMED: 20193335]

Yang 2010 *{published data only}*
Yang W, Zhang Y, Fu Z, Yu J, Sun X, Mu D, et al. Imaging of proliferation with 18F-FLT PET/CT versus 18F-FDG PET/CT in non-small-cell lung cancer. *European Journal of Nuclear Medicine and Molecular Imaging* 2010;**37**(7):1291–9. [3420845; PUBMED: 20309686]

Yi 2008 *{published data only}*
Yi CA, Shin, KM, Lee KS, Kim BT, Kim H, et al. Non-small cell lung cancer staging: efficacy comparison of integrated PET/CT versus 3.0-T whole-body MR imaging. *Radiology* 2008;**248**(2):632–42. [3420847; PUBMED: 18552311]

References to studies excluded from this review

Agraval 2012 *{published data only}*
Agraval J. Accuracy of FDG-PET/CT nodal staging for non-small-cell lung cancer in surgically resectable cases: a local experience. *Journal of Medical Imaging and Radiation Oncology* 2012;**Conference**:78. [3420849]

Ahmed 2010 *{published data only}*
Ahmed I, Stuart R, Muller M, Stamenkovic S. PET/CT has enabled the development of a non invasive strategy for mediastinal staging in non small cell lung cancer (NSCLC). *Lung Cancer* 2010;**Conference**:S16. [3420851; DOI: 10.1016/S0169-5002(10)70049-0]
Ahmed IM, Stuart R, Stamenkovic S. Impact of NICE guidelines on mediastinal staging strategies. *Lung Cancer* 2010;**Conference**:S16–7. [3420852; DOI: 10.1016/S0169-5002(10)70050-7]

Akpinar 2013 *{published data only}*
Akpinar D, Ceylan KC, Duman E, Unsal S, Kaya SO. [The accuracy of positron emission tomography in mediastinal staging of non-small cell lung cancer]. *Türk Göğüs Kalp Damar Cerrahisi Dergisi [Turkish Journal of Thoracic and Cardiovascular Surgery]* 2013;**21**(1):100–5. [3420854]

Al-Ibraheem 2012 *{published data only}*
Al-Ibraheem A. Performance of FDG PET/CT in mediastinal nodal staging in NSCLC; comparison with chest CE-CT scan and EBUS/TBNA. *European Journal of Nuclear Medicine and Molecular Imaging* 2012;**Conference**:S456. [3420856]

Allen-Auerbach 2006 *{published data only}*
Allen-Auerbach M, Yeom K, Park J, Phelps M, Czernin J. Standard PET/CT of the chest during shallow breathing is inadequate for comprehensive staging of lung cancer. *Journal of Nuclear Medicine* 2006;**47**(2):298–301. [3420858; PUBMED: 16455636]

Al-Sarraf 2008 *{published data only}*
Al-Sarraf N, Gately K, Lucey J, Wilson L, McGovern E, Young V. Lymph node staging by means of positron emission tomography is less accurate in non-small cell lung cancer patients with enlarged lymph nodes: analysis of 1,145 lymph nodes. *Lung Cancer* 2008;**60**(1):62–8. [3420860; PUBMED: 17920724]
Al-Sarraf N, Gately K, Lucey J, Wilson L, McGovern E, Young V. Mediastinal lymph node staging by means of positron emission tomography is less sensitive in elderly patients with non-small-cell lung cancer. *Clinical Lung Cancer* 2008;**9**(1):39–43. [3420861; PUBMED: 18282357]

An 2008 *{published data only}*
An YS, Sun JS, Park KJ, Hwang SC, Park KJ, Sheen SS, et al. Diagnostic performance of (18)F-FDG PET/CT for lymph node staging in patients with operable non-small-cell lung cancer and inflammatory lung disease. *Lung* 2008;**186**(5):327–36. [3420863; PUBMED: 18670805]

Antoch 2003 *{published data only}*
Antoch G, Stattaus J, Nemat AT, Marnitz S, Beyer T, Kuehl H, et al. Non-small cell lung cancer: dual-modality PET/CT in preoperative staging. *Radiology* 2003;**229**(2):526–33. [3420865; PUBMED: 14512512]

Aquino 2003 *{published data only}*
Aquino SL, Asmuth JC, Alpert NM, Halpern EF, Fischman AJ. Improved radiologic staging of lung cancer with 2-[18F]-flouro-2-deoxy-D-glucose-positron emission tomography and computed tomography registration. *Journal of Computer Assisted Tomography* 2003;**27**(4):479–84. [3420867; PUBMED: 12886128]

Balcil 2012 *{published data only}*
Balcil TA. The relationship between Lymph node staging and tumor SUVmax in NSCLC. *European Journal of Nuclear Medicine and Molecular Imaging* 2012;**Conference**:S510. [3420869]

Beyer 2010 *{published data only}*
Beyer F, Buerke B, Gerss J, Scheffe K, Puesken M, Weckesser M, et al. Prediction of lymph node metastases in NSCLC. Three dimensional anatomical parameters do not substitute FDG-PET-CT. *Nuklearmedizin. Nuclear medicine* 2010;**49**(1):41–8. [3420871; PUBMED: 20087533]

Bhatt 2012 *{published data only}*
Bhatt MK. Objective FDG-PET analysis of mediastinal and hilar lymph nodes sampled by endobronchial ultrasound transbronchial needle aspiration. *Internal Medicine Journal* 2012;**Conference**:10. [3420873]

Booth 2009 *{published data only}*
Booth K, Hanna G, McGonigle N, McGuigan J, McManus K, O'Sullivan J, et al. What is the shelf life of PET/CT staging of the mediastinum in non-small cell lung cancer? . *Interactive Cardiovascular and Thoracic Surgery* 2009; **Conference**:S14–5. [3420875]

Boulougouri 2012 *{published data only}*
Boulougouri K. Establishing a threshold SUVmax value on 18FDG-PET/CT for identifying metastatic lymph nodes in surgically resectable Non Small Cell Lung Cancer. *European Journal of Nuclear Medicine and Molecular Imaging* 2012; **Conference**:S517. [3420877]

Bryant 2006b *{published data only}*
Bryant AS, Cerfolio RJ. The clinical stage of non-small cell lung cancer as assessed by means of fluorodeoxyglucose-positron emission tomographic/computed tomographic scanning is less accurate in cigarette smokers. *The Journal of Thoracic and Cardiovascular Surgery* 2006;**132**(6):1363–8. [3420879; PUBMED: 17140957]

Carrillo 2012 *{published data only}*
Carrillo SA, Daniel, VC, Hall N, Hitchcock CL, Ross P Jr, Kassis ES. Fusion Positron emission/Computed tomography underestimates the presence of hilar nodal metastases in patients with resected non-small cell lung cancer. *The Annals of Thoracic Surgery* 2012;**93**(5):1621–4. [3420881; PUBMED: 22429676]

Cerfolio 2004 *{published data only}*
Cerfolio RJ, Ojha B, Bryant AS, Raghuveer V, Mountz JM, Bartolucci AA. The accuracy of integrated PET-CT compared with dedicated PET alone for the staging of patients with nonsmall cell lung cancer. *The Annals of Thoracic Surgery* 2004;**78**(3):1017–23. [3420883; PUBMED: 15337041]

Cerfolio 2006 *{published data only}*
Cerfolio, RJ, Bryant AS, Eloubeidi MA. Routine mediastinoscopy and esophageal ultrasound fine-needle

aspiration in patients with non-small cell lung cancer who are clinically N2 negative: aprospective study. *Chest* 2006; **130**(6):1791–95. [3420885; PUBMED: 17166998]

Cerfolio 2007 *{published data only}*
Cerfolio RJ, Bryant AS. Ratio of the maximum standardized uptake value on FDG-PET of the mediastinal (N2) lymph nodes to the primary tumor may be a universal predictor of nodal malignancy in patients with nonsmall-cell lung cancer. *The Annals of Thoracic Surgery* 2007;**83**(5):1826–9. [3420887; PUBMED: 17462407]

Cerfolio 2008 *{published data only}*
Cerfolio RJ, Bryant AS. Is palpation of the nonresected pulmonary lobe(s) required for patients with non-small cell lung cancer? A prospective study. *The Journal of Thoracic and Cardiovascular Surgery* 2008;**135**(2):261–68. [3420889; PUBMED: 18242247]

Cetinkaya 2011 *{published data only}*
Cetinkaya E, Seyhan EC, Ozgul A, Gencoglu A, Ozgul G, Cam E, Kamiloglu E. Efficacy of convex probe endobronchial ultrasound (CP-EBUS) assisted transbronchial needle aspiration for mediastinal staging in non-small cell lung cancer cases with mediastinal lymphadenopathy. *Annals of Thoracic and Cardiovascular Surgery* 2011;**17**(3):236–42. [3420891; PUBMED: 21697783]

Ceylan 2012 *{published data only}*
Ceylan N, Doğan S, Kocaçelebi K, Sava R, Çakan A, Çağrici U. Contrast enhanced CT versus integrated PET-CT in pre-operative nodal staging of non-small cell lung cancer. *Diagnostic and Interventional Radiology* 2012;**18**(5):435–40. [3420893; PUBMED: 22374706]

Chiba 2010 *{published data only}*
Chiba K, Isoda M, Chiba M, Kanematsu T, Eguchi S. Significance of PET/CT in determining actual TNM staging for patients with various lung cancers. *International Surgery* 2010;**95**(3):197–204. [3420895; PUBMED: 21066996]

Colville 2013 *{published data only}*
Colville D. Impact of F18-FDG PET/CT in suspected or confirmed early stage non-small cell lung cancer. *Nuclear Medicine Communications* 2013;**Conference**:4. [3420897]

Cömert 2012 *{published data only}*
Cömert SS, Caglayan B, Fidan A, Salepci B, Dogan C, Demirhan R, Ece D. [A comparison of endobronchial ultrasound-guided transbronchial needle aspiration and integrated positron emission tomography-computed tomography in the diagnosis of malignant mediastinal/hilar lymph nodes]. *Türk Göğüs Kalp Damar Cerrahisi Dergisi [Turkish Journal of Thoracic and Cardiovascular Surgery]* 2012;**20**(4):843–49. [3420899]

Delgado-Bolton 2010 *{published data only}*
Delgado-Bolton RC. Pre-treatment staging with 18F-FDG PET/CT in non-small cell lung cancer (NSCLC): Comparative analysis of the 6th and 7th edition of the TNM classification. *European journal of nuclear medicine and molecular imaging* 2010;**Conference**(var.pagings):S257. [3420901]

Duan 2012 *{published data only}*
Duan X-B. Analysis of risk factors for mediastinal lymph nodes metastases in non-small cell lung cancer patients with 18F-FDG PET/CT. *Chinese Journal of Medical Imaging Technology* 2012;**28**(6):1135–9. [3420903]
Duan X-B. Optimal threshold in diagnosing mediastinal lymph node metastasis of non-small cell lung cancer with maximum standardized uptake value. *Chinese Journal of Medical Imaging Technology* 2012;**28**(5):916–20. [3420904]

Faber 2011 *{published data only}*
Faber DL, Kremer R, Orlovsky M, Lapidot M, Guralnik L, Kagna O, et al. Does PET-CT based clinical staging of non-small cell lung cancer obviate invasive procedures?. *Lung Cancer* 2011;**Conference**:S33–4. [3420906]

Flechsig 2012 *{published data only}*
Flechsig P. 3D CT-Histogram analysis enables to distinguish affected and FDG-negative lymph nodes in patients with lung cancer. *European Journal of Nuclear Medicine and Molecular Imaging* 2012;**Conference**:S512–3. [3420908]

Gómez-Caro 2012 *{published data only}*
Gómez-Caro A, Boada M, Cabañas M, Sanchez M, Arguis P, Lomeña F, et al. False-negative rate after positron emission tomography/computer tomography scan for mediastinal staging in cI stage non-small-cell lung cancer. *European Journal of Cardio-Thoracic Surgery* 2012;**42**(1):93–100. [3420910; PUBMED: 22290911]

Gregory 2012 *{published data only}*
Gregory DL, Hicks RJ, Hogg A, Binns DS, Shum PL, Milner A, et al. Effect of PET/CT on management of patients with non-small cell lung cancer: results of a prospective study with 5-year survival data. *Journal of Nuclear Medicine* 2012;**53**(7):1007–15. [3420912; PUBMED: 22677701]

Günlüoğlu 2010 *{published data only}*
Günlüoğlu MZ, Melek H, Turna A, Medetoğlu B, Kara HV, Demir A, et al. Is there a need for invasive mediastinal staging in centrally located non-small cell lung cancer?. *Chest* 2010;**138**:664A. [3420914; DOI: 10.1378/chest.10138]

Günlüoğlu 2011b *{published data only}*
Günlüoğlu, MZ, Melek H, Demir A, Medetoğlu B, Kara HV, Dinçer Sİ . [The accuracy and cost of mediastinal staging strategies for non-small cell lung cancer]. *Türk Göğüs Kalp Damar Cerrahisi Dergisi [Turkish Journal of Thoracic and Cardiovascular Surgery]* 2011;**19**: 397–404. [3420916]

Halpern 2005 *{published data only}*
Halpern BS, Schiepers C, Weber WA, Crawford TL, Fueger BJ, Phelps ME, et al. Presurgical staging of non-small cell lung cancer: positron emission tomography, integrated positron emission tomography/CT, and software

image fusion. *Chest* 2005;**128**(4):2289–97. [3420918; PUBMED: 16236886]

Hong 2010 *{published data only}*
Hong SP, Song HC, Chong A, Oh JR, Kim JH, Yoo SU, et al. Diagnostic accuracy of preoperative intrathoracic lymph node staging on 18F-FDG PET/CT in patients with non-small cell lung cancer. *Molecular Imaging and Biology* 2010; **Conference**:S1624. [3420920]

Hu 2008 *{published data only}*
Hu M, Yu JM, Liu NB, Liu LP, Guo HB, Yang GR, et al. [Significance of dual-time-point 18F-FDG PET imaging in evaluation of hilar and mediastinal lymph node metastasis in non-small-cell lung cancer]. *Zhonghua Zhong Liu Za Zhi [Chinese Journal of Oncology]* 2008;**30**(4):306–9. [3420922; PUBMED: 18788639]

Huang 2012 *{published data only}*
Huang TW, Hsieh CM, Chang H, Cheng YL, Tzao C, Huang WS. Standard uptake value of positron emission tomography in clinical stage I lung cancer: clinical application and pathological correlation. *European Journal of Cardio-Thoracic Surgery* 2012;**41**(4):869–73. [3420924; PUBMED: 22219418]

Kasai 2010 *{published data only}*
Kasai T, Motoori K, Horikoshi T, Uchiyama K, Yasufuku K, Takiguchi Y, et al. Dual-time point scanning of integrated FDG PET/CT for the evaluation of mediastinal and hilar lymph nodes in non-small cell lung cancer diagnosed as operable by contrast-enhanced CT. *European Journal of Radiology* 2010;**75**(2):143–6. [3420926; PUBMED: 19446975]

Kelly 2006 *{published data only}*
Kelly A, Cachin F, de Freitas D, Geissler B, Bapt A, Karidioula I, et al. PET-CT in non-small-cell-lung cancer: Clermontoise experience. *Medecine Nucleaire* 2006;**30**(28): 97–106. [3420928]

Kim 2011 *{published data only}*
Kim HK, Choi YS, Kim K, Shim YM, Park K, Ahn YC, et al. Outcomes of mediastinoscopy and surgery with or without neoadjuvant therapy in patients with non-small cell lung cancer who are N2 negative on positron emission tomography and computed tomography. *Journal of Thoracic Oncology* 2011;**6**(2):336–342. [3420930; PUBMED: 21164366]

Kim 2012 *{published data only}*
Kim YN, Yi CA, Lee KS, Kwon O, Lee HY, Kim BT, et al. A proposal for combined MRI and PET/CT interpretation criteria for preoperative nodal staging in non-small-cell lung cancer. *European adiology* 2012;**22**(7):1537–46. [3420932; PUBMED: 22367469]

Kim 2012a *{published data only}*
Kim DW, Kim WH, Kim CG. Dual-time-point FDG PET/CT: Is it useful for lymph node staging in patients with non-small-cell lung cancer?. *Nuclear Medicine and Molecular Imaging* 2012;**46**(3):196–200. [3420934; PUBMED: 24900060]

Kim 2012b *{published data only}*
Kim D, Song B, Hong C, Lee S, Ahn B, Lee J. Prediction of occult lymph node metastasis in clinically N0 squamous cell lung carcinoma. *European Journal of Nuclear Medicine and Molecular Imaging* 2012;**Conference**:S568. [3420936]

Kommata 2011 *{published data only}*
Kommata S. Clinical value of 18F- FDG-PET/CT scan in nodal staging of patients with non small cell lung cancer. An explorative analysis of retrospective data of 401 patients. *Nuklearmedizin* 2011;**Conference**:6. [3420938]
Kommata S. Clinical value of 18F-FDG-PET/CT scan in nodal staging of patients with non small cell lung cancer. Evaluation of retrospective data of 401 patients. *European Journal of Nuclear Medicine and Molecular Imaging* 2012; **Conference**:S402–3. [3420939]

Krueger 2006 *{published data only}*
Krueger S, Buck A, Pauls S, Halter G, Schumann C, Kropf C, et al. Integrated FDG-PET/CT for improved T and N staging of non-small-cell lung cancer. *European Respiratory Journal* 2006;**28**:S784. [3420941]

Lapinska 2011 *{published data only}*
Lapinska GE, Fijolek-Warszewska A, Bryszewska M, Kozlowicz-Gudzinska I, Glogowski M, Zmijewski M, et al. Diagnostic value of PET/CT in preoperative staging in patients with non-small cell lung cancer. *European Journal of Nuclear Medicine and Molecular Imaging* 2011;**Conference**: S359. [3420943]

Lardinois 2003 *{published data only}*
Lardinois D, Weder W, Hany TF, Kamel EM, Korom S, Seifert B, et al. Staging of non-small-cell lung cancer with integrated positron-emission tomography and computed tomography. *The New England Journal of Medicine* 2003; **348**:2500–7. [3420945; PUBMED: 12815135]

Lasnon 2012 *{published data only}*
Lasnon C. Impact of point spread function reconstruction on pre-therapeutic nodal staging with 18 F -FDG PET in non-small cell lung cancer. *European Journal of Nuclear Medicine and Molecular Imaging* 2010;**Conference**:S318. [3420948]
Lasnon C, Hicks RJ, Beauregard JM, Milner A, Paciencia M, Guizard AV, et al. Impact of point spread function reconstruction on thoracic lymph node staging with 18F-FDG PET/CT in non-small cell lung cancer. *Clinical Nuclear Medicine* 2012;**37**(10):971–6. [3420947; PUBMED: 22899197]

Lebioda 2013 *{published data only}*
Lebioda A, Makarewicz R, Malkowski B, Dancewicz M, Kowalewski J, Windorbska W. Measurement of primary tumor volume by PET-CT to evaluate risk of mediastinal nodal involvement in NSCLC patients with clinically negative N2 lymph nodes. *Reports of Practical Oncology and Radiotherapy: Journal of Great Poland Cancer Center in Poznan and Polish Society of Radiation Oncology* 2013;**18**(2): 76–81. [3420950; PUBMED: 24416539]

Lee 2004 *{published data only}*
Lee EJ, Choi JY, Choi Y, Choe YS, Lee KH, Lee KS, et al. Improvement in the differentiation of malignant nodes from benign nodes preoperatively in non-small cell lung cancer (NSCLC) using FDG PET/CT. *Journal of Nuclear Medicine* 2004;**45**:374P. [3420952]

Lee 2008 *{published data only}*
Lee BE, Redwine J, Foster C, Abella E, Lown T, Lau D, et al. Mediastinoscopy might not be necessary in patients with non-small cell lung cancer with mediastinal lymph nodes having a maximum standardized uptake value of less than 5.3. *The Journal of Thoracic and Cardiovascular Surgery* 2008;**135**(3):615–9. [3420954; PUBMED: 18329480]

Lee 2009b *{published data only}*
Lee HJ, Kim YT, Kang WJ, Lee HJ, Kang CH, Kim JH. Integrated positron-emission tomography for nodal staging in lung cancer. *Asian Cardiovascular and Thoracic Annals* 2009;**17**(6):622–6. [3420956; PUBMED: 20026540]

Li 2009 *{published data only}*
Li M, Wu N, Liang Y, Zheng R, Liu Y, Zhang WJ, et al. [Value of (18)F-FDG PET-CT in the preoperative N staging of non-small cell lung cancer]. *Zhonghua Zhong Liu Za Zhi [Chinese Journal of Oncology]* 2009;**31**(4):288–92. [3420958; PUBMED: 19615286]

Li 2011 *{published data only}*
Li X. Mediastinal lymph nodes staging by 18F-FDG PET/CT for early-stage non-small cell lung cancer: A multicenter study. *International Journal of Radiation Oncology Biology Physics* 2011;**Conference**:2. [3420960]
Li X. Mediastinal lymph nodes staging by 18F-FDG PET/CT for early-stage non-small cell lung cancer: A multicenter study. *Journal of Clinical Oncology* 2011;**Conference**:15. [3420961]

Li 2012b *{published data only}*
Li X, Zhang H, Xing L, Ma H, Xie P, Zhang L, et al. Mediastinal lymph nodes staging by 18F--FDG PET/CT for early stage non-small cell lung cancer: a multicenter study. *Radiotherapy and Oncology* 2012;**102**(2):246–50. [3420963; 22100657]

Li 2012c *{published data only}*
Li SB, He JX, Li SY, Chen HZ, Yin WQ, Cheng XL, et al. [Value of endobronchial ultrasound-transbronchial needle aspiration biopsy for diagnosis of PET-CT positive mediastinal lymph nodes]. *Zhonghua Zhong Liu Za Zhi [Chinese Journal of Oncology]* 2012;**34**(8):613–5. [3420965; PUBMED: 23158997]

Lin 2012 *{published data only}*
Lin WY, Hsu WH, Lin KH, Wang SJ. Role of preoperative PET-CT in assessing mediastinal and hilar lymph node status in early stage lung cancer. *Journal of the Chinese Medical Association* 2012;**75**(5):203–08. [3420967; PUBMED: 22632985]

Liu 2009 *{published data only}*
Liu BJ, Dong JC, Xu CQ, Zuo CT, Le JJ, Guan YH, et al. Accuracy of 18F-FDG PET/CT for lymph node staging in non-small-cell lung cancers. *Chinese Medical Journal* 2009; **122**(15):1749–54. [3420969; PUBMED: 19781319]

Low 2006 *{published data only}*
Low SY, Eng P, Keng GH, Ng DC. Positron emission tomography with CT in the evaluation of non-small cell lung cancer in populations with a high prevalence of tuberculosis. *Respirology* 2006;**11**(1):84–9. [3420971; PUBMED: 16423207]

Ma 2011 *{published data only}*
Ma W, Xu W, Zhu X, Dai D, Song X, Zhu L, et al. Differential diagnosis of mediastinal lymph node metastasis in lung cancer patients using 18F-FDG PET/CT with double quantitative SUVmax and CT value. *Chinese Journal of Clinical Oncology* 2011;**38**:284–7. [3420973]
Ma W, Xu W, Zhu X, Wang J, Dai D, Song X, et al. Diagnostic value of qualitative and quantitative analysis by 18F-FDG PET/CT for mediastinal lymph nodes of lung cancer. *Chinese Journal of Clinical Oncology* 2011;**38**:512–5. [3420974]

Maeda 2009 *{published data only}*
Maeda R, Isowa N, Onuma H, Miura H, Harada T, Touge H, et al. The maximum standardized 18F-fluorodeoxyglucose uptake on positron emission tomography predicts lymph node metastasis and invasiveness in clinical stage IA non-small cell lung cancer. *Interactive Cardiovascular and Thoracic Surgery* 2009;**9**(1): 79–82. [3420976; PUBMED: 19366724]

Mariam 2009 *{published data only}*
Mariam J, Peng C, Clarke J, Thompson JC. Pre-operative radiological staging (CT and PET-CT) compared with pathological staging in patients with resected non-small cell lung cancer attending a regional thoracic centre. *European Journal of Cancer* 2009;**Conference**:S518. [3420978]

Meduoye 2009 *{published data only}*
Meduoye A, Black E, Duffy J, Beggs D, Majewski A. A comparison of CT/FDG-PET scan staging with final pathological staging in patients following lung resection. *Interactive Cardiovascular and Thoracic Surgery* 2009; **Conference**:S41. [3420980]

Mendez 2012 *{published data only}*
Mendez M, Maximiano C, Huelves M, Doger de Speville DG, Ibeas P, Ruiz-Valdepeñas A, et al. Mediastinal lymph node staging in patients with non-small cell lung cancer (NSCLC) with F18-fluorodeoxyglucose positron emission tomography/computed tomography (FDG-PET/CT). *Journal of Clinical Oncology* 2012;**Conference**:15. [3420982]

Mi 2012 *{published data only}*
Mi B, Wan W, Yu C, You X, Jiang F, You Q. [The value of extra-lung lesions on ^{18} F-FDG PET/CT in improving diagnosis of lung cancer].. *Zhongguo Fei Ai Za Zhi [Chinese Journal of Lung Cancer]* 2012;**15**(2):78–83. [3420984; PUBMED: 22336234]

Moodie 2009 *{published data only}*
Moodie K, Cherk MH, Lau E, Turlakow A, Skinner S, Hicks R, et al. Evaluation of pulmonary nodules and lung cancer with one-inch crystal gamma coincidence positron emission tomography/CT versus dedicated positron emission tomography/CT. *Journal of Medical Imaging and Radiation Oncology* 2009;**53**(1):32–39. [3420986; PUBMED: 19453526]

Moreno Garcia 2009 *{published data only}*
Moreno Garcia V, Castro JD, Feliu J, Belda C, Barriuso J, Marin MD, et al. Accuracy of integrated PET-CT for mediastinal lymph node metastases in non-small cell lung cancer. *European Journal of Cancer* 2009;**Conference**:S169. [3420988]

Morikawa 2011 *{published data only}*
Morikawa M. Clinical usefulness of positron emission tomography/computed tomography and magnetic resonance imaging for lung cancer. *American Journal of Respiratory and Critical Care Medicine* 2011;**Conference**:1. [3420990]

Nakajima 2011 *{published data only}*
Nakajima E, Sakata Y, Saji H, Kato Y, Suga Y, Uchino Y, et al. The consensus to evaluate mediastinal lymph node metastasis in primary lung cancer with integrated FDG PET/CT. *American Journal of Respiratory and Critical Care Medicine* 2011;**Conference**:1. [3420992]

Nakamura 2008 *{published data only}*
Nakamura H, Taguchi M, Kitamura H, Nishikawa J. Fluorodeoxyglucose positron emission tomography integrated with computed tomography to determine resectability of primary lung cancer. *General Thoracic and Cardiovascular Surgery* 2008;**56**(8):404–9. [3420994; PUBMED: 18696206]

Nomori 2008 *{published data only}*
Nomori H, Mori T, Ikeda K, Kawanaka K, Shiraishi S, Katahira K, et al. Diffusion-weighted magnetic resonance imaging can be used in place of positron emission tomography for N staging of non-small cell lung cancer with fewer false-positive results. *The Journal of Thoracic and Cardiovascular Surgery* 2008;**135**(4):816–22. [3420996; PUBMED: 18374761]

Ohno 2007 *{published data only}*
Ohno Y, Koyama H, Nogami M, Takenaka D, Yoshikawa T, Yoshimura M, et al. STIR turbo SE MR imaging vs. coregistered FDG-PET/CT: quantitative and qualitative assessment of N-stage in non-small-cell lung cancer patients. *Journal of Magnetic Resonance Imaging* 2007;**26**(4):1071–80. [3420998; PUBMED: 17896365]

Ozkan 2010 *{published data only}*
Ozkan ZG. How efficient is PET/CT in metabolic characterization of lung lesions?. *European Journal of Nuclear Medicine and Molecular Imaging* 2010;**Conference**: S423. [3421000]

Özta 2012 *{published data only}*
Özta S, Öztürk AV, Acartürk E, Tezel Y, Özdemir M, Ataç G, et al. The role of tumor SUVmax/lymph node SUVmax ratio viewed on PET-CT in the detection of mediastinal metastasis in patients with lung cancer. *Türk Göğüs Kalp Damar Cerrahisi Dergisi [Turkish Journal of Thoracic and Cardiovascular Surgery]* 2012;**20**(3):544–51. [3421002]

Pauls 2012 *{published data only}*
Pauls S, Schmidt SA, Juchems MS, Klass O, Luster M, Reske SN, et al. Diffusion-weighted MR imaging in comparison to integrated [^{18} F]-FDG PET/CT for N-staging in patients with lung cancer. *European Journal of Radiology* 2012;**81**(1):178–82. [3421004; PUBMED: 20932700]

Pfannenberg 2007 *{published data only}*
Pfannenberg AC, Aschoff P, Brechtel K, Müller M, Bares R, Paulsen F, et al. Low dose non-enhanced CT versus standard dose contrast-enhanced CT in combined PET/CT protocols for staging and therapy planning in non-small cell lung cancer. *European Journal of Nuclear Medicine and Molecular Imaging* 2007;**34**(1):36–44. [3421006; PUBMED: 16896664]

Pozo-Rodriguez 2005 *{published data only}*
Pozo-Rodriguez F, Martin de Nicolás JL, Sánchez-Nistal MA, Maldonado A, Garcia de Barajas S, Calero-Garcia R, et al. Accuracy of helical computed tomography and [18F] fluorodeoxyglucose positron emission tomography for identifying lymph node mediastinal metastases in potentially resectable non-small-cell lung cancer. *Journal of Clinical Oncology* 2005;**23**(33):8348–56. [3421008; PUBMED: 16219937]

Prévost 2009 *{published data only}*
Prévost A, Papathanassiou D, Jovenin N, Menéroux B, Cuif-Job A, Bruna-Muraille C, et al. [Comparison between PET (-FDG) and computed tomography in the staging of lung cancer. Consequences for operability in 94 patients]. *Revue de Pneumologie Clinique* 2009;**65**(6):341–9. [3421010; PUBMED: 19995654]

Quaia 2008 *{published data only}*
Quaia E, Tona G, Gelain F, Lubin E, Pizzolato R, Boscolo E, et al. Integrated fluorine-18 fluorodeoxyglucose (18F-FDG) PET/CT compared to standard contrast-enhanced CT for characterization and staging of pulmonary tumors eligible for surgical resection. *Acta Radiologica* 2008;**49**(9): 995–1004. [3421012; PUBMED: 18651256]

Salman 2009 *{published data only}*
Salman KA, Steinmann CH, Sukumar VP, Schulthess GK, Steinert HC. PET/CT staging of T1-stage non-small cell lung cancer. *Internal Medicine Journal* 2009;**39**(Suppl4): A110. [3421014]

Sánchez Sánchez 2011 *{published data only}*
Sánchez Sánchez R, Rodríguez Fernández A, Gómez Río M, Alkurdi Martínez A, Castellón Rubioc VE, Ramos Font C, et al. [Utility of PET/CT for mediastinal staging of non-small cell lung cancer in stage III (N2)]. *Revista Espanola de Medicina Nuclear* 2011;**30**(4):211–6. [3421016; PUBMED: 21514978]

Schiavariello 2012 *{published data only}*
Schiavariello S. Lymph-node mediastinal staging in patients with NSCLC: 18-FDG PET-TC versus TC MDC. *European Journal of Nuclear Medicine and Molecular Imaging* 2012;**Conference**:S454. [3421018]

Schreyögg 2010 *{published data only}*
Schreyögg J, Weller J, Stargardt T, Herrmann K, Bluemel C, Dechow T, et al. Cost-effectiveness of hybrid PET/CT for staging of non-small cell lung cancer. *Journal of Nuclear Medicine* 2010;**51**(11):1668-75. [3421020; PUBMED: 21051648]

Schwenzer 2012 *{published data only}*
Schwenzer NF, Schraml C, Mueller M, Brendle C, Sauter A, Spengler W, et al. Pulmonary lesion assessment: comparison of whole-body hybrid MR/PET and PET/CT imaging--pilot study. *Radiology* 2012;**264**(2):551–58. [3421022; PUBMED: 22653189]

Shim 2005 *{published data only}*
Shim SS, Lee KS, Kim BT, Chung MJ, Lee EJ, Han J, et al. Non-small cell lung cancer: prospective comparison of integrated FDG PET/CT and CT alone for preoperative staging. *Radiology* 2005;**236**(3):1011–19. [3421024; PUBMED: 16014441]

Sit 2010 *{published data only}*
Sit AK, Sihoe AD, Suen WS, Cheng LC. Positron-emission tomography for lung cancer in a tuberculosis-endemic region. *Asian Cardiovascular and Thoracic Annals* 2010;**18**(1):33–8. [3421026; PUBMED: 20124294]

Sivrikoz 2010 *{published data only}*
Sivrikoz MC. Mediastinoscopy and F-18 PET-CT in NSCLC. *European Journal of Nuclear Medicine and Molecular Imaging* 2010;**Conference**(var.pagings):S318. [3421028]

Sivrikoz 2012 *{published data only}*
Sivrikoz C, Ak I, Simsek FS, Döner E, Dündar E. Is Mediastinoscopy Still the Gold Standard to Evaluate Mediastinal Lymph Nodes in Patients with Non-Small Cell Lung Carcinoma?. *The Thoracic and Cardiovascular Surgeon* 2012;**60**(2):116–21. [3421030; PUBMED: 21692019]

Steinert 2010 *{published data only}*
Steinert H. Outcome of PET/CT imaging in staging of non-small-cell lung cancer <= 3cm in diameter. *Nuklearmedizin* 2010;**Conference**:2. [3421032]

Tamura 2012 *{published data only}*
Tamura, M, Oda M, Matsumoto I, Waseda R, Watanabe G. Pattern and predictors of false positive lymph node involvement on positron emission tomography in patients with non-small cell lung Cancer. *The Thoracic and Cardiovascular Surgeon* 2012;**60**(2):105–10. [3421034; PUBMED: 21789758]

Tsutani 2012 *{published data only}*
Tsutani Y, Miyata Y, Nakayama H, Okumura S, Adachi S, Yoshimura M, et al. Prediction of pathologic node-negative clinical stage IA lung adenocarcinoma for optimal candidates undergoing sublobar resection. *The Journal of Thoracic and Cardiovascular Surgery* 2012;**144**(6):1365–71. [3421036; PUBMED: 22883546]

Vaz 2012 *{published data only}*
Vaz AP, Fernandes G, Souto Moura C, Bastos P, Queiroga H, Hespanhol V. Integrated PET/CT in non-small cell lung cancer staging--clinical and pathological agreement. *Revista Portuguesa de Pneumologia* 2012;**18**(3):109–114. [3421038; PUBMED: 22405953]

Ventura 2010 *{published data only}*
Ventura E, Islam T, Gee MS, Mahmood U, Braschi M, Harisinghani MG. Detection of nodal metastatic disease in patients with non-small cell lung cancer: comparison of positron emission tomography (PET), contrast-enhanced computed tomography (CT), and combined PET-CT. *Clinical Imaging* 2010;**34**(1):20–8. [3421040; PUBMED: 20122515]

Wang 2012 *{published data only}*
Wang F, Ma S, Shen L, Li N, Yang Z, Chen K. [Application of ^{18} F-FDG PET/CT in pulmonary disease: a report of 419 cases]. *Zhongguo Fei Ai Za Zhi [Chinese Journal of Lung Cancer]* 2012;**15**(1):21–6. [3421042; PUBMED: 22237120]

Wiese 2012 *{published data only}*
Wiese MN. Preoperative staging of non-small-cell lung cancer: Comparison of whole body diffusion weighted magnetic resonance imaging and 18f-fluorodeoxyglucose positron emission tomography/computed tomography. *Interactive Cardiovascular and Thoracic Surgery* 2012; **Conference**:S57–8. [3421044]

Wu 2010 *{published data only}*
Wu CS. Preoperative lymph node staging in patients of non-small-cell lung cancer: accuracy of 18F-FDG PET/CT. *European Journal of Nuclear Medicine and Molecular Imaging* 2010;**Conference**:S424. [3421046]

Yi 2007 *{published data only}*
Yi CA, Lee KS, Kim BT, Shim SS, Chung MJ, Sung YM, et al. Efficacy of helical dynamic CT versus integrated PET/CT for detection of mediastinal nodal metastasis in non-small cell lung cancer. *American Journal of Roentgenology* 2007;**188**(2):318–25. [3421048; PUBMED: 17242237]

Yi 2011 *{published data only}*
Yi CA, Lee KS, Kim YN, Lee E, Kwon OJ, Kim BT, et al. Preoperative nodal staging in patients with non-small cell lung cancer: Comparison between multimodality MR imaging plus PET/CT and PET/CT alone. *Journal of Thoracic Imaging* 2011;**26**:W106. [3421050]

Yi 2013 *{published data only}*
Yi CA, Lee KS, Lee HY, Kim S, Kwon OJ, Kim H, et al. Coregistered whole body magnetic resonance imaging-positron emission tomography (MRI-PET) versus PET-computed tomography plus brain MRI in staging resectable lung cancer: Comparisons of clinical effectiveness in a randomized trial. *Cancer* 2013;**119**(10):1784-91. [3421052; PUBMED: 23423920]

Yu 2007 *{published data only}*
Yu L-J, Duan Y, Liang X-Y, Wang X. 18FDG PET/CT in diagnosis and metastasis detection of lung neoplasms. *Chinese Journal of Medical Imaging Technology* 2007;**23**: 605–7. [3421054]

Zhang 2006 *{published data only}*
Zhang X, Li TR, Chen ZS, Ouyang XN. [Comparing serum tumor antigen detection combined with CT scan with PET-CT for lung cancer diagnosis]. *Ai Zheng [Chinese Journal of Cancer]* 2006;**25**(1):66–8. [3421056; PUBMED: 16405752]

Zhang 2012 *{published data only}*
Zhang TM, Zhang LM, Liu Y, Zhang ZF, Wang CL. [Diagnostic value of (18)F-FDG PET/CT imaging plus serum tumor marker assays for pulmonary lesions and clinical significance of SUVmax]. *Zhonghua Yie Xue Za Zhi* 2012;**92**(41):2901–4. [3421058; PUBMED: 23328236]

Zsiray 2009 *{published data only}*
Zsiray M, Markoczy Z, Magyar M, Lengyel Z, Fekeshazy A, Borbely L. [The advantage of positron emission tomography combined with computer tomography (PET-CT) in the diagnosis of lung cancer, experience with 408 patients]. *Magyar Onkologia* 2009;**53**(1):17–21. [3421060; PUBMED: 19318322]

Zsiray 2011 *{published data only}*
Zsiray M, Markoczy Z, Lengyel Z, Fekeshazy A, Kasler M, Borbely K. Contribution of 18F-FDG PET/CT in the management of patients with lung cancer. *European Journal of Nuclear Medicine and Molecular Imaging* 2011; **Conference**:S125–5. [3421062]

Zsiray 2012 *{published data only}*
Zsiray M. Preoperative mediastinal staging of NSCLC: the clinical impact of 18F-FDG PET/CT. *European Journal of Nuclear Medicine and Molecular Imaging* 2012;**Conference**: S455–6. [3421064]

Additional references

Casali 2010
Casali C, Cucca M, Rossi G, Barbieri F, Iacuzio L, Bagni B, et al. The variation of prognostic significance of maximum standarized uptake value of [18F]-fluoro-2-deoxy-glucose positron emission tomography in different histological subtypes and pathological stages of surgically resected non-small cell lung carcinoma. *Lung Cancer* 2010;**69**(2): 187–93. [PUBMED: 19942313]

Cerfolio 2005
Cerfolio RJ, Bryant AS, Ojha B, Eloubeidi M. Improving the inaccuracies of clinical staging of patients with NSCLC: a prospective trial. *The Annals of Thoracic Surgery* 2005;**80** (4):1207–13. [PUBMED: 16181842]

Davidson 2009
Davidson JA, Wong V, Fraser R, Hirsch V. Comparison of primary tumor maximal standardized uptake value (SUVmax) on preoperative [18F]fluorodeoxyglucose positron emission tomography/computed tomography (PET/CT) and histological subtype in patients with non-small cell lung cancer (NSCLC). *Journal of Clinical Oncology* 2009;**27**(15s):abstr7571.

De Leyn 2014
De Leyn P, Dooms C, Kuzdzal J, Lardinois D, Passlick B, Rami-Porta R, et al. Revised ESTS guidelines for preoperative mediastinal lymph node staging for non-small-cell lung cancer. *European Journal of Cardiothoracic Surgery* 2014;**45**(5):787–98. [PUBMED: 24578407]

Department of Health 2011
Department of Health. Improving Outcomes: a strategy for cancer. http://www.dh.gov.uk/prod˙consum˙dh/groups/dh˙digitalassets/documents/digitalasset/dh˙123394.pdf 2011.

Detterbeck 2007
Detterbeck FC, Jantz MA, Wallace M, Vansteenkiste J, Silvestri GA. Invasive mediastinal staging of lung cancer. *Chest* 2007;**132 Suppl**:202S–220S. [DOI: 10.1378/chest.07-1362]

Harbord 2007
Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* 2007;**8**(2):239–51. [PUBMED: 16698768]

Jeong 2002
Jeong HJ, Min JJ, Park JM, Chung JK, Kim BT, Jeong JM, et al. Determination of the prognostic value of [18F]fluorodeoxyglucose uptake by using positron emission tomography in patients with non-small cell lung cancer. *Nuclear Medicine Communications* 2002;**23**(9):865–70. [PUBMED: 12195091]

Kim 2006
Kim B-T, Lee KS, Shim SS, Choi JY, Kwon OJ, Kim H, et al. Stage T1 non-small cell lung cancer: preoperative mediastinal nodal staging with integrated FDG PET/CT -- a prospective study. *Radiology* 2006;**241**(2):501–9. [PUBMED: 16966480]

Lee 2007a
Lee BE, von Haag D, Lown T, Lau D, Calhoun R, Follette D. Advances in positron emission tomography technology have increased the need for surgical staging in non-small cell lung cancer. *The Journal of Thoracic and Cardiovascular Surgery* 2007;**133**(3):746–52. [PUBMED: 17320577]

Lindholm 2014
Lindholm H, Staaf J, Jacobsson H, Brolin F, Hatherly R, Sânchez-Crespo A. Repeatability of the Maximum Standard Uptake Value (SUVmax) in FDG PET. *Molecular Imaging and Radionuclide Therapy* 2014;**23**(1):16–20. [PUBMED: 24653930]

Lu 2010
Lu P, Yu L, Li Y, Sun Y. A correlation study between maximum standardized uptake values and pathology and clinical staging in nonsmall cell lung cancer. *Nuclear Medicine Communications* 2010;**31**(7):646–51. [PUBMED: 20545045]

Macaskill 2004
Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *Journal of Clinical Epidemiology* 2004;**57**:925-32..

Maemondo 2010
Maemondo M, Inoue A, Kobayashi K, Sugawara S, Oizumi S, Isobe H, et al. Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. *The New England Journal of Medicine* 2010;**362**(25):2380–8. [PUBMED: 20573926]

Manser 2005
Manser R, Wright G, Hart D, Byrnes G, Campbell D, Wainer Z, et al. Surgery for local and locally advanced non-small cell lung cancer. *Cochrane Database of Systematic Reviews* 2005, Issue 1. [DOI: 10.1002/14651858.CD004699.pub2]

NICE 2005
National Institute for Health and Clinical Excellence (NICE). Referral guidelines for suspected cancer. http://guidance.nice.org.uk/CG27/Guidance 2005.

NICE 2011
National Institute for Health and Clinical Excellence (NICE). The diagnosis and treatment of lung cancer (update). http://guidance.nice.org.uk/CG121 2011.

O'Rourke 2010
O'Rourke N, Roqué i Figuls M, Farré Bernadó N, Macbeth F. Concurrent chemoradiotherapy in non-small cell lung cancer. *Cochrane Database of Systematic Reviews* 2010, Issue 6. [DOI: 10.1002/14651858.CD002140.pub3]

Reitsma 2005
Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* 2005; **58**(10):982–90. [PUBMED: 16168343]

Review Manager 2012 [Computer program]
The Nordic Cochrane Centre, The Cochrane Collaboration. Review Manager (RevMan). Version 5.2. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2012.

Rutter 2001
Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine* 2001;**20**(19):2865–84. [PUBMED: 11568945]

Shim 2005
Shim SS, Lee KS, Kim BT, Chung MJ, Lee EJ, Han J, et al. Non-small cell lung cancer: prospective comparison of integrated FDG PET/CT and CT alone for preoperative staging. *Radiology* 2005;**236**(3):1011–9. [PUBMED: 16014441]

Silvestri 2013
Silvestri GA, Gonzalez AV, Jantz MA, Margolis ML, Gould MK, Tanoue LT, et al. Methods for staging non-small cell lung cancer: diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest* 2013;**143**(5 Suppl): e211S–50S. [PUBMED: 23649440]

Tournoy 2007
Tournoy KG, Maddens S, Gosselin R, Van Maele G, van Meerbeeck JP, Kelles A. Integrated FDG-PET/CT does not make invasive staging of the intrathoracic lymph nodes in non-small cell lung cancer redundant: a prospective study. *Thorax* 2007;**62**(8):696–701. [PUBMED: 17687098]

Whiting 2011
Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al (the QUADAS-2 Group). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine* 2011;**155**(8): 529–36. [PUBMED: 22007046]

Yi 2007
Yi CA, Lee KS, Kim BT, Shim SS, Chung MJ, Sung YM, et al. Efficacy of helical dynamic CT versus integrated PET/CT for detection of mediastinal nodal metastasis in non-small cell lung cancer. *American Journal of Roentgenology* 2007;**188**(2):318–25. [PUBMED: 17242237]

* *Indicates the major publication for the study*

Discusión

We don't see things as they are,
we see them as we are

Anaïs Nin

6 Discusión

6.1 Discusión específica de las publicaciones

El presente trabajo de tesis doctoral es un trabajo metodológico, pero, del análisis de las RS incluidas en la tesis, se pueden extraer mensajes generalizables a las RS en su conjunto. Por ello, la discusión específica se organizará por las fortalezas y limitaciones de las publicaciones que incluye, así como las estrategias que se plantean para superarlas.

Por completitud en este capítulo se discuten, además, los resultados de otras dos publicaciones que, si bien no forman parte del compendio de publicaciones de la tesis, son RS que complementan el contenido conceptual de la tesis al abordar preguntas clínicas de factores pronóstico [41,42]. Las publicaciones 5 y 6 se presentan en el **anexo 2**.

6.1.1 Principales resultados

Las publicaciones que conforman esta tesis presentan resultados en distintos ámbitos, desde el metodológico hasta ámbitos de aplicación en salud. La publicación 1 presenta resultados en el ámbito metodológico al mostrar una selección de métodos y recursos para desarrollar revisiones sistemáticas de alta calidad que responden a la mayoría de los tipos de preguntas clínicas. Esta publicación ofrece una guía de recursos completa para revisiones que abordan preguntas de la prevalencia de una condición clínica, el pronóstico, el diagnóstico y la efectividad de intervenciones. Se trata de una herramienta muy útil para aquellos investigadores que deseen desarrollar RS o realizar trabajos de investigación metodológica en el campo de la investigación de síntesis [30].

Las restantes publicaciones compiladas en este trabajo de tesis presentan resultados en el ámbito clínico. La publicación 2 muestra en sus resultados de metanálisis que el ejercicio conlleva algunos beneficios en las personas mayores frágiles, aunque todavía existe incertidumbre con respecto a qué características del ejercicio (tipo, frecuencia, duración) son más efectivas. La evidencia que compara diferentes modalidades de ejercicio es escasa y heterogénea [31]. La publicación 3 muestra que las técnicas de fisioterapia torácica analizadas (técnicas espiratorias pasivas convencionales o técnicas espiratorias forzadas) no han demostrado una reducción en la gravedad de la bronquiolitis aguda en niños. La calidad de la evidencia es de baja a alta, según el tipo de fisioterapia [32]. Finalmente, la publicación 4 muestra que la exactitud diagnóstica de la prueba PET-CT es insuficiente para realizar un estadiaje mediastínico que sustente la toma de decisiones sobre el tratamiento de pacientes con NSCLC [40]. La evidencia obtenida presenta inconsistencia y falta de precisión. Estos hallazgos respaldan las recomendaciones de NICE sobre este tema, por las que la prueba PET-CT es solo una guía para los médicos al decidir el siguiente paso: ya sea la realización de una biopsia o, cuando los ganglios negativos son pequeños, directamente la realización de cirugía.

Las dos publicaciones adicionales presentadas como anexos presentan también resultados en el ámbito clínico. La publicación 5 muestra que la evidencia disponible no respalda que las mochilas escolares que pesan > 10% del peso corporal estén asociadas a una mayor prevalencia de dolor lumbar entre los escolares de 9 a 16 años. La certeza de la evidencia es baja. Se requiere más

investigación sobre la relación entre el peso de la mochila y el dolor lumbar, de carácter longitudinal, y que tenga en cuenta la duración del transporte y la capacidad física de cada sujeto [41]. La publicación 6 muestra que la evidencia del impacto del comportamiento sedentario sobre los biomarcadores es inconsistente. Cuando se encontraron resultados estadísticamente significativos, el sedentarismo se asoció de manera desfavorable a los biomarcadores, pero los resultados se derivaron principalmente de estudios transversales y, por lo tanto, deben interpretarse en consecuencia. La evidencia obtenida es de calidad variable y muy heterogénea [42].

6.1.2 Fortalezas

Este trabajo de tesis presenta diversas fortalezas generales, ya que se ha facilitado un recurso metodológico esencial para la realización de distintos tipos de RS, y se han desarrollado tres RS de alta calidad que proporcionan evidencias sobre eficacia de las intervenciones y exactitud diagnóstica en tres ámbitos de salud relevantes. A continuación, se explorarán estas fortalezas con más detalle. Se explorarán, desde un punto de vista general, las fortalezas metodológicas de las publicaciones del compendio así como las 2 RS incluidas en los anexos.

Identificación de retos metodológicos para el desarrollo de revisiones sistemáticas de prevalencia, pronóstico y diagnóstico

Una de las fortalezas de este trabajo de tesis es la identificación de retos metodológicos para el desarrollo de RS de prevalencia, pronóstico y diagnóstico. La recopilación de métodos para desarrollar los principales tipos de RS que configura la publicación 1, junto con las experiencias al desarrollar las RS incluidas en este trabajo de tesis, permiten identificar algunas lagunas metodológicas que se presentan en la **tabla 4** organizadas por etapa de revisión.

Tabla 4. Lagunas metodológicas para el desarrollo de RS

	Lagunas
Manual	No existe un manual específico monográfico para desarrollar RS de modelos pronósticos, y el manual Cochrane para el desarrollo de RS de precisión diagnóstica está todavía incompleto.
Pregunta	No hay laguna, ya que se dispone de guías para formular preguntas para todos los tipos de RS.
Búsqueda	Más difícil para estudios observacionales (prevalencia, pronóstico, diagnóstico). Faltan filtros específicos para estas cuestiones para limitar el número de títulos y resúmenes a cribar.
Valoración del sesgo	No existe una escala específica para valorar el riesgo de sesgo de estudios de pronóstico global (incidencia), y debería validarse la propuesta de Hoy 2012 para los estudios de prevalencia.
Síntesis estadística	No están bien desarrollados los métodos para la valoración de la heterogeneidad estadística en RS de factores y modelos pronóstico, y RS de exactitud diagnóstica.
Calidad de la evidencia	No existe un desarrollo GRADE específico para RS de prevalencia ni de modelos pronóstico. El desarrollo existente para RS de factores pronóstico debería ser validado de forma externa.
Informe	No existe una ampliación de la declaración PRISMA específica para RS de prevalencia ni de pronóstico (pronóstico global, factores pronóstico y modelos pronóstico).
Valoración de la RS	No existen escalas de calidad metodológica para RS que no sean de intervención.

Riesgo de sesgo y calidad metodológica de las revisiones sistemáticas de la tesis

Otra fortaleza de este proyecto de tesis es la elaboración de las RS de alta calidad y rigor, cuyas conclusiones son fiables. La calidad de las RS se ha evaluado con las herramientas ROBIS y AMSTAR-2 [43,44].

La herramienta ROBIS es una escala destinada a valorar el riesgo de sesgo en cuatro tipos de RS en salud: intervención, diagnóstico, pronóstico y etiología [43]. ROBIS sigue un enfoque basado en dominios de sesgo y comprende no solo la evaluación de la validez interna del proceso de revisión, sino también la relevancia de la pregunta de revisión para sus usuarios. Evalúa cuatro dominios específicos (criterios de elegibilidad, identificación y selección de estudios, recopilación de datos y evaluación de estudios, y síntesis y hallazgos) por los que el sesgo puede afectar a la revisión, y cada dominio se evalúa en función de un número de preguntas indicadores. A partir de las respuestas a las preguntas indicadoras se obtiene una valoración ponderada de las inquietudes (*concerns*) que plantea el proceso de revisión para cada dominio; la valoración puede ser de inquietud alta, baja o poco clara. A partir de las evaluaciones por dominios y de la forma en que la RS aborda las inquietudes identificadas, se genera una valoración general del riesgo de sesgo de la RS. ROBIS dispone de una guía de aplicación para realizar las valoraciones de los motivos de inquietud por dominio y la valoración del riesgo de sesgo. Por otra parte, la escala AMSTAR-2 es una herramienta de valoración de la calidad metodológica específica para RS de intervención, que, a partir de 16 preguntas, genera una valoración de la confianza en los resultados de la revisión. Esta confianza puede ser alta, moderada, baja o críticamente baja. AMSTAR-2 dispone de una guía muy detallada de aplicación para realizar las valoraciones de los distintos ítems [44].

En la tabla 5 se presentan las valoraciones de riesgo de sesgo de las 5 RS evaluadas, mientras que en la tabla 6 se presentan las valoraciones de calidad metodológica de AMSTAR-2 para las dos RS de intervención. Cuatro de las RS incluidas en esta tesis doctoral presentan un bajo riesgo de sesgo global medido con la herramienta ROBIS. La valoración de bajo riesgo de sesgo implica que la interpretación de los resultados y la formulación de las conclusiones de las RS han considerado adecuadamente los motivos de inquietud detectados en los dominios. Por tanto, los resultados de las publicaciones son probablemente fiables, ya que las conclusiones parecen debidamente fundamentadas en la evidencia y tienen en cuenta la relevancia de los estudios incluidos en la RS. Esto no obsta para que la calidad de la evidencia generada en cada caso pueda ser mayor o menor, ya que la misma depende de cómo fueron realizados los estudios incluidos y de cómo era la evidencia generada por los mismos. La última RS presenta un riesgo de sesgo poco claro, debido principalmente a la incertidumbre sobre el impacto en las conclusiones de las decisiones tomadas en identificación de estudios y síntesis narrativa de resultados.

Las dos RS de intervención de este trabajo de tesis presentan valoraciones dispares de su calidad, según la guía de interpretación de los resultados de la escala AMSTAR-2 [44]. La publicación 2 presenta una calidad metodológica críticamente baja, debido a que presenta diversas debilidades en ítems considerados críticos, como la preespecificación metodológica *a priori* (disponibilidad de un protocolo previo), el listado de los estudios excluidos o la valoración del sesgo de publicación [31]. Esto significa que, según la valoración AMSTAR-2, la revisión no proporciona un resumen preciso y completo de los estudios disponibles que abordan la cuestión de interés. Sin embargo, la publicación 3 presenta una calidad metodológica alta, debido a que no presenta ningún defecto crítico (es decir,

cumple con los ítems considerados críticos) [32]. Por ello, la RS proporciona un resumen preciso y completo de los resultados para la pregunta clínica de interés.

La discrepancia entre las valoraciones del riesgo de sesgo y la calidad metodológica de la publicación 2 puede explicarse por las diferencias entre sesgo y calidad metodológica. Sesgo y calidad metodológica son conceptos relacionados, aunque distintos. La calidad metodológica nos indica hasta qué punto una revisión fue diseñada, realizada, analizada, interpretada y reportada siguiendo los estándares más rigurosos. Por el contrario, el riesgo de sesgo indica en qué medida una revisión evita errores sistemáticos en la evidencia que genera [1]. Un estudio puede estar realizado con la mayor calidad metodológica posible en el tema que evalúa y, sin embargo, estar sesgado. Así, por ejemplo, al comparar intervenciones que no pueden enmascararse no incurriríamos en una limitación metodológica, pero estaríamos abiertos a un sesgo. Y, por el contrario, determinados elementos de calidad metodológica (como puede ser el cálculo del tamaño muestral), no están relacionados con una reducción del riesgo de sesgo. En el caso de la publicación 2, debería valorarse si las limitaciones identificadas en AMSTAR-2 pueden matizar la confianza en los resultados que nos indica ROBIS.

Tabla 5. Valoración del riesgo de sesgo de las RS con la herramienta ROBIS

	Inquietud				Riesgo de sesgo
	Eligibilidad de estudios	**Identificación y selección de estudios**	**Extracción de datos y valoración del estudio**	**Síntesis y resultados**	
Publicación 2	Alta	Baja	Baja	Alta	
Publicación 3	Baja	Baja	Baja	Baja	
Publicación 4	Baja	Baja	Baja	Baja	
Publicación 5	Alta	Baja	Baja	Baja	
Publicación 6	Baja	Sin información suficiente	Baja	Sin información suficiente	

 : Alto riesgo de sesgo;  : Bajo riesgo de sesgo;  : Riesgo de sesgo desconocido

Tabla 6. Resultados de valoración de calidad metodológica de las publicaciones mediante la escala AMSTAR-2

Ítem		Publicación 2	Publicación 3
1	PICO	Sí	Sí
*2**	*Métodos a priori*	No	Sí
3	Diseño del estudio	Sí	Sí
*4**	*Búsqueda*	Sí	Sí
5	Selección duplicada	Sí	Sí
6	Extracción duplicada	Sí	Sí
*7**	*Listado de estudios*	No	Sí
8	Características estudios	Sí	Sí
9*	RoB	Sí	Sí
10	Financiación	No	Sí
*11**	*Metanálisis*	Sí	NA
12	RoB en los resultados	No	NA
*13**	*RoB en la interpretación*	Sí	Sí
14	Heterogeneidad	Sí	Sí
*15**	*Sesgo de publicación*	No	Sí
16	Conflicto de intereses	Sí	Sí
* Items críticos. Los enunciados completos de los ítems se listan en el **anexo 3**			

Impacto de las revisiones sistemáticas de la tesis

Otra fortaleza de las RS que conforman la tesis doctoral es su utilidad e impacto directo en la práctica clínica y la toma de decisiones. Dado el contexto de investigación clínica actual, en el que con frecuencia existen numerosas evidencias dispersas sobre un mismo tema de salud, la síntesis de la evidencia se erige como una base sobre la que formular recomendaciones de práctica clínica e investigación. Las publicaciones 2, 3 y 4 de esta tesis han aportado conocimiento a tres áreas de práctica asistencial (geriatría, fisioterapia y oncología) mediante la aplicación de técnicas rigurosas de síntesis, lo que queda reflejado en su incorporación a guías de práctica clínica (GPC) de los respectivos ámbitos. La publicación 5 es todavía muy reciente para valorar adecuadamente su impacto, y la publicación 6 ha tenido un impacto directo en el diseño de un ECA. A continuación, se detalla el impacto observado de cada una de las publicaciones de esta tesis.

Publicación 2: La publicación 2 ha sido citada en una GPC reciente en identificación y manejo de la fragilidad física en personas mayores, desarrollada por la International Conference on Frailty and Sarcopenia Research [25]. La GPC formula una recomendación sobre programas multicomponente de ejercicio físico para personas mayores, basándose en la evidencia de 5 RS (entre ellas, la publicación 2) según la cual «*Older people with frailty should be offered a multi-component physical activity programme (or those with pre-frailty as a preventative component*)» [25]. La recomendación es fuerte, por lo que se interpreta que los beneficios de la intervención probablemente compensan cualquier riesgo asociado. Esta recomendación se basa en una calidad global de la evidencia moderada, por lo que se interpreta que podrían aparecer resultados de estudios adicionales que podrían tener un impacto importante en los estimadores del efecto obtenidos, y por esta razón la confianza en estos estimadores de la eficacia de las intervenciones analizadas es limitada.

Publicación 3: La RS Cochrane que da lugar a la publicación 3 ha tenido un impacto relevante en las recomendaciones de las GPC sobre bronquiolitis. En 2005, era práctica habitual usar las técnicas de vibración y percusión en el tratamiento de la bronquiolitis, pese a que están asociadas a un mayor riesgo de efectos adversos como, por ejemplo, las fracturas. La publicación de la RS en 2005 llevó a la American Association of Pediatrics a no recomendarlas en su GPC de 2006 por falta de eficacia [45]. Posteriormente, en países como Francia, se popularizaron las técnicas de espiración forzada, consideradas más seguras, pero que tampoco demostraron tener un beneficio en el control de la bronquiolitis, en la versión de 2012 de la publicación 3 [35]. Esta es la base de las recomendaciones de las actualizaciones de 2014 y 2019 de la GPC de la American Academy of Pediatrics [46] y de la GPC de la Canadian Pediatric Society [47]. En todas ellas, se mantiene la recomendación de no realizar fisioterapia respiratoria para el tratamiento de la bronquiolitis. La AAP emite su recomendación con una valoración de calidad de la evidencia B (se interpreta que la evidencia proviene de estudios con limitaciones menores), y una fuerza de recomendación moderada al considerar simultáneamente que los beneficios anticipados de seguir la recomendación claramente superan a los daños, y la calidad de la evidencia es buena, pero no excelente. Desafortunadamente, en el caso de la Canadian Pediatric Society, su recomendación no va acompañada de valoración de la calidad de la evidencia ni de la fuerza de la recomendación. Finalmente, la GPC de NICE [27], publicada inicialmente en 2015 y actualizada en 2019, se basa en la evidencia de la publicación 3 para recomendar no realizar fisioterapia respiratoria en niños con bronquiolitis, a menos que presenten otras comorbilidades relevantes que dificulten la expulsión de mucosidades, como atrofia musculo-espinal o traqueomalacia severa. No se da información de la fuerza de la recomendación ni de la certeza de evidencia.

Publicación 4: La prueba de imagen PET-CT para el estadiaje del NSCLC es una práctica habitual y ya era una recomendación en las GPC previas a la RS. La RS de la publicación 4 pretendía resolver determinados aspectos de su indicación. Esta publicación fue seleccionada por el editor de JAMA para un comentario corto en su sección Clinical Evidence Synopsis [48] y, además, ha sido citada en una guía clínica nacional de diagnóstico, estadiaje y manejo del cáncer de pulmón, desarrollada por el Health Service Executive de Irlanda [29]. A partir de la publicación 4 y de diversos estudios individuales, la GPC formula las recomendaciones «*In non-small cell lung cancer (NSCLC) patients with mediastinal and hilar adenopathy, PET-CT is recommended for mediastinal and hilar lymph node staging in patients with potentially radically treatable non-small cell lung cancer (NSCLC) prior to invasive staging*" y "*In non-small cell lung cancer (NSCLC) patients with mediastinal and hilar adenopathy, patients with PET activity in a mediastinal lymph node and normal appearing nodes by CT (and no distant metastases), sampling of the mediastinum is recommended over staging by imaging alone*» [29], ambas con grado C, por lo que se interpreta que la evidencia proviene de revisiones sistemáticas con homogeneidad de estudios de cohortes o casos-controles.

Publicación 5: No ha sido posible determinar el impacto de los resultados de esta publicación dada su reciente aparición en la literatura indexada.

Publicación 6: La RS de la publicación 6 tiene por objetivo sintetizar la evidencia de la asociación entre biomarcadores asociados al envejecimiento y el comportamiento sedentario. Los resultados permitieron identificar biomarcadores diana para ser evaluados en un subestudio del proyecto SITLESS, en el que el comportamiento sedentario medido objetivamente se correlacionó con biomarcadores y resultados de biopsias musculares para determinar la influencia bioquímica del

comportamiento sedentario sobre desenlaces de salud [40]. Esta RS ha tenido un impacto directo en la investigación, al proporcionar evidencia para guiar el análisis de biomarcadores del proyecto SITLESS, ECA multinacional financiado por el programa Horizonte 2020 de la UE [49].

Impacto de la política de actualización Cochrane en la vigencia y validez de las revisiones

Finalmente, otra fortaleza de este trabajo de tesis es su utilidad para ilustrar el impacto de la política de actualización Cochrane en la vigencia y validez de las revisiones. Las revisiones Cochrane siguen un proceso de actualización periódica, durante el cual se pueden incorporar mejoras metodológicas, pero también corregir errores o imprecisiones, e incorporar las aportaciones de investigadores externos a la revisión. Las nuevas versiones de las revisiones Cochrane reciben una nueva referencia bibliográfica si presentan cambios relevantes respecto a una versión anterior (por ejemplo, si se han modificado las conclusiones), de modo que en los buscadores bibliográficos se consideran dos publicaciones (dos revisiones) distintas, aunque son esencialmente la misma revisión.

La publicación 3 es un buen ejemplo del impacto de esta política de actualización sobre la calidad y validez de los resultados de las revisiones Cochrane. El proceso de actualización ha permitido mejorar la revisión a lo largo del tiempo, no sólo por la actualización de la búsqueda y los métodos, sino a partir de los comentarios recibidos de tres investigadores externos y los cambios que se derivaron en respuesta, que se resumen en la **tabla 7**.

El historial de la publicación 3 ilustra dos ideas: la primera es que ningún autor está libre de cometer errores de transcripción o interpretación que limiten la validez de su trabajo, y la segunda es que cualquier trabajo de investigación se beneficia de procesos de actualización y mejora de la calidad que los hacen vigentes y aumentan su validez. Además, introduce la necesidad de un nuevo concepto: las revisiones sistemáticas actualizadas continuamente (*living systematic reviews*), que se discutirán más adelante.

Tabla 7. Comentarios recibidos a las versiones publicadas de la publicación 3

Versión	Comentarios recibidos	Cambios respecto a la versión anterior publicada
Roqué 2012	Marzo de 2012 - Comentarios de Guy Postiaux, autor de uno de los estudios incluidos [37].	Cambio en la terminología de las técnicas de espiración pasiva, y clasificación de las mismas en los subgrupos de técnicas de flujo lento y espiración forzada. También se corrigieron pequeños errores en la descripción de los estudios, y se incorporó el concepto de gravedad de la enfermedad en la descripción e interpretación de los resultados de la revisión sistemática.
Roqué 2016	Abril de 2016 - Comentarios de Helen Main [38].	Corrección de un error en la transcripción de resultados de un estudio incluido Nueva versión publicada mayo de 2016 sin nueva cita bibliográfica
	Mayo de 2017 - Comentarios de Fernanda Remondini, autora de uno de los estudios incluidos [39].	Los comentarios incidían en la necesidad de reclasificar el estudio en función de la intervención evaluada, mejorar la descripción del comparador e incorporar datos mínimos sobre un desenlace secundario. Nueva versión publicada en junio de 2017, publicación incluida en esta tesis. Esta versión no generó una nueva cita bibliográfica, sin embargo es la versión que aparece al consultar The Cochrane Library.

6.1.3 Limitaciones

Este trabajo de tesis presenta también limitaciones relacionadas con su enfoque y realización. Por una parte, la tesis integra síntesis de la evidencia en problemas de salud muy dispares en lugar de considerar un único problema de salud y realizar las RS sobre distintos aspectos del abordaje del mismo en términos de prevalencia, pronóstico, exactitud diagnóstica y efecto de las intervenciones. Esta variabilidad representa una oportunidad perdida para realizar transferencia de conocimiento y experiencia clínica de un trabajo de síntesis a otro. Sin embargo, no limita la evidencia obtenida en cada RS ni limita el trabajo y las conclusiones metodológicas derivadas.

En segundo lugar, a pesar de que la tesis proporciona directrices metodológicas para los principales tipos de RS, la metodología de síntesis se ha aplicado solo a preguntas de efecto de las intervenciones y exactitud diagnóstica. La incorporación de trabajos de RS de prevalencia y modelos pronóstico al compendio de publicaciones hubiera permitido una discusión más transversal y completa de la metodología de revisión aplicada a distintas preguntas, pero esto superaba el alcance previsto de la tesis doctoral. Sin embargo, este trabajo de tesis doctoral incluye como anexos dos RS de factores pronóstico, que han permitido ampliar la discusión metodológica más allá de los efectos de las intervenciones y la exactitud diagnóstica.

Finalmente, cada publicación presenta limitaciones inherentes a su diseño de estudio y ejecución, que ya se han comentado en la discusión de las respectivas publicaciones. A continuación, se explorarán, desde un punto de vista general, las limitaciones metodológicas de las publicaciones del compendio así como las 2 RS complementarias.

Consideraciones metodológicas

En esta sección se comparan los métodos empleados en cada revisión con los recomendados en la publicación 1, en un ejercicio de consistencia interna para evaluar hasta qué punto las RS incluidas aplican los métodos más apropiados, y en qué aspectos podrían mejorar su metodología en futuras actualizaciones. Se realizará este ejercicio siguiendo las etapas de desarrollo de una RS: formulación de la pregunta de investigación, búsqueda bibliográfica, valoración del riesgo de sesgo de los estudios incluidos, síntesis de resultados, y valoración de la calidad de la evidencia.

Pregunta de investigación y registro del protocolo: Todas las publicaciones han especificado una pregunta de investigación y unos criterios de inclusión claros, y todas las publicaciones, excepto la publicación 2, registraron el protocolo de revisión en el registro PROSPERO o en la CDSR.

Búsquedas bibliográficas: Todas ellas han realizado búsquedas bibliográficas exhaustivas en 2 o más bases de datos y proporcionan una copia de la estrategia aplicada, lo que garantiza su reproducibilidad, así como el flujo de artículos identificados en la búsqueda y posteriormente cribados. Tan solo las dos RS Cochrane (publicaciones 3 y 4) realizaron búsquedas de estudios no publicados.

Valoración del riesgo de sesgo de los estudios incluidos: Todas las revisiones han valorado el riesgo de sesgo de los estudios incluidos. Las publicaciones 2, 3 y 4 han aplicado las escalas de valoración de riesgo de sesgo recomendadas para RS diagnósticas y de efecto de las intervenciones (QUADAS-2 y RoB). Por el contrario, las publicaciones 5 y 6 no han aplicado la escala de riesgo de sesgo QUIPS recomendada para RS de factores pronóstico, sino que han aplicado escalas de valoración de la calidad metodológica. La publicación 5 ha aplicado la escala de Newcastle-Ottawa para estudios longitudinales y una escala específica para estudios transversales de dolor lumbar [50]. La publicación 6 ha valorado los estudios incluidos a partir de una modificación de la escala CASP derivada dentro de un programa internacional de formación en habilidades para la lectura crítica de la literatura médica (http://www.redcaspe.org). Todas las revisiones, salvo una (publicación 5), han presentado de forma explícita los resultados de la valoración de la calidad para cada estudio incluido.

Síntesis de resultados: En los casos en que se ha realizado un metanálisis (publicaciones 2, 4 y 5), se han descrito los métodos aplicados, incluyendo la valoración de la heterogeneidad, métodos de síntesis y análisis de subgrupos y sensibilidad previstos y finalmente realizados. En los casos en que no se ha realizado un metanálisis (publicaciones 3 y 6), se han justificado los motivos por los que no se ha realizado dicho análisis estadístico y se han presentado sendas síntesis narrativas de los resultados. En las publicaciones en las que se realizó un metanálisis se evaluó la heterogeneidad estadística además de la heterogeneidad clínica, en las publicaciones 2 y 5 mediante el indicador I^2 y en la publicación 4 mediante la incorporación de los posibles factores de heterogeneidad como covariables en el modelo estadístico bivariado de metanálisis. Además, en la publicación 5, se realizó un metanálisis de datos de pacientes individuales, que permitió recalcular las medidas de asociación de cada estudio, ajustando por los mismos factores, a fin de eliminar eventuales fuentes de confusión que pudieran inducir variabilidad en los resultados. Todas las publicaciones describieron los análisis de sensibilidad y subgrupos previstos, y, en caso de no poder realizarlos (publicaciones 2 y 3), se describieron los motivos.

Calidad global de la evidencia: Dos de las revisiones publicadas (publicaciones 3 y 5) realizaron una valoración de la calidad de la evidencia, en ambos casos mediante el sistema GRADE. La publicación 5 aplicó una modificación específica para RS de factores pronóstico.

Reporte de las RS: De las publicaciones que forman parte de esta tesis, solo la publicación 6 refiere explícitamente adherirse a los estándares PRISMA [51], pero el cumplimiento con los estándares o sus extensiones es muy elevado en todas ellas. Los incumplimientos de los estándares de reporte corresponden generalmente a información disponible para los autores pero que no se reportó en la publicación (como la presentación de los datos individuales de los estudios), y, por tanto, pueden ser subsanados. En contadas ocasiones el incumplimiento de los estándares se debe a limitaciones metodológicas en la realización de las revisiones (como no registrar el protocolo de la RS), que no son subsanables.

De los párrafos previos se deriva que las publicaciones que integran este compendio no siempre han aplicado metodologías congruentes con las recomendaciones metodológicas de la publicación 1. Esto puede explicarse parcialmente porque estas publicaciones anteceden a la publicación 1 y a varios de los desarrollos metodológicos presentados allí. Así, por ejemplo, la publicación 3 era previa a las recomendaciones SWiM para síntesis narrativa, y la publicación 4 antecedía a la publicación de las herramientas para evaluar la calidad de la evidencia en RS diagnósticas [52-54]. Además, en el caso de las publicaciones 4, 5 y 6, que corresponden a RS de diagnóstico y pronóstico, los investigadores no disponían de manuales completos que guiaran su desarrollo, o no estaban suficientemente familiarizados con los materiales disponibles. Cuando no existe un consenso claro sobre los mejores métodos a aplicar, la elección de los métodos depende de la experiencia, conocimientos y preferencias de los investigadores que desarrollan la RS, así como de los consensos que alcancen respecto a la metodología a aplicar. Un ejemplo es la aplicación de una escala de valoración metodológica distinta a QUIPS en la publicación 5, o la aplicación de una herramienta basada en CASP en la publicación 6. Finalmente, la elección de los métodos depende de las particularidades de cada RS. Así, por ejemplo, la publicación 6 es una RS de pronóstico de carácter exploratorio, en la que se debía sintetizar un volumen de evidencia muy amplio y heterogéneo, y en la que se aplicó una estrategia de «recuento de votos» de los estudios, clasificados siguiendo una regla de decisión basada en la significación de los resultados [55]; pero todo ello sin dar resúmenes numéricos de los resultados, que serían poco interpretables y comparables al ser estudios pronósticos que se ajustan por diferentes factores.

Propuestas de mejora de cara a futuras actualizaciones

Las limitaciones identificadas en estos trabajos podrían ser subsanadas en futuras actualizaciones de las RS. La actualización de una revisión debe considerar cuatro ámbitos de mejora: actualización de la búsqueda para incorporar nueva evidencia, actualización de los objetivos para asegurar que la pregunta que se responde sigue siendo pertinente, actualización metodológica para incorporar métodos más apropiados y mejorar la calidad, y actualización de la implicación (*engagement*) de los usuarios finales para incrementar la usabilidad de la RS y su grado de transferencia del conocimiento al público [56]. El plan de actualización y mejora de las revisiones que conforman la tesis debería incorporar los aspectos presentados en la **tabla 8**.

Tabla 8. Propuestas de mejora de las RS incluidas

	Búsqueda	Alcance/objetivo	Métodos	Implementación
Publicación 2	La fecha de búsqueda es 2015, por lo que debería realizarse una nueva búsqueda.	La definición de fragilidad centrada en déficits físicos podría ser reconsiderada, y se podría ampliar a una definición que incluya déficits sociales, cognitivos y psicológicos.	Incorporar la valoración de la calidad de la evidencia y una tabla de resumen de los hallazgos Incorporar elementos de SWiM para la síntesis narrativa de los estudios no metanalizados	Incorporar el punto de vista directo de pacientes en el planteamiento de la revisión, interpretación de resultados y formulación de conclusiones.
Publicación 3	La fecha de búsqueda es 2015, y se identificaron estudios en curso, por lo que debería realizarse una nueva búsqueda.	El alcance de la revisión incluye la fisioterapia ambulatoria, por lo que deberían incorporarse desenlaces y perspectivas ambulatorias.	Si en la siguiente actualización la síntesis es todavía narrativa, incorporar elementos de SWiM.	
Publicación 4	La fecha de búsqueda es 2013, por lo que debería realizarse una nueva búsqueda.	--	Incorporar la valoración de la calidad de la evidencia y una tabla de resumen de los hallazgos	
Publicación 5	La fecha de búsqueda es 2019, por lo que todavía se considera actualizada.	--	Evaluar el riesgo de sesgo con la escala QUIPS.	
Publicación 6	La fecha de búsqueda es 2015, por lo que debería realizarse una nueva búsqueda.	El alcance debería ser menos amplio y exploratorio y centrarse en los biomarcadores específicos para los que exista más evidencia.	Incorporar la valoración de la calidad de la evidencia y una tabla de resumen de los hallazgos.	Incorporar los resultados de los proyectos en marcha sobre comportamiento sedentario para delimitar los biomarcadores de mayor interés.

6.2 Discusión en el contexto del conocimiento actual

6.2.1 Consideraciones en la aplicación de ROBIS y AMSTAR

Las escalas ROBIS y AMSTAR, descritas anteriormente, son dos herramientas clave para el investigador que debe evaluar la solidez de una RS de intervención, o que realiza un *overview* de intervención y debe evaluar las RS incluidas. Algunos investigadores han realizado estudios de validación comparando la aplicación de ambas herramientas en muestras de RS de intervención. Pieper y colaboradores destacan una buena concordancia entre las respuestas en ítems comparables de ambas escalas [57]. Por otro lado, Gates y colaboradores destacan la dificultad que implica para los autores

menos experimentados valorar los ítems de ROBIS, debido a cierta ambigüedad en los documentos de guía y el grado de subjetividad implícito en las valoraciones [58].

Los investigadores que realizan *overviews* de intervención tienen la opción de evaluar de forma complementaria tanto la calidad metodológica como el riesgo de sesgo, aplicando las dos herramientas a las RS incluidas. Así, por ejemplo, un *overview* Cochrane concluye que las RS incluidas son de bajo riesgo de sesgo por ROBIS y de alta calidad metodológica por AMSTAR, aunque lamentablemente no informa de cómo se habrían interpretado posibles resultados discrepantes entre las herramientas [59].

Sin embargo, los autores que deben evaluar la calidad metodológica de RS que no son de intervención se enfrentan a la disyuntiva de valorar el riesgo de sesgo con ROBIS como una aproximación de la calidad metodológica, o aplicar la escala AMSTAR a pesar de que no sea una herramienta específica para el tipo de RS considerado. Esta situación correspondería, por ejemplo, a un *overview* específico de cuestiones clínicas de prevalencia (o de pronóstico o de exactitud diagnóstica), pero también a un *overview* transversal que incluya RS de diversos tipos, por ejemplo, por estar centrado en la detección, pronóstico e intervención de una patología determinada.

Una búsqueda exploratoria de la literatura permite ver cómo han afrontado la disyuntiva los autores de *overviews* o *umbrella reviews*. Algunos autores de *overviews* transversales optan por presentar la evidencia de RS de prevalencia, pronóstico, diagnóstico e intervención, pero solo dan la valoración metodológica de AMSTAR para las RS incluidas de intervención [60,61]. Otros han resuelto aplicar AMSTAR a todos los tipos de RS [62,63], con las limitaciones que esto conlleva. En el caso de los *overviews* específicos que incluyen un único tipo de RS, coexisten trabajos que optan por la aplicación abusiva de AMSTAR [64,65], con otras publicaciones que optan por la opción más correcta de valorar el riesgo de sesgo aplicando ROBIS [66,67]. La búsqueda exploratoria no permitió identificar ningún *overview* de prevalencia o diagnóstico que aplicara ROBIS.

La aplicación de la herramienta AMSTAR-2 a tipos de RS que no son de intervención plantea problemas de consistencia y estructurales, porque algunos de los ítems en AMSTAR-2 no son aplicables o relevantes en el contexto de una RS de prevalencia, pronóstico o diagnóstico. Así, por ejemplo, el ítem 3 sobre la justificación de incluir estudios no aleatorizados (ver **anexo 3**), es irrelevante en RS diagnósticas o pronósticas, donde los diseños que proporcionan evidencia de calidad son precisamente no aleatorizados [30, 68]. El ítem 7 de adecuada identificación de los estudios excluidos es difícil de aplicar a las revisiones no publicadas en revistas electrónicas, y a los tipos de RS en los que no existen filtros precisos para la identificación de estudios, y que, por tanto, deben excluir un elevado número de referencias. El ítem 9 de valoración del riesgo de sesgo en los estudios no aleatorizados indica que debe contemplar el dominio de riesgo de sesgo de confusión, sesgo de selección de la muestra, medición de las exposiciones y desenlaces, e informe selectivo de desenlaces y análisis. Sin embargo, aunque estos elementos serían totalmente relevantes para las RS pronósticas, para las RS diagnósticas el sesgo de confusión es poco relevante (ya que todos los participantes reciben el test de interés y el estándar de referencia), igual que la medición de las exposiciones. Finalmente, el ítem 15 de sesgo de publicación corresponde a un aspecto que está poco explorado en el ámbito diagnóstico y pronóstico, y para el que no hay recomendaciones claras de cómo llevarlo a cabo en estas RS [54,69].

Por todo ello, se puede concluir que existe una necesidad clara de desarrollar una herramienta para valorar la calidad metodológica de las RS de prevalencia, pronósticas y diagnósticas que evite el uso inadecuado de AMSTAR en estos ámbitos. Dicha herramienta complementaría a ROBIS en la evaluación de estos tipos de RS. La herramienta ROBIS es todavía insuficientemente conocida y se percibe difícil de aplicar, por lo que parece conveniente darla a conocer y promover su uso dada la relevancia fundamental de la valoración del riesgo de sesgo. Finalmente, sería interesante establecer pautas para los investigadores que ayudaran a interpretar las posibles discrepancias entre las valoraciones de riesgo de sesgo y calidad metodológica, e incidieran en aquellos aspectos que se pueden solventar (por ejemplo, aplicando métodos más rigurosos en la actualización de la RS) y en aquellos aspectos que dependen exclusivamente del cuerpo de evidencia de base (y que requerirían de la realización de nuevos estudios individuales, menos sesgados).

6.2.2 Otros tipos de revisiones sistemáticas en salud

Esta tesis se ha centrado en cuatro tipos principales de RS, pero existen otros tipos, también relevantes, como pueden ser las RS metodológicas, cualitativas, económicas, psicométricas o de intervenciones complejas [68,70]. El grado de desarrollo metodológico para realizar estas RS es muy variable para los distintos tipos, y el conocimiento se encuentra a menudo fragmentado y es incompleto. Algunas publicaciones han realizado un esfuerzo de compilación metodológica similar al de la publicación 1 para un abanico de tipos de RS. Así, por ejemplo, Pollock y colaboradores identifican los manuales disponibles y las herramientas de valoración de calidad para tipos de RS que incluyen las RS cualitativas o las psicométricas [70]. Munn y colaboradores revisan la estructura del título y las guías de informe, manuales y herramientas de valoración de la calidad para 10 tipos de RS, entre las que se incluyen las revisiones metodológicas [68].

Las revisiones metodológicas tienen por objeto examinar aspectos metodológicos específicos relacionados con el diseño, la realización y la revisión de estudios de investigación y síntesis de evidencia. Aunque llevan realizándose desde hace tiempo, la variabilidad en sus temáticas dificulta el desarrollo de guías metodológicas y manuales, y no hay ningún manual completo disponible [68]. Trabajos recientes van en la dirección de clasificar adecuadamente los estudios metodológicos, con el fin de sentar las bases para las RS metodológicas [71,72]. Las revisiones cualitativas se centran en el análisis de las experiencias humanas y los fenómenos sociales y culturales. Una de las fuentes más solventes en este tipo de RS es el Joanna Briggs Institute, institución de referencia en el desarrollo de RS metodológicas que publica un manual de referencia para llevarlas a cabo [73].

Las RS de evaluación económica revisan los costes de una determinada intervención, proceso o procedimiento, a menudo en relación con su efectividad. El investigador van Mastrigt y sus colaboradores realizaron un esfuerzo similar al de la publicación 1, integrando las recomendaciones formuladas por distintas entidades, que estaban previamente fragmentadas, no siempre eran específicas para estudios económicos, o no eran suficientemente detalladas [74].

Las RS psicométricas evalúan los instrumentos de medición en base a, entre otras, sus características de validez, fiabilidad y respuesta. La iniciativa COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) mantiene actualizada una base de datos de RS psicométricas

(https://www.cosmin.nl/database/), así como un repositorio de recursos metodológicos para desarrollar RS psicométricas, como pueden ser filtros de búsqueda o herramientas de valoración de la calidad.

Las RS de intervenciones complejas analizan intervenciones con múltiples componentes, que a menudo están influidas por interacciones complejas de las características individuales, determinantes sociales, elementos del sistema de salud y las propias intervenciones [75]. La Agency for Healthcare and Quality desarrolló una serie de publicaciones sobre el desarrollo de RS de intervenciones complejas, disponibles de forma centralizada en su web.

6.2.3 Las revisiones sistemáticas del futuro

No es fácil hacer predicciones de cómo serán las RS del futuro, ya que las RS abarcan campos y preguntas muy diversos, el desarrollo metodológico es constante, y la investigación primaria tiene una altísima velocidad de desarrollo. Sin embargo, se pueden hacer algunos apuntes de los cambios que se están produciendo ahora mismo.

¿Qué formato tendrán las revisiones sistemáticas?

Ioannidis y colaboradores plantean el futuro de las RS en términos de RS prospectivas, RS de datos individuales, revisiones de revisiones (*overview* o *umbrella reviews*), y metanálisis en red (*network metanalysis*) [76,77]. Todas estas opciones presentan ventajas y limitaciones sobre el formato de una RS tradicional, y cada una de ellas presenta sus propias condiciones. Así, por ejemplo, tanto las revisiones prospectivas, en las que los ensayos se diseñan con el propósito explícito y predefinido de integrar la RS, como las revisiones de datos individuales requieren colaboración con los investigadores de los ensayos, lo que no siempre es posible para un autor de revisiones sistemáticas. Por otra parte, las revisiones de revisiones y los metanálisis en red se interesan en una constelación o familia de intervenciones. La opción más apropiada dependerá del uso que se le quiera dar a la RS como síntesis de la evidencia.

Sin embargo, una controversia más actual surge como crítica a la complejidad metodológica de las RS tradicionales y a la importante inversión de recursos y tiempo necesarios para desarrollarlas. Como alternativa a este diseño tradicional se han propuesto las revisiones rápidas, que sintetizan el conocimiento simplificando los procesos de la revisión sistemática sin disminuir su rigor [3]. Acelerar la síntesis de la evidencia es esencial para mejorar los tiempos de respuesta de los sistemas sanitarios y los gestores de salud. Especialmente en situaciones de emergencia y crisis, las revisiones rápidas pueden ser cruciales para la toma de decisiones y la formulación de respuestas rápidas por parte de los sistemas de salud [78]. En la situación de emergencia global creada por el virus SARS-CoV-2, la iniciativa COVID-END ha llegado a proponer un modelo para el desarrollo de informes de síntesis en un máximo de 3 horas (https://www.mcmasterforum.org/networks/covid-end). Sin embargo, las revisiones rápidas dependen aún más de la fiabilidad de la investigación primaria que las RS tradicionales, y una síntesis prematura de evidencia poco sólida o insuficientemente contrastada puede dar lugar a conclusiones sesgadas. Diversos estudios, que comparan la fiabilidad de las conclusiones de las revisiones rápidas con la de las RS tradicionales, muestran que no siempre alcanzan las mismas conclusiones y señalan la importancia de valorar en qué situaciones las revisiones rápidas son un sustituto adecuado de las RS tradicionales [79].

Otro hito que marcará el futuro de las RS son las revisiones actualizadas continuamente (*living systematic reviews*), como respuesta a la necesidad imperiosa de disponer constantemente de la mejor evidencia actualizada [80,81]. Si el ejemplo de la publicación 3 mostraba cómo el paso del tiempo influye en la vigencia y calidad de la evidencia de una RS, la situación creada por la Covid-19 muestra la velocidad a la que se puede llegar a publicar información y la necesidad de evidencia inmediata para la práctica clínica asistencial. Por ello, las revisiones actualizadas continuamente representan una evolución necesaria a las RS tradicionales. Se caracterizan por una monitorización continua y activa de la nueva evidencia (mensual e incluso semanal), la inclusión inmediata de cualquier nueva evidencia relevante identificada (estudios o datos), y la comunicación y difusión periódica del estatus de la revisión y la nueva evidencia incorporada. Aunque la metodología de estas revisiones no es fundamentalmente distinta a la de una RS, sí que debe incorporar el informe transparente y explícito de las decisiones tomadas en cuanto a la frecuencia de búsqueda de nueva evidencia, y cómo esta nueva evidencia es incorporada a la RS.

¿Quién desarrollará revisiones sistemáticas?

Los equipos de investigadores contarán cada vez más con recursos informáticos basados en inteligencia artificial y automatización de procesos que permitirán semiautomatizar distintos pasos del desarrollo de RS [82]. Además, se abandonará progresivamente la concepción de la RS como un simple proyecto de investigación secundaria realizado por un grupo reducido de investigadores de élite, para concebir las RS como proyectos transversales que implican a los autores de la revisión pero también a los autores de los estudios individuales, agentes interesados como gestores sanitarios o consumidores, y a redes de investigadores externos que pueden realizar tareas específicas como el cribado de referencias, a menudo de forma transversal en distintas revisiones. Esta concepción de las RS como una iniciativa colaborativa y menos personalista se refleja en la respuesta a la pandemia por SARS-CoV-2, en que investigadores, instituciones y revistas aunaron esfuerzos y generaron registros abiertos de revisiones, ensayos y publicaciones sobre la pandemia como el Covid-19 Study Register (https://covid-19.cochrane.org/) o Covid Reviews (https://covidreviews.cochrane.org/).

¿En qué ámbito se desarrollarán?

Las RS surgieron en el campo de la salud y la atención sanitaria, pero como herramienta de investigación son aplicables a cualquier otro campo en que se desee obtener síntesis de la evidencia. Actualmente, ya se desarrollan RS en campos como las ciencias sociales [83], las ciencias políticas [84], la veterinaria [85] o la agricultura [86]. Incluso, existen colaboraciones similares a Cochrane para la síntesis de evidencia en algunos de estos campos, como puede ser la Campbell Collaboration (https://campbellcollaboration.org/), red internacional que tiene por objeto desarrollar síntesis de evidencia de alta calidad, gratuitas y relevantes para la formulación de políticas sociales. Otras iniciativas similares incluyen la Collaboration for Environmental Evidence (https://www.environmentalevidence.org/), red internacional que promueve y disemina síntesis de evidencia en temas de entorno global sostenible y conservación de la biodiversidad; el Centre for Evidence-Based Agriculture (https://www.harper-adams.ac.uk/research/ceba/), que sintetiza la evidencia en agricultura y alimentación para apoyar la toma de decisiones en políticas, industria, práctica e investigación, y, finalmente, la iniciativa Systematic Reviews for Animals & Food (http://www.syreaf.org/), que ofrece materiales formativos y un repositorio de protocolos de RS en veterinaria y seguridad alimentaria.

Conclusiones

És quan dormo que hi veig clar

JV Foix

7 Conclusiones

7.1 Conclusiones para la práctica

- Los programas multicomponente de ejercicio físico deberían ofrecerse a las personas mayores frágiles para mejorar su capacidad física, dado que han mostrado tener un impacto clínicamente relevante sobre la velocidad de la marcha normal, la velocidad de la marcha rápida y la escala SPPB de función física.
- En niños hospitalizados con bronquiolitis aguda, moderada o severa, no es recomendable realizar fisioterapia respiratoria con técnicas convencionales (vibración y percusión) o espiración pasiva forzada. Las técnicas de espiración pasiva lenta son aparentemente seguras (calidad de la evidencia muy baja) pero ineficaces en la mejora de la enfermedad (calidad de la evidencia baja).
- En pacientes con NSCLC potencialmente resecable, la prueba PET-CT es útil para el estadiaje ganglionar, como paso previo al estadiaje invasivo por biopsia.

7.1 Conclusiones para la investigación

- La metodología para desarrollar RS de prevalencia y pronóstico es incompleta y se requiere investigación para desarrollar un manual específico para RS de modelos pronóstico, métodos y herramientas para la valoración del riesgo de sesgo en RS de prevalencia, guías para la valoración de la calidad de la evidencia en RS de prevalencia y modelos pronóstico, y guías para el reporte de RS de prevalencia y pronósticas. Asimismo, sería conveniente desarrollar herramientas de valoración de la calidad metodológica específica para las RS de prevalencia, pronóstico y exactitud diagnóstica.
- Las RS son iniciativas complejas y costosas en tiempo y esfuerzo, que deben realizarse por equipos multidisciplinarios que apliquen metodología rigurosa. Para evitar la duplicación de esfuerzos, todas las RS deberían tener un plan de actualización y de incorporación de comentarios, que permita la incorporación periódica de nueva evidencia o de innovaciones metodológicas después de su publicación inicial.
- Es necesario desarrollar más evidencia sobre los criterios que deben guiar las políticas de actualización de RS, ya sea por priorización o por criterios temporales.

Bibliografía

8 Bibliografía

1. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated August 2019). Cochrane, 2019. Disponible en www.training.cochrane.org/handbook.
2. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Info Libr J. 2009;26(2):91-108.
3. Khangura S, Konnyu K, Cushman R, Grimshaw J, Moher D. Evidence summaries: the evolution of a rapid review approach. Syst Rev. 2012;1:10.
4. Tricco, A.C., Antony, J., Zarin, W. et al. A scoping review of rapid review methods. BMC Med 13, 224 (2015).
5. Lizarondo L, Stern C, Carrier J, et al. Chapter 8: Mixed methods systematic reviews. In: Aromataris E, Munn Z (Editors). JBI Manual for Evidence Synthesis. JBI, 2020. Disponible en https://synthesismanual.jbi.global.
6. Peters MDJ, Godfrey C, McInerney P, Munn Z, Tricco AC, Khalil, H. Chapter 11: Scoping Reviews (2020 version). In: Aromataris E, Munn Z (Editors). JBI Manual for Evidence Synthesis, JBI, 2020. Disponible en https://synthesismanual.jbi.global.
7. Small S, Porr C, Swab M, Murray C. Experiences and cessation needs of Indigenous women who smoke during pregnancy: a systematic review of qualitative evidence. JBI Database System Rev Implement Rep. 2018;16(2):385-452.
8. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline BMJ 2020; 368 :l6890.
9. Siriwardhana DD, Hardoon S, Rait G, et al. Prevalence of Frailty and Prefrailty Among Community-Dwelling Older Adults in Low-Income and Middle-Income Countries: A Systematic Review and Meta-Analysis. BMJ Open. 2018; 8(3): e018195.
10. Hemingway H, Croft P, Perel P, et al. Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes BMJ 2013; 346 :e5595.
11. Roehr S, Pabst A, Luck T, et al. Is dementia incidence declining in high income countries? A systematic review and meta-analysis. Clin Epidemiol. 2018; 10: 1233–1247.
12. Lara E, Martín-María N, De la Torre-Luque A, et al. Does loneliness contribute to mild cognitive impairment and dementia? A systematic review and meta-analysis of longitudinal studies. Ageing Res Rev. 2019 Jul;52:7-16.
13. Lindroth H, Bratzke L, Purvis S, et al. Systematic review of prediction models for delirium in the older adult inpatient. BMJ Open. 2018 Apr 28;8(4):e019223.
14. Ambagtsheer RC, Thompson MQ, Archibald MM, et al.: Diagnostic test accuracy of self-reported frailty screening instruments in identifying community-dwelling older people at risk of frailty and pre-frailty: a systematic review protocol. JBI Database System Rev Implement Rep. The Joanna Briggs Institute, 2017; 15(10): 2464–2468.
15. Ellis G, Gardner M, Tsiachristas A, et al.: Comprehensive geriatric assessment for older adults admitted to hospital. Cochrane Database Syst Rev. 2017; 9(1): CD006211.
16. Haynes RB, McKibbon KA, Wilczynski NL, Walter SD, Werre SR. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. BMJ. 2005 May 13; 330 :1179.

17. Cumpston M, Chandler J. Chapter IV: Updating a review. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated August 2019). Cochrane, 2019.
Disponible en www.training.cochrane.org/handbook.
18. Ioannidis JP. The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. Milbank Q. 2016 Sep;94(3):485-514.
19. Page MJ, Shamseer L, Altman DG, et al. Epidemiology and reporting characteristics of systematic reviews of biomedical research: a crosssectional study. PLoS Med. 2016;13(5):e1002028.
20. Shea B, Moher D, Graham I, Pham B, Tugwell P. A comparison of the quality of Cochrane reviews and systematic reviews published in paper-based journals. Eval Health Prof. 2002 Mar;25(1):116-29.
21. Windsor B, Popovich I, Jordan V, Showell M, Shea B, Farquhar C. Methodological quality of systematic reviews in subfertility: a comparison of Cochrane and non-Cochrane systematic reviews in assisted reproductive technologies. Hum Reprod. 2012 Dec;27(12):3460-6.
22. DID-NHS. Diagnostic Imaging Dataset Annual Statistical Release 2017/18. Version number: 1.0 First published: 22nd November 2018. Prepared by: Operational Information for Commissioning. Accesible en: https://www.england.nhs.uk/statistics/wp-content/uploads/sites/2/2018/11/Annual-Statistical-Release-2017-18-PDF-1.6MB-1.pdf.
23. Smith-Bindman R, Kwan ML, Marlow EC, et al. Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016. JAMA. 2019 Sep 3;322(9):843-856.
24. Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: Analysing and Presenting Results. In: Deeks JJ, Bossuyt PM, Gatsonis C (editors), Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. The Cochrane Collaboration, 2010. Disponible en: http://srdta.cochrane.org/.
25. Dent E, Morley JE, Cruz-Jentoft AJ, et al. Physical Frailty: ICFSR International Clinical Practice Guidelines for Identification and Management. J Nutr Health Aging. 2019;23(9):771-787.
26. Gobbens RJ, Luijkx KG, Wijnen-Sponselee MT, Schols JM. Toward a conceptual definition of frail community dwelling older people. Nurs Outlook 2010;58:76-86.
27. NG9 - National Institute for Health and Care Excellence. Bronchiolitis in children: diagnosis and management [Internet]. [London]: NICE; 2015 [updated 2019 Aug; cited 2020 Oct 20]. (NICE guideline [NG9]). Disponible en: www.nice.org.uk/guidance/ng9.
28. González J, Ochoa C, Grupo Investigador del Proyecto aBREVIADo (BRonquiolitis-Estudio de Variabilidad, Idoneidad y ADecuación). Study of variability in the management of acute bronchiolitis in Spain in relation to age of patients. National multicenter study (aBREVIADo project) [Estudio de variabilidad en el abordaje de la bronquiolitis aguda en España en relación con la edad de los pacientes]. Anales de Pediatria (Barcelona) 2010;72(1): 4–18.
29. NCG 16 - Department of Health (2017). Diagnosis, staging and treatment of lung cancer (NCEC National Clinical Guideline No. 16). Disponible en: http://health.gov.ie/national-patient-safety-office/ncec/national-clinical-guidelinesIn text citation (Department of Health 2017) https://assets.gov.ie/11571/078289ce86b848d4a2147cf7f73aba9d.pdf [Consultado el 11/09/2020].
30. Roqué M, Martínez-García L, Solà I, et al. Toolkit of methodological resources to conduct systematic reviews [version 3; peer review: 2 approved]. F1000Research 2020, 9:82.
Disponible en: https://doi.org/10.12688/f1000research.22032.3.
31. Giné-Garriga M, Roqué-Fíguls M, Coll-Planas L, Sitjà-Rabert M, Salvà A. Physical exercise interventions for improving performance-based measures of physical function in community-

dwelling, frail older adults: a systematic review and meta-analysis. Arch Phys Med Rehabil. 2014 Apr; 95(4):753-769.e3.

32. Roqué i Figuls M, Giné-Garriga M, Granados Rugeles C, Perrotta C, Vilaró J. Chest physiotherapy for acute bronchiolitis in paediatric patients between 0 and 24 months old. Cochrane Database Syst Rev. 2016;2(2):CD004873.
33. Perrotta C, Ortiz Z, Roque M. Chest physiotherapy for acute bronchiolitis in paediatric patients between 0 and 24 months old. Cochrane Database Syst Rev. 2007;(1):CD004873.
34. Perrotta C, Ortiz Z, Roque M. Chest physiotherapy for acute bronchiolitis in paediatric patients between 0 and 24 months old. Cochrane Database Syst Rev. 2005;(2):CD004873.
35. Roqué i Figuls M, Giné-Garriga M, Granados Rugeles C, Perrotta C. Chest physiotherapy for acute bronchiolitis in paediatric patients between 0 and 24 months old. Cochrane Database Syst Rev. 2012;(2):CD004873.
36. Garner P, Hopewell S, Chandler J, et al. When and how to update systematic reviews: consensus and checklist. BMJ 2016; 354: i3507
37. Postiaux G, Louis J. Comment on: Chest physiotherapy for acute bronchiolitis in paediatric patients between 0 and 24 months old. 5 March 2012. Disponible en: https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD004873.pub5/detailed-comment/en?messageId=265543688.
38. Main E. Comment on: Chest physiotherapy for acute bronchiolitis in paediatric patients between 0 and 24 months old. 29 April 2016. Disponible en: https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD004873.pub5/detailed-comment/en?messageId=265543679.
39. Remondini R. Comment on: Chest physiotherapy for acute bronchiolitis in paediatric patients between 0 and 24 months old. 4 May 2017. Disponible en: https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD004873.pub5/detailed-comment/en?messageId=265542477.
40. Schmidt-Hansen M, Baldwin DR, Hasler E, Zamora J, Abraira V, Roqué i Figuls M. PET-CT for assessing mediastinal lymph node involvement in patients with suspected resectable non-small cell lung cancer. Cochrane Database of Systematic Reviews 2014, Issue 11. Art. No.: CD009519.
41. Wirth K, Klenk J, Brefka S, et al; SITLESS consortium. Biomarkers associated with sedentary behaviour in older adults: A systematic review. Ageing Res Rev. 2017 May;35:87-111.
42. Calvo-Muñoz I, Kovacs FM, Roqué M, Seco-Calvo J. The association between the weight of schoolbags and low back pain among schoolchildren. A systematic review, meta-analysis and individual patient data meta-analysis. Eur J Pain. 2019;00:1–19.
43. Whiting P, Savović J, Higgins JP, et al; ROBIS group. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol. 2016 Jan;69:225-34.
44. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017 Sep 21;358:j4008.
45. American Academy of Pediatrics Subcommittee on Diagnosis and Management of Bronchiolitis. Diagnosis and management of bronchiolitis. Pediatrics. 2006 Oct;118(4):1774-93.
46. Ralston SL, Lieberthal AS, Meissner HC, et al. Clinical Practice Guideline: The Diagnosis, Management, and Prevention of Bronchiolitis. [published correction appears in Pediatrics. 2015 Oct;136(4):782]. Pediatrics. 2014;134(5):e1474-e1502.

47. Friedman JN, Rieder MJ, Walton JM; Canadian Paediatric Society, Acute Care Committee, Drug Therapy and Hazardous Substances Committee. Bronchiolitis: Recommendations for diagnosis, monitoring and management of children one to 24 months of age. Paediatr Child Health. 2014;19(9):485-498.
48. Schmidt-Hansen M, Baldwin DR, Zamora J. FDG-PET/CT Imaging for Mediastinal Staging in Patients With Potentially Resectable Non–Small Cell Lung Cancer. JAMA. 2015;313(14):1465–1466.
49. Giné-Garriga M, Coll-Planas L, Guerra M, et al. The SITLESS project: exercise referral schemes enhanced by self-management strategies to battle sedentary behaviour in older adults: study protocol for a randomised controlled trial. Trials. 2017 May 18;18(1):221.
50. Wells, G., Shea, B., O'Connell, D., Peterson, J., Welch, V., Losos, M., & Tugwell, P. (2009). The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa, ON: Ottawa Hospital Research Institute. Disponible en: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.
51. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(6): e1000097.
52. Campbell K, Coleman-Haynes T, Bowker K, Cooper SE, Connelly S, Coleman T. Factors influencing the uptake and use of nicotine replacement therapy and e-cigarettes in pregnant women who smoke: a qualitative evidence synthesis. Cochrane Database Syst Rev. 2020;5(5):CD013629.
53. Schünemann HJ, Mustafa RA, Brozek J, et al. GRADE guidelines: 21 part 1. Study design, risk of bias, and indirectness in rating the certainty across a body of evidence for test accuracy. J Clin Epidemiol. 2020;122:129-141.
54. Schünemann HJ, Mustafa RA, Brozek J, et al. GRADE guidelines: 21 part 2. Test accuracy: inconsistency, imprecision, publication bias, and other domains for rating the certainty of evidence and presenting it in evidence profiles and summary of findings tables. J Clin Epidemiol. 2020;122:142-152.
55. CADTH - Canadian Agency for Drugs and Technology in Health.
Disponible en: https://www.cadth.ca/interventions-directed-professionals.
56. Waddington H, Masset E, Jimenez E. What have we learned after ten years of systematic reviews in international development?. Journal of Development Effectiveness. 2018; 10:1, 1-16.
57. Pieper D, Puljak L, González-Lorenzo M, Minozzi S. Minor differences were found between AMSTAR 2 and ROBIS in the assessment of systematic reviews including both randomized and nonrandomized studies. J Clin Epidemiol. 2019;108:26-33.
58. Gates M, Gates A, Duarte G, et al. Quality and risk of bias appraisals of systematic reviews are inconsistent across reviewers and centers. J Clin Epidemiol. 2020;125:9-15.
59. Martis R, Crowther CA, Shepherd E, Alsweiler J, Downie MR, Brown J. Treatments for women with gestational diabetes mellitus: an overview of Cochrane systematic reviews. Cochrane Database of Systematic Reviews 2018, Issue 8. Art. No.: CD012327.
60. Livingston G, Huntley J, Sommerlad A, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. Lancet. 2020;396(10248):413-446.
61. Dietz P, Reichel JL, Edelmann D, et al. A Systematic Umbrella Review on the Epidemiology of Modifiable Health Influencing Factors and on Health Promoting Interventions Among University Students. Front Public Health. 2020 Apr 28;8:137.
62. Yu, Y, Shi, Q, Zheng, P, et al. Assessment of the quality of systematic reviews on COVID-19: A comparative study of previous coronavirus outbreaks. J Med Virol. 2020; 92: 883–890.

63. Santaguida PL, Keshavarz H, Carlesso LC, et al; ICON Working Group. A description of the methodology used in an overview of reviews to evaluate evidence on the treatment, harms, diagnosis/classification, prognosis and outcomes used in the management of neck pain. Open Orthop J. 2013 Sep 20;7:461-72.
64. Lucaroni F, Cicciarella Modica D, Macino M, et al. Can risk be predicted? An umbrella systematic review of current risk prediction models for cardiovascular diseases, diabetes and hypertension. BMJ Open 2019;9:e030234.
65. Gao Y, Liu M, Shi S, et al. Cancer Biomarker Assessment Working Group. Diagnostic value of seven biomarkers for breast cancer: an overview with evidence mapping and indirect comparisons of diagnostic test accuracy. Clin Exp Med. 2020 Feb;20(1):97-108.
66. Gast A, Mathes T. Medication adherence influencing factors-an (updated) overview of systematic reviews. Syst Rev. 2019 May 10;8(1):112.
67. Rodrigues BS, Alves M, Duarte GS, Costa J, Pinto FJ, Caldeira D. The impact of influenza vaccination in patients with cardiovascular disease: An overview of systematic reviews. Trends Cardiovasc Med. 2020 Jun 12:S1050-1738(20)30082-7.
68. Munn Z, Stern C, Aromataris E, et al.: What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. BMC Med Res Methodol. 2018; 18(1): 5.
69. Riley RD, Moons KGM, Snell KIE, et al.: A guide to systematic review and meta-analysis of prognostic factor studies. BMJ. 2019; 364: k4597.
70. Pollock A, Berge E: How to do a systematic review. Int J Stroke. 2018; 13(2): 138–56
71. Mbuagbaw L, Lawson DO, Puljak L,et al. A tutorial on methodological studies: the what, when, how and why. BMC Med Res Methodol 20, 226 (2020).
72. Lawson DO, Leenus A, Mbuagbaw L. Mapping the nomenclature, methodology, and reporting of studies that review methods: a pilot methodological review. Pilot Feasibility Stud. 2020 Jan 30;6:13.
73. Lockwood C, Porrit K, Munn Z, et al. Chapter 2: Systematic reviews of qualitative evidence. In: Aromataris E, Munn Z (Editors). JBI Manual for Evidence Synthesis. JBI, 2020.
Disponible en https://synthesismanual.jbi.global.
74. Vandvik PO, Brignardello-Petersen R, Guyatt GH. Living cumulative network meta-analysis to reduce waste in research: A paradigmatic shift for systematic reviews? BMC Med. 2016 Mar 29;14:59.
75. Guise JM, Chang C, Butler M, Viswanathan M, Tugwell P. AHRQ series on complex intervention systematic reviews-paper 1: an introduction to a series of articles that provide guidance and tools for reviews of complex interventions. J Clin Epidemiol. 2017 Oct;90:6-10.
76. Ioannidis J. Next-generation systematic reviews: prospective meta-analysis, individual-level data, networks and umbrella reviews. British Journal of Sports Medicine 2017;51:1456-1458.
77. Møller MH, Ioannidis JPA, Darmon M. Are systematic reviews and meta-analyses still useful research? We are not sure. Intensive Care Med. 2018 Apr;44(4):518-520.
78. Tricco AC, Langlois EV, Straus SE, editors. Rapid reviews to strengthen health policy and systems: a practical guide. Geneva: World Health Organization; 2017. Licence: CC BY-NC-SA 3.0 IGO.
79. Marshall IJ, Marshall R, Wallace BC, Brassey J, Thomas J. Rapid reviews may produce different results to systematic reviews: a meta-epidemiological study. J Clin Epidemiol. 2019 May;109:30-41. Epub 2018 Dec 25.

80. Elliott JH, Synnot A, Turner T, et al. Living Systematic Review Network. Living systematic review: 1. Introduction-the why, what, when, and how. J Clin Epidemiol. 2017 Nov;91:23-30.
81. van Mastrigt G, Hiligsmann M, Arts J, et al. How to prepare a systematic review of economic evaluations for informing evidence-based healthcare decisions: a five-step approach (part 1/3), Expert Review of Pharmacoeconomics & Outcomes Research. 2016. 16:6, 689-704.
82. Marshall IJ, Wallace BC: Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019; 8(1): 163.
83. Noonan E, Bjørndal A. The Campbell Collaboration. Cochrane Database Syst Rev. 2010 Sep 8;2011:ED000011.
84. Dacombe R. Systematic Reviews in Political Science: What Can the Approach Contribute to Political Research?. Political Studies Review. 2018;16(2):148-157.
85. Sargeant JM, O'Connor AM. Scoping Reviews, Systematic Reviews, and Meta-Analysis: Applications in Veterinary Medicine. Front Vet Sci. 2020 Jan 28;7:11.
86. Koutsos TM, Menexes GC, Dordas CA. An efficient framework for conducting systematic literature reviews in agricultural sciences. Science of The Total Environment. 2019;682: 106-117.
87. Ciapponi A. AMSTAR-2: herramienta de evaluación crítica de revisiones sistemáticas de estudios de intervenciones de salud. Evid Act Pract Ambul. 2018; 21(1):4-13 Traducido, resumido y comentado de: Shea BJ, y col. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. 2017; 358.j4008.

Anexos

9 Anexos

Anexo 1: Abreviaturas

DM	Diferencia de medias
ECA	Ensayos clínicos aleatorizados
FI	Factor de impacto
GPC	Guías de práctica clínica
GRADE	Grading of Recommendations Assessment, Development and Evaluation
IC	Intervalo de confianza
NICE	Instituto Nacional de Salud y Atención de Excelencia
NSCLC	Cáncer de pulmón no microcítico
PET-CT	Tomografía con emisión de positrones combinada con tomografía computada
RR	razón de riesgo
RS	Revisión sistemática
SUV	Valores de captación estándar

Anexo 2. Publicaciones complementarias

Se presentan dos publicaciones anexas que complementan los resultados de este trabajo de tesis. Estas publicaciones son:

Publicación 5

Calvo-Muñoz I, Kovacs FM, Roqué M, Seco-Calvo J. The association between the weight of schoolbags and low back pain among schoolchildren. A systematic review, meta-analysis and individual patient data meta-analysis. Eur J Pain. 2019;00:1–19.

FI: 3.492 (2019). Puntuación de atención Altmetric: 86

Esta publicación puede consultarse al completo y de forma libre en
https://onlinelibrary.wiley.com/doi/10.1002/ejp.1471

Received: 5 January 2019 | Revised: 24 July 2019 | Accepted: 11 August 2019
DOI: 10.1002/ejp.1471

REVIEW ARTICLE

EJP European Journal of Pain WILEY

The association between the weight of schoolbags and low back pain among schoolchildren: A systematic review, meta-analysis and individual patient data meta-analysis

Inmaculada Calvo-Muñoz[1,2] | **Francisco M. Kovacs**[2,3] | **Marta Roqué**[2,4,5] | **Jesús Seco-Calvo**[2,6,7]

[1]Faculty of Health Sciences, Catholic University San Antonio, UCAM, Murcia, Spain
[2]Spanish Back Pain Research Network, Madrid, Spain
[3]Unidad de la Espalda Kovacs, Hospital Universitario de Moncloa, Madrid, Spain
[4]Iberoamerican Cochrane Centre, Biomedical Research Institute Sant Pau (IIB Sant Pau), Barcelona, Spain
[5]CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain
[6]Institute of Biomedicine (IBIOMED), University of León, León, Spain
[7]University of the Basque Country, Leioa, Spain

Correspondence
Francisco M. Kovacs, Unidad de la Espalda, Hospital Universitario HLA-Moncloa, Avda. Menéndez Pelayo, 67, Madrid 28006, Spain.
Email: fmkovacs@kovacs.org

Funding information This study did not receive any funding. The authors were the only parties responsible for; designing and conducting the study, collecting and managing data, analysing and interpreting the data, preparing, reviewing and approving the manuscript, and deciding to submit it for publication.

Abstract

Background: The objective of this study was to determine whether carrying a heavy schoolbag is associated to a higher prevalence of low back pain (LBP).

Methods: A systematic review and meta-analysis was conducted (PROSPERO, CRD42018077839). Observational studies analysing the relationship between schoolbag weight and LBP, were searched for in 20 electronic databases and 12 specialized journals until February 28th, 2019, without date or language restrictions. All studies which included ≥ 50 subjects aged 9 to 16, were reviewed. Methodological quality was assessed by two reviewers separately, using validated tools. A meta-analysis and an individual patient data (IPD) meta-analysis were conducted to examine the relationship between schoolbag weight and LBP. Certainty of evidence was assessed using an adapted GRADE methodology.

Results: 5,524 citations were screened, 21 studies (18,296 subjects) were reviewed and 11 studies (9,188 subjects) were included in the meta-analysis. The IPD meta-analysis included 9,188 subjects from seven studies. Among the 21 studies reviewed, the mean score for methodological quality was 78.3 of 100. Only one study suggested an association between heavier schoolbags and LBP. Neither the meta-analysis nor the IPD meta-analysis found an association between carrying schoolbags weighing > 10% of bodyweight, and LBP. No differences based on age, gender or sport activity were found.

Discussion: Available evidence does not support that schoolbags weighing > 10% of bodyweight are associated with a higher prevalence of LBP among schoolchildren aged 9–16. The certainty of evidence is low. Further research is required on the relationship between schoolbag weight and LBP.

Significance: This systematic review, with a meta-analysis and an IPD meta-analysis, failed to find a link between schoolbags weighing ≥ 10% of body weight and LBP among schoolchildren aged 9 to 16. Further longitudinal studies, with large samples, long follow-up periods, and rigorous methods taking into account duration of carry and the physical capacity of each subject, are required in this field.

Abbreviations: CI 95%, confidence interval 95%; HKSJ, Hartung-Knapp-Sidik-Jonkman method; IPD meta-analysis, individual patient data meta-analysis; LBP, low back pain; OR, odds ratio.

1 | BACKGROUND

Common low back pain (LBP) is defined as pain between the costal margins and the inferior gluteal folds, which is usually accompanied by painful limitation of movement, may be associated with pain referred down to the leg, and is not related to fracture, direct trauma or systemic diseases, such as neoplastic, infectious, vascular, metabolic or endocrine-related processes (Bardin, King, & Maher, 2017; Hoy et al., 2012; Maher, Underwood, & Buchbinder, 2017).

LBP represents a major health, social and economic burden. It is the main cause of years lived with disability worldwide (Hoy et al., 2014) and, only in the United States, the yearly costs associated with the condition have been estimated at 100 billion dollars (Dieleman et al., 2016).

LBP is more common among schoolchildren than previously believed (Calvo-Muñoz, Gómez-Conesa, & Sánchez-Meca, 2013; Kamper, Yamato, & Williams, 2016), with a lifetime prevalence of 47% at 14 years (Swain et al., 2014). Moreover, reporting LBP during adolescence is a risk factor for suffering it in adulthood (Hestbaek, Leboeuf-Yde, & Kyvik, 2006).

Many factors have been suggested to be associated with a higher risk of LBP among schoolchildren (Calvo-Muñoz, Kovacs, Roqué, Gago Fernández, & Seco Calvo, 2018), including biological (e.g., body weight, weight bearing, muscle strength or ergonomics) (Fairbank, Pynsent, Poortvliet, & Phillips, 1984; Sano et al., 2015; Yamato, Maher, Traeger, Wiliams, & Kamper, 2018), psychosocial (e.g., family and social relations or satisfaction with school) (Dianat, Alipour, & Asghari Jafarabadi, 2017; Mikkonen et al., 2016) and lifestyle related variables (e.g., physical activity, participation in sports or smoking) (Kovacs et al., 2003; Wedderkopp, Leboeuf-Yde, Bo Andersen, Froberg, & Hansen, 2003).

A recent systematic review has found that the evidence assessing the association between these factors and LBP in childhood and adolescence, is inconsistent (Calvo-Muñoz et al., 2018). It also suggested that the weight of the schoolbags schoolchildren carry to school is one of the factors which should be researched further. In fact, over 80% of schoolchildren reporting LBP blame the excessive weight of the schoolbag for their pain (Skaggs, Early, D'Ambra, Tolo, & Kay, 2006), the percentage of body weight that the schoolbag represents among schoolchildren exceeds the limit recommended for adults carrying weight in the work environment, which is usually established at 10% (Alghadir, Gabr, & Al-Eisa, 2017; Erne & Elfering, 2011; Grimmer & Williams, 2000; Negrini, Carabalona, & Sibilla, 1999; Spiteri et al., 2017), and several studies have suggested that excessively heavy schoolbags may be a risk factor for LBP (Dockrell, Simms, & Blake, 2013; Goodgold & Nielsen, 2003; Mackenzie, Sampath, Kruse, & Sheir-Neiss, 2003; Moore, White, & Moore, 2007; Rateau, 2004; Siambanes, Martinez, Butler, & Haider, 2004; Viry, Creveuil, & Marcelli, 1999).

However, most of the studies assessing the relationship between schoolbag weight and LBP among school students, are exploratory studies of low methodological quality (Huguet et al., 2013), and their heterogeneity makes it difficult to perform systematic reviews and meta-analysis (Dretzke et al., 2014; Riley et al., 2013). Individual Participant Data (IPD) meta-analysis offers advantages over meta-analysis of summary data, such as the opportunity to standardize the categorization of exposure variables and to explore heterogeneity through subgroup analysis (Stewart & Tierney, 2002).

Hence, IPD meta-analysis would be appropriate to analyse the existing data on the potential relationship between carrying schoolbags weighing more than 10% of bodyweight, and LBP among schoolchildren (Abo-Zaid, Sauerbrei, & Riley, 2012).

Therefore, the objective of this study was to perform a systematic review, coupled with a meta-analysis and IPD meta-analysis, to estimate whether carrying a heavier schoolbag, and specifically one weighing > 10% of bodyweight, is associated with a higher prevalence of LBP among schoolchildren aged 9–16.

2 | METHODS

The protocol of this meta-analysis was registered in an international register (PROSPERO, CRD42018077839).

2.1 | Search and study selection

An electronic search was conducted up to February 28th, 2019, in the following databases: CINAHL,Current Contents, EMBASE, Family health database, FSTA (Food Science and Technology Abstracts), ISI Web of Knowledge, LILACS, MEDLINE, NNNConsult, OvidMD, PEDro, ProQuest Central, PubMed, SciFinder Scholar, Science Direct, Scopus, SPORTDiscus, The Cochrane Library, Web of Science, Wiley Online Library. The search strategy used both MeSH and terms in "all fields", and was designed to ensure maximum sensitivity. It was conducted in seven successive phases, as shown in Appendix 1, adding references retrieved at each phase to those identified in the previous ones.

Additionally, an electronic search was conducted in the Websites of the Journals which were considered more likely to publish high quality studies on LBP in children. These Journals were: "Pain", "European Journal of Pain", "Clinical Journal of Pain", "Spine", "Spine Journal", "European Spine Journal", "Open Journal of Pediatrics", "European Journal of Pediatrics", "European Journal of Public Health", "Scandinavian J Public Health", "Ergonomics" and "Applied Ergonomics". This second electronic search combined the terms “adolescent”, “children”, “schoolchildren”, “young”,

"pediatric", "back pain", "low back pain", "lumbar pain", "prevalence", "epidemiology", "risk factors", "schoolbags", "school bag", "backpack", "carrying bag" and "bag".

Finally, references in the reviewed studies were manually tracked to identify additional studies.

All the references identified were listed and crosschecked to delete redundancies. The title and abstract of each study were screened by two authors separately (ICM and JSC). The full texts of those which were eligible were assessed for inclusion criteria by two authors separately (ICM and JSC). Disagreements at the screening and assessment stages were resolved through consensus with a third author (FMK).

Studies were included in this review if they: (a) were published, observational studies focusing on risk factors for LBP, either cross-sectional or longitudinal, including case–control studies and cohort studies; (b) included ≥ 50 subjects aged 9–16; (c) explored the relationship between the schoolbag weight (as a proportion of bodyweight) and LBP. No date or language restrictions were applied.

Among the studies included in the review, those which also provided estimates of the association between schoolbag weight (as a proportion of body weight) and LBP, with the corresponding 95% CI, were included in the meta-analysis.

The authors of the studies included in the meta-analysis were contacted and requested to provide the datasets of their original studies, with the names of the subjects deleted and substituted by codes. An individual patient data meta-analysis (IPD meta-analysis) was conducted including all data from all the studies whose authors had provided the data requested. In order to assess the completeness and consistency of these data, the analyses performed in each original study were reproduced, and consistency of results with those published was assessed.

2.2 | Variables

For the systematic review and the (non-IPD) meta-analysis, the weight of the schoolbag was categorized as "hot heavy" or "heavy" based on the definition implemented in the original studies. For the IPD meta-analysis, >10% of body weight was used as the cut-off value (Devroey, Jonkers, Becker, Lenaerts, & Spaepen, 2007; Mackie & Legg, 2008).

Covariates were; age, gender and sport activity outside the school (yes/no), since previous studies suggest that the prevalence of LBP increases with age, being female and practicing sports on a competitive level, and results from a previous systematic review on factors associated with a higher prevalence of LBP among children aged 9–16, supported this (Calvo-Muñoz et al., 2018).

2.3 | Quality assessment of the studies

The Ottawa-Newcastle Scale was selected to assess the methodological quality of longitudinal studies (Wells et al., 2009). This scale scores the quality of each study from 0 to 12 points (from worst to best). The score is composed of a maximum of 9 points assigned to studies meeting eight specific methodological criteria (one of these criteria is scored with up to 2 points), and 3 additional points attributed to studies complying with three criteria which are specific for LBP.

A tool used previously in systematic reviews and meta-analyses on the prevalence of LBP (Calvo-Muñoz et al., 2018; Loney & Stratford, 1999; Louw, Morris, & Grimmer-Somers, 2007; Walker, 2000), was selected to assess the methodological quality of the cross-sectional studies included in this review (Table S1). This tool includes 12 questions and assesses; how detailed the definition of LBP was (precision of anatomic description, gathering of data on frequency, duration, severity, etc.), validity of the sample (representativeness, description and management of missing data, etc.) and data quality (data gathered directly from subjects -without intermediaries-, methods for data gathering, etc.). The score ranges, from worst to best, from 0% to 100%.

Two independent authors conducted assessments separately (ICM and JSC), and disagreements were solved through consensus with a third author (FMK).

2.4 | Analyses

Odds Ratios and 95% CI were calculated to assess the association between reporting LBP and using a schoolbag weighing > 10% of body weight, versus a schoolbag weighing ≤ 10%.

A meta-analysis was conducted under the random effects model, in order to calculate a combined OR based on the individual OR for each study. An IPD meta-analysis was subsequently conducted. For both meta-analyses, the combined OR and corresponding 95% CI were converted post-hoc to the Hartung-Knapp-Sidik-Jonkman method (IntHout, Ioannidis, & Borm, 2014).

For all the studies for which adjusted ORs were available, the ORs adjusted in the most saturated models were included in the meta-analysis. When studies only presented unadjusted ORs, the latter were included. In the IPD meta-analysis, ORs were adjusted for age and gender.

In the meta-analysis, the same definition of "heavy" schoolbag which had been used in the original studies, was maintained. In the IPD meta-analysis, the cut-off value to define a bag as "heavy" was > 10% of bodyweight.

Heterogeneity was assessed through the I^2 (Higgins & Thompson, 2002; Higgins, Thompson, Deeks, & Altman, 2003). I^2 values between 30% and 60% were considered to be indicative of moderate heterogeneity, and values > 60% were considered to indicate considerable heterogeneity.

Two sub-group analyses were conducted in order to explore possible differences in the relationship between

schoolbag weight (≤10% vs.> 10% of bodyweight) and LBP, depending on age and sport activity. The first sub-group analysis categorized the schoolchildren in "children" (≤ 12 years of age) and "teenagers" (≥ 13 years), and combined risk estimates adjusted for gender. The second sub-group analysis categorized schoolchildren depending on whether they performed any sport activities outside the school, and combined estimates adjusted for age and gender.

2.5 | Assessment of certainty of evidence

Certainty of the evidence in the review was assessed using an adaptation of the GRADE system for studies on prognostic factors (Huguet et al., 2013).

Certainty of evidence was assessed based on the extent to which users can be confident that the estimated prognostic association reflects the item being evaluated (Guyatt et al., 2008), taking into account limitations in the methodology of the studies included, the inconsistency, indirectness and imprecision of results, and the potential for publication bias in the review (Huguet et al., 2013).

The certainty of evidence for prognosis was initially considered to be high, given that this review focuses on Phase 2 explanatory studies "aimed to confirm independent associations between potential prognostic factor and the outcome" (Huguet et al., 2013). This initial assessment was later downgraded due to limitations in five factors (study limitations, inconsistency, indirectness, imprecision and publication bias) and/or upgraded due to impact factors (moderate or large estimated effects and dose–response effects) (Huguet et al., 2013).

Assessment of certainty of evidence was conducted by two independent authors (ICM and JSC), and a consensus was reached through discussion with a third author (MR).

3 | RESULTS

The search strategies provided a total of to 5,524 records, which were reduced to 5,502 after duplicates were removed, and led to 84 full-text articles being assessed for eligibility.

Among these 84 studies, 21 representing a total sample of 18,296 complied with the criteria to be included in the systematic review (Akbar et al., 2019; Alghadir et al., 2017; Alghamdi, Nafee, El-Sayed, & Alsaadi, 2018; Angarita-Fonseca et al., 2019; Chiang, Jacobs, & Orsmond, 2006; de Oliveira, Chinaglia, & Lima, 2017; Dianat et al., 2017; Dianat, Sorkhi, Pourhossein, Alipour, & Asghari-Jafarabadi, 2014; Grimmer & Williams, 2000; Johnson, Adeniji, Mbada, Obembe, & Akosile, 2011; Korovessis, Koureas, & Papazisis, 2004; Martínez-Crespo et al., 2009; Minghelli, Oliveira, & Nunes, 2016; Mohseni-Bandpei, Bagheri-Nesami, & Shayesteh-Azar, 2007; Mwaka, Munabi, Buwembo, Kukkiriza, & Ochieng, 2014; Noormohammadpour et al., 2019; Oka, Ranade, & Kulkarni, 2019; Trevelyan & Legg, 2011; Vidal, Borràs, Ponseti, Gili, & Palou, 2010; Watson et al., 2003; Young, Haig, & Yamakawa, 2006), 11 with 9,188 subjects were included in the meta-analysis (Akbar et al., 2019; Alghadir et al., 2017; Angarita-Fonseca et al., 2019; de Oliveira et al., 2017; Dianat et al., 2017, 2014; Martínez-Crespo et al., 2009; Minghelli et al., 2016; Mohseni-Bandpei et al., 2007; Trevelyan & Legg, 2011; Vidal et al., 2010), and data from 8,218 subjects from 7 studies were included in the IPD meta-analysis (Akbar et al., 2019; de Oliveira et al., 2017; Martínez-Crespo et al., 2009; Minghelli et al., 2016; Mohseni-Bandpei et al., 2007; Trevelyan & Legg, 2011; Vidal et al., 2010). Figure 1 shows the PRISMA flow diagram.

3.1 | Systematic review

All of the 21 studies included in the review were cross-sectional, and in all of them the outcome was the prevalence of LBP. Six studies focused on the point prevalence (Alghamdi et al., 2018; de Oliveira et al., 2017; Johnson et al., 2011; Korovessis et al., 2004; Noormohammadpour et al., 2019; Young et al., 2006), 4 on the 1- or 2-week prevalence (Chiang et al., 2006; Grimmer & Williams, 2000; Mwaka et al., 2014; Vidal et al., 2010), 10 on the 1-month prevalence (Akbar et al., 2019; Angarita-Fonseca et al., 2019; Dianat et al., 2017, 2014; Martínez-Crespo et al., 2009; Mohseni-Bandpei et al., 2007; Noormohammadpour et al., 2019; Oka et al., 2019; Trevelyan & Legg, 2011; Watson et al., 2003), 2 on the 1-year prevalence (Alghadir et al., 2017; Minghelli et al., 2016) and 3 on the lifetime prevalence (Akbar et al., 2019; Noormohammadpour et al., 2019; Vidal et al., 2010). Table 1 shows the main characteristics of the studies included in the review.

Sample size of the studies ranged between 55 and 4,813. Three studies included only children (Angarita-Fonseca et al., 2019; de Oliveira et al., 2017; Vidal et al., 2010), and three studies included only teenagers (Akbar et al., 2019; Alghamdi et al., 2018; Noormohammadpour et al., 2019), while all the others included both children and teenagers. Two studies included only girls (Alghamdi et al., 2018; Noormohammadpour et al., 2019) while all the others included both boys and girls.

Among these 21 studies, 19 used standardized self-report questionnaires to determine the prevalence of LBP (Akbar et al., 2019; Alghadir et al., 2017; Alghamdi et al., 2018; Angarita-Fonseca et al., 2019; Chiang et al., 2006; de Oliveira et al., 2017; Dianat et al., 2017, 2014; Grimmer & Williams, 2000; Johnson et al., 2011; Martínez-Crespo et al., 2009; Minghelli et al., 2016; Mohseni-Bandpei et al., 2007; Mwaka et al., 2014; Noormohammadpour et al., 2019; Oka et al., 2019; Trevelyan & Legg, 2011; Vidal et al., 2010; Watson et al., 2003). One study used a non-standardized, ad hoc self-questionnaire (Young et al., 2006), and the last one verbally asked the subjects a non-standardized question (Korovessis et al., 2004).Twelve of these 21 studies (including the two studies which did not use standardized self-report

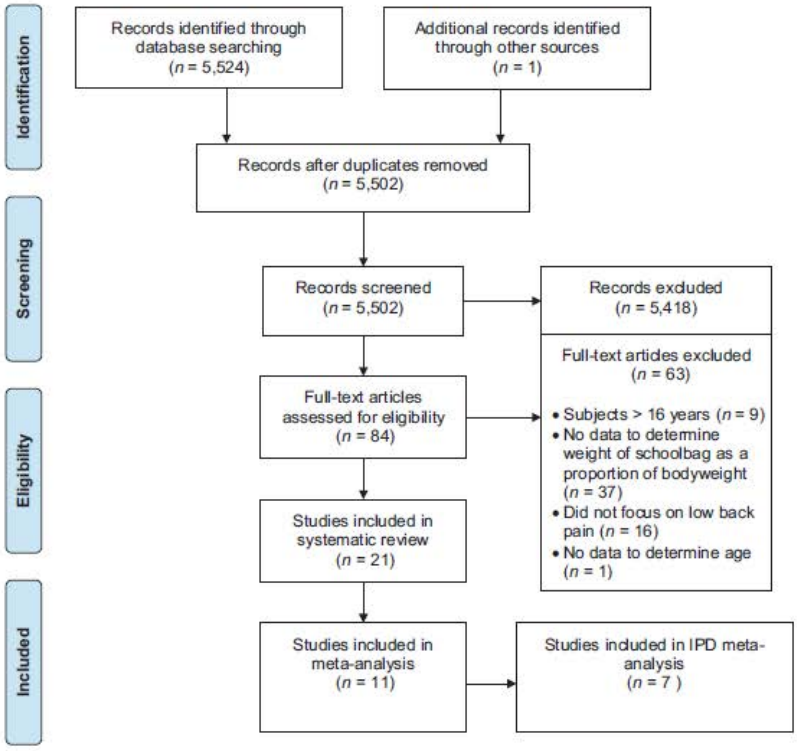


FIGURE 1 PRISMA flow diagram of the study

questionnaires), also used some form of physical examination (Akbar et al., 2019; Alghadir et al., 2017; Chiang et al., 2006; de Oliveira et al., 2017; Johnson et al., 2011; Minghelli et al., 2016; Mohseni-Bandpei et al., 2007; Mwaka et al., 2014; Noormohammadpour et al., 2019; Trevelyan & Legg, 2011).

Among the studies included in the systematic review, the relationship between the weight of the schoolbag and body weight was assessed quantitatively in six studies (Chiang et al., 2006; Korovessis et al., 2004; Noormohammadpour et al., 2019; Oka et al., 2019; Vidal et al., 2010; Young et al., 2006). The other 15 studies classified this relationship into categories. The cut-off values for establishing these categories varied across studies, as follows:">6.6% of bodyweight" in one study (Alghadir et al., 2017),">10% of bodyweight" in 10 studies (Akbar et al., 2019; de Oliveira et al., 2017; Dianat et al., 2017, 2014; Grimmer & Williams, 2000; Martínez-Crespo et al., 2009; Minghelli et al., 2016; Mohseni-Bandpei et al., 2007; Mwaka et al., 2014; Trevelyan & Legg, 2011),"≥12% of bodyweight" in one study (Angarita-Fonseca et al., 2019), two categories ("10% to 15%", ">15%") in one study (Alghamdi et al., 2018), five categories ("2.2% to 6.6%"; "6.7% to 8.8%"; "8.9% to 10.5%"; "10.6% to 13.5%" and "13.6% to 32.1%") in one study (Watson et al., 2003), and "normal and abnormal" (without defining the classification criteria in these categories) in one study (Johnson et al., 2011).

Ten studies gathered data on whether subjects participated in sports outside the school hours (Akbar et al., 2019; Alghadir et al., 2017; Angarita-Fonseca et al., 2019; Dianat et al., 2017; Johnson et al., 2011; Martínez-Crespo et al., 2009; Noormohammadpour et al., 2019; Oka et al., 2019; Vidal et al., 2010; Young et al., 2006).

Among the studies included in the review, the mean score for methodological quality was 78.3% of 100%, with all scores ranging between 55% and 100% (Table 2). Six studies scored ≤ 70%, five scored between 71% and 80%, five scored between 81% and 90% and three over 90%. Table 2 shows the methodological strengths and weaknesses of each study.

Among the 10 studies which were not included in the meta-analyses, only one suggested that LBP was more prevalent among the children who carried schoolbags representing a higher percentage of their bodyweight, especially among boys (vs. girls), with the percentage of the bodyweight associated with a higher risk for LBP being smaller when the children were younger (Grimmer & Williams, 2000). None of the other nine studies which were not included in the meta-analysis, found such an association (Alghamdi et al., 2018; Chiang et al., 2006; Johnson et al., 2011; Korovessis et al., 2004; Mwaka et al., 2014; Noormohammadpour et al., 2019; Oka et al., 2019; Watson et al., 2003; Young et al., 2006).

3.2 | Meta-analysis

Seven studies were included in the IPD meta-analysis (Akbar et al., 2019; de Oliveira et al., 2017; Martínez-Crespo et al.,

TABLE 1 Main characteristics of the studies included in the systematic review

Study	*N*	Age (years)	Prevalence of LBP (D: Days, M: Month, Y: Years, P: Point prevalence)	Method for assessing LBP	Weight of schoolbag	OR (95% CI) for LBP among subjects with heavier schoolbags	Adjusted analysis	Included in meta-analysis	Included in the IPD Meta-analysis
Alghadir et al., 2017	250	12–16	1 y: 130/250 (52.0%)	STD + PE	>6.6% of body weight: 136/250 (54.4%)	1.1 (0.86–1.32)	Logistic regression adjusted for age and gender	Yes	No
Alghamdi et al., 2018	300	13–15	P: 75/300 (25%)	STD	>15% of body weight: 289/300 (96.3%)	NA	NA	No	No
Akbar et al., 2019	950	14–19 (16.7)	L: 668/950 (70.3%) 6 m: 466/950 (49.1%) 1 m: 293/950 (30.8%)	STD + PE	>10% of body weight: 150/299 (50.17%)	1.21 (0.92–1.59	Unconditional logistic regression.	Yes	Yes
Angarit-Fonseca et al., 2019	73	10–12	1 m: 15/73 39.7%	STD	Among subjects with LBP; 12%–20% of body weight (16/55, *SD* 3.75) Among those without LBP; 12%–20% of body weight (19/55, *SD* 3.75)	PR (prevalence ratio) Bivariate analysis 1.33 (0.76–2.36) Bivariate analysis 1.33 (0.76–2.36) Multivariate analysis 1.88 (1.04–3.39)	Bivariate analysis and Multivariate analysis	Yes	No
Chiang et al., 2006	55	13–14	14 d: 19/55 (34.5%)	STD + PE	Among subjects with LBP; 10% of body weight (*SD* 3.75) Among those without LBP: 8% (*SD* 3.86)	NA	NA	No	No
de Oliveira et al., 2017	217	6–10	PC: 11/217 (5.07%)	STD + PE	>10% of body weight: 49/74 (66.22%)	0.72 (0.11–4.70)	Logistic regression	Yes	Yes
Dianat et al., 2014	586	12–14	1 m: 193/586 (32.9%)	STD	Not reported	1.50 (0.59–3.81)	Univariate logistic regression	Yes	No
Dianat et al., 2017	1611	11–14	1 m: 553/1611 (34.3%)	STD	>10% of body weight: 33/64 (51.6%)	1.14 (0.69–1.87)	Univariate logistic regression	Yes	No
Grimmer & Williams, 2000	1,193	12–17	14 d: 293/1193 (24.6%)	STD	>10% of body weight 387/1,193 (32.4%)	NA	NA	No	No
Johnson et al., 2011	381	10–16	P: 70/381 (18.3%)	STD + PE	"Abnormal" (not defined): 163/381 (42.8%)	NA	Univariate logistic regression	No	No
Korovessis et al., 2004	3,441	9–15	P: 671/3441 (19.5%)	NSTD + PE	Average 4,6% of body weight (*SD* 12)	NA	NA	No	No
Martínez-Crespo et al., 2009	849	12–16	1 m: 560/849 (66.0%)	STD	>10% of body weight: 580/848 (68.4%)	0.898 (0.65–1.24)	Logistic regression adjusted for age and gender	Yes	Yes

(Continues)

TABLE 1 (Continued)

Study	*N*	Age (years)	Prevalence of LBP (D: Days, M: Month, Y: Years, P: Point prevalence)	Method for assessing LBP	Weight of schoolbag	OR (95% CI) for LBP among subjects with heavier schoolbags	Adjusted analysis	Included in meta-analysis	Included in the IPD Meta-analysis
Minghelli et al., 2016	966	10–16	1 y: 456/966 (47.2%)	STD + PE	>10% of body weight: 397/966 (41.1%)	0.97 (0.74–1,27)	Logistic regression adjusted for age and gender	Yes	Yes
Mohseni-Bandpei et al., 2007	4,813	11–14	1 m: 695/4813 (14.4%)	STD + PE	>10% of body weight: 13/4813 (0,3%)	1.29 (0.28–5.92)	Logistic regression adjusted for age and gender	Yes	Yes
Mwaka et al., 2014	532	10–21	14 d (pain for ≥ 1 d):[a] 201/532 (37.8%)	STD + PE	>10% of body weight: 164/532 (30.8%)	Not reported	Univariate logistic regression	No	No
Oka et al., 2019	163	12–16	1 m 22/163 (13.50%)	STD	>10% of body weight: 124/163 (76.1%)	NA	NA	No	No
Noormohammadpour et al., 2019	372	13–18	L: 172/372 (46.2%) 3 m: 43/372 (11.6%) 1 m: 116/372 (31.2%) P: 84/372 (22.6%)	STD + PE	Not reported	NA	NA	No	No
Trevelyan & Legg, 2011	245	7–14	1 m (pain for > 1 d):[a] 75/245 (30.6%)[b]	STD + PE	>10% of body weight: 8/233 (3.4%)	0.77 (0.15–0.94)	Logistic regression adjusted for age and gender	Yes	Yes
Vidal et al., 2010	178	10–12	L: 38/178 (21.2%) 7 d: 16/178 (9.0%)	STD	>10% of body weight 108/178 (60.7%)	1.18 (0.40–3.49)	Logistic regression adjusted for age and gender	Yes	Yes
Watson et al., 2003	1,446	11–14	1 m: 330/1446 (23.9%)	STD	>10% of body weight 419/1048	NA	Logistic regression adjusted for age and gender	No	No
Young et al., 2006	125	11–14	P: 57/125 (45.6%)	NSTD + PE	Subjects with LBP: mean 10.9% of body weight (*SD* 4.8) Subjects without LBP: mean: 11.8% (*SD* 4.9)	NA	NA	No	No

Abbreviations: D, days; IPD, individual patient data; LBP, low back pain; M, months; NSTD, non-standardized self-report; P: point prevalence; PC, point prevalence: pain when carrying the schoolbag; PE, physical examination; *SD*, Standard Deviation; STD, standardized self-reported questionnaire; Y, years.

[a]In Mwaka 2014 and Trevelyan 2011, data on prevalence relates to pain lasting longer than the specified duration (e.g., in Trevelyan 2011, 30.6% of the sample reported having suffered from low back pain lasting for over 1 day, during the month before data gathering).

[b]Figures based on IPD data. They compute all the subjects included in the original study, 75 of which reported LBP (the published paper only reports 67 subjects with LBP).

(Continues)

2009; Minghelli et al., 2016; Mohseni-Bandpei et al., 2007; Trevelyan & Legg, 2011; Vidal et al., 2010).

Four additional studies were included in the meta-analysis, but could not be included in the IPD meta-analysis because their authors did not provide IPD data (Alghadir et al., 2017; Angarita-Fonseca et al., 2019; Dianat et al., 2017, 2014). Among these four studies, three defined "heavy" schoolbag using the > 10% bodyweight threshold (Angarita-Fonseca et al., 2019; Dianat et al., 2017, 2014), while the last one had used > 6.6% (Alghadir et al., 2017).

The methodological scores of studies included in the meta-analysis ranged between 55 and 91, and those included in the IPD meta-analysis ranged between 70 and 91 (Tables 1 and 2).

The meta-analysis failed to identify a significant association between carrying schoolbags representing > 10% of bodyweight, and the prevalence of LBP (OR = 1.06 [95% HKSJ CI]: 0.94; 1.20; $I^2 = 0\%$; 11 studies, 10,087 participants; "low" certainty of evidence) (Figure 2; Table 3).

The forest plots (Figures 2–4) do not distinguish between adjusted and non-adjusted ORs, because there was no heterogeneity between them.

Results were homogeneous ($I^2 = 0\%$) and no significant differences were found among studies, both in the meta-analysis of unadjusted results (OR = 1,13 [95% HKSJ CI]; 0.91; 1.40) and in the meta-analysis of adjusted results (OR = 0.95 [95% HKSJ CI]: 0.86 to 1.05).

The reanalysis of data from studies included in the IPD meta-analysis, led to results which were consistent with those from the original studies except in one case (Trevelyan & Legg, 2011), in which the database provided by the authors identified 75 of the included 245 subjects as reporting LBP, whereas the publication only mentioned 67. In the IPD meta-analysis, the 75 subjects identified in the database as reporting LBP, were treated as such.

The sub-group meta-analyses were conducted with individual participant data (IPD meta-analysis). In the first one, on age, there were non-significant differences between children (i.e., ≤ 12 years) and teenagers (i.e., ≥ 13 years) with regards to the value of the weight of the schoolbag (as a proportion of bodyweight) to predict LBP (test for differences between subgroups: $Chi^2 = 2.88$, $df = 1$ [$p = .09$], $I^2 = 65.2\%$). Moderate heterogeneity was found in the group of teenagers (four studies; 4,982 participants; $I^2 = 26\%$; "very low" certainty of evidence), while no heterogeneity was found in the group of children (six studies; 2,421 participants; $I^2 = 0\%$; "very low" certainty of evidence).

In the second subgroup analysis, on sport activity, schoolbag weight (as a proportion of bodyweight) was not found to predict LBP regardless of whether the subjects did and did not do sports (test for differences between subgroups: $Chi^2 = 0.52$, $df = 1$ [$p = .47$], $I^2 = 0\%$). No heterogeneity was found in the group practicing sports (three studies; 2,889 participants; $I^2 = 0\%$; "very low" certainty of evidence), while moderate heterogeneity was found in the group not practicing sports (two studies; 3,005 participants; $I^2 = 35\%$; "very low" certainty of evidence).

4 | DISCUSSION

The notion that the excessive weight of schoolbags carried by school students can trigger LBP, makes clinical sense. Moreover, the cut-off point of > 10% of bodyweight defining an "excessive" weight to be carried, is indirectly supported by some evidence and has elicited consensus among clinicians, researchers and some professional associations (Alghadir et al., 2017; American Chiropractic Association, 2018; American Occupational Therapy Association, 2017; American Physical Therapy Association, 2017; Devroey et al., 2007; Erne & Elfering, 2011; Grimmer & Williams, 2000; Mackie & Legg, 2008; Negrini et al., 1999; Sahli et al., 2013; Spiteri et al., 2017).

However, results from this meta-analysis reflect that the available evidence does not support this notion. In fact, the prevalence of LBP was the same among subjects carrying schoolbags weighing > 10% of their body weight, and those carrying lighter ones, even after adjusting for age, gender or sport activity. These results are consistent with data from most of the previous studies which have analysed the relationship between the weight of the schoolbag and the prevalence of LBP (Table 2) (Calvo-Muñoz et al., 2018).

Therefore, the "take home" message from this study is that the available evidence does not show a relationship between schoolbag weight and LBP among children and teenagers aged 9–16, despite having been assumed for years and sounding plausible. Nevertheless, only cross-sectional, observational studies could be included in this systematic review and meta-analysis, and the certainty of evidence is low. An IPD meta-analysis was designed to allow reanalysis of the data gathered by the previous original studies, but only seven studies could be included and some of their limitations, such as their cross-sectional design, could not be overcome. Therefore, the available evidence does not completely rule out the possibility that excessive weight of the schoolbags may indeed influence the presence, recurrence or persistence of LBP among school students, or that this may be the case among some school students facing specific circumstances (e.g., those needing to carry their schoolbags for long distances or periods). In fact, the level of certainty of the conclusion on the irrelevance of schoolbag weight to predict LBP is very low, implying that further research may modify it.

Establishing 10% of the bodyweight as the cut-off point to consider the weight of the schoolbag as "excessive" may appear reasonable; it has some data supporting it and is equivalent to the threshold most commonly used for adults

TABLE 2 Methodological quality of the studies included in the systematic review

	Methodological criteria[a]												
	1	2	3	4	5	6	7	8	9	10	11	12	Score[a]
Alghadir et al., 2017	YES	YES	NO	YES	YES	YES	YES	NA	YES	NO	YES	YES	82
Alghamdi et al., 2018	YES	NO	NO	YES	YES	YES	YES	NA	NA	NO	YES	YES	70
Akbar et al., 2019	YES	YES	NO	YES	YES	YES	YES	NA	YES	YES	YES	YES	91
Angarita-Fonseca et al., 2019	YES	YES	YES	YES	YES	YES	YES	NA	NA	YES	YES	YES	100
Chiang et al., 2006	YES	YES	YES	YES	YES	YES	YES	NA	YES	NO	NO	YES	82
de Oliveira et al., 2017	YES	YES	NO	YES	YES	YES	YES	NA	YES	NO	NO	YES	73
Dianat et al., 2014	YES	NO	NO	YES	YES	YES	YES	NA	NA	NO	NO	YES	60
Dianat et al., 2017	YES	YES	YES	YES	YES	YES	YES	NA	NA	NO	YES	YES	90
Grimmer & Williams, 2000	YES	NO	YES	YES	YES	YES	YES	NA	NA	NO	NO	YES	70
Johnson et al., 2011	YES	NO	NO	YES	YES	YES	YES	NA	YES	NO	NO	YES	64
Korovessis et al., 2004	YES	YES	YES	YES	YES	YES	NA	NO	YES	YES	NO	YES	82
Martínez-Crespo et al., 2009	YES	YES	NO	YES	YES	YES	YES	NA	NA	NO	YES	YES	80
Minghelli et al., 2016	YES	YES	NO	YES	YES	YES	YES	NA	YES	YES	YES	YES	91
Mohseni-Bandpei et al., 2007	YES	NO	YES	YES	YES	YES	YES	NA	YES	YES	NO	YES	82
Mwaka et al., 2014	YES	YES	YES	YES	YES	YES	YES	NA	YES	YES	NO	YES	91
Oka et al., 2019	YES	YES	NO	YES	YES	YES	YES	NA	NA	YES	YES	YES	90
Noormohammadpour et al., 2019	YES	NO	NO	YES	YES	YES	YES	NA	YES	YES	NO	YES	73
Trevelyan & Legg, 2011	YES	NO	NO	YES	YES	YES	YES	NA	YES	NO	YES	YES	73
Vidal et al., 2010	YES	YES	YES	YES	NO	YES	YES	NA	NA	YES	YES	YES	70
Watson et al., 2003	YES	YES	YES	YES	YES	YES	YES	NA	NA	NO	YES	NO	80
Young et al., 2006	YES	NO	YES	YES	NO	YES	NO	NA	YES	NO	YES	NO	55

Note: Score range (from worst to best): 0–100.
Abbreviation: NA, not applicable.
[a]The methodological criteria and the scoring procedure are described in Table S1.

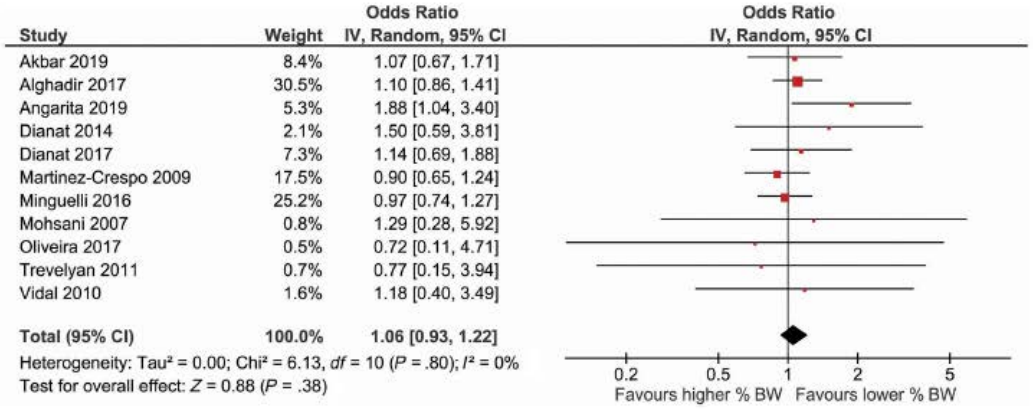


FIGURE 2 Meta-analysis of the relationship between carrying schoolbags representing >10% of bodyweight, and low back pain

TABLE 3 Summary of findings and certainty of evidence

Prognostic factors	Number of participants	Number of studies	Number of cohorts	Estimated effect size (95% confidence interval)	Phase	Factors to downgrade certainty					Factors to upgrade certainty		Overall certainty
						Study limitations	Inconsistency	Indirectness	Imprecision	Publication bias	Moderate/large effect size	Dose effect	
LBP (children and adolescents) [*1]	9,188	11	11	1.06 [0.94–1.20]	++	x[a]	✓	X[b]	✓	✓	X[d]	X[e]	++ LOW
LBP children [*2]	1,475	6	6	1.41 [1.8–1.84]	++	x[a]	✓	X[b]	X[c]	✓	X[d]	X[e]	+ VERY LOW
LBP adolescents [*3]	6,816	4	4	0.98 [0.60–1.60]	++	x[aOTNOTES:	✓	X[b]	X[c]	✓	X[d]	X[e]	+ VERY LOW
LBP sport practice [*4]	2,889	3	3	1,06 [0,74, 1,52]	++	x[a]	✓	X[b]	X[c]	✓	X[d]	X[e]	+ VERY LOW
LBP no sport practice [*4]	3,005	2	2	0,66 [0,19, 2,29]	++	x[a]	✓	X[b]	X[c]	✓	X[d]	X[e]	+ VERY LOW

Note: For GRADE factors: ✓, no serious limitations; X, serious limitations (or not present for moderate/large effect size, dose effect); unclear, unable to rate item based on available information. For overall quality of evidence: +, very low; ++, low; +++, moderate; ++++, high.

Abbreviation: LBP, low back pain.

[a]Due to risk of bias of included studies.

[b]Outcome of interest was measured at different time points

[c]Confidence intervals for measure of association include both clinically relevant and non-relevant values.

[d]No moderate or large effect was estimated.

[e]Studies did not allow assessment of potential dose–response.

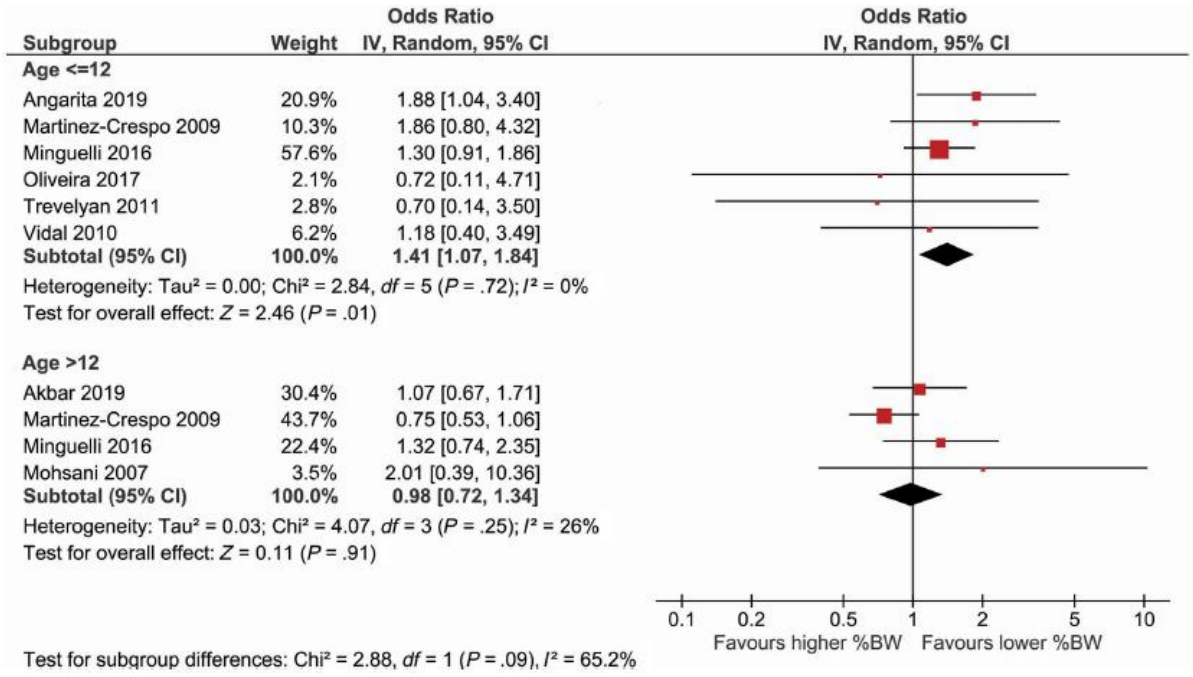


FIGURE 3 Subgroup meta-analysis on age, adjusting by gender

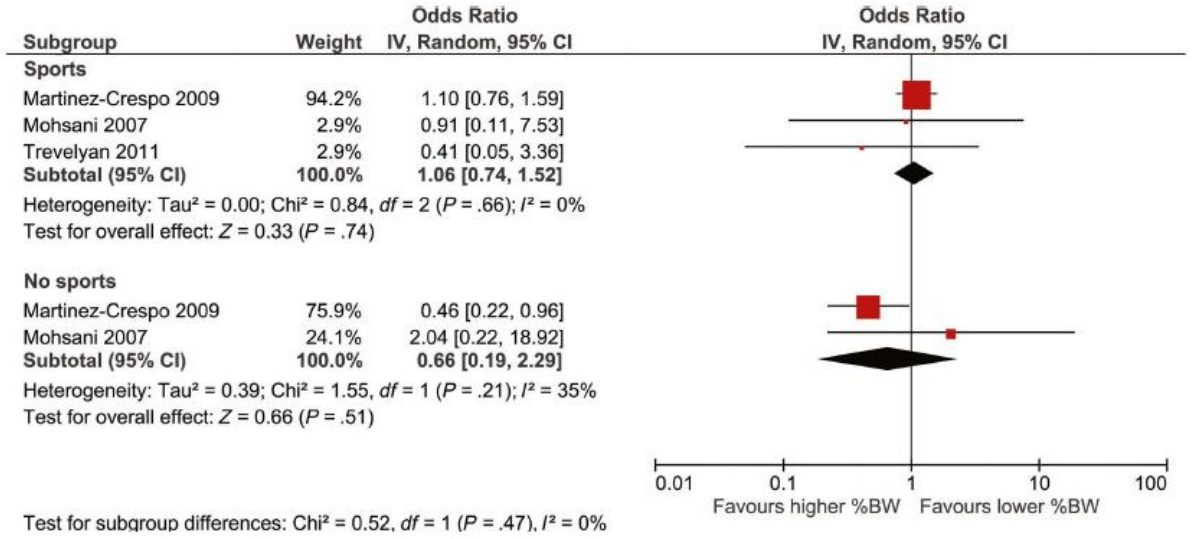


FIGURE 4 Subgroup meta-analysis on sport activity, adjusting by age and gender

in the work-related environment. This has led to most recommendations establishing that the "acceptable" weight for the schoolbag should be below this threshold (Alghadir et al., 2017; American Chiropractic Association, 2018; American Occupational Therapy Association, 2017; American Physical Therapy Association, 2017; Devroey et al., 2007; Erne & Elfering, 2011; Grimmer & Williams, 2000; Mackie & Legg, 2008; Negrini et al., 1999; Sahli et al., 2013; Spiteri et al., 2017). However, other cut-off points have been proposed, as low as 5% or as high as 20% (Dockrell et al., 2013; Rateau, 2004). A different cut-off limit for schoolbag weight might alter the conclusions from this study but this could not be analysed in the IPD meta-analysis because, due to the data available, it was impossible to explore other thresholds.

The weight of the schoolbag has been shown to vary significantly from one day to another (Negrini & Carabalona,

2002).Therefore, the data on weight gathered in a cross-sectional study, may not accurately reflect the load the school student carries most of the days. The reporting of LBP on one day (i.e., point prevalence) may be more influenced by the load carried previously than on the weight of the schoolbag on that very day. Moreover, most data on prevalence of LBP for longer periods (e.g., 14 days, 1 month or 1 year), were based on children's memory, as opposed to data included in a registry. All of the above may have diluted the potential relation between "weight of the schoolbag" and "LBP". Ideally, future studies should use previously validated methods, register data on LBP and on the weight of the schoolbag for a period long enough to capture daily variations in the weight of the schoolbag, and determine whether there is a minimum exposure to the presumed risk factor for it to increase the risk of LBP.

The method used to carry the schoolbag (e.g., cross-body bag vs. backpack vs. hung on one vs. two straps, etc.), has not shown to be significantly associated with the prevalence of LBP among schoolchildren, probably because the vast majority use backpacks strapped on both shoulders (Calvo-Muñoz et al., 2018). Nevertheless, this and other mechanical factors, such as the consistence between the size of class furniture and the students' size, could be further explored in future studies.

Some studies suggest that non-mechanical factors can also increase the risk of schoolchildren reporting LBP. These factors include students' perception that the weight of the schoolbag is excessive for them, irrespective of its actual weight (Dockrell, Blake, & Simms, 2015; Dockrell, Simms, & Blake, 2015), and psycho-social factors, such as a conflictive relationship with parents and schoolmates, behavioural or emotional problems, hyperactivity, and lack of attention (Dianat et al., 2017; Mikkonen et al., 2016; Sjolie, 2002; Watson et al., 2003). Therefore, future studies should also gather valid data on these factors.

Furthermore, it is likely that the key mechanical factor triggering LBP when carrying a schoolbag is not its weight per se, but whether it exceeds the physical capacity of the subject, how often and severely this occurs, and for how many years the overexertion persists or recurs. This would explain results from studies showing that LBP and other musculoskeletal complaints are higher among children who have to carry their schoolbags for longer periods, (Delele, Janakiraman, Bekele Abebe, Tafese, & Water, 2018; Dockrell, Blake, et al., 2015; Dockrell, Simms, et al., 2015) and may suggest that the weight of schoolbags is less relevant in environments where the children have to walk carrying their schoolbag short distances, than in those where they have to carry them for miles every day to attend class (Delele et al., 2018).

Therefore, further studies should gather valid and reliable data on the usual duration of carry and on the physical capacity of the children, such as muscle balance, strength and resistance. In this study, "sport activity" was included in the meta-analysis as a surrogate for "physical capacity", assuming that LBP could be more prevalent among children who, carrying schoolbags of equal weight to those of their peers, had a lower level of physical fitness, and factoring in previous studies which suggest that LBP is more prevalent among schoolchildren involved in competitive sports (Kovacs et al., 2003; Wedderkopp et al., 2003). Results did not confirm this assumption, although this may be due to the fact that gathering valid and reliable quantitative data on sport activity is difficult, and that the methods used to this end in the original studies included in the meta-analysis were inconsistent.

Future studies on this topic should be longitudinal, include large samples, implement follow-up periods long enough for the potential effects of a heavy schoolbag to appear, and use reliable and valid methods to gather data on presence, recurrence and duration of LBP, psychosocial factors, daily variations in the weight of the schoolbag, duration of carry, bodyweight, as well as variables reflecting the balance, strength and resistance of muscles involved in spine function.

In conclusion, data analysed in this study do not support the notion that carrying schoolbags weighing > 10% of bodyweight, is associated with a higher prevalence of LBP among schoolchildren aged 9–16. However, this conclusion is based only on cross-sectional studies, and future research may modify it.

ACKNOWLEDGEMENTS

The authors are grateful to Inés Gago Fernández, PT, for her help in conducting the search and obtaining the manuscripts, and to Abdullah Al-Taia, PhD, Artur Herbst de Oliveira, PT, Masoumeh Bagheri-Nesami, PhD, Gracia Martínez Crespo, MD, Josep Vidal Conti, PhD, Stephen Legg, PhD and Beatriz Minghelli, PhD, for providing access to the databases of their original studies.

CONFLICTS OF INTEREST

This study does not discuss off-label or investigational use of any drugs or devices. No benefits in any form have been or will be received from a commercial party related directly or indirectly to the subject of this article. The authors do not have any financial or personal relationships with third parties that could influence this work inappropriately. The authors have no conflicts of interest to report.

REFERENCES

Abo-Zaid, G., Sauerbrei, W., & Riley, R. D. (2012). Individual participant data meta-analysis of prognostic factor studies: State of the art? *BMC Medical Research Methodology*, *12*, 56. https://doi.org/10.1186/1471-2288-12-56

Akbar, F., AlBesharah, M., Al-Baghli, J., Bulbul, F., Mohammad, D., & Qadoura, B., & Al-Taiar, A. (2019). Prevalence of lowBack pain among adolescents in relation to the weight of school bags.

BMC Musculoskeletal Disorders, *20*:37. https://doi.org/10.1186/s12891-019-2398-2

Alghadir, A. H., Gabr, S. A., & Al-Eisa, E. S. (2017). Mechanical factors and vitamin D deficiency in schoolchildren with low back pain: Biochemical and cross-sectional survey analysis. *Journal of Pain Research*, *10*, 855–865. https://doi.org/10.2147/JPR.S124859

Alghamdi, R. S., Nafee, H. M., El-Sayed, A., & Alsaadi, S. M. (2018). A study of school bag weight and back pain among intermediate female students in Dammam City, Kingdom of Saudi Arabia. *Journal of Nursing Education and Practice*, *8*, 105–111. https://doi.org/10.5430/jnep.v8n12p105

American Chiropractic Association. (2018). Backpack misuse leads to chronic back pain, doctors of chiropractic say. Available from: https://www.acatoday.org/Patients/Health-Wellness-Information/Backpack-Safety. Accessed on November 9th, 2018.

American Occupational Therapy Association. Backpack facts: what's all the flap about? 2017. Available from: https://www.aota.org/~/media/Corporate/Files/Backpack/Backpack%20Strategies%20for%20Parents%20%20Students.pdf, Accessed on November 9th, 2018.

American Physical Therapy Association. Backpack safety: American. 2017. Available from: http://www.moveforwardpt.com/Resources/Detail/backpack-safety, Accessed on November 9th, 2018.

Angarita-Fonseca, A., Boneth-Collante, M., Ariza-Garcia, C. L., Parra-Patiño, J., Corredor-Vargas, J. D., & Villamizar-Niño, A. P. (2019). Factors associated with non-specific low back pain in children aged 10–12 from Bucaramanga, Colombia: A cross-sectional study. *Journal of Back and Musculoskeletal Rehabilitation*, 1–9. https://doi.org/10.3233/BMR-160561

Bardin, L. D., King, P., & Maher, C. G. (2017). Diagnostic triage for low back pain: A practical approach for primary care. *Medical Journal of Australia*, *206*, 268–273. https://doi.org/10.5694/mja16.00828

Calvo-Muñoz, I., Gómez-Conesa, A., & Sánchez-Meca, J. (2013). Prevalence of low back pain in children and adolescents: A meta-analysis. *BMC Pediatrics*, *13*, 14. https://doi.org/10.1186/1471-2431-13-14

Calvo-Muñoz, I., Kovacs, F. M., Roqué, M., Gago Fernández, I., & Seco Calvo, J. (2018). Risk factors for low back pain in childhood and adolescence: A systematic review. *Clinical Journal of Pain*, *34*, 468–484. https://doi.org/10.1097/AJP.0000000000000558

Chiang, H. Y., Jacobs, K., & Orsmond, G. (2006). Gender-age environmental associates of middle school students' low back pain. *Work*, *26*, 19–28.

de Oliveira, A. H., Chinaglia, C. G., & Lima, M. C. (2017). [O peso da mochila escolar não possui relação com dores musculoesqueléticas de estudantes do ensino fundamental] Backpack weight has no relation with musculoskeletal pain in first grade school students. *Journal of the Health Sciences Institute*, *35*, 117–121. https://www.unip.br/presencial/comunicacao/publicacoes/ics/edicoes/2017/02_abrjun/V35_n2_2017_p117a121.pdf

Delele, M., Janakiraman, B., Bekele Abebe, A., Tafese, A., & van de Water, A. T. M. (2018). Musculoskeletal pain and associated factors among Ethiopian [redacted] *Musculoskeletal Disorders*, *19*, 276. https://doi.org/10.1186/s12891-018-2192-6

Devroey, C., Jonkers, I., de Becker, A., Lenaerts, G., & Spaepen, A. (2007). Evaluation of the effect of backpack load and position during standing and walking using biomechanical, physiological and subjective measures. *Ergonomics*, *50*, 728–742. https://doi.org/10.1080/00140130701194850

Dianat, I., Alipour, A., & Asghari Jafarabadi, M. (2017). Prevalence and risk factors of low back pain among school age children in Iran. *Health Promotion Perspectives*, *7*, 223–229. https://doi.org/10.15171/hpp.2017.39

Dianat, I., Sorkhi, N., Pourhossein, A., Alipour, A., & Asghari-Jafarabadi, M. (2014). Neck, shoulder and low back pain in secondary schoolchildren in relation to school bag carriage: Should the recommended weight limits be gender-specific? *Applied Ergonomics*, *45*, 437–442. https://doi.org/10.1016/j.apergo.2013.06.003

Dieleman, J. L., Baral, R., Birger, M., Bui, A. L., Bulchis, A., & Chapin, A., … Murray, C. J. (2016). Us spending on personal health care and public health, 1996–2013. *JAMA*, *316*, 2627–2646. https://doi.org/10.1001/jama.2016.16885

Dockrell, S., Blake, C., & Simms, C. (2015). Guidelines for schoolbag carriage: An appraisal of safe load limits for schoolbag weight and duration of carriage. *Work*, *53*, 679–688. https://doi.org/10.3233/WOR-162260

Dockrell, S., Simms, C., & Blake, C. (2013). Schoolbag weight limit: Can it be defined? *Journal of School Health*, *83*, 368–377. https://doi.org/10.1111/josh.12040

Dockrell, S., Simms, C., & Blake, C. (2015). Schoolbag carriage and schoolbag-related musculoskeletal discomfort among primary school children. *Applied Ergonomics*, *51*, 281–290. https://doi.org/10.1016/j.apergo.2015.05.009

Dretzke, J., Ensor, J., Bayliss, S., Hodgkinson, J., Lordkipanidzé, M., Riley, R. D., … Moore, D. (2014). Methodological issues and recommendations for systematic reviews of prognostic studies: An example from cardiovascular disease. *Systematic Reviews*, *3*, 140. https://doi.org/10.1186/2046-4053-3-140

Erne, C., & Elfering, A. (2011). Low back pain at school: Unique risk deriving from unsatisfactory grade in maths and school-type recommendation. *European Spine Journal*, *20*, 2126–2133. https://doi.org/10.1007/s00586-011-1803-9

Fairbank, J. C., Pynsent, P. B., Van Poortvliet, J. A., & Phillips, H. (1984). Influence of anthropometric factors and joint laxity in the incidence of adolescent back pain. *Spine*, *9*, 461–464. https://doi.org/10.1097/00007632-198407000-00007

Goodgold, S. A., & Nielsen, D. (2003). Effectiveness of a school-based backpack health promotion program: Backpack Intelligence. *Work*, *21*, 113–123.

Grimmer, K., & Williams, M. (2000). Gender-age environmental associates of adolescent low back pain. *Applied Ergonomics*, *31*, 343–360. https://doi.org/10.1016/S0003-6870(00)00002-8

Guyatt, G. H., Oxman, A. D., Kunz, R., Vist, G. E., Falck-Ytter, Y., & Schunemann, H. J. (2008). What is "quality of evidence" and why is it important to clinicians? *BMJ*, *336*, 995–998. https://doi.org/10.1136/bmj.39490.551019.BE

Hestbaek, L., Leboeuf-Yde, C., & Kyvik, K. O. (2006). Is comorbidity in adolescence a predictor for adult low back pain? A prospective study of a young population. *BMC Musculoskeletal Disorders*, *7*, 29. https://doi.org/10.1186/1471-2474-7-29

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*, 1539–1558. https://doi.org/10.1002/sim.1186

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*, 557–560. https://doi.org/10.1136/bmj.327.7414.557

Hoy, D., Bain, C., Williams, G., March, L., Brooks, P., Blyth, F., … Buchbinder, R. (2012). A systematic review of the global prevalence of low back pain. *Arthritis and Rheumatism*, *64*, 2028–2037. https://doi.org/10.1002/art.34347

Hoy, D., March, L., Brooks, P., Blyth, F., Woolf, A., Bain, C., … Buchbinder, R. (2014). The global burden of low back pain: Estimates from the Global Burden of Disease 2010 study. *Annals of the Rheumatic Diseases*, *73*, 968–974. https://doi.org/10.1136/annrheumdis-2013-204428

Huguet, A., Hayden, J. A., Stinson, J., McGrath, P. J., Chambers, C. T., Tougas, M. E., & Wozney, L. (2013). Judging the quality of evidence in reviews of prognostic factor research: Adapting the GRADE framework. *Systematic Reviews*, *2*, 71. https://doi.org/10.1186/2046-4053-2-71

IntHout, J., Ioannidis, J. P. A., & Borm, G. F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology*, *14*, 25. https://doi.org/10.1186/1471-2288-14-25

Johnson, O. E., Adeniji, O. A., Mbada, C. E., Obembe, A. O., & Akosile, C. O. (2011). Percent of body weight carried by secondary school students in their bags in a nigerian school. *Journal of Musculoskeletal Research*, *14*, 1250003. https://doi.org/10.1142/S0218957712500030

Kamper, S. J., Yamato, T. P., & Williams, C. M. (2016). The prevalence, risk factors, prognosis and treatment for back pain in children and adolescents: An overview of systematic reviews. *Best Practice & Research Clinical Rheumatology*, *30*, 1021–1036. https://doi.org/10.1016/j.berh.2017.04.003

Korovessis, P., Koureas, G., & Papazisis, Z. (2004). Correlation between backpack weight and way of carrying, sagittal and frontal spinal curvatures, athletic activity, and dorsal and low back pain in schoolchildren and adolescents. *Journal of Spinal Disorders & Techniques*, *17*, 33–40. https://doi.org/10.1097/00024720-200402000-00008

Kovacs, F. M., Gestoso, M., Gil del Real, M. T., López, J., Mufraggi, N., & Méndez, J. I. (2003). Risk factors for non-specific low back pain in schoolchildren and their parents: A population based study. *Pain*, *103*, 259–268. https://doi.org/10.1016/S0304-3959(02)00454-2

Loney, P. L., & Stratford, P. W. (1999). The prevalence of low back pain in adults: A methodological review of the literature. *Physical Therapy*, *79*, 384–396. https://doi.org/10.1093/ptj/79.4.384

Louw, Q. A., Morris, L. D., & Grimmer-Somers, K. (2007). The prevalence of low back pain in Africa: A systematic review. *BMC Musculoskeletal Disorders*, *8*, 105. https://doi.org/10.1186/1471-2474-8-105

Mackenzie, W. G., Sampath, J. S., Kruse, R. W., & Sheir-Neiss, G. J. (2003). Backpacks in children. *Clinical Orthopaedics and Related Research*, *409*, 78–84. https://doi.org/10.1097/01.blo.0000058884.03274.d9

Mackie, H. W., & Legg, S. J. (2008). Postural and subjective responses to realistic schoolbag carriage. *Ergonomics*, *51*, 217–231. https://doi.org/10.1080/00140130701565588

Maher, C., Underwood, M., & Buchbinder, R. (2017). Non-specific low back pain. *The Lancet*, *389*, 736–747. https://doi.org/10.1016/S0140-6736(16)30970-9

Martínez-Crespo, G., Rodríguez-Piñero Durán, M., López-Salguero, A. I., Zarco-Periñan, M. J., Ibáñez-Campos, T., & Echevarría-Ruiz de Vargas, C. (2009). [Dolor de espalda en adolescentes: Prevalencia y factores asociados] Back pain among adolescents: Prevalence and associated factors. *Rehabilitacion*, *43*, 72–80. https://doi.org/10.1016/S0048-7120(09)70773-X

Mikkonen, P., Heikkala, E., Paananen, M., Remes, J., Taimela, S., Auvinen, J., & Karppinen, J. (2016). Accumulation of psychosocial and lifestyle factors and risk of low back pain in adolescence: A cohort study. *European Spine Journal*, *25*, 635–642. https://doi.org/10.1007/s00586-015-4065-0

Minghelli, B., Oliveira, R., & Nunes, C. (2016). Postural habits and weight of backpacks of Portuguese adolescents: Are they associated with scoliosis and low back pain? *Work*, *54*, 197–208. https://doi.org/10.3233/WOR-162284

Mohseni-Bandpei, M. A., Bagheri-Nesami, M., & Shayesteh-Azar, M. (2007). Nonspecific low back pain in 5000 Iranian school-age children. *Journal of Pediatric Orthopedics*, *27*, 126–129. https://doi.org/10.1097/BPO.0b013e3180317a35

Moore, M. J., White, G. L., & Moore, D. L. (2007). Association of relative backpack weight with reported pain, pain sites, medical utilization, and lost school time in children and adolescents. *Journal of School Health*, *77*, 232–239. https://doi.org/10.1111/j.1746-1561.2007.00198.x

Mwaka, E. S., Munabi, I. G., Buwembo, W., Kukkiriza, J., & Ochieng, J. (2014). Musculoskeletal pain and school bag use: A cross-sectional study among Ugandan pupils. *BMC Research Notes*, *7*, 222. https://doi.org/10.1186/1756-0500-7-222

Negrini, S., & Carabalona, R. (2002). Backpacks on! Schoolchildren's perceptions of load, associations with back pain and factors determining the load. *Spine*, *27*, 187–195. https://doi.org/10.1097/00007632-200201150-00014

Negrini, S., Carabalona, R., & Sibilla, P. (1999). Backpack as a daily load for schoolchildren. *The Lancet*, *354*, 1974. https://doi.org/10.1016/S0140-6736(99)04520-1

Noormohammadpour, P., Borghei, A., Mirzaei, S., Mansournia, M. A., Ghayour-Najafabadi, M., Kordi, M., & Kordi, R. (2019). The risk factors of low back pain in female high school students. *Spine*, *15*, E357–E365. https://doi.org/10.1097/BRS.0000000000002837

Oka, G. A., Ranade, A. S., & Kulkarni, A. A. (2019). Back pain and school bag weight - a study on Indian children and review of literature. *Journal of Pediatric Orthopaedics B*, *28*, 397–404. https://doi.org/10.1097/BPB.0000000000000602

Rateau, M. R. (2004). Use of backpacks in children and adolescents. A potential contributor of back pain. *Orthopaedic Nursing*, *23*, 101–105. https://doi.org/10.1097/00006416-200403000-00004

Riley, R. D., Hayden, J. A., Steyerberg, E. W., Moons, K. G., Abrams, K., & Kyzas, P. A., … … PROGRESS Group. (2013). Prognosis Research Strategy (PROGRESS) 2: Prognostic factor research. *PLoS Medicine*, *10*, e1001380. https://doi.org/10.1371/journal.pmed.1001380

Sahli, S., Rebai, H., Ghroubi, S., Yahia, A., Guermazi, M., & Elleuch, M. H. (2013). Theeffectsof backpackloadandcarryingmethodon thebalanceofadolescentidiopathicscoliosissubjects. *Spine Journal*, *13*, 1835–1842. https://doi.org/10.1016/j.spinee.2013.06.023

Sano, A., Hirano, T., Watanabe, K., Endo, N., Ito, T., & Tanabe, N. (2015). Body mass index is associated with low back pain in childhood and adolescence: A birth cohort study with a 6-year follow-up in Niigata City, Japan. *European Spine Journal*, *24*, 474–481. https://doi.org/10.1007/s00586-014-3685-0

Siambanes, D., Martinez, J. W., Butler, E. W., & Haider, T. (2004). Influence of school backpacks on adolescent back pain. *Journal of Pediatric Orthopedics*, *24*, 211–217. https://doi.org/10.1097/01241398-200403000-00015

Sjolie, A. N. (2002). Psychosocial correlates of low-back pain in adolescents. *European Spine Journal*, *11*, 582–588. https://doi.org/10.1007/s00586-002-0412-z

Skaggs, D. L., Early, S. D., D'Ambra, P., Tolo, V. T., & Kay, R. M. (2006). Back pain and backpacks in school children. *Journal of Pediatric Orthopedics*, *26*, 358–363. https://doi.org/10.1097/01.bpo.0000217723.14631.6e

Spiteri, K., Busuttil, M. L., Aquilina, S., Gauci, D., Camilleri, E., & Grech, V. (2017). Schoolbags and back pain in children between 8 and 13 years: A national study. *British Journal of Pain*, *11*, 81–86. https://doi.org/10.1177/2049463717695144

Stewart, L. A., & Tierney, J. F. (2002). To IPD or not to IPD?: Advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & the Health Professions*, *25*(1), 76–97. https://doi.org/10.1177/0163278702025001006

Swain, M. S., Henschke, N., Kamper, S. J., Gobina, I., Ottová-Jordan, V., & Maher, C. G. (2014). An International survey of pain in adolescents. *BMC Public Health*, *14*, 447. https://doi.org/10.1186/1471-2458-14-447

Trevelyan, F. C., & Legg, S. J. (2011). Risk factorsassociated with back pain in New Zealand schoolchildren. *Ergonomics*, *54*, 257–262. https://doi.org/10.1080/00140139.2010.547608

Vidal, J., Borràs, P. A., Ponseti, X., Gili, M., & Palou, P. (2010). Factores de riesgo asociados al dolor de espalda en escolares de entre 10 y 12 años de Mallorca. [Risk factors associated with low back pain among schoolchildren aged 10–12 years in Majorca]. Retos. Nuevas Tendencias En Educación Física, Deporte Y Recreación, *17*, 10–14. http://www.redalyc.org/articulo.oa?xm-l:id=34573228300.

Viry, P., Creveuil, C., & Marcelli, C. (1999). Nonspecific back pain in children. A search for associated factors in 14 year old schoolchildren. *Revue Du Rhumatisme*, *66*, 381–388.

Walker, B. F. (2000). The prevalence of low back pain: A systematic review of the literature from 1966 to 1998. *Journal of Spinal Disorders*, *13*, 205–217. https://doi.org/10.1097/00002517-200006000-00003

Watson, K. D., Papageorgiou, A. C., Jones, G. T., Taylor, S., Symmons, D. P., Silman, A. J., & Macfarlane, G. J. (2003). Low back pain in schoolchildren: The role of mechanical and psychosocial factors. *Archives of Disease in Childhood*, *88*, 12–17. https://doi.org/10.1136/adc.88.1.12

Wedderkopp, N., Leboeuf-Yde, C., Bo Andersen, L., Froberg, K., & Hansen, H. S. (2003). Back pain in children: No association with objectively measured level of physical activity. *Spine*, *28*, 2019–2024. https://doi.org/10.1097/01.BRS.0000083238.78155.31

Wells, G., Shea, B., O'Connell, D., Peterson, J., Welch, V., Losos, M., & Tugwell, P. (2009). *The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses*. Ottawa, ON: Ottawa Hospital Research Institute. Available: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.

Yamato, T. P., Maher, C. G., Traeger, A. C., Wiliams, C. M., & Kamper, S. J. (2018). Do schoolbags cause back pain in children and adolescents? A systematic review. *British Journal of Sports Medicine*, *52*, 1241–1245. https://doi.org/10.1136/bjsports-2017-098927

Young, I. A., Haig, A. J., & Yamakawa, K. S. (2006). The association between backpack weight and low back pain in children. *Journal of Back and Musculoskeletal Rehabilitation*, *19*, 25–33. https://doi.org/10.3233/BMR-2006-19104

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Calvo-Muñoz I, Kovacs FM, Roqué M, Seco-Calvo J. The association between the weight of schoolbags and low back pain among schoolchildren: A systematic review, meta-analysis and individual patient data meta-analysis. *Eur J Pain*. 2020;24:91–109. https://doi.org/10.1002/ejp.1471

APPENDIX 1

Search strategy

PHASE 1

"adolescent"[MeSH Terms] OR "adolescent"[All Fields]
"child"[MeSH Terms] OR "child"[All Fields]
"young"[All Fields]
"schools"[MeSH Terms] OR "schools"[All Fields] OR "school"[All Fields]
"pediatrics"[MeSH Terms] OR "pediatrics"[All Fields] OR "pediatric"[All Fields]
("adolescent"[MeSH Terms] OR "adolescent"[All Fields]) OR ("child"[MeSH Terms] OR "child"[All Fields]) OR ("Young"[Journal] OR "young"[All Fields]) OR ("schools"[MeSH Terms] OR "schools"[All Fields] OR "school"[All Fields]) OR ("pediatrics"[MeSH Terms] OR "pediatrics"[All Fields] OR "pediatric"[All Fields])
"back pain"[MeSH Terms] OR ("back"[All Fields] AND "pain"[All Fields]) OR "back pain"[All Fields]
"low back pain"[MeSH Terms] OR ("low"[All Fields] AND "back"[All Fields] AND "pain"[All Fields]) OR "low back pain"[All Fields]
"low back pain"[MeSH Terms] OR ("low"[All Fields] AND "back"[All Fields] AND "pain"[All Fields]) OR "low back pain"[All Fields] OR ("lumbar"[All Fields] AND "pain"[All Fields]) OR "lumbar pain"[All Fields]
("back pain"[MeSH Terms] OR ("back"[All Fields] AND "pain"[All Fields]) OR "back pain"[All Fields]) OR ("low back pain"[MeSH Terms] OR ("low"[All Fields] AND "back"[All Fields] AND "pain"[All Fields]) OR "low back pain"[All Fields]) OR ("low back pain"[MeSH Terms] OR ("low"[All Fields] AND "back"[All Fields] AND "pain"[All Fields]) OR "low back pain"[All Fields] OR ("lumbar"[All Fields] AND "pain"[All Fields]) OR "lumbar pain"[All Fields])
waist[All Fields] AND ("pain"[MeSH Terms] OR "pain"[All Fields])

Publicación 6

Wirth K, Klenk J, Brefka S, Dallmeier D, Faehling K, Roqué I Figuls M, et al; SITLESS consortium. Biomarkers associated with sedentary behaviour in older adults: A systematic review. Ageing Res Rev. 2017 May;35:87-111. PubMed PMID: 28025174.

FI: 10.616 (2019). Puntuación de atención Altmetric: 28

Esta publicación puede consultarse al completo y de forma libre en
https://www.sciencedirect.com/science/article/pii/S1568163716301763?via%3Dihub

Ageing Research Reviews 35 (2017) 87–111

Contents lists available at ScienceDirect

Ageing Research Reviews

journal homepage: www.elsevier.com/locate/arr

Review

Biomarkers associated with sedentary behaviour in older adults: A systematic review

Katharina Wirth[a,b,c,*], Jochen Klenk[c,d], Simone Brefka[a,b], Dhayana Dallmeier[a,b], Kathrin Faehling[a,b,c], Marta Roqué i Figuls[e], Mark A. Tully[f], Maria Giné-Garriga[g], Paolo Caserotti[h,i], Antoni Salvà[e], Dietrich Rothenbacher[c], Michael Denkinger[a,b,c], Brendon Stubbs[j,k,l], on behalf of the SITLESS consortium[1]

[a] Agaplesion Bethesda Hospital, Geriatric Medicine Ulm University, Ulm, Germany
[b] Geriatric Center Ulm/Alb-Donau, Ulm, Germany
[c] Department of Epidemiology and Medical Biometry, Ulm University, Ulm, Germany
[d] Department of Clinical Gerontology, Robert-Bosch-Hospital, Stuttgart, Germany
[e] Fundació Salut i Envelliment – Universitat Autònoma de Barcelona, Biomedical Research Institute Sant Pau (IIB-Sant Pau), Barcelona, Spain
[f] UKCRC Centre of Excellence for Public Health (NI), Centre for Public Health, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, United Kingdom
[g] Faculty of Psychology, Education and Sport Sciences Blanquerna, Ramon Llull University, Barcelona, Spain
[h] Department of Sports Science and Clinical Biomechanics, SDU Muscle Research Cluster (SMRC), University of Southern Denmark, Odense, Denmark
[i] National Institutes of Health, National Institute on Aging, Laboratory of Epidemiology and Population Sciences (LPES), Bethesda, MD, USA
[j] Physiotherapy Department, South London and Maudsley NHS Foundation Trust, Denmark Hill, London SE5 8AZ, United Kingdom
[k] Health Service and Population Research Department, Institute of Psychiatry, Psychology and Neuroscience, King's College London, De Crespigny Park, London, Box SE5 8AF, United Kingdom
[l] Faculty of Health, Social Care and Education, Anglia Ruskin University, Chelmsford, United Kingdom

ARTICLE INFO

Article history:
Received 1 August 2016
Received in revised form 30 October 2016
Accepted 12 December 2016
Available online 23 December 2016

Keywords:
Older adults
Sedentary behaviour
Biomarker

ABSTRACT

Objective: Pathomechanisms of sedentary behaviour (SB) are unclear. We conducted a systematic review to investigate the associations between SB and various biomarkers in older adults.
Methods: Electronic databases were searched (MEDLINE, EMBASE, CINAHL, AMED) up to July 2015 to identify studies with objective or subjective measures of SB, sample size ≥50, mean age ≥60 years and accelerometer wear time ≥3 days. Methodological quality was appraised with the CASP tool. The protocol was pre-specified (PROSPERO CRD42015023731).
Results: 12701 abstracts were retrieved, 275 full text articles further explored, from which 249 were excluded. In the final sample (26 articles) a total of 63 biomarkers were detected. Most investigated markers were: body mass index (BMI, n = 15), waist circumference (WC, n = 15), blood pressure (n = 11), triglycerides (n = 12) and high density lipoprotein (HDL, n = 15). Some inflammation markers were identified such as interleukin-6, C-reactive protein or tumor necrosis factor alpha. There was a lack of renal, muscle or bone biomarkers. Randomized controlled trials found a positive correlation for SB with BMI, neck circumference, fat mass, HbA1C, cholesterol and insulin levels, cohort studies additionally for WC, leptin, C-peptide, ApoA1 and Low density lipoprotein and a negative correlation for HDL.
Conclusion: Most studied biomarkers associated with SB were of cardiovascular or metabolic origin. There is a suggestion of a negative impact of SB on biomarkers but still a paucity of high quality investigations exist. Longitudinal studies with objectively measured SB are needed to further elucidate the pathophysiological pathways and possible associations of unexplored biomarkers.

* Corresponding author at: Agaplesion Bethesda Hospital Ulm, Geriatric Research Unit, Ulm University and Geriatric Centre Ulm/Alb-Donau, Zollernring 26, 89073 Ulm, Germany.
E-mail address: katharina.wirth@bethesda-ulm.de (K. Wirth).
[1] List of consortium members is given in "Acknowledgement".

http://dx.doi.org/10.1016/j.arr.2016.12.002

Contents

1. Introduction

According to the National Institute of Health (NIH) Biomarkers Definitions Working group (Biomarkers Definitions Working Group, 2001), a biomarker is a characteristic that is objectively measured as an indicator of normal biological processes, pathogenic processes, or a pharmacological response to a therapeutic intervention. Therefore, biomarkers can be very helpful as surrogate markers for diseases or pathophysiological links between exposure and disease; or as intermediate measures of the effectiveness of interventions on disease processes. Within the past few decades, a considerable amount of literature has clearly demonstrated that physical activity (PA) has a range of benefits on the health (Nelson et al., 2007; Castaneda et al., 2002) and wellbeing of older adults (McAuley et al., 2000). Recently, there has been an interest in understanding the biomarkers underlying the response to PA. For example, in a cohort of community dwelling older adults, levels of N-terminal pro-brain natriuretic peptide (NT-proBNP) and high-sensitive troponin T have been associated with objectively monitored PA and showed a more beneficial profile with increasing PA, suggesting a dose response relationship (Jefferis et al., 2014; Klenk et al., 2013). To date, most of the PA biomarker research has focussed on cardiovascular risk factors (Gabriel et al., 2012; Reaven et al., 1991; Jefferis et al., 2014), but there are many other biological systems with associated biomarkers which may be affected by PA or especially SB. Recent examples include β-amyloid burden and glucose metabolism as markers of neurodegeneration (Okonkwo et al., 2014), interleukin-6 (IL-6) and C-reactive protein (CRP) as markers for systemic inflammation (Jarvie et al., 2014) or DNA-repair as a marker for cell homeostasis (Brocklebank et al., 2015).

An emerging evidence base has started to demonstrate that sedentary behaviour (SB), over and above time spent in PA, is independently associated with several important detrimental health outcomes, including endpoints such as mortality, frailty, sarcopenia, dementia, and cardiovascular diseases (Biswas et al., 2015). According to the Sedentary Behaviour Research Network, SB is defined as any waking behaviour characterised by an energy expenditure ≤1.5 metabolic equivalents (METs) whilst in a sitting or reclining posture (Sedentary Behaviour Research Network, 2012). The emerging research highlighting the deleterious impact of SB on health is of particular concerns as adults spend on average 5 h of their time in sedentary behaviour (Loyen et al., 2016). Indeed, some studies have demonstrated that on the population-level sedentary time (ST) increased over the decades from 1960 to 2010 (Church et al., 2011). Especially older people spent most of their time in SB. A recent meta-analysis illustrated that older people were sedentary for 65–80% of their waking time (Wullems et al., 2016), other sources mentioned ST with an average of 9 h (Dunlop et al., 2015) to 13.8 h per day (Cawthon et al., 2013). Older people are seen as the age group engaging in the highest level of SB (Wullems et al., 2016) and thus could benefit most from changing their daily habits. The developing evidence on the harms associated with SB has illustrated that it is not only the absence of daily or weekly moderate-to-vigorous physical activity (MVPA), but rather, SB is a separate category of behaviour with unique determinants, consequences and sequences for possible intervention (Owen et al., 2010). Considering the physiological changes occurring with age in several organ systems (Boss and Seegmiller, 1981), results from middle-aged adults can't be simply transferred to older adults. Therefor the EU study SITLESS investigates how SB can be reduced sustainably and how sedentariness effects biomarkers especially in older adults. In this framework the interest on outcomes of studies performed in elderly, assessing SB and its impact on biomarkers was the focus. In addition, biomarker studies are important to

further understand the link between SB, PA and adverse health outcomes like total mortality and harmful phenotypes like Metabolic Syndrome (MES) (Gardiner et al., 2011), frailty or sarcopenia. Perhaps it can help to understand the role of biomarkers as possible mediators of the association between SB and adverse phenotypes or aging-related diseases. Therefore, the aims of this systematic review were to provide a comprehensive overview of aging-related biomarkers associated with SB and report on the strength of the observed associations in community-dwelling older adults.

2. Methods

2.1. Study design

This systematic review adhered to the PRISMA guidelines (Moher et al., 2009) and followed a predetermined published protocol (PROSPERO No. CRD42015023731) (Stubbs et al., 2015).

2.2. Condition or domain being studied

SB, as defined by the Sedentary Behaviour Research Network (2012) (waking behaviour with an energy expenditure ≤ 1.5 METs whilst in a sitting or reclining posture), represented our exposure of interest. We also considered studies which did not fully comply with this definition (e.g. television watching time, SB identified by other questionnaires or accelerometer data that do not allow for disentangling posture issues or clearly indicate METs) but are highly relevant to SB.

With respect to the biomarkers we were interested in any inflammatory, renal and cardiac biomarkers, lipids and metabolic markers, genetic and metabolomics markers, endocrine markers and markers of muscle strength, body composition, as well as of specific physical performance measures (e.g. gait speed and balance).

2.3. Information sources and searches

Two authors (KW, BS) searched the electronic databases: MEDLINE (PubMed), EMBASE, CINAHL (via EBSCO), AMED (via Ovid/EBSCO) from inception to 15 July 2015. We used search terms described in Appendix A. Appropriate search strategies and MESH-terms were selected (see Appendix A).

2.4. Study selection and eligibility criteria

Studies meeting the following criteria were included:

a Explicitly measured SB using objective accelerometer wear time ≥ 3 days (to follow the recommendations of good clinical practice (Ward et al., 2005)) or self-report instruments. Studies defining SB purely as a lack of PA were excluded.
b Including community dwelling, older adults (mean age of sample ≥ 60 years).
c Sample size of n ≥ 50 participants, to ensure adequate power.
d Quantitative study design including randomized controlled trials (RCTs), controlled clinical trials (CCTs), pre- and post-intervention measurement studies, prospective observational studies (POS), or studies (only prospective trials) that examined an association of any biomarker with SB. We also considered cross-sectional studies (CSS) but present them separately because of their descriptive nature due to the inability to clearly establish the temporal sequence between SB and biomarkers.

2.5. Participants and population

We selected studies, with the above mentioned characteristics that included older adults (mean age ≥ 60 years) conducted in the community.

When we encountered studies with a large age range and a mean age below 60 years, indicating the study included some older adults (>60 years), we attempted to contact the authors to acquire the variables of interest for all participants with an age of 60 years and older. Populations with specific co-morbidity (e.g. diabetes mellitus type 2 (DM-II)), peripheral artery disease (PAD), chronic obstructive pulmonary disease (COPD) were included, but critically evaluated and highlighted as such.

2.6. Data extraction

All results of the searches were inserted in a bibliographic database. A data extraction form was created and amended to the requirements of the review. Two authors piloted (KW; SB) the data extraction form in a random sample of 3 studies that employ different study designs. This ensured that the relevant information was selected to assess the effectiveness and study quality.

All data were extracted by these two reviewers. Data extraction included: first author, country, setting, population, aims of the study, type of the study (RCT, POS or CSS), number of studies and participants included in the article, details of the intervention (including duration), inclusion criteria, type of recruitment, type and definition of SB or PA used, biomarkers analysed and results, details of control condition, overall study quality (internal risk of bias), association statistics, acknowledged limitations by authors, the authors' conclusions and other notes.

Any disagreements in data extraction were resolved through discussion between the reviewers.

2.7. Risk of bias and quality assessment

Assessment of studies followed the PRISMA (Moher et al., 2009) guidelines. Two authors conducted the methodological quality appraisal of all included studies using a modified Critical Appraisal Skills Programme (CASP) tool, adapted for each study design (CASP, 2016):

- RCTs (max. CASP score = 6) were assessed for risk of bias in the following domains: clearly focused issue, randomization, performance (blinding, personnel), comparability (treatment, groups at baseline) and attrition (participants accounted for at its conclusion).
- POS (max. CASP score = 8) were assessed for risk of bias in the following domains: clearly focused issue, selection and recruitment (random approach or representative for a defined population, accuracy of measurement (exposure, outcome), identification of important confounding factors, adjustment for confounding factors and follow-up (period, completion).
- CCS (max. CASP score = 6) were assessed for risk of bias in: clearly focused issue, selection and recruitment (random approach or representative for defined population), accuracy of measurement (exposure, outcome), identification of important confounding factors and adjustment for confounding factors.

In an attempt to assess the potential effect and direction of the effect of SB on specific biomarkers, additional information related to statistical evidence of an association, as adapted from the Canadian Agency for Drugs and Technology in Health (CADTH) (CADTH, 2016), was included in Table 2 for the high quality studies. The fol-

Table 1a
Descriptive overview of randomized controlled trials (RCTs) and prospective observational studies (POS).

	Author, Year, Country	Setting, country, study	Follow-up	No. of participants	male	Age, mean ±SD [years]	sedentary behaviour assessment (method: measure)	measured biomarkers	CASP score	remarks
RCTs	Kallings et al. (2009), Sweden	CD, "Move study", efficacy of PA on prescription to reduce CMRF	6 months	CG: 54 IG: 47	43% 43%	all 68 years	IPAQ questionnaire: total sitting time in hours/day	BMI, WC, AD, NC, BF%, fat mass, BF% in trunk, fat mass in trunk, glucose, HbA_1c, s/d BP, cholesterol, HDL, LDL, LDL/HDL, triglycerides, ApoA1, ApoB, ApoB/ApoA1	5 of 6	
	Kirk et al. (2009), UK	CD+RP,"Time2Act study" in DM-II patients, compare 2 methods of PA promotion to standard care	6 & 12 months	CG: 35 IG1: 47 IG2: 52	51% 53% 42%	59.2 ± 10.4 60.9 ± 9.6 63.2 ± 10.6	ActiGraph GT1M (waist) for 7 days (≥10 h/d and ≥4days), SB not defined	BMI, WC, s/d BP, cholesterol, HDL, HbA_1c	4 of 6	Groups were not equally balanced
	Suboc et al. (2014), USA	CD, investigate if reduction of SB improves vascular endothelial function and specific biomarkers	12 weeks	CG: 35 IG1: 32 IG2: 29	76% 61% 60%	62 ± 7 64 ± 7 63 ± 8	ActiGraph GT3X for 7 days, (≥600 min/d and ≥4 days); SB defined as ≤ 1.5 METS or < 100 counts/min	BMI, WC, glucose, insulin, QUICKI, HOMA-IR, cholesterol, HDL, LDL, triglycerides, CRP, s/d BP, HR, brachial artery diameter, peak shear, hyperemic peak shear, Nitroglycerin mediated dilation, carotid-femoral pulse wave velocity, augmentation index, aortic s/d BP	6 of 6	
	Aadahl et al. (2014), Denmark	CD, Health2010 study, effect of motivational counseling at reducing sitting time	6 months	total: 66 CG: 28 IG: 38	nr	63 IG: 64.1 CG: 63.8	ActivPAL 3TM (triaxial, right thigh) for 7 days (at least 2 days for analysis), SB not defined	WC, BF%, glucose, insulin, HbA_1c, cholesterol, HDL, LDL, triglycerides	5 of 6	Only 2 days of accelerometer but 94% of total group n = 166 had 5 days or more; special sub-analysis for us with people > 60 years; Biomarker only at follow-up → no change available
POS	Fung (2000), USA	CD, "Health Professionals Follow-up Study", television watching and biomarkers	1986–1994	466	100%	60	Questionnaire: hours of television watching/week	BMI, cholesterol, HDL, LDL, triglycerides, ApoA1, Lp(a), Leptin Fibrinogen, Insulin, C-peptide, HbA_1c	5 of 8	
	Cooper et al. (2012), UK	CD+RP, "Early ACTivity in Diabetes study", with DM-II patients, RCT but results treated as a cohort	6 months	528 cross sectional; 380 longitudinal data	65%	59.8 ± 10.0	ActiGraph GT1M (waist) for 7 days (≥600 min/d and ≥3 valid days) removed for sleeping; non-wear time ≥20 min with 0 counts, SB (h/day) defined as < 100 counts/minute	WC, HbA_1c, HDL, glucose, insulin, HOMA-IR	5 of 8	

AD = abdominal diameter; Apo = Apo lipoprotein; BF% = percent of body fat; BMI = body mass index; BP = blood pressure; CD = community-dwelling; CG = control group; CMRF = cardio-metabolic risk factors; CRP = C-reactive protein; DM-II = diabetes mellitus type 2; HbA_1c = glycated hemoglobin; HDL = high density lipoprotein; HOMA-IR = homeostatic model assessment of insulin resistance; HOMA%B = homeostatic model assessment of B-cell function; HR = heart rate; IG = intervention group; MET = metabolic equivalent; LDL = low density lipoprotein; Lp(a) = lipoprotein a; NC = neck circumference; nr = not reported; s/d BP = systolic/diastolic blood pressure; PAS = physical activity scale; POS = prospective observational studies; QUICKI = quantitative insulin sensitivity check index; RCT = randomized controlled trials; RP = risk population; WC = waist circumference.

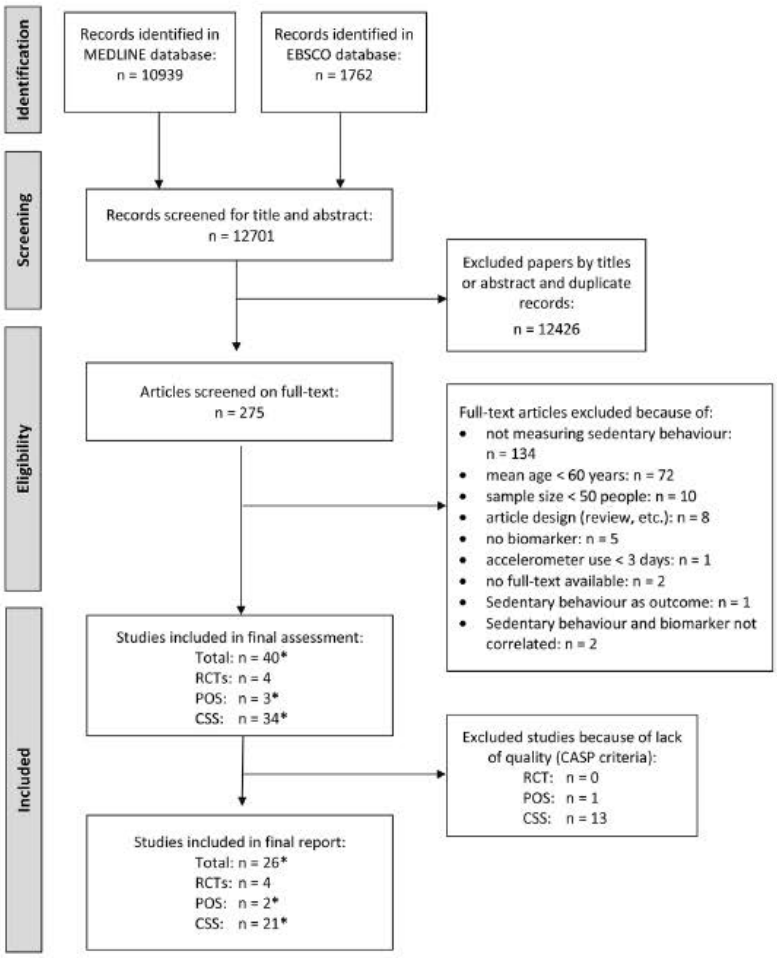


Fig. 1. Flow chart for article selection of randomized controlled trials (RCT), prospective observational studies (POS) and cross-sectional studies (CSS).
*Remark: One study of Cooper et al. (2012) was a cohort study with prospective data as well as cross sectional data and thus counted as POS and CS.

lowing decision rules as suggested by CADTH (CADTH, 2016), were used for standardised statements about the statistical significance:

- 0% of studies showed statistically significant results = no evidence for any association
- 1% to 33% of studies showed statistically significant results = generally no evidence for any association.
- 34% to 66% of studies showed statistically significant results = mixed evidence for association
- 67% or more studies showed statistically significant results = generally evidence for association

Due to the few studies of high quality, we decided to apply this method of categorisation, although often less than 5 studies with statistically significant results were found. To ensure a minimum level of validity, we applied this tool in all biomarkers measured in ≥3 studies (RCT and/or POS).

2.8. Strategy for data synthesis and subgroup analysis

We tabulated the single study results and grouped them according to comparable biomarkers. All results were stratified with appropriate subgroup analyses, for instance according to exposure type (SB and PA separately), type of SB/PA assessment (questionnaire- versus sensor-based), biomarker type and study design (RCT and CCT separately). We anticipated conducting a meta-analysis if sufficient homogeneity was evident across the study types and outcomes of interest and enough studies could be identified in comparable areas.

3. Results

3.1. Results of the literature search

Our initial searches identified 12,701 hits. After the exclusion at title level, removing of duplicates and the matching of results from the two independent reviewers (including removing duplicates), a final list of 275 full-text articles was scrutinised. 235 articles were

subsequently excluded according to our in- and exclusion criteria (full details in Fig. 1). 3 studies included people with a large age range in their sample, yet a mean age below 60 years. Upon 3 attempts to contact the authors, 1 group (Aadahl et al., 2014)) provided additional data, whilst 2 authors did not respond and were subsequently excluded due to age <60 years (Knight et al., 2014; Mohri et al., 2013).

After exclusions, 40 studies were considered eligible, however after further revision and evaluation, another 14 articles (1 POS, 13 CSS) were excluded (for more details see "risk of bias (quality) appraisal"), thus leading to a total of 26 articles (4 RCT, 2 POS and 21 CSS). The study from Cooper et al. (2012) was included as a POS and CSS due to longitudinal and cross-sectional analysis of the data reported by the study authors.

3.2. Definition of sedentary behaviour

We found a highly heterogeneous definition of SB, which was often misclassified as simply the absence of PA and therefor 134 papers were excluded. The most frequent definition of SB was total time spent at less than 100 counts per minute using data from an accelerometer (Gabriel et al., 2012; Cooper et al., 2012; Bann et al., 2015; Gennuso et al., 2013; Lee et al., 2015; Lynch et al., 2010, 2011; Santos et al., 2012; Sardinha et al., 2015; Stamatakis et al., 2012). Henson et al. (2013) defined SB in a similar way but with smaller epochs of less than 25 counts per 15 s. Other authors used the same definition of SB as used in this review with less than 1.5 MET (Bann et al., 2015; Cooper et al., 2014; Suboc et al., 2014). Some studies did not define SB at all (Aadahl et al., 2014; Kirk et al., 2009).

A total of 14 studies (3 RCT, 1 POS, 11 CSS, whereas Cooper et al. (2012) were included as POS and CSS) measured SB with sensors or accelerometers, another 2 CSS (Bann et al., 2015; Stamatakis et al., 2012) with both, accelerometer and questionnaire, while 10 studies measured SB by questionnaires only. Of these 10 questionnaires, 6 enquired about TV watching time (1 POS, 5 CSS), whilst the remainders included more detailed questions about SB (1 RCT, 3 CSS).

3.3. Characteristics of included studies

3.3.1. Randomized controlled trials (RCTs)

An overview of the RCTs is listed in Table 1a. Overall, a total of 397 participants in the RCTs were represented (Intervention Group, (IG): 245; Control Group, (CG): 152). Although SB was evaluated, the primary aims of 2 RCTs were to increase PA but not reduce SB. 3 RCTs (Aadahl et al., 2014; Suboc et al., 2014; Ewald et al., 2010) captured SB objectively with an accelerometer (ActiGraph), whereas Kallings et al. (2009) evaluated SB with the International Physical Activity Questionnaire (IPAQ), which consists of 2 questions on the amount of sitting time in the last 7 days; one for average weekday and one for average weekend day.

Intervention was mainly focused on increasing habitual PA. This was triggered through different processes: an intervention with pedometer-use plus a weekly visit on an interactive website with the aim of increasing PA level by 10% each week up to 10,000 steps/day (Suboc et al., 2014); written PA prescription with the aim of increasing PA level to 30 min MVPA per day (Kallings et al., 2009); a written PA pack with a self-instructional workbook based on a *trans*-theoretical model of behaviour change (Kirk et al., 2009). Only 1 study focused on decreasing SB with 4 main aims, such as decreasing daily TV viewing time, substitute sitting with standing, break up prolonged sitting time and a maximum of 30 min of sitting per episode (Aadahl et al., 2014).

3.3.2. Prospective observational studies (POS)

An overview of the POS is listed in Table 1a. In the 2 cohort studies (Cooper et al., 2012; Fung, 2000) a total of 846 participants were represented. Fung (2000) used a questionnaire focusing on the number of hours of television watching to measure SB in men, whereas Cooper et al. (2012) used objectively measured SB time by accelerometer measurements (ActiGraph).

3.3.3. Cross-sectional studies (CSS)

A total of 41,816 participants were included across the 21 CSS. The characteristics of these CCS are listed in Table 1b. Most of the studies focused on SB and its association to biomarkers (16/21 studies), from which 5 focused on TV watching time. 2 other studies investigated both SB and PA as exposure (Lynch et al., 2010; Santos et al., 2012), 1 study (Larsen et al., 2014a) calculated sitting time, whereas 2 studies evaluated SB as secondary outcome (Gabriel et al., 2012; Reaven et al., 1991).

The majority of the studies used objectively measured time of SB by accelerometer (11/21 studies). 2 studies evaluated both objectively measured SB and SB measured by questionnaire (Bann et al., 2015; Stamatakis et al., 2012). 5 studies focused on time spent with TV watching and 3 used the following questionnaires: Reaven et al. (1991) adapted a questionnaire from the Health Interview Survey, which measured 17 different leisure time activities in the last 2 weeks; Larsen et al. (2014a) measured daily SB by asking about time spent being sedentary on a typical weekday; Allison et al. (2012) evaluated the time spent being sedentary by using the "Typical Week Physical Activity Survey" which measures SB in the last 7 days.

3.4. Sedentary behaviour and biomarkers

3.4.1. Overview of biomarkers explored in the literature

Table 2 provides an overview of the associations between SB and each biomarker system including: anthropometric parameters, systemic parameters, blood lipids, glycaemic parameters, performance biomarkers, inflammatory biomarkers and others. A total of 63 biomarkers were evaluated (counting ratios of different biomarkers separately). Table 3 considers the specific biomarker results within each study design. There was insufficient homogenous data to perform a meta-analysis. Therefore, we describe the number of studies that explored each biomarker and the summary statistics reporting the overall proportion of these studies that found a statistical association. Only the statistically significant results from the multivariable analyses are shown. If significant, biomarkers showed evidence for an unfavourable association with higher ST. Body mass index (BMI, 9 of 15 of the studies significant), waist circumference (WC, more than 8 of 15 of the studies significant), insulin (4 of 8 studies significant) and high density lipoprotein (HDL, 6 of 15 studies significant) were examined in a lot of studies and demonstrated the most reliable results. For a more detailed description see Tables 2 and 3.

We identified 4 "risk population" studies. 1 POS of Cooper et al. (2012), performed in diabetes type 2 patients, showed a (statistically significant) positive correlation for SB with WC, HDL, insulin and HOMA-IR. The CSS from Cooper et al. (2014), also performed in diabetes type 2 patients, revealed a positive association for SB with WC, too. The study from Lee et al. (2015), performed in the high risk osteoarthritis population showed lower gait speed and lower chair stand rate associated with higher levels of SB. There were no statistically significant results for the study of Lynch et al. (2010) investigating the association between SB with BMI, WC and insulin in a breast cancer survivor cohort.

Table 1b
Descriptive overview of 21 cross-sectional studies (CSS).

Author, year, country	Setting, study, aim	No. of partici-pants	Male (%)	Age, mean (±SD/range) [years]	Sedentary behaviour (SB) assessment (method: measure)	Analyzed biomarkers	CASP score	Remarks
Allison et al. (2012), USA	CD, Multi-Ethnic Study of Atherosclerosis (MESA), association of SB with adiposity associated measures of inflammation	1543	49.8	64.3 (9.6) (45–84)	Questionnaire "Typical Week Physical Activity Survey" (TWPAS), which measures also SB (TV, computer, reading) (min/week: continuous and in tertiles) during a typical week	adiponectin, leptin, TNF-a, resistin, adiponectin/leptin	4 of 6	Multivariate adjusted means and coefficients of multivariate linear regression models, three models with different level of adjustment;
Anuradha et al. (2011), USA McAuley et al. (2000)	CD, Multi-Ethnic Study of Atherosclerosis (MESA), association of TV watching time and retinal vascular caliber	5893	48	63.1 (9.9) (45–84)	Questionnaire about TV watching time (quartiles: hours/week) during a typical week	central retinal artery equivalent, central retinal vein equivalent	4 of 6	Least square means of multivariate linear regression models, two models with different level of adjustment
Bankoski et al. (2011), USA	CD, National Examination Survey (NHANES); association between SB and MES	1367	51.8	71 (7.8) (≥60)	ActiGraph AM-7164 (uniaxial, waist) for 7 days (≥4 valid days) removed for bathing and sleeping; non-wear time >60 min with 0 counts; SB (hours/day) defined as <100 counts/minute	dichotomized: WC, HDL, triglycerides, glucose, BP	5 of 6	Means adjusted for age and sex
Bann et al. (2015), USA	CD, Lifestyle Interventions and Independence for Elders (LIFE) Study, association of SB with BMI and grip strength	1130	33	NR (70–89) m: 79.3 (5.3) w: 78.5 (5.3)	ActiGraph GT3X (triaxial, waist) for 7 days (≥600 min/d and ≥3 valid days) removed for bathing and sleeping; non-wear time ≥90 min with 0 counts; SB (min/day) defined as <100 counts/minute CHAMPS questionnaire about a typical week; SB (min/day) was defined as time ≤1.5 METs	BMI, grip strength	4 of 6	Coefficients of multivariate linear regression models, two models with different level of adjustment; no numbers regarding men and women although all analyses were stratified according to sex
Cooper et al. (2012), UK	RP (type 2 diabetes), Early Activity in Diabetes (Early-ACTID), association between SB & CMRF	528 m: 344 w: 184	65	59.8 (10) (30–80) m: 60.7 (9.7) w: 58.1 (10.4)	ActiGraph GT1M (waist) for 7 days (≥600 min/d and ≥3 valid days) removed for bathing and sleeping; non-wear time ≥20 min with 0 counts, SB (h/day) defined as <100 counts/minute	WC, HDL, insulin, HOMA-IR	5 of 6	Subsample in POS table Coefficients of multivariate linear regressions
Cooper et al. (2014), UK	RP (type 2 diabetes), ADDITION-Plus Study, associations of SB and PA with metabolic risk	394 m: 250 w: 144	63	60.3 (7.4) m: 60.2 (7.4) w: 60.5 (7.4)	Actiheart for 4 days; SB was defined as activity <1.5 MET	WC, systolic BP, HbA_1c, triglycerides, HDL	6 of 6	Coefficients of multivariate linear regression models, three models with different level of adjustment; adjusted to possible confounders, limitations mentioned; difficult to distinguish between sitting and standing by using Actiheart
Jakes et al. (2003), UK	CD, European Prospective Investigation into Cancer (EPIC) study, association between TV watching and vigorous activity with obesity and CVD risk profile	14189	42	60.3 (45–74) m: 61(9) w: 59.9 (8.9)	Self-reported TV watching time (four groups: hours/day), separated for weekday and weekend day,	BMI, WC, HC, WHR, BF%, s/d BP, HbA_1c, triglycerides, cholesterol, HDL, LDL	4 of 6	Means adjusted for age and multivariate adjusted; rational for grouping of TV watching is unclear, numbers per group are not given
Gabriel et al. (2012), USA	CD, Healthy Women Study (HWS), association of PA with coronary artery calcification progression	148	0	73.2 (1.7)	ActiGraph GT1M (uni-axial, waist) for 7 days for 24 h (≥600 min/d and ≥4 days); non-wear time ≥60 min with 0 counts; SB (min/day) was defined as <100 counts per minute	BMI, WC, s/d BP, cholesterol, LDL, HDL, triglycerides, glucose, insulin	5 of 6	Correlation coefficients between SB and biomarkers; different methods used during different FU state

Table 1 b (*Continued*)

Author, year, country	Setting, study, aim	No. of partici-pants	Male (%)	Age, mean (±SD/range) [years]	Sedentary behaviour (SB) assessment (method: measure)	Analyzed biomarkers	CASP score	Remarks
Gao et al. (2007), USA	CD, association of TV watching time and prevalence of MES	455	40	68.8 (≥60)	Self-reported TV watching time (quartiles: hours/day)	BMI, dichotomized: WC, triglycerides, HDL, BP, glucose, cholesterol/HDL, WHR	5 of 6	Means for BMI, proportions and multivariate adjusted ORs for all dichotomized variables; only Hispanics with Puerto Rican or Dominican origin;
Gardiner et al. (2011), Australia	CD, Australian Diabetes, Obesity and Lifestyle (AusDiab) study, relation between TV watching and sitting time with MES	1958	46	69 (≥60) m: 69.6 w: 69	Questionnaire about TV and sitting time (quartiles: hours/day)	Dichotomized: WC, triglycerides, HDL, BP, glucose	4 of 6	OR from multivariate adjusted models; TV time and sitting time measured separately. Discrepancy in numbers, given in the paper, detected: Females: No MES (n = 643) and MES (n = 460) –> n = 1103 does not match total number of 1062;
Gennuso et al. (2013), USA	CD, National Examination Survey (NHANES) subsample, association between SB and CMRF	1914	52	74.6 (6.5) (≥65)	ActiGraph AM-7164 (uniaxial, waist) for 7 days (≥600 min/day and ≥1 valid day) removed for bathing and sleeping; non-wear time ≥60 min with 0 counts; SB (quartiles: hours/day) defined as <100 counts/minute	BMI, WC, s/d BP, cholesterol, HDL, triglycerides, LDL, glucose, HbA_1c, CRP	5 of 6	Least square means multivariate adjusted; 1 day enough for getting included in analysis; analysis of triglycerides, LDL and glucose only on subsample of 809 people – we already excluded papers because of accel. only 1 day
Hamer et al. (2013), UK	CD, English Longitudinal Study of Ageing (ELSA), association between TV watching time. CRP and depressive symptoms	4964	45.1	64.5 (8.9)	Questionnaire about TV watching time on 5 weekdays and weekend separately (4 groups: hours/day)	CRP	4 of 6	Mean change in relation to a reference group, two models with different level of adjustment
Henson et al. (2013), UK	CD, "Walking Away from Type 2 Diabetes study", association between SB and inflammation and adiposity	558	65	63.6 (7.7)	ActiGraph GT3X (tri axial) for 7 days (≥600 min/d and ≥4 valid days); non-wear time ≥60 min with 0 counts; SB (hours/day) defined as <25 counts/15 secs;	CRP, leptin, IL-6, adiponectin, leptin/adiponectin	5 of 6	Coefficients of multivariate linear regression models, three models with different level of adjustment; not mentioned where accelerometer got attached to;
Larsen et al. (2014a), USA	CD, Rancho Bernardo Study (RBS); associations of sitting time with regional fat and abdominal muscle	539 m: 135 w: 404	25	64.6 (7.4) (≥55)	Single item about time spent in leisure time sitting activities on a typical weekday (tertiles: hours/day)	BMI, TNF-a, adiponectin, leptin, IL-6, HDL, LDL, triglycerides; pericardial-, intra-thoracic-, visceral-, intermuscular- and subcutaneous fat, abdominal and psoas muscle	4 of 6	Unadjusted means; low sensitivity with measuring SB by a single self-report item for 1 day
Lee et al. (2015), USA	RP (adults with or at high risk for knee osteoarthritis), osteoarthritis initiative (OAI), association between SB and physical function	1168	45	66 (45–79)	ActiGraph GT1M (uniaxial, waist) for 7 days (≥600 min/d and ≥4 days) removed for bathing and sleeping; non-wear time by ≥90 min with 0 counts; SB (quartiles: % of day) was defined as <100 counts per min	BMI (3 categories), gait speed, chair stand rate	5 of 6	Unadjusted means for BMI and multivariate adjusted mean differences between categories of SB; adjusted to confounders but only arthritis patients

Lynch et al. (2010), USA	RP (breast cancer survivors), National Examination Survey (NHANES), association of PA and SB with adiposity	111	0	69.2 (13)	ActiGraph 7164 (uniaxial, waist) for 7 days (≥600 min/d) removed for bathing and sleeping; non-wear time ≥60 min with 0 counts; SB (hours/day) was defined as <100 counts/min;	BMI, WC, insulin	5 of 6	Coefficients of multivariate linear regression models, three models with different level of adjustment; due to missing values the number of subjects varied (BMI: 106, WC: 100, insulin: 35); not mentioned how many days of accelerometer were necessary to get included in study
Lynch et al. (2011), USA	CD, postmenopausal women of National Examination Survey (NHANES),	1024	0	63.0 (9.4)	ActiGraph 7164 (uniaxial, waist) for 7 days (≥600 min/d) removed for bathing and sleeping; non-wear time ≥60 min with 0 counts; SB (hours/day) was defined as <100 counts/min;	BMI, WC, CRP, fasting glucose, insulin, HOMA-IR	5 of 6	Coefficients of multivariate linear regression models, three models with different level of adjustment; data not following normal distribution were transformed by natural logarithm
Reaven et al. (1991), USA	CD, relation between leisure time PA and BP	641	0	66.5 (50–89)	Questionnaire adapted from Health Interview Survey with 17 leisure time activities, (2 weeks)	HR, BMI, s/d BP, fasting insulin, 2 h insulin	3 of 6	Means adjusted for age (all), means multivariate adjusted (s/d BP)
Santos et al. (2012), Portugal	CD, association of PA and SB with functional fitness	312 m: 117 w: 195	37.5	74.3(6.6) (≥65) m: 74.2 (6.2) w: 74.3(6.9)	ActiGraph, GT1M (waist) for 4 days (2 weekdays and 2 weekend days) (≥10 h/d and ≥3 days with ≥1 weekend day); non-wear time ≥60 min 0 counts; SB (min/day) was defined as <100 counts per minute	Chair stand repetitions, arm curl, 6MWT, 8 foot up and go, chair sit and reach, back scratch	4 of 6	Coefficients of multivariate linear regression models, four models with different level of adjustment; nothing said about exclusion criteria, not mentioned if it was performed in the same centre of same examiners, not medication or comorbidities got evaluated
Sardinha et al. (2015), Portugal	CD, association of SB with physical function	215 m: 87 w: 128	40	73.3 (5.9) (65–94) m: 73.7 (6.2) w: 73.0 (5.7)	ActiGraph, GT1M (waist) for 4days (2 weekdays and 2 weekend days) (≥ 10h/d and≥3 days with ≥1 weekend day); non-wear time ≥60 min 0 counts; SB (min/day) was defined as <100 counts per minute	6MWT, 8 foot up and go, arm curl, chair stand, chair sit and reach, back scratch	6 of 6	Good adjustment for possible confounders
Stamatakis et al. (2012), UK	CD, Health Survey for England (HSE), association between SB and CMRF	2765 with self-report; 649 with accelerom-eter	45	70 (≥60)	ActiGraph GT1 M for 7 days (≥600 min/d and ≥1 valid day), non-wear time ≥60 min with 0 count, SB (tertiles: min/day) defined as <100 counts/minute; self-reported leisure-time SB (tertiles: min/day)	BMI, WC, cholesterol, HDL, $HbA_{1}c$, cholesterol/HDL ratio	5 of 6	Unadjusted means; 1 valid day included in analysis, but 91% had >5 days; different sample sizes of accelerometer measured and self-report sample (sensitivity analysis showed that this difference might contribute to differential associations; sample for blood biomarker was considerably smaller (1354/333)

6MWT = 6 m walk test; ABI = ankle brachial index; AD = abdominal diameter; Apo = Apo lipoprotein; BP = blood pressure; BF% = percent of body fat; BFMI = body fat mass index (kg/m²); BMD = bone mineral density; BMI = body mass index; CCS = cross-sectional study; CD = community-dwelling; CG = control group; CMRF = cardio-metabolic risk factors; CRP = C-reactive protein; CV = cardiovascular; CVBM = cardiovascular biomarker; CVD = cardiovascular disease; DM-II = diabetes mellitus type 2; FEV1 = forced expiratory volume within 1 s; FFMI = fat free mass index (kg/m2); FVC = forced vital capacity; HC = hip circumference; HDL = high density lipoprotein; HOMA-IR = homeostatic model assessment of insulin resistance; HOMA% B = homeostatic model assessment of B-cell function; HR = heart rate; IG = intervention group; IL-6 = interleukin 6; LDL = low density lipoprotein; MES = metabolic syndrome; MET = metabolic equivalent; NC = neck circumference; NHANES = National Health and Nutrition Examination Survey; NR = not reported; OR = odds ratio; PA = physical activity; PAD = peripheral artery disease; PAI = plasminogen activator inhibitor 1; PAL = physical activity level; PAS = physical activity scale; PASE = physical activity scale for the elderly; RCT = randomized controlled trials; RP = "risk population" defined as population with specific illness such as diabetes or peripheral artery disease; s/d = systolic/diastolic; SB = sedentary behaviour; t-PA = tissue plasminogen activator; TNF-a = tumor necrosis factor a; WC = waist circumference; WHR = waist to hip ratio.

Table 2
Overview of biomarkers evaluated in the systematic review articles.

			Study results			Interpretation of the statistical significance level in high quality papers, adapted from CADTH (CADTH, 2016)		
Category of biomarker	Biomarker type	Number of studies by study type (RCT/POS/CSS)	Statistically not significant studies (n)	$p < 0.05$ in un-adjusted results (n)	$p < 0.05$ fully adjusted (n, study type, direction of association +/−)	High quality studies (n); participants (n)	Results (%)	Interpretation
Anthropo-metric parameters	BMI	3/1/11	5 (2 RCT (Kirk et al., 2009; Suboc et al., 2014), 3 CSS (Larsen et al., 2014a; Lee et al., 2015; Gao et al., 2007))	10	9 (1 RCT+ (Kallings et al., 2009), 1 POS+ (Fung, 2000), 7 CSS+ (Gennuso et al., 2013; Jakes et al., 2003; Reaven et al., 1991; Bann et al., 2015; Gabriel et al., 2012; Stamatakis et al., 2012; Lynch et al., 2011))	4 studies; 797 participants	50% significant	Mixed evidence for association
	WC	4/1/10	5 (4 RCT (Kirk et al., 2009; Suboc et al., 2014; Aadahl et al., 2014; Kallings et al., 2009), 1 CSS (Gao et al., 2007))	10	8 (1 POS+ (Cooper et al., 2012), 7 CSS+ (Gabriel et al., 2012; Gennuso et al., 2013; Jakes et al., 2003; Lynch et al., 2011; Bankoski et al., 2011; Cooper et al., 2014; Stamatakis et al., 2012))	5 studies; 777 participants	20% significant	Generally no evidence for association
	HC	0/0/1	0	1	1 CSS+ (Jakes et al., 2003)	Not applicable		
	WHR	0/0/2	0	2	2 CSS+ (Gao et al., 2007; Jakes et al., 2003)	Not applicable		
	neck circumference	1/0/0	0	1	1 RCT+ (Kallings et al., 2009)	Not applicable		
	abdominal diameter	1/0/0	1 RCT (Kallings et al., 2009)	0	0	Not applicable		
	BF%	2/0/1	2 RCT (Aadahl et al., 2014; Kallings et al., 2009)	1	1 CSS+ (Jakes et al., 2003)	Not applicable		
	fat mass	1/0/0	0	1	1 RCT+ (Kallings et al., 2009)	Not applicable		
	BF in trunk	1/0/0	1 RCT (Kallings et al., 2009)	0	0	Not applicable		
	fat mass in trunk	1/0/0	1 RCT (Kallings et al., 2009)	0	0	Not applicable		
	pericardial fat	0/0/1	0	1	1 CSS+ (Larsen et al., 2014a)	Not applicable		
	intra-thoracic fat	0/0/1	0	1	0	Not applicable		
	visceral fat	0/0/1	0	1	0	Not applicable		
	intermuscular fat	0/0/1	0	1	0	Not applicable		
	subcutaneous fat	0/0/1	0	1	0	Not applicable		
	abdominal muscle	0/0/1	1 CSS (Larsen et al., 2014a)	0	0	Not applicable		

	psoas muscle	0/0/1	1 CSS (Larsen et al., 2014a)	0		Not applicable		
Systemic parameters	systolic BP	3/0/8	7 (3 RCT (Kallings et al., 2009; Kirk et al., 2009; Suboc et al., 2014), 4 CSS (Gennuso et al., 2013; Cooper et al., 2014; Bankoski et al., 2011; Gardiner et al., 2011))	4	3 CSS+ (Jakes et al., 2003; Reaven et al., 1991; Gabriel et al., 2012)	3 studies; 331 participants	0% significant	No evidence for association
	diastolic BP	3/0/7	7 (3 RCT (Kallings et al., 2009; Kirk et al., 2009; Suboc et al., 2014), 4 CSS (Gennuso et al., 2013; Gabriel et al., 2012; Bankoski et al., 2011; Gardiner et al., 2011))	3	1 CSS+ (Jakes et al., 2003)	3 studies; 331 participants	0% significant	No evidence for association
	HR	1/0/1	1 RCT (Suboc et al., 2014)	1	1 CSS+ (Reaven et al., 1991)	Not applicable		
	brachial artery diameter	1/0/0	1 RCT (Suboc et al., 2014)	0	0	Not applicable		
	peak shear	1/0/0	1 RCT (Suboc et al., 2014)	0	0	Not applicable		
	hyperemic peak shear	1/0/0	1 RCT (Suboc et al., 2014)	0	0	Not applicable		
	nitroglycerin mediated dilation	1/0/0	1 RCT (Suboc et al., 2014)	0	0	Not applicable		
	carotid-femoral pulse wave velocity	1/0/0	1 RCT (Suboc et al., 2014)	0	0	Not applicable		
	augmentation index	1/0/0	1 RCT (Suboc et al., 2014)	0	0	Not applicable		
	aortic s/d BP	1/0/0	1 RCT (Suboc et al., 2014)	0	0	Not applicable		
	central retinal artery equivalent	0/0/1	1 CSS (Anuradha et al., 2011)	0	0	Not applicable		
	central retinal vein equivalent	0/0/1	0	1	1 CSS+ (Anuradha et al., 2011)	Not applicable		
Blood lipids	total cholesterol	4/1/4	7 (3 RCT (Aadahl et al., 2014; Kirk et al., 2009; Suboc et al., 2014), 1 POS (Fung, 2000), 3 CSS (Gennuso et al., 2013; Stamatakis et al., 2012; Gabriel et al., 2012))	2	2 (1 RCT+ (Kallings et al., 2009), 1 CSS+ (Jakes et al., 2003))	Not applicable		

Table 2 (*Continued*)

Category of biomarker	Biomarker type	Number of studies by study type (RCT/POS/CSS)	Study results			Interpretation of the statistical significance level in high quality papers, adapted from CADTH (CADTH, 2016)		
			Statistically not significant studies (n)	p<0.05 in un-adjusted results (n)	p<0.05 fully adjusted (n, study type, direction of association +/−)	High quality studies (n); participants (n)	Results (%)	Interpretation
	HDL	4/2/9	7 (4 RCT (Aadahl et al., 2014; Kirk et al., 2009; Suboc et al., 2014; Kallings et al., 2009), 3 CSS (Gennuso et al., 2013; Larsen et al., 2014a; Gabriel et al., 2012))	8	6 (2 POS- (Cooper et al., 2012; Fung, 2000), 4 CSS- (Jakes et al., 2003; Stamatakis et al., 2012; Bankoski et al., 2011; Gao et al., 2007))	6 studies; 1243 participants	33% significant	Generally no evidence for association
	LDL	3/1/4	6 (3 RCT (Aadahl et al., 2014; Suboc et al., 2014; Kallings et al., 2009), 3 CSS (Gennuso et al., 2013; Larsen et al., 2014a; Gabriel et al., 2012))	2	2 (1 POS+ (Fung, 2000), 1 CSS+ (Jakes et al., 2003))	4 studies; 729 participants	25% significant	Generally no evidence for association
	LDL/HDL	1/0/0	1 RCT (Kallings et al., 2009)	0	0	Not applicable		
	cholesterol/HDL	0/0/2	0	2	2 CSS+ (Gao et al., 2007; Stamatakis et al., 2012)	Not applicable		
	triglycerides	3/1/8	8 (3 RCT (Aadahl et al., 2014; Suboc et al., 2014; Kallings et al., 2009), 1 POS (Fung, 2000), 4 CSS (Gennuso et al., 2013; Larsen et al., 2014a; Gabriel et al., 2012; Gao et al., 2007))	4	2 CSS+ (Bankoski et al., 2011; Jakes et al., 2003)	4 studies; 729 participants	0% significant	No evidence for association
	ApoA1	1/1/0	1 RCT (Kallings et al., 2009)	1	1 POS- (Fung, 2000)	Not applicable		
	ApoB	1/0/0	1 RCT (Kallings et al., 2009)	0	0	Not applicable		
	ApoB/ApoA1	1/0/0	1 RCT (Kallings et al., 2009)	0	0	Not applicable		
	Lp(a)	0/1/0	1 POS (Fung, 2000)	0	0	Not applicable		

Glycemic parameters	HbA1c	3/1/4	5 (2 RCT (Aadahl et al., 2014; Kirk et al., 2009), 1 POS (Fung, 2000), 2 CSS (Gennuso et al., 2013; Cooper et al., 2014))	3	2 (1 RCT+ (Kallings et al., 2009), 1 CSS+ (Stamatakis et al., 2012))	4 studies; 767 participants	25% significant	Generally no evidence for association
	glucose	3/0/6	6 (3 RCT (Aadahl et al., 2014; Suboc et al., 2014; Kallings et al., 2009), 3 CSS (Lynch et al., 2011; Gabriel et al., 2012; Gao et al., 2007))	3	2 CSS+ (Gennuso et al., 2013; Bankoski et al., 2011)	3 studies; 263 participants	0% significant	No evidence for significant association
	insulin (fasting)	2/2/4	2 (1 RCT (Suboc et al., 2014), 1 POS (Fung, 2000))	6	4 (1 RCT+ (Aadahl et al., 2014), 1 POS+ (Cooper et al., 2012), 2 CSS+ (Gabriel et al., 2012; Reaven et al., 1991))	4 studies; 1008 participants	50%	Mixed evidence for association
	insulin (after 2 h)	0/0/1	0	1	1 CSS- (Reaven et al., 1991)	Not applicable		
	HOMA-IR	1/1/1	1 RCT (Suboc et al., 2014)	2	1 POS+ (Cooper et al., 2012)	Not applicable		
	QUICKI	1/0/0	1 RCT (Suboc et al., 2014)	0	0	Not applicable		
	C-peptide	0/1/0	0	1	1 POS+ (Fung, 2000)	Not applicable		
Performance biomarkers	6MWT	0/0/2	2 (1 CSS (Sardinha et al., 2015), NR* (Santos et al., 2012))	0	0	Not applicable		
	8 foot up and go	0/0/2	2 (1 CSS (Sardinha et al., 2015), NR* (Santos et al., 2012))	0	0	Not applicable		
	grip strength	0/0/1	1 CSS (Bann et al., 2015)	0	0	Not applicable		
	gait speed	0/0/1	0	1	1 CSS- (Lee et al., 2015)	Not applicable		
	arm curl	0/0/2	NR* (Santos et al., 2012)	1	0	Not applicable		
	Chair stand rate	0/0/3	NR* (Santos et al., 2012)	2	1 CSS- (Lee et al., 2015)	Not applicable		
	Chair sit and reach	0/0/2	NR* (Santos et al., 2012)	1	0	Not applicable		
	Back scratch	0/0/2	1 (CSS (Sardinha et al., 2015), NR* (Santos et al., 2012))	0	0	Not applicable		

Table 2 (*Continued*)

Category of biomarker	Biomarker type	Number of studies by study type (RCT/POS/CSS)	Study results			Interpretation of the statistical significance level in high quality papers, adapted from CADTH (CADTH, 2016)		
			Statistically not significant studies (n)	$p < 0.05$ in un-adjusted results (n)	$p < 0.05$ fully adjusted (n, study type, direction of association +/−)	High quality studies (n); participants (n)	Results (%)	Interpretation
Inflammatory biomarkers	CRP	1/0/4	1 RCT (Suboc et al., 2014)	4	2 CSS+ (Gennuso et al., 2013; Hamer et al., 2013)	Not applicable		
	Fibrinogen	0/1/0	1 POS (Fung, 2000)	0	0	Not applicable		
	IL-6	0/0/2	1 CSS (Larsen et al., 2014a)	1	1 CSS+ (Henson et al., 2013)	Not applicable		
Others	Leptin	0/1/3	1 CSS (Larsen et al., 2014a)	3	2 (1 POS+ (Fung, 2000), 1 CSS+ (Allison et al., 2012))	Not applicable		
	Adiponectin	0/0/3	2 CSS (Allison et al., 2012; Henson et al., 2013)	1	1 CSS+ (Larsen et al., 2014a)	Not applicable		
	leptin/adiponectin ratio	0/0/1	0	1	0	Not applicable		
	adiponectin/leptin ratio	0/0/1	0	1	1 CSS− (Allison et al., 2012)	Not applicable		
	TNF-a	0/0/2	1 CSS (Larsen et al., 2014a)	1	1 CSS+ (Allison et al., 2012)	Not applicable		
	Resistin	0/0/1	1 CSS (Allison et al., 2012)	0	0	Not applicable		

Remark: Results from Cooper et al. (2012) are only listed in POS results, not additionally in CSS column; *Santos et al. (Santos et al., 2012) calculated a composite Z-score, but didn't report separate associations for each biomarker with SB.

6MWT = 6 m walk test; adj. = adjusted; Apo = Apolipoprotein; BF% = percent of body fat; BMI = body mass index (kg/m^2); BP = blood pressure; CSS = cross-sectional study; CRP = C-reactive protein; FFMI = fat free mass index (kg/m^2); HbA_1c = specific glycated hemoglobin; HC = hip circumference; HDL = high density lipoprotein; HOMA-IR = homeostatic model assessment of insulin resistance; HR = heart rate; IL = interleukin; LDL = low density lipoprotein; Lp(a) = lipoprotein a; NR = not reported; PA = physical activity; POS = prospective observational studies; QUICKI = quantitative insulin sensitivity check test; RCT = randomized controlled trials; reg. coeff. = regression coefficient; s/d = systolic/diastolic; sig. = significant; SB = sedentary behaviour; ST = sedentary time; TNF-a = tumor necrosis factor a; unadj. = unadjusted; WC = waist circumference; WHR = waist to hip ratio.

Table 3
Details of randomized controlled trials (RCTs), prospective observational studies (POS) and cross-sectional studies (CSS) including significant associations of biomarkers with Sedentary Behaviour.

	Author, Year	No. of participants	What was analysed? What was measured?	Measured biomarkers	Main results (95%CI) or [SD] or (SE)	P-value	CASP score	Remarks
RCTs	Kallings et al. (2009)	CG: 54 IG: 47	Significant differences between IG and CG in mean change (MC) of biomarker from baseline (B) to follow up (FU); reducing SB was measured, thus changes are negative	BMI (kg/m²)	MC IG: −0.6 (−0.9 to −0.3) vs. MC CG: −0.2 (−0.4 to 0.0)	0.02	5/6	MC from B to FU in sitting time (hours/day) in CG with −1 h/d ($p<0.001$) and IG with −2 h/d ($p<0.0.01$)
				NC (cm)	MC IG: −1.2 (−1.6 to −0.8) vs. MC CG: −0.6 (−1.0 to −0.2)	0.01		
				Fat mass (kg)	MC IG: −1.7 (−2.5 to −0.9) vs. MC CG: −0.6 (−1.2 to −0.1)	0.03		
				HbA_1c (%)	MC IG: −0.1 (−0.2 to 0.0) vs. MC CG: 0.2 (0.1 to 0.3)	0.001		
				Cholesterol (mmol/l)	MC IG: −0.3 (−0.6 to 0.0) vs. MC CG: 0.1 (−0.1 to 0.1)	0.04		
	Aadahl et al. (2014)	Total: 66 IG: 38 CG: 28	Mean difference (MD) in change of fasting serum insulin from baseline (B) to follow up (FU) between IG and CG for **reducing SB**	Fasting insulin (pmol/l)	−0.51 (−0.01 to −1.00)	0.04	5/6	CG means [SD] of sitting time in B = 9.8 [2.0] and FU = 10.2 [1.9]; IG means [SD] of sitting time in B = 9.27 [1.9] and FU = 8.7 [1.5]
POS	Fung (2000)	466	Pearson correlation coefficient (PCC) of television hours and biomarker; linear regression coefficient (lrc) for 1994 TV hours[1] or average TV hours in 1988–1994[2]	BMI (kg/m²)	PCC: 0.13	<0.01	5 of 8	Lrc calculated for increment of 14 h television watching per week
				Leptin (ng/ml)	PCC: 0.15 lrc: 1.3 (0.5)[2], adj. to BMI 0.8 (0.4)[2]	<0.01; <0.01,		
				C-peptide (ng/dl)	PCC: 0.12	<0.05		
				ApoA1 (mg/dl)	lrc: −5.3 (2.0)[1]	<0.05		
					adj. to BMI −4.9 (2.0)[1]	<0.05		
				HDL (mg/dl)	lrc: −3.9 (1.2)[1]	<0.01		
					adj. to BMI −3.4 (1.2)[1]	<0.01		
				LDL (mg/dl)	lrc: 6.1 (2.9)[1]	<0.05		
					adj. to BMI 6.1 (2.9)[1]	<0.05		
	Cooper et al. (2012)	528/380	Mean change (MC) in biomarker from baseline (B) to follow-up; cross-sectional regression coefficient (csrc) for baseline sample (bs) and longitudinal sample (ls) or longitudinal linear regression coefficient (lirc); additionally adj. to WC[3]	WC (cm)	MC: −1.9 (−2.3 to −1.4)	<0.001	5 of 8	Csrc and lirc calculated for ST in hours/day
					B csrc: 1.8 (0.9–2.8)	<0.001		
					ls csrc: 1.8 (0.6–2.9)	0.002		
				HDL (mmol/l)	bs B csrc: −0.03 (−0.06 to −0.01)	0.005		
					bs B csrc[3]: −0.03, (−0.05 to −0.004)	0.02		
					ls B csrc: −0.04 (−0.076 to −0.01)	0.006		
					ls B csrc[3]: −0.05 (−0.07 to −0.00)	0.01		
					ls FU csrc: −0.05 (−0.088 to −0.020)	0.002		
					ls FU csrc[3]: −0.05 (−0.08 to −0.01)	0.003		
				Insulin (pmol/l)	lirc: −0.04 (−0.08 to −0.01)	0.007		
					MC: −9.4 (−14.4 to −4.4)	<0.001		
					bs b csrc: 8.2 (2.8 to 13.6)	0.003		
					ls B csrc: 12.0 (5.0 to 19.1)	0.001		
					ls B csrc[3]: 8.5 (1.8 to 15.2)	0.01		
					lirc: 8.1 (1.5 to 14.7)	0.01		
				HOMA-IR	MC: −0.36 (−0.6 to −0.0)	0.03		
					bs B csrc: 0.4 (0.1 to 0.7)	0.004		
					ls B csrc: 0.6 (0.2 to 0.9)	0.001		
					ls B csrc[3]: 0.4 (0.1 to 0.8)	0.009		
					ls lirc: 0.4 (0.0 to 0.9)	0.02		
	Allison et al. (2012)	1543	Linear regression coefficient (lrg) calculated for natural logarithm of biomarker and increment of SB (790 MET-minutes/week) adj. for confounders or additionally to BMI and more conf.[4], or add. to WC[5]	Leptin (ng/ml)	0.15 (0.10 to 0.20)	<0.05	4 of 6	
					0.07 (0.04 to 0.11)[4]	<0.05		
					0.07 (0.03 to 0.10)[5]	<0.05		
				TNF-a (pg/ml)	0.04 (0.01 to 0.06)	<0.05		
					0.03 (0.01 to 0.06)[4]	<0.05		
					0.03 (0.00 to 0.06)[5]	<0.05		

Table 3 (*Continued*)

Author, Year	No. of participants	What was analysed? What was measured?	Measured biomarkers	Main results (95%CI) or [SD] or {SE}	P-value	CASP score	Remarks
			Adiponectin / Leptin ratio	−0.17 (−0.23 to −0.11) −0.08 (−0.12 to −0.03)[4] 0.07 (−0.11 to −0.02)[5]	<0.05 <0.05 <0.05		
Anuradha et al. (2011)	5893	Multivariate-adjusted mean difference for highest compared to lowest quartile of TV viewing time (>3 h/d to<1 h/d)	Central retinal vein equivalent (μm)	1.8 (0.4 to 3.2)	<0.05	4 of 6	
Bankoski et al. (2011)	1367	Age and sex adjusted means of biomarker and MES risk profile (according to the ATP III guidelines) for duration of ST[6] (in hours/day) and % of ST of total wear time[7]	Large WC*	no risk 9.2 {0.1}[6] to risk 9.6 {0.1}[6]	0.04	5 of 6	
			Low HDL*	no risk 62.6 {0.4}[7] to risk 66.6 {0.6}[7]	<0.01		
			High Triglycerides*	no risk 64.2 {0.5}[7] to risk 67.7 {0.6}[7]	<0.01		
			High glucose	no risk 9.2 {0.1}[6] to risk 9.6 {0.1}[6]	0.05		
			BMI (kg/m^2)	no risk 63.8 {0.5}[7] to risk 67.0 {0.6}[7]	<0.01		
				no risk 63.6 {0.5}[7] to risk 66.0 {0.6}[7]	<0.01		
Bann et al. (2015)	1130	Mean differences in BMI per hour/day increase in ST measured by accelerometer (ACC)[8] and self-report (SR)[9]	BMI (kg/m^2)	ACC minimally adj.: 0.42 (0.13 to 0.71) ACC fully adj.: 0.44 (0.14 to 0.74) SR minimally adj.: 0.37 (0.06 to 0.67) SR fully adj.: 0.51 (0.19 to 0.82)	0.005 0.004 0.01 0.002	4 of 6	
Cooper et al. (2014)	394 m: 250 w: 144	Linear reg. coeff. for ST and WC, crude, adj. to confounders and add. adj. to PA	WC (cm)	crude: 0.80 (0.31 to 1.29) adj.: 0.97 (0.46 to 1.48) adj. + PA: 0.68 (0.01 to 1.35)	<0.05 <0.05 <0.05	6 of 6	Linear regression analyses performed for ST in hours/day
Jakes et al. (2003)	14189	Age adjusted mean value of biomarker in relation to television viewing h/d (<2 h, 2-2.9, 3-3.9, >4) in men and women, add. adj. to confounders[10]	BMI (kg/m^2)	m: 26.1 [3.1], 26.4 [3.0], 26.7 [3.2], 27.1 [3.2] w: 25.3 [3.9], 26.0 [4.1], 26.2 [4.0], 26.9 [4.4]	<0.001 <0.001	4 of 6	
			WC (cm)	m: 94.0 [9.1], 94.9 [8.8], 95.7 [9.2], 97.0 [9.3] w: 79.6 [9.5], 80.7 [10.0], 81.5 [9.9], 82.9 [10.4]	<0.001 <0.001		
			HC (cm)	m: 102.3 [5.7], 102.8 [5.7], 103.0 [6.2], 103.7 [6.6] w: 102.0 [8.1] 103.2 [8.9], 103.5 [8.4], 104.7 [9.3]	<0.001 <0.001		
			WHR	m: 0.92 [0.06], 0.92 [0.05], 0.93 [0.05], 0.93 [0.05] w: 0.78 [0.05], 0.78 [0.06], 0.79 [0.06], 0.79 [0.06]	<0.001 <0.001		
			BF (%)	m: 22.5 [6.6], 23.4 [6.2], 24.3 [7.5], 25.3 [7.0] w: 37.1 [8.6], 38.9 [8.8], 40.0 [8.6], 41.6 [9.3]	<0.001 <0.001		
			s BP (mmHg)	m: 135.0 [16.8], 136.4 [17.7], 138.1 [17.7], 138.4 [19.0] m[10]: 135.6 [15.1], 136.7 [15.4], 137.7 [14.9], 137.8 [15.6] w: 130.9 [15.1], 132.3 [11.6], 133.4 [11.2], 134.0 [11.7] w[10]: 131.6 [16.8], 132.6 [13.3], 133.3 [13.3], 133.1 [14.3]	<0.001 <0.001 <0.001 <0.01		

			d BP (mmHg)	m: 81.2 [12.6], 83.7 [13.3], 84.9 [13.3], 85.6 [14.3]	<0.001		
				m[10]: 83.6 [11.3], 83.9 [11.6], 84.6 [11.2], 85.1 [11.7]	<0.001		
				w: 79.2 [7.5], 80.1 [7.7], 80.7 [7.5], 81.1 [7.8]	<0.001		
				w[10]: 79.7 [8.4], 80.3 [8.9], 80.7 [8.9], 80.5 [9.5]			
			Triglycerides (mmol/l)	m: 1.70 (0.8–3.6), 1.80 (0.8–3.8), 1.82 (0.9–3.8), 1.92 (0.9–4.1)	<0.001		
				m[10]: 1.73 (0.8–3.6), 1.82 (0.9–3.9), 1.80 (0.9–3.7), 1.88 (0.9–4.0)	<0.01		
				w: 1.38 (0.6–3.1), 1.40 (0.6–3.3), 1.49 (0.6–3.6), 1.54 (0.6–3.9)	<0.001		
				w[10]: 1.42 (0.6–3.2), 1.42 (0.6–3.4), 1.48 (0.6–3.5), 1.49 (0.6–3.8)	<0.001		
			Cholesterol (mmol/l)	m: 5.91 [1.26], 5.93 [1.33], 5.96 [1.33], 6.05 [1.43]	<0.001		
				m[10]: 5.92 [1.13], 5.93 [1.16], 5.96 [1.12], 6.04 [1.17]	<0.01		
				w: 6.17 [1.13], 6.22 [0.77], 6.27 [0.75], 6.28 [0.78]	0.001		
				w[10]: 6.19 [1.26], 6.23 [0.89], 6.27 [0.89], 6.26 [0.95]	<0.01		
			HDL (mmol/l)	m: 1.28 (0.6–2.7), 1.23 (0.6–2.6), 1.22 (0.6–2.5), 1.20 (0.6–2.6)	<0.001		
				m[10]: 1.27 (0.6–2.7), 1.22 (0.6–2.6), 1.22 (0.6–2.5), 1.21 (0.6–2.6)	<0.01		
				w: 1.63 (0.7–3.7), 1.58 (0.7–3.8), 1.57 (0.7–3.7), 1.51 (0.6–3.8)	<0.001		
				w[10]: 1.60 (0.7–3.6), 1.57 (0.7–3.7), 1.57 (0.7–3.7), 1.54 (0.6–3.9)	<0.001		
			LDL (mmol/l)	m: 3.75 [0.84], 3.77 [0.89], 3.82 [0.89], 3.87 [0.95]	<0.001		
				m[10]: 3.75 [0.75], 3.77 [0.77], 3.82 [0.75], 3.87 [0.78]	<0.01		
				w: 3.82 [0.75], 3.90 [0.77], 3.93 [0.75], 3.97 [0.78]	<0.001		
				w[10]: 3.85 [0.84], 3.91 [0.89], 3.93 [0.89], 3.95 [0.95]	<0.01		
			BMI (kg/m^2)	PCC: 0.18	<0.05		
		Pearson cc between biomarker and accelerometer measured ST (in min/d)	WC (cm)	PCC: 0.21	<0.01	5 of 6	
			SBP (mmHg)	PCC: 0.17	<0.05		
			Insulin (mU/dl)	PCC: 0.24	<0.01		
Gabriel et al.(2012)	148	OR of unfavourable biomarker profile by quartiles of TV viewing time (0–1.5 h = reference, 1.6-3.4 h, 3.5–5.5 h, 5.6–18 h), adj. for confounders[11], add. for dietary habits[12], add. for ADJ[13]	Low HDL*	1 (ref), 0.9 (0.4–2.0), 1.2 (0.5–2.7), 2.5 (1.0–5.9)	0.02[11], 0.02[12], 0.01[13]	5 of 6	P-values for linear trends
Gao et al. (2007)	455		High cholesterol/HDL ratio*	1 (ref), 1.2 (0.7–2.1), 1.3 (0.7–2.4), 2.0 (1.1–3.7)	0.01[11], 0.03[12], 0.04[13]		
			High WHR*	1 (ref), 1.6 (0.8–3.1), 2.3 (1.1–4.8), 3.9 (1.8–8.4)	0.0003[11], 0.0008[12], 0.0006[13]		

Table 3 (*Continued*)

Author, Year	No. of participants	What was analysed? What was measured?	Measured biomarkers	Main results (95%CI) or [SD] or {SE}	P-value	CASP score	Remarks
Gennuso et al. (2013)	1914	Association of least square means of biomarkers with quartiles of sedentary hours (0–7.92, 7.93–8.17, 8.18-10.63, >10.64)	BMI (kg/m^2)	26.6 [0.6], 27.4 [0.5], 27.8 [0.5], 28.8 [0.4]	0.01	5 of 6	P-values for linear trends
			WC (cm)	98.2 [1.6], 100.2 [1.3], 101.9 [1.4], 104.4 [1.0]	<0.01		
			Glucose (mg/dl)	115.0 [1.2], 114.8 [1.2], 119.2 [1.2], 119.8 [1.2]	0.04		
			CRP (mg/dl)	0.24 [1.15], 0.24 [1.12], 0.26 [1.12], 0.34 [1.14]	<0.01		
Harner et al. (2013)	4964	Dose-response association for TV viewing (<2 = Ref., 2–4, 4–6, >6 h/d) and log transformed mean CRP values, adj. for age, sex[14]; further adj. to PA, BMI[15]	CRP (log transformed)	Ref., 0.11 (0.04 to 0.18), 0.27 (0.2 to 0.34), 0.29 (0.22 to 0.36)[14]	<0.001	4 of 6	
				0.04 (−0.03 to 0.1), 0.12 (0.06 to 0.19), 0.11 (0.04 to 0.17)[15]	<0.001		
Henson et al. (2013)	558	Regression coeff. for ST (in h/day) with biomarker, adj. to confounders, add. to PA[16], add. adj. to BMI and HbA_1c[17]	IL-6 (pg/ml)	0.242 {0.056}, 0.231 {0.073}[16], 0.212 {0.072}[17]	<0.001, 0.002[16], 0.003[17]	5 of 6	
Larsen et al. (2014a)	539 m: 135 w: 404	Variance (V) in mean values of biomarker and ST tertiles or cross-sectional regression coefficient (csrc) of biomarker to ST tertiles (<2.5, 2.5–4, >4 sitting hours/day), unadj., adj. to demographics[18], to CVD RF[19], to BMI[20], to inflammatory markers[21]	Adiponectin (μg/ml)	V: 10.4 [6.0], 9.4 [4.9], 10.8 [6.6], 10.8 [5.9]	0.032	4 of 6	Pos. assoc. for V of intra-thoracic fat (p = 0.018), intermuscular fat (p = 0.001) and subcutaneous fat (p = 0.034) but not sig. in csrc
			Intra-thoracic fat (cm^2)	V: 71.8 [64.1], 61.2 [50.5], 75.8 [70.9], 80.0 [66.1]	0.018		
			Intermuscular fat (cm^2)	V: 21.4 [11.0], 19.4 [8.8], 23.5 [12.1], 21.4 [11.3]	0.001		
			Subcutaneous fat (cm^2)	V: 253.8 [122.7], 243.4 [106.3], 273.2 [131.4], 246.6 [126.2]	0.034		
			Pericardial fat (cm^2)	csrc: 3.19 (0.45 to 5.92)	0.022		
				csrc[18]: 3.19 (0.45 to 5.92)	0.022[18]		
				csrc[19]: 3.32 (0.84 to 5.81)	0.009[19]		
				csrc[20]: 2.39 (0.07 to 4.72)	0.044[20]		
				csrc[21]: 2.45 (0.12 to 4.77)	0.039[21]		
Lee et al. (2015)	1168	Unadj.[22] as well as adj.[23] average differences (AD) in function (as biomarker) between SB quartiles (Q2 vs. Q1, Q3 vs. Q1 and Q4 vs. Q1)	Gait speed (feet/s)	AD[22]: 0.35 [0.08], 0.44 [0.08], 0.44 [0.08]	<0.001[22]	5 of 6	
				AD[23]: 0.20 [0.07], 0.21 [0.08], 0.21 [0.08]	<0.001[23]		
			Chair stand rate (stands/min)	AD[22]: 3.00 [0.95], 3.28 [0.98], 5.30 [0.95]	<0.001[22]		
				AD[23]: 1.85 [0.90], 1.46 [0.96], 3.43 [0.98]	0.0016[23]		
Lynch et al. (2011)	1024	Association of SB quartiles (<7.74, 7.74-<8.8, 8.8-<9.84, ≥9.84 h/d), adj. in model 1[24] to age; model 2 for BMI: ethnicity, alcohol intake, age at first birth, age at menarche; model 2[25] for WC: ethnicity, educational attainment, marital status, annual family income, alcohol intake, age at first birth	BMI	model 1[24]: 26.7 (25.9 to 27.5), 27.6 (26.8 to 28.5), 27.6 (26.6 to 28.6), 29.9 (28.6 to 31.2)	<0.001[24]	5 of 6	CRP, insulin, HOMA-IR showed also sig. pos. trend with SB quartiles, sig. after multivariate adj., but ns after adj. to WC; all results as marginal means for each quartile, back-transformed for all log-transformed outcomes
				model 2[25]: 27.2 (26.4 to 27.9), 27.7 (26.9 to 28.6), 27.5 (26.6 to 28.4), 29.3 (28.1 to 30.5)	0.02[25]		
			WC	model 1[24]: 91.9 (89.7 to 94.2), 94.7 (92.9 to 96.5), 95.7 (93.3 to 98.1), 102.1 (99.4 to 104.8)	<0.001[24]		
				model 2[25]: 93.2 (90.8 to 95.7), 95.1 (93.1 to 97.1), 95.5 (93.2 to 97.9), 100.5 (97.9 to 103.1)	0.003[25]		

Reaven et al. (1991)	641	Mean values of age adj. biomarker by exercise category (none, light, moderate, heavy); s BP additionally adj. to age[26], age + BMI[27], age + BMI + alcohol + estrogen[28], age + BMI + fasting insulin[29], age + BMI + 2 h insulin[30]	HR (beats/min)	66.5, 64.8, 63.9, 61.4	0.01	3 of 6		P-values for linear trends; values for DBP were ns when unadj. but linear trend adj. to same confounders as SBP were sig (p = 0.006[24], p = 0.044[25], p = 0.049[26], p = 0.034[27], p = 0.025[28])
			BMI (kg/m²)	26.3, 24.1, 25.1, 23.4	0.05			
			S BP (mmHg)	143.3, 136.8, 130.3, 122.6	<0.001			
				142.1, 135.5, 133.0, 130.3[26]	0.003			
				140.8, 135.6, 132.5, 131.3[27]	0.012			
				140.7 135.6 132.5 131.4[28]	0.013			
				140.7 135.5 132.5 131.4[29]	0.014			
				140.9 134.9 131.0 131.3[30]	0.010			
			Fasting Insulin (μU/ml)	16.9, 13.7, 12.4, 11.2	0.002			
			2 h Insulin (μU/ml)	15.0, 88.5, 79.2, 66.2	0.001			
Stamatakis et al. (2012)	2765 (SR) 649 (accel.)	Mean values of biomarker and tertiles of self-reported (SR; <291, 291–394, > 394 min/d) or accelerometer measured (accel; <507, 507–571, > 571 min/d) ST	BMI (kg/m²)	SR: 27.4 [4.5], 27.9 [4.6], 28.5 [5.1]	<0.01	5 of 6		P-value for one-way ANOVA test; a sig. pos. multivariate reg. coeff. was calculated for SR SB with BMI and HbA_1c and similar for accel. measured SB with cholesterol and HbA_1c, which was ns after further adj.
				accel.: 27.1 [4.0], 28.6 [4.9], 28.5 [4.7]	<0.01			
			WC (cm)	SR: 94.8 [13.1], 96.0 [12.8], 98.3 [13.4]	<0.01			
				accel.: 93.1 [12.7], 96.5 [13.7], 99.6 [12.8]	<0.01			
			HDL (mmol/l)	SR: 1.6 [0.4], 1.6 [0.4], 1.5 [0.4]	<0.01			
				accel.: 1.7 [0.4], 1.6 [0.4], 1.5 [0.4]	<0.01			
			HbA_1c (%)	SR: 5.8 [0.7], 5.8 [0.6], 6.0 [0.9]	<0.01			
				accel.: 5.8 [0.6], 5.8 [0.6], 6.0 [0.8]	0.01			
			Cholesterol/HDL ratio	SR: 3.9 [1.0], 4.0 [1.0], 4.1 [1.2]	0.01			

accel. = accelerometer; AD = average differences; adj. = adjusted; ADL = activities of daily living; B = baseline; BF% = percent of body fat; BMI = body mass index; BP = blood pressure; bpm = beats per minute; cc = correlation coefficient; CI = confidence interval; coeff. = coefficient; CG = control group; CMRF = cardio-metabolic risk factors; CRP = C-reactive protein; csrc = cross-sectional regression coefficient; FU = follow up; HbA1c = glycated hemoglobin; HC = hip circumference; HDL = high density lipoprotein; HOMA-IR = homeostatic model assessment of insulin resistance; HR = heart rate; IC = intervention group; IL = interleukin; LDL = low density lipoprotein; lrc = linear regression coefficient; MC = mean change; MD = mean difference; MES = metabolic syndrome; NC = neck circumference; neg. = negative; NR = not reported; OR = odds ratio; PCC = Pearson correlation coefficient; pos. = positive; POS = prospective observational studies; RCT = randomized controlled trials; ref = reference; S/D = systolic/diastolic; SB = sedentary behaviour; SD = standard deviation; SE = standard error; sig. = significant; SR = self report; ST = sedentary time; TNF-a = tumor necrosis factor alpha; unadj. = unadjusted; V = Variance; WC = waist circumference; WHR = waist to hip ratio.

* According to the definitions of Adult Treatment Panel III (ATP-III): Cholesterol/HDL ratio > 4.5 was considered as high.

3.5. Risk of bias (quality) appraisal

After revising the 40 articles by CASP criteria (CASP, 2016) and general quality criteria (correctness of data illustration, selection or reporting bias, misclassification etc.) we excluded 13 CSS for the following quality linked issues [CASP score] and 1 POS:

a SB was not sufficiently measured: Ewald et al. (2010) [3 of 6] and Bianchi et al. (2008) [2 of 6] evaluated time spent being sedentary with the Physical Activity Scale for the Elderly (PASE) and Kaino et al. (2013) [2 of 6] used the Japan Arteriosclerosis Longitudinal Study Physical Activity Questionnaire (JALSPAQ) which are good instruments to measure PA but weak in calculating SB; Calderon-Garcia et al. (2013) [4 of 6] calculated ST by asking "How much do you exercise or strain yourself physically in your leisure time?", which had poor validity.
b Missing information about recruitment or cohort characterstics (Azzabou et al., 2015);
c SB defined as simply being the opposite of PA, as in Gába et al. (2012) [3 of 6] or unclear definition of sedentariness (Elkan et al., 2011) [1 of 6], Belza et al. (2001) [3 of 6]);
d Poor quality of exposure or outcome assessment; e.g. Inoue et al. (2012) [3 of 6] calculated BMI by self- reported weight and height, among other issues;
e Missing evaluation or lack of adjustment for important confounding factors (comorbidities, medication status etc.), e.g. Li et al. (2009) [2 of 6], Babaroutsi et al. (2005) [3 of 6] and others (Azzabou et al., 2015; Elkan et al., 2011; Belza et al., 2001; Inoue et al., 2012);
f Evidence of selection bias – like in Knight and Bermingham (1999) [2 of 6] who compared a cohort from Day Care Centre to a cohort from a bowling club;
g Implausible or irreproducible data – e.g. implausible data of sample size and sample origin (Azzabou et al., 2015) [3 of 6]);
h Biomarker calculated by self-report, like BMI from self-reported weight and height or SB and biomarkers weren't measured at the same point in time (Scott et al., 2015));
i We excluded the POS from Wijndaele et al. (2009) [3 of 8] because BMI was calculated by self-report weight.

After the exclusion of these studies only 2 articles (Reaven et al., 1991; Chase et al., 2014) with a low CASP-score ≤ 3 remained. The mean CASP score of RCTs was 5 out of 6, for cohort studies 5 out of 8 and for CSS 4.5 out of 6.

3.6. Relation of SB and biomarkers

3.6.1. Sedentary behaviour and anthropometric and systemic biomarkers

Of the 15 studies exploring this biomarker, 9 demonstrated a positive association, including 1 RCT (Kallings et al., 2009) and 1 POS (Fung, 2000) study (Table 2 and 3), whereas 2 RCTs (Suboc et al., 2014; Kirk et al., 2009) didn't show statistically significance, thus there is mixed evidence for the association of SB to BMI. WC was also positively associated with SB in 1 POS (Cooper et al., 2012) and 7 CSS, but were not statistically significant in 4 RCTs. Relationships between SB and both systolic BP (3 of 11 studies reporting this biomarker found positive association) and diastolic BP (1 out of 10 studies found a positive associations) were found, whereas the majority showed non-significant results. Neck circumference and fat mass were positively correlated to SB but were investigated in only one RCT. There was only limited or no evidence for the other anthropometric biomarkers (see Table 2 and 3).

3.6.2. Sedentary behaviour and blood lipids

Total cholesterol, HDL, low density lipoprotein (LDL) and triglycerides were the main focus in the investigated studies. If statistically significant association was prevalent, it was in an unfavourable direction. For total cholesterol positive association was found in 1 RCT (Kallings et al., 2009), whereas the 3 RCTs (Aadahl et al., 2014; Suboc et al., 2014; Kirk et al., 2009) and 1 POS (Fung, 2000) didn't show any statistically significant association. HDL was statistically significant negatively associated with SB in 2 POS (Cooper et al., 2012; Fung, 2000) but results in 4 RCTs (Aadahl et al., 2014; Suboc et al., 2014; Kirk et al., 2009; Kallings et al., 2009) were statistically not significant. Similar results were detected for the other blood lipids (see Tables 2 and 3). Most RCTs didn't show statistically significant results, hence there is generally no evidence for an association of SB and blood lipids. Results linking SB and blood lipids mostly derived from CSS studies and thus should be interpreted accordingly.

3.6.3. Sedentary behaviour and glycaemic biomarkers

There was some indication found of an unfavourable impact of SB on fasting insulin levels, with statistically significant associations in 1 RCT (Aadahl et al., 2014) and 1 POS (Cooper et al., 2012). However 1 RCT (Suboc et al., 2014) and 1 POS (Fung, 2000) didn't show any association, which lead to mixed evidence for a possible impact of SB on insulin levels. For HbA_1c, only 1 RCT (Kallings et al., 2009) was statistically significant. HOMAR-IR (Cooper et al., 2012) and C-peptide (Fung, 2000) were positively correlated to SB in 1 POS. Glucose levels did not appear to be related to SB in 3 RCTs (Aadahl et al., 2014; Suboc et al., 2014; Kallings et al., 2009). Initially equivocal results in 2 CSS, with 1 positive (Gennuso et al., 2013) and 1 negative association (Bankoski et al., 2011) were clarified by contacting the author. In both studies SB was associated with higher blood glucose levels. The impact of SB on glycaemic biomarkers was limited and largely restricted to CSS (Tables 2 and 3), precluding definitive conclusion.

3.6.4. SB and muscle or physical performance biomarkers

Muscle tissue, performance, strength or other performance components were measured in 5 CSS (Bann et al., 2015; Santos et al., 2012; Sardinha et al., 2015; Lee et al., 2015; Larsen et al., 2014a). 4 CSS also evaluated the association of SB and some performance biomarkers. Lee et al. (2015) found a statistically significant negative correlation for SB with gait speed and chair stand rate. Santos et al. (2012) constructed a composite Z-score of 6 performance biomarkers (6 min walk test, 8 foot up and go, arm curl, chair stand rate, chair sit and reach or back scratch) which association with SB was significant negative, but he did not list the results separately. Bann et al. (2015) and Sardinha et al. (2015) did not find a significant correlation for SB and performance biomarkers.

3.6.5. SB and inflammatory biomarkers

There was a relative paucity of studies investigating inflammatory biomarkers and SB. CRP was investigated most frequently, although restricted to 4 CSS studies and 1 RCT, with only 2 CSS studies demonstrating that SB was positively associated with CRP. Only 2 CSS studies investigated IL-6 and SB, with 1 CSS finding a positive association. Given the limited number of studies and over reliance on CSS, the evidence base is inconclusive concerning the relationship between SB and inflammatory markers.

3.6.6. SB and other biomarkers

There was a distinct lack of studies investigating renal or bone biomarkers and SB. Only 1 study measured Vitamin D status (Scott et al., 2015), but it was considered as too low in quality ((see 9) in 'Risk of bias appraisal'), because different points of time exposure and outcome were measured.

Leptin, which can be seen as adiposity-associated inflammation marker or regulation marker of hunger and fat metabolism, was

higher with a higher amount of time spent sedentary (2 of 4 studies significant, 1 POS).

We could not identify any study investigating renal, cellular, respiratory, signal transduction or genetic biomarkers and SB meeting our inclusion criteria. None of the included studies evaluated the impact of SB on biomarkers of the gastrointestinal or peripheral/central nervous system, neither focused on steroid or hormone biomarkers.

4. Discussion

Within our comprehensive systematic review, findings from high quality papers showed mixed evidence for the association of SB and biomarkers. When statistically significant results were prominent, SB was associated in an unfavourable direction, especially in anthropometric (BMI, WC, neck circumference, fat mass), blood lipid (cholesterol, HDL, LDL), glycaemic (HbA_1c, insulin, HOMA-IR, C-peptide) and hormonal (leptin) biomarkers. However several statistically non-significant study results were detected, many of which were of high quality. Some results of lower quality studies may be incidental findings or point to the existence of additional confounders, which are unaccounted so far.

Despite the relative paucity and equivocal nature of SB and biomarkers in older age, studies performed in younger cohorts strengthen the hypothesis that SB has harmful effects on biomarker levels. For instance, Healy et al. (2008a,b) found an inverse relation of breaks in ST and BMI (Healy et al., 2008a) and WC (Healy et al., 2008a,b) or Zhou et al. (2016) revealed an increased risk for developing Metabolic Syndrome (MES) with higher ST support those findings. Fasting insulin levels, another MES risk factor, improved with reducing ST (Aadahl et al., 2014; Cooper et al., 2012). Similar results for glycaemic biomarkers, such as postprandial glucose and insulin levels were detected in other RCTs (Duvivier et al., 2013; Peddie et al., 2013) or CSS (Yates et al., 2012) performed in younger cohorts. Considering results from Krogh-Madsen et al. (2010), showing a decrease in insulin-stimulated muscle activity phosphorylation and decreased peripheral insulin sensitivity by reducing daily activity for only 2 weeks, there appears to be a strong connection between SB and impaired glucose and insulin metabolism in younger age.

Our review identified some studies that evaluated the association between change in ST and systemic parameters, including blood pressure (Gabriel et al., 2012; Reaven et al., 1991; Jakes et al., 2003), or heart rate (Reaven et al., 1991). Surprisingly and contrary to our expectation we identified no association between change of SB with blood pressure in 3 included RCTs (Suboc et al., 2014; Kirk et al., 2009; Kallings et al., 2009). Investigations in younger cohorts demonstrated a clear trend of significantly improving BP levels by reducing ST (Christofaro et al., 2015) or by breaking up prolonged sitting periods (Larsen et al., 2014b). Already the advice of increasing PA levels seems to have a positive effect coming along with lower BP levels (Figueira et al., 2014). Possible explanations for no effects in older cohorts could be confounding by antihypertensive medication, increased arterial stiffness or reduced heart rate variability (Bonnemeier et al., 2003) in older age. Similar effects were detected for blood lipids, with better profiles associated to less ST (Marsh et al., 2014). As underlying mechanism Hamilton et al. (2007, 2004) suggested a poor lipid metabolism with inactivity by suppression of skeletal muscle lipoprotein lipase activity. SB has also been associated with chronic low-grade inflammation in younger cohorts (Yates et al., 2012; Falconer et al., 2014). When looking in elderly people we only identified few, mainly CSS (Gennuso et al., 2013; Henson et al., 2013; Fung, 2000; Allison et al., 2012; Hamer et al., 2013), showing higher levels of CRP, IL-6 and leptin in those with less physical activity. Besides missing longitudinal data a higher low-grade inflammation (Franceschi and Campisi, 2014) in older age could distort or reduce the effect size of these outcomes.

Over and above preserving autonomy in older age is important in order to maintain independence and quality of life. Dunlop et al. (2015) reported a 46% greater odds of ADL disability for each hour spent sedentary. Muscle function (Conley et al., 2013) also appears to be negatively affected by SB suggesting that macroscopic/performance (Cawthon et al., 2013) and microscopic/biochemical parameters (Conley et al., 2013) would change depending on ST. The results found for our systematic review were few. Results from Santos et al. (2012), who constructed a composite Z-score out of different performance biomarkers, suggests a negative association for performance biomarkers with SB, but no longitudinal data of performance or muscle biomarkers is available and thus drawing of causal conclusions is not possible. Given this, future prospective studies should prioritise functional assessments like the short physical performance battery (SPPB), grip strength and dynamic muscle function. Such measures are easy to ascertain with an evaluated predictive profile and can serve as modifiable surrogates of autonomy in later life.

The highlighted results of the four "risk population" studies showed associations for SB with biomarkers in the same direction as the studies performed in non-risk populations. The results from Lee et al. (2015), performed in the high risk osteoarthritis population with lower gait speed and lower chair stand rate associated with higher levels of SB can be argued over. This is the only study, which showed (remaining) statistically significant results for performance parameters. Even if SB measurements were adjusted for osteoarthritis pain index, osteoarthritis symptoms and other comorbidity indices, there could be still another unknown confounder, related to osteoarthritis triggering this biomarker outcome.

Surprisingly, there was an absence of studies (meeting our inclusion criteria) investigating SB and its possible impact on renal, muscle or bone biomarkers performed in the elderly. There is however good reason to believe that especially bone and muscle metabolism is influenced from SB due to multifactorial processes. Prioreschi et al. (2015) results from a smaller cohort revealed low bone mass for higher levels of SB and a possible protective effect for bone mineral density with breaking up ST more frequently. Even in younger cohorts, ST has been implicated as being negatively related to changes in whole-body bone mineral density, lumbar spine bone mineral content, lumbar spine bone area and femoral neck (Ivuškāns et al., 2015).

A large number of studies were excluded from our review because they specifically measured PA rather than focus on the distinct construct of SB. For instance, several studies focussed on a lack of PA rather than SB (Kirk et al., 2009; Kallings et al., 2009). Recently there is a rising interest of SB consequences and the idea of clearly differentiating between the distinct behaviours of SB and PA. In this direction, Barone Gibbs et al. (2016) demonstrated a higher effectiveness for improving the SPPB score by reducing SB compared to increasing moderate to vigorous PA. For that reason, biomarkers should be evaluated for both, PA and SB. Former investigations have shown that SB effects on biomarkers are independently of MVPA levels (Cooper et al., 2014; Healy et al., 2008b; Yates et al., 2012). Additionally reducing inactivity often has a higher effectiveness on the biomarker level, than the amount of physical activity itself (Duvivier et al., 2013; Peddie et al., 2013). For that reason new studies should investigate biomarkers and health outcomes with focusing on reducing SB.

Whilst our comprehensive review provides novel insights, some limitations should be mentioned. First, we identified relatively few high quality or longitudinal studies investigating SB and biomarkers specifically in older adults. Therefore, we were not able to

conduct a meta-analysis as we anticipated. Additionally the CADTH tool (CADTH, 2016), used for standardised statements about the statistical significance, was adapted, so we were able to apply it to fewer studies available. This should be considered, when rating the state of evidence. Second, there were no stratified analyses assessing the question if age or gender is a possible effect modifier. Both, age and gender were often added into the analysis as confounders, but there is still the necessity to evaluate the possible presence of interaction in the association between sedentary behaviour and different biomarkers. Third there was considerable heterogeneity in the definitions of SB and the high diversity of reported outcome-parameters, again a pertinent factor making meta-analysis impossible. SB was often misclassified as simply a lack of PA. With respect to the performed analyses some studies measured the mean change (Bann et al., 2015; Kallings et al., 2009), others calculated odds ratios (Gao et al., 2007) or Pearson correlation coefficients (Gabriel et al., 2012), whereas others calculated a linear or multiple regression coefficient (Henson et al., 2013; Cooper et al., 2014; Allison et al., 2012). Strict definitions focussing specifically on SB are necessary to allow comparison of results from different studies. There are currently several initiatives attempting to harmonize these approaches such as the standardised definition of SB published in 2012 (Sedentary Behaviour Research Network, 2012), the 2011 launched online "Sedentary Behaviour Research Network" (SBRN) (SBRN, 2016) or the SIT project from Chastin and Skelton (in press). There are some initiatives aimed tackling SB. A large Canadian organization called ParticipACTION (ParticipACTION, 2016) is trying to help Canadians to sit less by offering age adjusted activity programs. Similar intentions are given in the multi-centre EU study SITLESS (SITLESS, 2016) with the aim of reducing SB in elderly by a PA intervention enhanced by self-management-strategies. Objectively measured SB will be correlated with several biomarkers and muscle biopsy results to further elucidate the biochemical influence of SB on health outcomes.

Currently, we have limited understanding of the impact of SB on different biomarker systems in older age. The current knowledge base in this regard is overwhelmingly based upon CSS. Given our findings, there is an urgent need for adequately representative, prospective cohort and randomized controlled studies to investigate the impact of SB on various biomarkers in order to ascertain a better understanding of the pathophysiological and also to test the hypothesis for causality. Besides the majority of studies were of moderate to high quality, the presence of reporting bias should still be considered. Some effects of selection bias could be present as well, regarding that some studies focused on participants of a high risk population, such as diabetes mellitus patients (Cooper et al., 2012, 2014) or breast cancer survivors (Lynch et al., 2010). Additionally 17 of our 26 studies calculated SB by subjective methods, which are less accurate than objective methods, since people tend to underestimate their time spend in SB, due to simple uncertainty or social desirability (Harvey et al., 2013). In future research objectively measured SB should be preferred to better calculate the real time spent sedentary.

5. Conclusion

There is a paucity of studies investigating the impact of sedentariness in older people. Currently there is mixed evidence for the impact of SB and biomarkers. When statistically significant results were found, SB was associated in an unfavourable way to biomarkers, but results were mostly derived from cross sectional studies and thus should be interpreted accordingly. Due to a broad definition and misclassification of sedentary behaviour as simple lack of physical activity there is still a deficiency of evident, causal relations. There is a need for high quality studies to better understand the underlying pathophysiological pathways and finally the burden between sedentary behaviour and the biomarkers implicated. Broad investigations are necessary to evaluate possible impact of sedentary behaviour on biomarkers, including those with an absence of data such as bone and muscles biomarkers. Future research should utilise an official definition of sedentary behaviour, clearly disentangle the relationships between each biomarker and sedentary behaviour and physical activity and use objective or at the least use standardised self-report measures for assessing sedentary time.

Acknowledgement

The work described in this publication was part of the SITLESS project, supported and funded by the European Union program Horizon 2020 (H2020-Grant 634270). Consortium members of the participating organisations of the SITLESS project: Antoni Salvà Casanovas, Health and Ageing Foundation of the Autonomous University of Barcelona, Spain; Laura Coll-Planas, Health and Ageing Foundation of the Autonomous University of Barcelona, Spain; Miriam Guerra-Balic, Faculty of Psychology, Education and Sport Sciences Blanquerna, Ramon Llull University, Barcelona. Carme Martin-Borràs, Faculty of Psychology, Education and Sport Sciences Blanquerna, Ramon Llull University, Barcelona. Javier Jerez-Roig, Faculty of Psychology, Education and Sport Sciences Blanquerna, Ramon Llull University, Barcelona. Guillermo Oviedo, Faculty of Psychology, Education and Sport Sciences Blanquerna, Ramon Llull University, Barcelona. Marta Santiago-Carrés, Faculty of Psychology, Education and Sport Sciences Blanquerna, Ramon Llull University, Barcelona; Mathias Skjoedt, Department of Sport Science and Clinical Biomechanics, University of South Denmark, Denmark; Frank Kee, Centre for Public Health, Queen's University Belfast, Belfast, North-Ireland; Jason J. Wilson, Centre for Public Health, Queen's University Belfast, Belfast, North-Ireland; Emma McIntosh, Health Economics and Health Technology Assessment, University of Glasgow, Scotland; Dietrich Rothenbacher, Institute of Epidemiology and Medical Biometry, Ulm University, Germany; Paolo Caserotti, Department of Sport Science and Clinical Biomechanics, University of South Denmark, Denmark; Guillaume Lefebvre, SIEL, Sport initiative et Loisir Bleu association, Straßbourg, France.

Appendix A.

Table A1

Table A1
Search strategy.

concepts	search terms
sedentariness	seden* OR television OR accelerometer OR pedometer
age	age OR aging OR elderly OR older
biomarkers	bone biomarker OR biomarker OR CRP OR interleukin OR endocrine OR diabetes OR insulin OR cardiovascular OR CNS OR central nervous system OR neurological OR hormones OR inflammation OR hematology OR blood OR liquor OR epigenetic OR genetic OR DNA OR RNA or ultrasound OR BIA or bioelectrical OR caliper OR stem cell OR cerebrovascular OR cancer OR cytokine OR mitochondr* OR immune OR protein OR urine OR muscle OR gait OR factor OR transcription OR strength OR handgrip OR oncology OR nephrology OR men health OR women health OR COPD OR pulmonary OR lung OR asthma OR glucose OR GID OR gastrointestinal OR gastric OR lipoprotein OR anabol OR katabol OR thyroid OR steroid OR metabolic OR testosterone OR estrogen

Appendix B.

Table A2

Table A2
Definitions.

Accelerometer	An instrument for measuring the acceleration of a moving body
Biomarker	A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention
Moderate-vigorous physical activity	Physical activity performed at intensity ≥3 Metabolic Equivalents (METs)
Sedentary behaviour	Activities with an energy expenditure ≤ 1.5 Metabolic Equivalents (METs) while in a sitting or reclining posture during waking hours; not simply the absence of physical activity

References

Aadahl, M., Linneberg, A., Møller, T.C., et al., 2014. Motivational counseling to reduce sitting time. Am. J. Prev. Med. 47 (5), 576–586, http://dx.doi.org/10.1016/j.amepre.2014.06.020.

Allison, M.A., Jensky, N.E., Marshall, S.J., Bertoni, A.G., Cushman, M., 2012. Sedentary behavior and adiposity-associated inflammation. Am. J. Prev. Med. 42 (1), 8–13, http://dx.doi.org/10.1016/j.amepre.2011.09.023.

Anuradha, S., Healy, G.N., Dunstan, D.W., et al., 2011. Physical activity, television viewing time, and retinal microvascular caliber: the multi-ethnic study of atherosclerosis. Am. J. Epidemiol. 173 (5), 518–525, http://dx.doi.org/10.1093/aje/kwq412.

Azzabou, N., Hogrel, J.-Y., Carlier, P.G., 2015. NMR based biomarkers to study age-related changes in the human quadriceps. Exp. Gerontol. 70, 54–60, http://dx.doi.org/10.1016/j.exger.2015.06.015.

Babaroutsi, E., Magkos, F., Manios, Y., Sidossis, L.S., 2005. Lifestyle factors affecting heel ultrasound in Greek females across different life stages. Osteoporos. Int. 16 (5), 552–561, http://dx.doi.org/10.1007/s00198-004-1720-4.

Bankoski, A., Harris, T.B., McClain, J.J., et al., 2011. Sedentary activity associated with metabolic syndrome independent of physical activity. Diabetes Care 34 (2), 497–503, http://dx.doi.org/10.2337/dc10-0987.

Bann, D., Hire, D., Manini, T., et al., 2015. Light intensity physical activity and sedentary behavior in relation to body mass index and grip strength in older adults: cross-Sectional findings from the lifestyle interventions and independence for elders (LIFE) study. PLoS One 10 (2), e0116058, http://dx.doi.org/10.1371/journal.pone.0116058, In: Tranah, G. (ed.).

Barone Gibbs, B., Brach, J.S., Byard, T., et al., 2016. Reducing sedentary behavior versus increasing moderate-to-vigorous intensity physical activity in older adults: a 12-week randomized, clinical trial. J. Aging Health, http://dx.doi.org/10.1177/0898264316635564.

Belza, B., Steele, B.G., Hunziker, J., Lakshminaryan, S., Holt, L., Buchner, D.M., 2001. Correlates of physical activity in chronic obstructive pulmonary disease. Nurs. Res. 50 (4), 195–202.

Bianchi, G., Rossi, V., Muscari, A., Magalotti, D., Zoli, M., Pianoro Study Group, 2008. Physical activity is negatively associated with the metabolic syndrome in the elderly. QJM 101 (9), 713–721, http://dx.doi.org/10.1093/qjmed/hcn084.

Biomarkers Definitions Working Group, 2001. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin. Pharmacol. Ther. 69 (3), 89–95, http://dx.doi.org/10.1067/mcp.2001.113989.

Biswas, A., Oh, P.I., Faulkner, G.E., et al., 2015. Sedentary time and its association with risk for disease incidence, mortality, and hospitalization in adults: a systematic review and meta-analysis. Ann. Intern. Med. 162 (2), 123, http://dx.doi.org/10.7326/M14-1651.

Bonnemeier, H., Richardt, G., Potratz, J., Wiegand, U.K., Brandes, A., Kluge, N., Katus, H.A., 2003. Circadian profile of cardiac autonomic nervous modulation in healthy subjects: differing effects of aging and gender on heart rate variability. J. Cardiovasc. Electrophysiol. 14 (August (8)), 791–799.

Boss, G.R., Seegmiller, J.E., 1981. Age-related physiological changes and their clinical significance. West. J. Med. 135 (6), 434–440.

Brocklebank, L.A., Falconer, C.L., Page, A.S., Perry, R., Cooper, A.R., 2015. Accelerometer-measured sedentary time and cardiometabolic biomarkers: a systematic review. Prev. Med. 76, 92–102, http://dx.doi.org/10.1016/j.ypmed.2015.04.013.

Canadian Agency for Drugs and Technology in Health (CADTH), https://www.cadth.ca/interventions-directed-professionals.

Critical Appraisal Skills Programme (CASP). http://www.casp-uk.net/.

Calderon-Garcia, J.F., Lavado-Garcia, J.M., Martin, R.R., Moran, J.M., Canal-Macias, M.L., Pedrera-Zamorano, J.D., 2013. Bone ultrasound and physical activity in postmenopausal Spanish women. Biol. Res. Nurs. 15 (4), 416–421, http://dx.doi.org/10.1177/1099800412459800.

Castaneda, C., Layne, J.E., Munoz-Orians, L., et al., 2002. A randomized controlled trial of resistance exercise training to improve glycemic control in older adults with type 2 diabetes. Diabetes Care 25 (12), 2335–2341.

Cawthon, P.M., Blackwell, T.L., Cauley, J.A., et al., 2013. Objective assessment of activity, energy expenditure, and functional limitations in older men: the osteoporotic fractures in men study. J. Gerontol. A. Biol. Sci. Med. Sci. 68 (12), 1518–1524, http://dx.doi.org/10.1093/gerona/glt054.

Chase, J.M., Lockhart, C.K., Ashe, M.C., Madden, K.M., 2014. Accelerometer-based measures of sedentary behavior and cardio-metabolic risk in active older adults. Clin. Invest. Med. 37 (2), E108–116.

Chastin, S., Skelton, D., http://www.sedentarybehaviourclassification.net.

Christofaro, D.G.D., De Andrade, S.M., Cardoso, J.R., Mesas, A.E., Codogno, J.S., Fernandes, R.A., 2015. High blood pressure and sedentary behavior in adolescents are associated even after controlling for confounding factors. Blood Press. 24 (5), 317–323, http://dx.doi.org/10.3109/08037051.2015.1070475.

Church, T.S., Thomas, D.M., Tudor-Locke, C., et al., 2011. Trends over 5 decades in U. S. occupation-related physical activity and their associations with obesity. PLoS One 6 (5), e19657, http://dx.doi.org/10.1371/journal.pone.0019657, In: Lucia, A. (ed.).

Conley, K.E., Amara, C.E., Bajpeyi, S., et al., 2013. Higher mitochondrial respiration and uncoupling with reduced electron transport chain content *in vivo* in muscle of sedentary versus active subjects. J. Clin. Endocrinol. Metab. 98 (1), 129–136, http://dx.doi.org/10.1210/jc.2012-2967.

Cooper, A.R., Sebire, S., Montgomery, A.A., et al., 2012. Sedentary time, breaks in sedentary time and metabolic variables in people with newly diagnosed type 2 diabetes. Diabetologia 55 (3), 589–599, http://dx.doi.org/10.1007/s00125-011-2408-x.

Cooper, A.J.M., Brage, S., Ekelund, U., Wareham, N.J., Griffin, S.J., Simmons, R.K., 2014. Association between objectively assessed sedentary time and physical activity with metabolic risk factors among people with recently diagnosed type 2 diabetes. Diabetologia 57 (1), 73–82, http://dx.doi.org/10.1007/s00125-013-3069-8.

Dunlop, D.D., Song, J., Arnston, E.K., et al., 2015. Sedentary time in US older adults associated with disability in activities of daily living independent of physical activity. J. Phys. Act. Health 12 (1), 93–101, http://dx.doi.org/10.1123/jpah.2013-0311.

Duvivier, B.M.F.M., Schaper, N.C., Bremers, M.A., Blanc, S., et al., 2013. Minimal intensity physical activity (Standing and walking) of longer duration improves insulin action and plasma lipids more than shorter periods of moderate to vigorous exercise (Cycling) in sedentary subjects when energy expenditure is comparable. PLoS One 8 (2), e55542, http://dx.doi.org/10.1371/journal.pone.0055542.

Elkan, A.-C., Hakansson, N., Frostegard, J., Hafstrom, I., 2011. Low level of physical activity in women with rheumatoid arthritis is associated with cardiovascular risk factors but not with body fat mass – a cross sectional study. BMC Musculoskelet Disord. 12 (1), 13, http://dx.doi.org/10.1186/1471-2474-12-13.

Ewald, B., McEvoy, M., Attia, J., 2010. Pedometer counts superior to physical activity scale for identifying health markers in older adults. Br. J. Sports Med. 44 (10), 756–761, http://dx.doi.org/10.1136/bjsm.2008.048827.

Falconer, C.L., Cooper, A.R., Walhin, J.P., et al., 2014. Sedentary time and markers of inflammation in people with newly diagnosed type 2 diabetes. Nutr. Metab. Cardiovasc. Dis. 24 (9), 956–962, http://dx.doi.org/10.1016/j.numecd.2014.03.009.

Figueira, F.R., Umpierre, D., Cureau, F.V., et al., 2014. Association between physical activity advice only or structured exercise training with blood pressure levels in patients with type 2 diabetes: a systematic review and meta-analysis. Sports Med. 44 (11), 1557–1572, http://dx.doi.org/10.1007/s40279-014-0226-2.

Franceschi, C., Campisi, J., 2014. Chronic inflammation (inflammaging) and its potential contribution to age-associated diseases. J. Gerontol. A. Biol. Sci. Med. Sci. 69 (January (Suppl. 1)), S4–9, http://dx.doi.org/10.1093/gerona/glu057, Review. PubMed PMID: 24833586.

Fung, T.T., 2000. Leisure-Time Physical Activity, Television Watching, and Plasma Biomarkers of Obesity and Cardiovascular Disease Risk. Am. J. Epidemiol. 152 (12), 1171–1178, http://dx.doi.org/10.1093/aje/152.12.1171.

Gába, A., Kapuš, O., Pelclová, J., Riegerová, J., 2012. The relationship between accelerometer-determined physical activity (PA) and body composition and bone mineral density (BMD) in postmenopausal women. Arch. Gerontol. Geriatr. 54 (3), e315–e321, http://dx.doi.org/10.1016/j.archger.2012.02.001.

Gabriel, K.P., Matthews, K.A., Pérez, A., et al., 2012. Self-reported and accelerometer-derived physical activity levels and coronary artery calcification progression in older women: results from the Healthy Women Study. Menopause, 1, http://dx.doi.org/10.1097/gme.0b013e31826115af.

Gao, X., Nelson, M.E., Tucker, K.L., 2007. Television viewing is associated with prevalence of metabolic syndrome in hispanic elders. Diabetes Care 30 (3), 694–700, http://dx.doi.org/10.2337/dc06-1835.

Gardiner, P.A., Healy, G.N., Eakin, E.G., et al., 2011. Associations between television viewing time and overall sitting time with the metabolic syndrome in older men and women: the australian diabetes obesity and lifestyle study: sedentary behavior and metabolic syndrome. J. Am. Geriatr. Soc. 59 (5), 788–796, http://dx.doi.org/10.1111/j.1532-5415.2011.03390.x.

Gennuso, K.P., Gangnon, R.E., Matthews, C.E., Thraen-Borowski, K.M., Colbert, L.H., 2013. Sedentary behavior, physical activity, and markers of health in older adults. Med. Sci. Sports Exerc. 45 (8), 1493–1500, http://dx.doi.org/10.1249/MSS.0b013e318288a1e5.

Hamer, M., Poole, L., Messerli-Bürgy, N., 2013. Television viewing, C-reactive protein, and depressive symptoms in older adults. Brain Behav. Immun. 33, 29–32, http://dx.doi.org/10.1016/j.bbi.2013.05.001.
Hamilton, M.T., Hamilton, D.G., Zderic, T.W., 2004. Exercise physiology versus inactivity physiology: an essential concept for understanding lipoprotein lipase regulation. Exerc. Sport Sci. Rev. 32 (4), 161–166.
Hamilton, M.T., Hamilton, D.G., Zderic, T.W., 2007. Role of low energy expenditure and sitting in obesity, metabolic syndrome, type 2 diabetes, and cardiovascular disease. Diabetes 56 (11), 2655–2667, http://dx.doi.org/10.2337/db07-0882.
Harvey, J., Chastin, S., Skelton, D., 2013. Prevalence of sedentary behavior in older adults: a systematic review. Int. J. Environ. Res. Public Health 10 (12), 6645–6661, http://dx.doi.org/10.3390/ijerph10126645.
Healy, G.N., Dunstan, D.W., Salmon, J., et al., 2008a. Breaks in sedentary time: beneficial associations with metabolic risk. Diabetes Care 31 (4), 661–666, http://dx.doi.org/10.2337/dc07-2046.
Healy, G.N., Wijndaele, K., Dunstan, D.W., et al., 2008b. Objectively measured sedentary time, physical activity, and metabolic risk: the australian diabetes, obesity and lifestyle study (AusDiab). Diabetes Care 31 (2), 369–371, http://dx.doi.org/10.2337/dc07-1795.
Henson, J., Yates, T., Edwardson, C.L., et al., 2013. Sedentary time and markers of chronic low-grade inflammation in a high risk population. PLoS One 8 (10), e78350, http://dx.doi.org/10.1371/journal.pone.0078350.
Inoue, S., Sugiyama, T., Takamiya, T., Oka, K., Owen, N., Shimomitsu, T., 2012. Television viewing time is associated with overweight/obesity among older adults, independent of meeting physical activity and health guidelines. J. Epidemiol. 22 (1), 50–56, http://dx.doi.org/10.2188/jea.JE20110054.
Ivuškāns, A., Mäestu, J., Jürimäe, T., et al., 2015. Sedentary time has a negative influence on bone mineral parameters in peripubertal boys: a 1-year prospective study. J. Bone Miner. Metab. 33 (1), 85–92, http://dx.doi.org/10.1007/s00774-013-0556-4.
Jakes, R.W., Day, N.E., Khaw, K.-T., et al., 2003. Television viewing and low participation in vigorous recreation are independently associated with obesity and markers of cardiovascular disease risk: EPIC-Norfolk population-based study. Eur. J. Clin. Nutr. 57 (9), 1089–1096, http://dx.doi.org/10.1038/sj.ejcn.1601648.
Jarvie, J.L., Whooley, M.A., Regan, M.C., Sin, N.L., Cohen, B.E., 2014. Effect of physical activity level on biomarkers of inflammation and insulin resistance over 5 years in outpatients with Coronary heart disease (from the heart and soul study). Am. J. Cardiol. 114 (8), 1192–1197, http://dx.doi.org/10.1016/j.amjcard.2014.07.036.
Jefferis, B.J., Whincup, P.H., Lennon, L.T., Papacosta, O., Goya Wannamethee, S., 2014. Physical activity in older men: longitudinal associations with inflammatory and hemostatic biomarkers, N-terminal pro-brain natriuretic peptide, and onset of coronary heart disease and mortality. J. Am. Geriatr. Soc. 62 (4), 599–606, http://dx.doi.org/10.1111/jgs.12748.
Kaino, W., Daimon, M., Sasaki, S., et al., 2013. Lower physical activity is a risk factor for a clustering of metabolic risk factors in non-obese and obese Japanese subjects: the Takahata study. Endocr. J. 60 (5), 617–628, http://dx.doi.org/10.1507/endocrj.EJ12-0351.
Kallings, L.V., Johnson, J.S., Fisher, R.M., et al., 2009. Beneficial effects of individualized physical activity on prescription on body composition and cardiometabolic risk factors: results from a randomized controlled trial. Eur. J. Cardiovasc. Prev. Rehabil. 16 (1), 80–84, http://dx.doi.org/10.1097/HJR.0b013e32831e953a.
Kirk, A., Barnett, J., Leese, G., Mutrie, N., 2009. A randomized trial investigating the 12-month changes in physical activity and health outcomes following a physical activity consultation delivered by a person or in written form in Type 2 diabetes: Time2Act. Diabet. Med. 263, 293–301, http://dx.doi.org/10.1111/j.1464-5491.2009.02675.x.
Klenk, J., Denkinger, M., Nikolaus, T., et al., 2013. Association of objectively measured physical activity with established and novel cardiovascular biomarkers in elderly subjects: every step counts. J. Epidemiol. Community Health 67 (2), 194–197, http://dx.doi.org/10.1136/jech-2012-201312.
Knight, S., Bermingham, M.A., Mahajan, D., 1999. Regular non-vigorous physical activity and cholesterol levels in the elderly. Gerontology 45 (4), 213–219, http://dx.doi.org/10.1159/000022090.
Knight, E., Stuckey, M.I., Petrella, R.J., 2014. Prescribing physical activity through primary care: does activity intensity matter? Phys. Sportsmed. 42 (3), 78–79, http://dx.doi.org/10.3810/psm.2014.09.2079.
Krogh-Madsen, R., Thyfault, J.P., Broholm, C., et al., 2010. A 2-wk reduction of ambulatory activity attenuates peripheral insulin sensitivity. J. Appl. Physiol. (Bethesda, MD, 1985) 108 (5), 1034–1040, http://dx.doi.org/10.1152/japplphysiol.00977.2009.
Larsen, B.A., Allison, M.A., Kang, E., et al., 2014a. Associations of physical activity and sedentary behavior with regional fat deposition. Med. Sci. Sports Exerc. 46 (3), 520–528, http://dx.doi.org/10.1249/MSS.0b013e3182a77220.
Larsen, R.N., Kingwell, B.A., Sethi, P., Cerin, E., Owen, N., Dunstan, D.W., 2014b. Breaking up prolonged sitting reduces resting blood pressure in overweight/obese adults. Nutr. Metab. Cardiovasc. Dis. 24 (9), 976–982, http://dx.doi.org/10.1016/j.numecd.2014.04.011.
Lee, J., Chang, R.W., Ehrlich-Jones, L., et al., 2015. Sedentary behavior and physical function: objective evidence from the osteoarthritis initiative: evidence on sedentary behavior from the OAI. Arthritis Care Res. 67 (3), 366–373, http://dx.doi.org/10.1002/acr.22432.
Li, C., Aronsson, C.A., Hedblad, B., Gullberg, B., Wirfält, E., Berglund, G., 2009. Ability of physical activity measurements to assess health-related risks. Eur. J. Clin. Nutr. 63 (12), 1448–1451, http://dx.doi.org/10.1038/ejcn.2009.69.
Loyen, A., van der Ploeg, H.P., Bauman, A., Brug, J., Lakerveld, J., 2016. European sitting championship: prevalence and correlates of self-reported sitting time in the 28 european union member states. PLoS One 11 (3), e0149320, http://dx.doi.org/10.1371/journal.pone.0149320, In: Buchowski, M. (ed.).
Lynch, B.M., Dunstan, D.W., Healy, G.N., Winkler, E., Eakin, E., Owen, N., 2010. Objectively measured physical activity and sedentary time of breast cancer survivors, and associations with adiposity: findings from NHANES (2003–2006). Cancer Causes Control 21 (2), 283–288, http://dx.doi.org/10.1007/s10552-009-9460-6.
Lynch, B.M., Friedenreich, C.M., Winkler, E.A.H., et al., 2011. Associations of objectively assessed physical activity and sedentary time with biomarkers of breast cancer risk in postmenopausal women: findings from NHANES (2003–2006). Breast Cancer Res. Treat. 130 (1), 183–194, http://dx.doi.org/10.1007/s10549-011-1559-2.
Marsh, S., Foley, L.S., Wilks, D.C., Maddison, R., 2014. Family-based interventions for reducing sedentary time in youth: a systematic review of randomized controlled trials: family-based sedentary time interventions. Obes. Rev. 15 (2), 117–133, http://dx.doi.org/10.1111/obr.12105.
McAuley, E., Blissmer, B., Marquez, D.X., Jerome, G.J., Kramer, A.F., Katula, J., 2000. Social relations, physical activity, and well-being in older adults. Prev. Med. 31 (5), 608–617, http://dx.doi.org/10.1006/pmed.2000.0740.
Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J. Clin. Epidemiol. 62 (10), 1006–1012, http://dx.doi.org/10.1016/j.jclinepi.2009.06.005.
Mohri, M., Motohama, R., Sato, N., 2013. Home-based cardiac rehabilitation decreases red cell distribution width in chronic heart failure. Acta Cardiol., 615–619, http://dx.doi.org/10.2143/AC.68.6.8000009.
Nelson, M.E., Rejeski, W.J., Blair, S.N., et al., 2007. Physical activity and public health in older adults: recommendation from the american college of sports medicine and the american heart association. Circulation 116 (9), 1094–1105, http://dx.doi.org/10.1161/CIRCULATIONAHA.107.185650.
Okonkwo, O.C., Schultz, S.A., Oh, J.M., et al., 2014. Physical activity attenuates age-related biomarker alterations in preclinical AD. Neurology 83 (19), 1753–1760, http://dx.doi.org/10.1212/WNL.0000000000000964.
Owen, N., Healy, G.N., Matthews, C.E., Dunstan, D.W., 2010. Too much sitting: the population health science of sedentary behavior. Exerc. Sport Sci. Rev. 38 (3), 105–113, http://dx.doi.org/10.1097/JES.0b013e3181e373a2.
ParticipACTION – It's time for Canada to sit less and move more.
Peddie, M.C., Bone, J.L., Rehrer, N.J., Skeaff, C.M., Gray, A.R., Perry, T.L., 2013. Breaking prolonged sitting reduces postprandial glycemia in healthy, normal-weight adults: a randomized crossover trial. Am. J. Clin. Nutr. 98 (2), 358–366, http://dx.doi.org/10.3945/ajcn.112.051763.
Prioreschi, A., Makda, M.A., Tikly, M., McVeigh, J.A., 2015. Habitual physical activity, sedentary behaviour and bone health in rheumatoid arthritis. Int. J. Sports Med. 36 (12), 1021–1026, http://dx.doi.org/10.1055/s-0035-1550049.
Reaven, P.D., Barrett-Connor, E., Edelstein, S., 1991. Relation between leisure-time physical activity and blood pressure in older women. Circulation 83 (2), 559–565, http://dx.doi.org/10.1161/01.CIR.83.2.559.
Sedentary Behaviour Research Network (SBRN), http://www.sedentarybehaviour.org/.
SITLESS. http://sitless.eu/.
Santos, D.A., Silva, A.M., Baptista, F., et al., 2012. Sedentary behavior and physical activity are independently related to functional fitness in older adults. Exp. Gerontol. 47 (12), 908–912, http://dx.doi.org/10.1016/j.exger.2012.07.011.
Sardinha, L.B., Santos, D.A., Silva, A.M., Baptista, F., Owen, N., 2015. Breaking-up sedentary time is associated with physical function in older adults. J. Gerontol. A. Biol. Sci. Med. Sci. 70 (1), 119–124, http://dx.doi.org/10.1093/gerona/glu193.
Scott, D., Ebeling, P.R., Sanders, K.M., Aitken, D., Winzenberg, T., Jones, G., 2015. Vitamin D and physical activity status: associations with five-year changes in body composition and muscle function in community-dwelling older adults. J. Clin. Endocrinol. Metab. 100 (2), 670–678, http://dx.doi.org/10.1210/jc.2014-3519.
Sedentary Behaviour Research Network, 2012. Letter to the Editor: standardized use of the terms sedentary and sedentary behaviours. Appl. Physiol. Nutr. Metab. 37 (3), 540–542, http://dx.doi.org/10.1139/h2012-024.
Stamatakis, E., Davis, M., Stathi, A., Hamer, M., 2012. Associations between multiple indicators of objectively-measured and self-reported sedentary behaviour and cardiometabolic risk in older adults. Prev. Med. 54 (1), 82–87, http://dx.doi.org/10.1016/j.ypmed.2011.10.009.
Stubbs, B., Brefka, S., Wirth, K., et al., 2015. Aging-related biomarkers associated with sedentary behaviour in older adults: a systematic review (and meta-analysis). BMC Public Health (Accessed 26 April 2016) http://dx.doi.org/10.15124/CRD42015023731.
Suboc, T.B., Strath, S.J., Dharmashankar, K., et al., 2014. Relative importance of step count, intensity, and duration on physical activity's impact on vascular structure and function in previously sedentary older adults. J. Am. Heart Assoc. 3 (1), e000702, http://dx.doi.org/10.1161/JAHA.113.000702.
Ward, D.S., Evenson, K.R., Vaughn, A., Rodgers, A.B., Troiano, R.P., 2005. Accelerometer use in physical activity: best practices and research recommendations. Med. Sci. Sports Exerc. 37 (Supplement), S582–S588, http://dx.doi.org/10.1249/01.mss.0000185292.71933.91.

Wijndaele, K., Lynch, B.M., Owen, N., Dunstan, D.W., Sharp, S., Aitken, J.F., 2009. Television viewing time and weight gain in colorectal cancer survivors: a prospective population-based study. Cancer Causes Control 20 (8), 1355–1362, http://dx.doi.org/10.1007/s10552-009-9356-5.
Wullems, J.A., Verschueren, S.M.P., Degens, H., Morse, C.I., Onambélé, G.L., 2016. A review of the assessment and prevalence of sedentariness in older adults, its physiology/health impact and non-exercise mobility counter-measures. Biogerontology, http://dx.doi.org/10.1007/s10522-016-9640-1.
Yates, T., Khunti, K., Wilmot, E.G., et al., 2012. Self-reported sitting time and markers of inflammation, insulin resistance, and adiposity. Am. J. Prev. Med. 42 (1), 1–7, http://dx.doi.org/10.1016/j.amepre.2011.09.022.
Zhou, Z., Xi, Y., Zhang, F., et al., 2016. Sedentary behavior predicts changes in cardiometabolic risk in professional workers: a one-year prospective study. J. Occup. Environ. Med. 58 (4), e117–e123, http://dx.doi.org/10.1097/JOM.0000000000000673.

Anexo 3. Escala AMSTAR-2: herramienta de evaluación crítica de revisiones sistemáticas de estudios de intervenciones de salud.

AMSTAR-2 item
1. ¿Las preguntas de investigación y los criterios de inclusión para la revisión incluyen los componentes PICO?
2. ¿El reporte de la revisión contiene una declaración explícita de que los métodos de la revisión fueron establecidos con anterioridad a su realización y justifica cualquier desviación significativa del protocolo?
3. ¿Los autores de la revisión explicaron su decisión sobre los diseños de estudio a incluir en la revisión?
4. ¿Los autores de la revisión usaron una estrategia de búsqueda bibliográfica exhaustiva?
5. ¿Los autores de la revisión realizaron la selección de estudios por duplicado?
6. ¿Los autores de la revisión realizaron la extracción de datos por duplicado?
7. ¿Los autores de la revisión proporcionaron una lista de estudios excluidos y justificaron las exclusiones?
8. ¿Los autores de la revisión describieron los estudios incluidos con suficiente detalle?
9. ¿Los autores de la revisión usaron una técnica satisfactoria para evaluar el riesgo de sesgo de los estudios individuales incluidos en la revisión?
10. ¿Los autores de la revisión reportaron las fuentes de financiación de los estudios incluidos en la revisión?
11. Si se realizó un metanálisis, ¿los autores de la revisión usaron métodos apropiados para la combinación estadística de resultados?
12. Si se realizó un metanálisis, ¿los autores de la revisión evaluaron el impacto potencial del riesgo de sesgo en estudios individuales sobre los resultados del meta-análisis u otra síntesis de evidencia?
13. ¿Los autores de la revisión consideraron el riesgo de sesgo de los estudios individuales al interpretar / discutir los resultados de la revisión?
14. ¿Los autores de la revisión proporcionaron una explicación satisfactoria y discutieron cualquier heterogeneidad observada en los resultados de la revisión?
15. Si se realizó síntesis cuantitativa ¿los autores de la revisión llevaron a cabo una adecuada investigación del sesgo de publicación (sesgo de estudio pequeño) y discutieron su probable impacto en los resultados de la revisión?
16. ¿Los autores de la revisión informaron de cualquier fuente potencial de conflicto de intereses, incluyendo cualquier financiamiento recibido para llevar a cabo la revisión?

Versión en español por Ciapponi y colaboradores [87]