CAPÍTULO 14

Lectura crítica de revisiones sistemáticas de pruebas diagnósticas

M.ª Nieves Plana Farrás ■ Javier Zamora Romero

OBJETIVOS DEL CAPÍTULO

- Valorar la pregunta de investigación que se aborda en una revisión sistemática de exactitud diagnóstica.
- Valorar críticamente la metodología empleada en estas revisiones sistemáticas.
- Entender las distintas alternativas de análisis estadístico de estas revisiones sistemáticas.
- Valorar los resultados de estos metaanálisis y su alcance para aplicarlos en un caso concreto.

Introducción

La utilidad de una prueba diagnóstica se fundamenta, entre otros aspectos, en su validez o exactitud, es decir, en su capacidad para discriminar, en los pacientes con sospecha de una determinada condición clínica, entre aquellos que realmente presentan dicha condición y los que no. Como ya se ha visto en el capítulo 6 de esta obra, las evidencias acerca de esta exactitud diagnóstica provienen de estudios con un diseño de investigación clínica particular y un análisis estadístico de los resultados que proporciona los conocidos índices de exactitud diagnóstica de sensibilidad, especificidad, valores predictivos y los cocientes de probabilidad junto con sus intervalos de confianza. Para poder abordar la lectura de este capítulo, recomendamos la lectura previa del mencionado capítulo de estudios primarios de exactitud diagnóstica, al igual que el capítulo relativo a las revisiones sistemáticas en general y de ensayos clínicos aleatorios en particular (v. capítulos 11 y 12, respectivamente). La lectura crítica de revisiones de diagnóstico comparte muchos aspectos con lo referido en esos capítulos previos.

Las revisiones sistemáticas de exactitud diagnóstica se conciben como un instrumento para integrar toda la evidencia disponible sobre la exactitud de una prueba diagnóstica. Los métodos para realizar estas revisiones paralelizan los métodos empleados para realizar sus homónimas revisiones de eficacia de intervenciones. Como lectores, deberíamos prestar atención a cómo los autores de la revisión han realizado los procesos de búsqueda de literatura, la selección de estudios para la revisión, la extracción de datos y el análisis de la calidad metodológica y el riesgo de sesgo de los artículos incluidos. El análisis crítico de estos aspectos determinará la validez de la revisión sistemática que estemos leyendo.

El presente capítulo pretende servir de guía para la realización de una lectura crítica de este tipo de artículos. Abordaremos inicialmente los aspectos que permiten evaluar cuán válidos son los resultados de la revisión. En segundo lugar, valoraremos críticamente la metodología estadística empleada para hacer el metaanálisis e interpretaremos los resultados. Por último, valoraremos

la aplicabilidad o validez externa de los resultados de la revisión. Para ilustrar estos apartados se utilizará como ejemplo una revisión publicada recientemente en el ámbito de la salud mental (1).

Escenario

Estás discutiendo con tus compañeros del centro de salud acerca de los múltiples cuestionarios disponibles de cribado cognitivo para el diagnóstico precoz de demencia. Os estáis planteando poner en marcha una estrategia de despistaje de demencia entre los pacientes de edad avanzada asintomáticos que tenéis en cupo. Hay cierta controversia entre los profesionales sobre la conveniencia o no de establecer una estrategia generalizada o ya dirigida a pacientes con quejas cognitivas y de cuál sería la mejor herramienta para identificar deterioro cognitivo de manera temprana desde atención primaria. Aunque algunos estáis familiarizados con el Mini-Mental State Examination (MMSE) y su versión española normalizada, el Mini-Examen Cognoscitivo (MEC), pensáis que sería bueno revisar en la literatura qué otros instrumentos de cribado podríais utilizar en el contexto de atención primaria y cuáles tienen mejor rendimiento para detectar deterioro cognitivo en fases tempranas.

En la búsqueda encuentras la siguiente revisión: Tsoi KKF, Chan JYC, Hirai HW, Wong SYS, Kwok TCY. Cognitive Tests to Detect Dementia: A Systematic Review and Meta-analysis. JAMA Intern Med. 2015;175(9):1450-1458 (1).

Te planteas las siguientes preguntas:

- ¿Cuál es el rendimiento diagnóstico del Mini-Examen Cognoscitivo (MEC) en población asintomática?
- ¿Aplicaría en mi medio el MEC para el cribado de deterioro cognitivo leve?

Puntos clave de la lectura crítica

DEFINICIÓN DE LA PREGUNTA DE LA REVISIÓN

¿Se hizo la revisión sobre un tema claramente definido?

El primer aspecto relevante que se debe considerar en la lectura de una revisión, y quizás el aspecto más crucial, es determinar qué pregunta de exactitud diagnóstica trata de responder la revisión sistemática que se tiene entre manos. Tener claro este punto es vital para poder valorar adecuadamente el resto de los aspectos de la revisión. Sobre una definición adecuada de la pregunta de la revisión pivotan el resto de los aspectos que iremos desgranando en este capítulo. Es importante notar que el rendimiento de una prueba diagnóstica depende no solo de sus capacidades técnicas, sino que también viene condicionado por las características de la población, el ámbito en el que se aplica la prueba y la estrategia diagnóstica previa que ha ido seleccionando la población de los distintos estudios incluidos en la revisión. Ya estamos familiarizados con la metodología de formulación de preguntas estructuradas. En el caso de las revisiones de diagnóstico, los componentes genéricos del acrónimo inglés PICO se han reformulado para adaptarse a una pregunta sobre la exactitud de una prueba diagnóstica *(P: population; I: index test; CO: comparator test; T: target condition)* (2). Se recomienda incluir además dos componentes adicionales *(P: prior testing; P: purpose)* para tener un acrónimo algo más complejo, pero más rico en detalles (PPPICOT). Con estas adiciones se pretende delimitar bien la población de estudio (el espectro de la enfermedad, nivel de riesgo y su severidad, etc.) y además delimitar el rol de la prueba en la ruta diagnóstica donde se empleará. Así se podrá distinguir si la prueba que se evalúa pretende sustituir a otra, o pretende añadirse al proceso diagnóstico como paso previo a otras pruebas más agresivas o costosas *(triage)* o como un paso posterior tras otras pruebas *(add-on)* (3).

Por último, la revisión debe especificar qué tipo de estudios va a incluir para responder a la pregunta planteada. Este aspecto se puede abordar vigilando que la revisión sistemática haya excluido diseños de estudio no adecuados para contestar a una pregunta sobre rendimiento diagnóstico. Ya sabemos que el mejor diseño de estudio para evaluar la exactitud de una prueba diagnóstica

es un estudio observacional con diseño transversal donde, a una serie consecutiva de pacientes con sospecha de la enfermedad que se diagnostica, de forma ciega e independiente, se les aplica la prueba que se evalúa y una prueba de referencia o patrón de oro *(gold standard)* que diferencia indiscutiblemente entre aquellos que tienen la enfermedad y los que no. A pesar de la aceptación unánime de este diseño transversal como el óptimo para esta tarea, es frecuente encontrar en la literatura estudios de casos y controles para responder a preguntas de exactitud diagnóstica. En ellos se selecciona un grupo de sujetos con la enfermedad objeto de estudio y otro grupo de controles sanos y a ambos se les aplica la prueba evaluada. Se ha demostrado empíricamente que este diseño de casos y controles sobreestima enormemente el rendimiento diagnóstico de la prueba que se evalúa y por tanto es práctica común, o debería serlo, que los autores de las revisiones sistemáticas excluyan este tipo de estudios de sus revisiones (4,5).

Determinar con precisión la pregunta de investigación que se trata de contestar en la revisión sistemática que estamos leyendo es crucial para valorarla críticamente y para determinar la aplicabilidad de sus resultados al escenario clínico en cuestión.

IDENTIFICACIÓN DE EVIDENCIAS: ESTRATEGIAS DE BÚSQUEDA

¿Crees que estaban incluidos los estudios importantes y pertinentes?

Al igual que en sus homólogas revisiones de ensayos clínicos, la exhaustividad de la búsqueda es el pilar básico del valor de una revisión sistemática de pruebas diagnósticas. Esta búsqueda debe ser exhaustiva para reducir el riesgo de sesgo de publicación e identificar la totalidad de estudios relevantes. Es obligado que los autores de la revisión describan las fuentes de información en las que realizaron la búsqueda y las plataformas de acceso empleadas, incluyendo tanto la estrategia de búsqueda como las fechas en las que se realizó. Con toda esta información es posible valorar el esfuerzo realizado y el riesgo de que haya estudios relevantes que no se encuentren en los resultados de la búsqueda.

Como en cualquier revisión sistemática, debe valorarse si la búsqueda de estudios primarios ha sido objetiva y si es reproducible. Una búsqueda solo en Medline se considera inadecuada. Los estándares metodológicos de la colaboración internacional Cochrane (www.editorial-unit.cochrane.org/mecir) establecen como obligatorio realizar una búsqueda al menos en Medline y Embase, y es altamente recomendable buscar en otras bases de datos al igual que realizar búsquedas manuales en las listas de referencias bibliográficas de los artículos incluidos, en los resúmenes de congresos relevantes (BIOSIS Database [http://www.biosis.org/], Mediconf [www.mediconf.com]), hacer consultas con investigadores destacados, etc.

Existen bases de datos específicas de estudios de diagnóstico como ARIF (www.arif.bham.ac.uk), HTA Database y DARE (www.york.ac.uk), que recogen un buen número de revisiones publicadas de estudios de diagnóstico y de cribado. Existen bases de datos de temas específicos (CINAHL de enfermería, BIOSIS de biología, PsycINFO, etc.) que pueden utilizarse también como fuentes de estudios primarios.

La identificación de artículos de diagnóstico presenta más dificultades que la búsqueda de ensayos clínicos. No existe como tal un término MeSH (Medical Subject Heading) o EMTREE (el equivalente a MeSH en Embase) específico que sea comparable al término *randomized controlled trial.* El término *sensitivity and specificity* podría ser el más adecuado, pero no en todas las bases de datos los artículos de diagnóstico están bien indexados. Muchos de los estudios de diagnóstico se realizan alrededor de la propia práctica clínica sin la existencia de un protocolo de investigación registrado en bases como clinicaltrials.gov, con lo que se dificulta su seguimiento. No existe, por el momento, una base de datos centralizada de estudios de diagnóstico equivalente a la de ensayos clínicos, aunque Cochrane está ultimando un registro similar al registro de ensayos clínicos (CENTRAL) para estudios de validez diagnóstica (CRDTAS). Hay que prestar atención al hecho de que el uso de filtros metodológicos para restringir y focalizar la búsqueda no está recomendado (6).

Esto se traduce en que normalmente la cantidad de títulos y resúmenes localizados en las búsquedas de las revisiones de diagnóstico sobrepasan con creces a los encontrados en las búsquedas de las revisiones de ensayos clínicos, dado que estas últimas son más fácil de enfocar empleando filtros metodológicos que restringen enormemente las búsquedas.

La estructura de una estrategia de búsqueda debe incluir términos referentes a la condición clínica que se trata de diagnosticar y términos para identificar la prueba diagnóstica que se evalúa. Esta estructura básica puede adaptarse y hacerse más específica incluyendo la prueba de referencia definida en la pregunta estructurada.

Los criterios de elegibilidad de estudios que se van a incluir en la revisión deben haber sido bien descritos en el artículo y se debe comprobar que efectivamente se ajustan a lo planteado en la pregunta estructurada (PPPICOT). Dado que el proceso de selección de estudios tiene una elevada carga de subjetividad, para evitar resultados sesgados es importante que los autores hayan realizado el proceso por duplicado (por parejas de revisores), haber determinado un método para resolver discordancias e idealmente haber valorado la reproducibilidad del proceso mediante un análisis de concordancia.

Finalmente, el flujo de estudios desde la búsqueda inicial hasta la realización del metaanálisis debería presentarse siguiendo las recomendaciones de las guías PRISMA (7); es decir, mediante una figura o diagrama de flujo donde consten los estudios en las distintas fases del proceso, con las exclusiones y los motivos de exclusión bien descritos (fig. 14.1). Este diagrama nos permite valorar todo el proceso realizado y entender bien qué estudios han sido finalmente analizados.

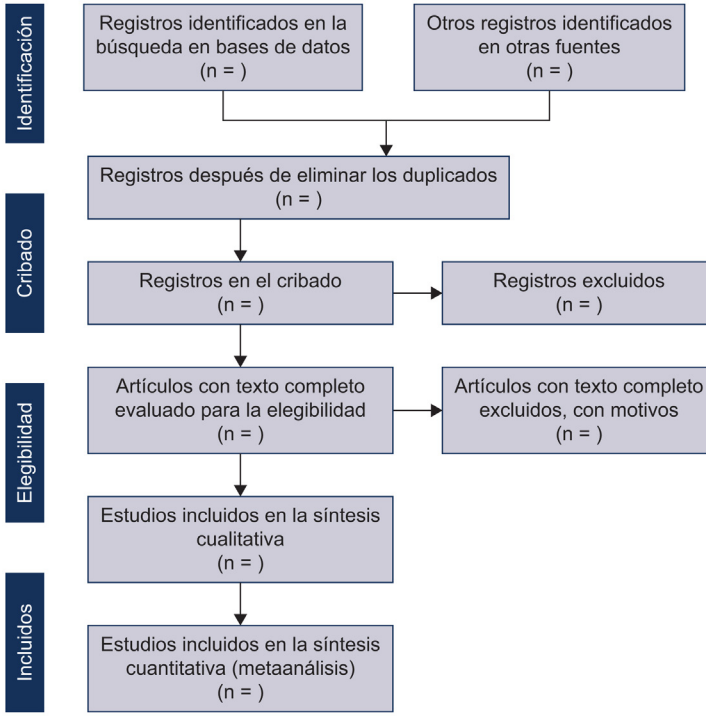

Figura 14.1 Diagrama de flujo propuesto por la declaración PRISMA para ilustrar el proceso de búsqueda y selección de estudios para la revisión. (Tomado de Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred Reporting Items for Systematic Reviews and MetaAnalyses: The PRISMA Statement. PLoS Med. 2009;6[7]:e1000097. Para más información visitar www.prisma-statement.org.)

EVALUACIÓN DE LA CALIDAD

¿Crees que los autores de la revisión han hecho suficiente esfuerzo para valorar la calidad de los estudios incluidos?

Un aspecto clave en toda revisión sistemática es la evaluación de la calidad metodológica de los estudios incluidos con la finalidad de identificar posibles riesgos de sesgo. El cuestionario QUADAS-2 es una herramienta diseñada específicamente para la evaluación de la calidad metodológica y el riesgo de sesgo de los estudios primarios incluidos en una revisión sistemática de pruebas diagnósticas (8). El cuestionario se organiza en cuatro dominios: 1) selección de pacientes; 2) prueba en evaluación; 3) prueba de referencia, y 4) flujo de pacientes y tiempo entre pruebas. Cada dominio se evalúa en términos del eventual riesgo de sesgo en el que se pudiera incurrir y adicionalmente se evalúan problemas de aplicabilidad a la pregunta de revisión de acuerdo con esos dominios. Este aspecto de aplicabilidad lo retomaremos más adelante en el último apartado de este capítulo. Es importante atender a cómo los autores han personalizado y señalizado las rúbricas específicas para cada dominio de la herramienta QUADAS-2 para así adaptarlo a las peculiaridades de la revisión.

Este ejercicio de evaluación del riesgo de sesgo no es un mero ejercicio intelectual que conduce a una gráfica más en el artículo para la descripción de esta calidad metodológica de los estudios incluidos en la revisión (fig. 14.2). Los resultados de este análisis deben influir tanto en la realización del metaanálisis como en la interpretación de los resultados del mismo. Para lo primero, los autores pueden haber hecho análisis de sensibilidad comparando los resultados que se obtienen al excluir determinados estudios en función de su riesgo de sesgo en determinados dominios del QUADAS-2 que se entienden críticos para la revisión en cuestión. Para lo segundo, es preciso entender los resultados del metaanálisis en el contexto de la calidad de evidencia disponible y discutir bajo esta óptica sus limitaciones y fortalezas.

EVALUACIÓN DE LA HETEROGENEIDAD

Si los resultados de los diferentes estudios han sido mezclados para obtener un resultado «combinado», ¿era razonable hacer eso?

Una limitación inherente a cualquier metaanálisis es la presencia de variabilidad (heterogeneidad) entre los resultados de los estudios primarios incluidos en la revisión. La heterogeneidad en las revisiones de pruebas diagnósticas suele ser mayor que la encontrada en las revisiones de eficacia de intervenciones y, también con frecuencia, mayor de lo que sería esperable por azar (variabilidad aleatoria). Las fuentes habituales de heterogeneidad son las mismas que en otras revisiones sistemáticas. La heterogeneidad puede venir de los diferentes métodos empleados en los estudios primarios (heterogeneidad metodológica) o de las diferencias en las poblaciones de pacientes o los ámbitos donde se han realizado los estudios (heterogeneidad clínica). Sin embargo, en el caso de

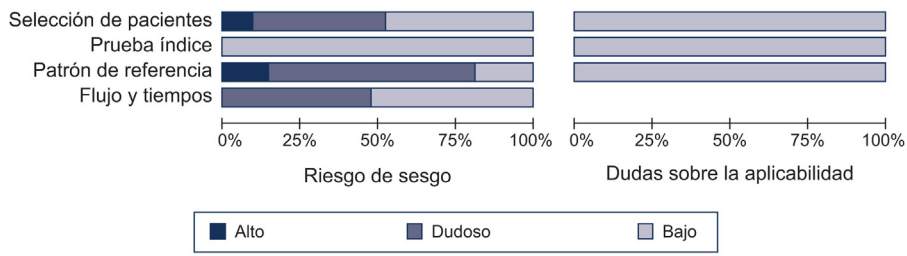

Figura 14.2 Representación gráfica del resultado de la evaluación del riesgo de sesgo de los estudios incluidos en una revisión sistemática de pruebas diagnósticas.

las revisiones de pruebas diagnósticas existe una fuente adicional e importante de heterogeneidad. Se trata de lo que se conoce como el **efecto umbral.** Esto es, el umbral para determinar cuándo un resultado de una prueba es positivo o negativo puede haber variado en los estudios incluidos en la revisión. Esta variación puede ser explícita (distintos puntos de corte para una medida cuantitativa, como puede ser el antígeno prostático) o puede ser una variabilidad en umbrales implícitos (por ejemplo, la que existiría entre radiólogos con distinto entrenamiento o experiencia y que se reflejaría en que tendrían un distinto umbral de detección de anormalidad al interpretar una prueba de imagen). Si los estudios emplean diferentes criterios de positividad, esto hace que la sensibilidad y la especificidad cambien y lo hagan en direcciones opuestas: un menor umbral puede incrementar la sensibilidad a expensas de perder especificidad. La presencia de este efecto umbral explicaría parte de las diferencias que vemos entre los resultados de los estudios individuales. Se verá más adelante que la presencia de este efecto umbral hace que el metaanálisis deba considerar simultáneamente ambos índices y deba estimar la correlación entre ellos. Además, y muy importante como veremos en el siguiente apartado, si los estudios incluidos han empleado umbrales de positividad diferentes, el análisis estadístico deberá abordarse de una determinada manera, distinta a si todos los estudios incluidos han empleado el mismo umbral de positividad.

Aparte de la heterogeneidad metodológica, la clínica y la debida al efecto umbral, existe también la heterogeneidad que viene por el propio proceso de muestreo y medición. Esta heterogeneidad viene como consecuencia de que los distintos estudios han reclutado distintos sujetos y también del propio proceso de realizar las mediciones de la prueba diagnóstica (obtención de la muestra, procesamiento, interpretación, etc.). Ambas fuentes de variabilidad se integran en lo que se conoce como heterogeneidad estadística.

Para valorar la heterogeneidad estadística se deben explorar las estimaciones de los índices de rendimiento diagnóstico de los estudios primarios incluidos (sensibilidad y la especificidad), viendo los resultados numéricos o bien evaluando su representación gráfica en forma de *forest plots* emparejados (9) (fig. 14.3). Estos *forest plots* idealmente pueden haber sido construidos presentando los estudios ordenados de mayor a menor sensibilidad (o especificidad). Esta representación ordenada puede ayudar a analizar la consistencia entre estudios y la eventual correlación entre sensibilidad y especificidad como indicio de la presencia de efecto umbral. Sin embargo, la mejor forma de ilustrar la correlación entre los índices es representar estas parejas estimadas en cada estudio individual en un plano Receiver Operating Characteristic (ROC), en el que se representa en el eje de abscisas la tasa de falsos positivos (1-especificidad) y en el eje de ordenadas la tasa de verdaderos positivos (sensibilidad). Cuando existan indicios de efecto umbral y por tanto de correlación entre sensibilidad y especificidad, esta representación gráfica resultará en un patrón característico con forma curvilínea de hombro (fig. 14.4).

Se han propuesto pruebas estadísticas específicas para poner a prueba la homogeneidad de los índices de validez diagnóstica por separado. También se ha propuesto cuantificar la proporción de variación total entre los estudios que va más allá de lo esperable por azar mediante el índice I^2 de Higgins (10). Sin embargo, estas aproximaciones deben interpretarse con suma cautela, dado que ignoran la correlación entre estos índices y están afectadas por el tamaño muestral. Las pruebas estadísticas para determinar la significación estadística de esta heterogeneidad pueden carecer de la potencia necesaria cuando en el metaanálisis se incluye un número reducido de pacientes. O de forma contraria, si los estudios han incluido tamaños muestrales muy grandes, ligeras diferencias interestudio pueden resultar en valores de estas pruebas altamente significativos. Por último, estos test no son útiles para detectar la heterogeneidad proveniente de fenómenos como los vistos de «efecto umbral».

Tan importante, o más, que identificar la heterogeneidad es la exploración de sus posibles fuentes. Esta exploración debe planificarse *a priori* antes de iniciar el análisis de los datos para evitar

Estudio	VP	FP	FN	VN	Sensibilidad (IC 95%)	Especificidad (IC 95%)
Nishimura 2016	24	0	40	156	0,38 (0,26; 0,50)	1,00 (0,98; 1,00)
Boettger 2017	43	6	49	112	0,47 (0,36; 0,57)	0,95 (0,89; 0,98)
Frenette 2016	25	15	16	42	0,61 (0,45; 0,76)	0,74 (0,60; 0,84)
Aljuaid 2018	88	4	48	74	0,65 (0,56; 0,73)	0,95 (0,87; 0,99)
Karlicic 2016	119	35	47	843	0,72 (0,64; 0,78)	0,96 (0,94; 0,97)
Gusmao-Flores 2011	29	3	11	76	0,72 (0,56; 0,85)	0,96 (0,89; 0,99)
Toro 2010	27	2	7	93	0,79 (0,62; 0,91)	0,98 (0,93; 1,00)
Luetz 2010	195	14	49	445	0,80 (0,74; 0,85)	0,97 (0,95; 0,98)
Pipanmekaporn 2014	12	1	3	54	0,80 (0,52; 0,96)	0,98 (0,90; 1,00)
Koga 2015	29	5	7	123	0,81 (0,64; 0,92)	0,96 (0,91; 0,99)
Tobar 2010	65	2	15	48	0,81 (0,71; 0,89)	0,96 (0,86; 1,00)
Vreeswijk 2010	9	0	2	18	0,82 (0,48; 0,98)	1,00 (0,81; 1,00)
Chanques 2018	34	0	7	67	0,83 (0,68; 0,93)	1,00 (0,95; 1,00)
Selim 2018	45	12	9	50	0,83 (0,71; 0,92)	0,81 (0,69; 0,90)
Adamis 2012	40	10	8	81	0,83 (0,70; 0,93)	0,89 (0,81; 0,95)
Barman 2018	238	21	42	319	0,85 (0,80; 0,89)	0,94 (0,91; 0,96)
Guenther 2010	45	0	5	58	0,90 (0,78; 0,97)	1,00 (0,94; 1,00)
Danzeng 2019	77	13	7	95	0,92 (0,84; 0,97)	0,88 (0,80; 0,93)
Wang 2013	113	14	9	116	0,93 (0,86; 0,97)	0,89 (0,83; 0,94)
Ely 2001	64	1	3	115	0,96 (0,87; 0,99)	0,99 (0,95; 1,00)

Figura 14.3 Representación gráfica en forma de *forest plot* de los resultados de sensibilidad y especificidad emparejados.

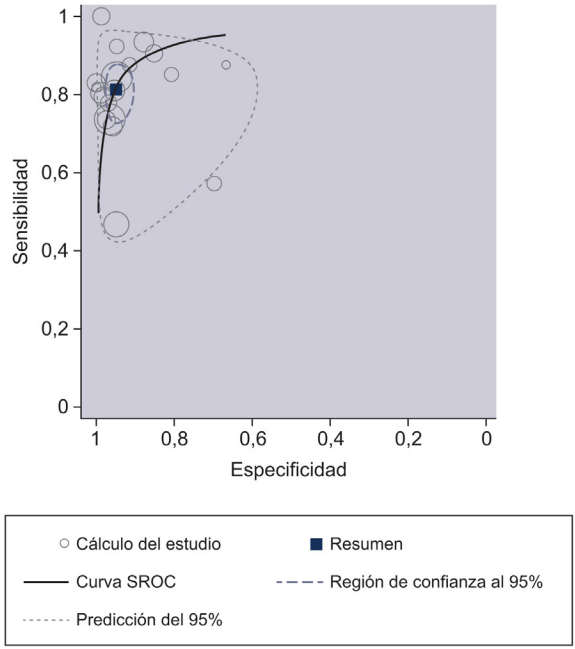

Figura 14.4 Representación en un plano ROC de los resultados de exactitud diagnóstica de los estudios individuales en una revisión sistemática. Se representa el punto promedio de sensibilidad y especificidad junto con las elipses de confianza y predicción y la curva SROC resumen.

hallazgos espurios. Se pueden hacer análisis de subgrupos o, más formalmente, emplear técnicas de metarregresión para probar si la heterogeneidad disminuye de forma significativa cuando se separan los estudios de acuerdo a características clínicas y/o metodológicas. Hay que mostrar cautela con las conclusiones que se deriven a nivel causal de estos análisis, dado que están sujetas al mismo sesgo de confusión como cualquier otro estudio observacional y están afectadas también por problemas de potencia estadística cuando el número de estudios es reducido.

METAANÁLISIS

¿Cuál es el resultado global de la revisión?, ¿cuál es la precisión del resultado?

Como siempre, es importante recordar que el metaanálisis solo debería realizarse si los estudios son clínica y metodológicamente homogéneos, si se han realizado con pacientes clínicamente similares, han evaluado pruebas comparables y han usado pruebas de referencia *(gold standard)* también comparables. Si existiera excesiva heterogeneidad entre los resultados de los estudios, podría ser más adecuado, en lugar de calcular los índices combinados, investigar las causas de dichas diferencias. En cualquier caso, la presencia de variabilidad, y concretamente del efecto umbral comentado anteriormente, va a condicionar las distintas alternativas de análisis que los autores de la revisión pudieran haber empleado.

En general, todo metaanálisis es un proceso en dos etapas. En un primer paso se estiman los resultados de cada estudio, que, en el caso de la evaluación de pruebas diagnósticas, cada estudio es resumido no por un índice, sino por la conocida pareja de índices sensibilidad y la especificidad, aunque también pueden ser los valores predictivos positivo y negativo o los cocientes de probabilidad positivo y negativo. También podemos encontrar resúmenes del rendimiento diagnóstico global de

una prueba en forma de un único índice, el *Odds Ratio* Diagnóstico (ORD) (11). Este índice es el producto de las proporciones de aciertos de la prueba (verdaderos positivos y verdaderos negativos) dividido por el producto de las proporciones de errores (falsos positivos y falsos negativos). Por lo tanto, cuanto mayor sea este índice mejor es una prueba diagnóstica. Sin embargo, su utilidad es limitada porque se pierde el carácter dual de las pruebas y no permite saber cuál es la probabilidad de tener la enfermedad si el resultado es positivo o negativo, ni determinar si la prueba es más útil para confirmar la presencia de una enfermedad o para descartarla.

Los métodos estadísticos empleados para sintetizar la revisión sistemática deben tener en cuenta esta dualidad y, por lo tanto, en lugar de combinar uno solo deben combinar ambos índices simultáneamente, teniendo en cuenta su correlación y la distinta precisión con la que se han estimado dependiendo del número de enfermos y no enfermos en cada estudio.

La vía de realización del metaanálisis depende en primera instancia de si existe heterogeneidad de umbrales de positividad entre los estudios (efecto umbral, sea este explícito o implícito). Si fuera así, el análisis estadístico debería haberse orientado en la dirección de estimar la curva ROC resumen que subyace entre los estudios incluidos (Summary Receiver Operating Characteristic [SROC] curve). En esta situación, los autores deben haber evitado la tentación de sintetizar la exactitud diagnóstica de los estudios incluidos proporcionando un valor de sensibilidad y especificidad promedio, pues estos valores no serían interpretables porque no sabríamos a qué punto de corte de la prueba diagnóstica corresponderían. Si el umbral de positividad es el mismo para todos los estudios, entonces sí tiene sentido acometer el metaanálisis con el objetivo de obtener la sensibilidad y la especificidad promedio. Empecemos por este último caso.

Estimación del rendimiento promedio de la prueba diagnóstica cuando no hay variabilidad en los umbrales de positividad, es decir, se asume que no hay efecto umbral

La primera alternativa de análisis que puede emplearse es aquella en la que los autores asumen que los estudios individuales incluidos en su revisión no muestran variabilidad en los umbrales de positividad explícitos o, si no hubiera umbrales explícitos, asumen que los umbrales implícitos son despreciables y de escaso impacto sobre el rendimiento diagnóstico. En este caso se tiene como objetivo el cálculo del punto promedio de sensibilidad y especificidad (el conocido como punto resumen o, en inglés, *summary point*) promedio para el perfil de umbrales incluidos en la revisión. A partir de esta pareja se pueden derivar los cocientes de probabilidad. La metodología para obtener estos promedios de sensibilidad y especificidad es compleja y se suele recurrir a modelos estadísticos avanzados (modelos jerárquicos de efectos aleatorios). Los más frecuentemente empleados son el modelo bivariante y el modelo Hierarchical Summary Receiver Operating Characteristic (HSROC) (12,13). Estos modelos jerárquicos permiten cuantificar además la heterogeneidad presente en los resultados y la correlación entre ambos parámetros. Es habitual representar los resultados de estos análisis mediante un plano ROC con los puntos de los estudios individuales generalmente representados con un tamaño proporcional al tamaño del estudio, junto con el punto promedio de sensibilidad y especificidad calculado, que estará rodeado por una elipse de confianza al 95%. Esta elipse de confianza representa la precisión de esta estimación. Es el intervalo de confianza bivariante que contendría el verdadero valor de sensibilidad y especificidad que se está estimando con una confianza del 95%. También se suele representar como medida de heterogeneidad una elipse de predicción al 95% alrededor del *summary point.* Esta elipse representa los posibles resultados de sensibilidad y especificidad que se podrían obtener en nuevos estudios que se realizasen (v. fig. 14.4).

El cálculo de los promedios de los valores predictivos positivo y negativo no se recomienda, pues es bien conocida su dependencia de la prevalencia de la condición que se diagnostica y, presumiblemente, esta magnitud varía de estudio a estudio, constituyéndose en una fuente extra de heterogeneidad.

Estimación del rendimiento promedio global de la prueba diagnóstica cuando hay variabilidad en los umbrales (curva SROC)

La segunda alternativa de análisis corresponde con la situación más frecuente en la que se observa una gran variabilidad entre los índices de validez diagnóstica de los estudios individuales como consecuencia del efecto de los distintos umbrales para definir los resultados positivo y negativo empleados en los distintos estudios. En esta circunstancia no es adecuado obtener un resumen de la validez diagnóstica mediante los índices de sensibilidad y especificidad porque, como decíamos anteriormente, este resumen no sería interpretable. En su lugar, el análisis estima una curva ROC resumen (*summary* ROC o SROC) que represente la relación de la validez diagnóstica con el umbral de positividad. Los modelos jerárquicos bivariante y HSROC vistos en el apartado anterior pueden servir para estimar los parámetros de esta curva SROC.

Se han propuesto distintos estadísticos útiles para resumir una curva SROC. El más habitual es el área bajo la curva (AUC), que, como el *Odds Ratio* Diagnóstico (ORD), resume el rendimiento diagnóstico de la prueba en un solo número (14): las pruebas que discriminan bien tienen un AUC mayor que las pruebas con peor rendimiento. Finalmente, puede usarse el área para comparar el rendimiento de distintas pruebas diagnósticas. Los autores también podrían haber empleado la curva SROC estimada para extrapolar una sensibilidad a partir de una especificidad dada o viceversa.

Evaluación del sesgo de publicación

La valoración del sesgo de publicación en los estudios de diagnóstico es compleja. Los mecanismos que pueden originar la presencia de un sesgo de publicación en este tipo de estudios no son claros. A diferencia de lo que ocurre con los estudios de tratamiento, no parece plausible que la magnitud estimada para la sensibilidad o la especificidad, ni su significación estadística, puedan influir en que el estudio se publique o no. En consecuencia, los gráficos de embudo o de chimenea *(funnel plots)* y demás métodos estadísticos basados en la asimetría de estas gráficas utilizados para evaluar dicho sesgo de publicación en las revisiones de tratamiento están discutidos para las revisiones de diagnóstico. El manual Cochrane para revisiones sistemáticas de diagnóstico desaconseja la utilización de los habituales test utilizados en revisiones de tratamiento y en su lugar recomienda la utilización de test alternativos (15).

Aplicabilidad de los resultados

Tras la lectura crítica de cualquier estudio de investigación, analizar la aplicabilidad de sus resultados no es otra cosa que determinar el grado en el que estos se corresponden con la pregunta que había motivado la lectura del artículo, es decir, el escenario o la situación concreta en la que queremos aplicar los resultados del estudio.

En el caso de las revisiones sistemáticas de pruebas diagnósticas, la capacidad de transferir sus resultados a un escenario u otro depende de aspectos críticos como el ámbito donde se aplicará la prueba, el espectro de los pacientes, la propia prueba que se evalúa y el patrón de oro o estándar de referencia empleado. Todas estas características afectarán a las estimaciones de exactitud diagnóstica de la prueba y por tanto deberán tenerse en cuenta las eventuales diferencias entre la revisión y el escenario concreto donde se aplicarán los resultados.

Se pueden distinguir dos situaciones diferentes a la hora de evaluar la aplicabilidad de los resultados de una revisión sistemática de pruebas diagnósticas. Estas se distinguen por la amplitud o estrechez de la pregunta de la revisión que, en última instancia, se reflejará en los criterios de inclusión de estudios en la revisión. Si la pregunta es amplia, la revisión debería haber explorado si el rendimiento de la prueba varía en distintos ámbitos o en diferentes grupos de pacientes o varía fruto de variaciones en la propia prueba evaluada. En cambio, si la revisión ha planteado una

pregunta restringida, aunque esto favorece que los resultados sean más homogéneos, la capacidad de transferencia a otros escenarios distintos es más limitada.

El cuestionario QUADAS-2, del que hemos hablado anteriormente, incluye varias preguntas concretas sobre la aplicabilidad de los resultados en tres ámbitos distintos: el ámbito de la selección de pacientes, el ámbito de la realización de la prueba diagnóstica que se evalúa y por último en el ámbito del estándar de referencia.

SELECCIÓN DE PACIENTES

Se sabe que el rendimiento diagnóstico de una prueba varía según el espectro clínico de los participantes incluidos en el estudio (16). Por tanto, se debe valorar si el espectro de participantes de la revisión es similar al de la población donde se aplicará la prueba. Pistas para analizar esta similitud se deben buscar en las características demográficas, las comorbilidades, la gravedad de la enfermedad que se pretende diagnosticar, su sintomatología y las pruebas previas realizadas en estos pacientes. Es esperable que la sensibilidad de una prueba aumente ante la presencia de poblaciones con enfermedad de mayor gravedad. Del mismo modo, la presencia de una gran variabilidad en los diagnósticos diferenciales en las poblaciones no enfermas es esperable que disminuya la especificidad de la prueba.

PRUEBA EVALUADA

También se deben valorar eventuales diferencias que pudiera haber en las pruebas utilizadas en los estudios incluidos en la revisión respecto a la prueba que se aplicará realmente en el escenario de interés. Estas diferencias pueden ser fruto de versiones diferentes de la prueba, las muestras evaluadas, los observadores y su entrenamiento, el umbral de positividad empleado, sea de forma explícita o de forma implícita. Otro aspecto fundamental para la aplicabilidad ya mencionado en el capítulo 6 es la consideración de la reproducibilidad de la prueba, su calibración y las necesidades técnicas y humanas para su correcta aplicación. Estos aspectos podrían penalizar la transferencia de los resultados de la revisión a nuestro medio.

PRUEBA DE REFERENCIA

La aplicabilidad de los resultados de la revisión a nuestro medio puede verse penalizada si la definición de enfermedad en los estudios incluidos difiere de la definición en nuestro escenario, sea por diferencias de criterio (por ejemplo, qué se considera una cardiopatía congénita grave) o sea por diferencias prácticas debidas por ejemplo a los umbrales empleados en el estándar de referencia para considerar como patológica una función renal o una insuficiencia cardíaca. Además de las connotaciones que estas diferencias pudieran tener sobre el espectro de pacientes incluido en la revisión, los cambios en las definiciones de la enfermedad pueden hacer que los resultados de la revisión nos sean ajenos, dado que nuestro interés bien podría ser diagnosticar una entidad nosológica distinta a la analizada en la revisión.

Artículo

Tsoi KKF, Chan JYC, Hirai HW, Wong SYS, Kwok TCY. Cognitive Tests to Detect Dementia: A Systematic Review and Meta-analysis. JAMA Intern Med. 2015;175(9):1450-1458. Disponible en: https://pubmed.ncbi.nlm.nih.gov/26052687/.

Plantilla CASPe contestada para este artículo concreto

En el cuadro 14.1 se muestra la plantilla CASPe contestada para este artículo concreto.

CUADRO 14.1 ■ Evaluación crítica del artículo propuesto (plantillas CASPe)

A) ¿Los resultados de la revisión son válidos?

Preguntas «de eliminación»

1. ¿Se hizo la revisión sobre un tema claramente definido?

PISTA: un tema debe ser definido en términos de:

- *Población de estudio.*
- *Prueba que se evalúa y de referencia o* gold standard.
- *Condición clínica.*
- *Resultados* (outcomes) *considerados.*

Sí ✓ No sé No

En la revisión sistemática se establecen los criterios de inclusión de estudios. En cuanto a la población, debían ser estudios de pacientes reclutados en cualquier entorno clínico o comunitario **(población de estudio),** a los que se realiza un test de cribado cognitivo en una entrevista presencial con el paciente o cuidador **(prueba en evaluación)** y se compara con unos criterios estandarizados de diagnóstico *(Diagnostic and Statistical Manual of Mental Disorders, International Classification of Diseases, National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer Disease and Related Disorders Association, National Institute of Neurological and Communicative Disorders and Stroke y la Association Internationale pour la Recherche et L'Enseignement en Neuroscience criteria, o juicio clínico tras series de evaluación completa)* **(pruebas de referencia)** para la detección de demencia asociada a alzhéimer, demencia vascular o enfermedad de Parkinson **(condiciones clínicas).** Los resultados de exactitud diagnóstica que consideraron fueron: sensibilidad, especificidad y cocientes de probabilidad positivo y negativo. Los autores excluyeron los estudios que evaluaban test aplicados a pacientes con discapacidad visual o con tiempo de administración mayor a 20 min.

2. ¿Buscaron los autores el tipo de artículos adecuado?

PISTA: el mejor tipo de estudio es el que:

- *Se dirige a la pregunta objeto de revisión.*
- *Tiene un diseño apropiado para la pregunta.*

Sí ✓ No sé No

Los autores incluyen **estudios transversales,** que es el mejor diseño de estudio para evaluar la exactitud de una prueba diagnóstica. Aunque los autores lo señalan como criterio de inclusión, debemos comprobar después que los estudios incluidos finalmente tengan el diseño adecuado.

Preguntas detalladas

3. ¿Crees que estaban incluidos los estudios importantes y relevantes?

PISTA: busca:

- *¿Qué bases de datos se han usado?*
- *Si se han utilizado bases específicas de estudios de diagnóstico.*
- *Seguimiento de las referencias.*
- *Contacto personal con expertos.*
- *Búsqueda de estudios no publicados.*
- *Búsqueda de estudios en idiomas distintos del inglés.*

Sí ✓ No sé No

Los autores han realizado la búsqueda en Medline, Embase y PsycINFO. Se incluyen las fechas de búsqueda (hasta 1-9-2014) y términos de búsqueda (Alzheimer, Parkinson, vascular, *stroke, cognitive impairment* y *dementia*), pero no la estrategia utilizada ni las plataformas de acceso a las bases de datos. La búsqueda se complementó con búsquedas manuales en las listas de referencias bibliográficas de los artículos incluidos y en Google Scholar. La búsqueda se restringió a estudios publicados en inglés.

4. ¿Crees que los autores de la revisión han hecho suficiente esfuerzo para valorar la calidad de los estudios incluidos?

PISTA: revisa:

- *¿Qué herramienta han utilizado para valorar la calidad?*
- *Si han evaluado distintos dominios en cuanto a selección de pacientes, prueba a evaluar, prueba de referencia, flujo de pacientes y tiempo entre pruebas.*
- *Si se ha adaptado la herramienta a las peculiaridades de la revisión.*
- *Si los resultados han influido en la realización del metaanálisis o en la interpretación de los resultados.*

Sí | No sé ✓ | No

Los autores dicen que utilizan la herramienta QUADAS-2 para evaluar la calidad de los estudios incluidos en la revisión. En la tabla 2 de características de los estudios se incluye para cada uno de los test y en cada uno de los dominios (selección de pacientes, prueba a evaluar, prueba de referencia y flujo de pacientes) el porcentaje de estudios con alto riesgo de sesgo. Se echa en falta más detalle de cómo han adaptado la herramienta QUADAS-2 a la revisión con los criterios para definir en cada dominio si el riesgo de sesgo es alto, bajo o no claro.

Los autores también incluyen en la tabla 2 una valoración de la calidad en una escala de 8 puntos diseñada de acuerdo con las recomendaciones STARD (Standards for Reporting of Diagnostic Accuracy statement). Aquí los autores confunden calidad de los estudios con calidad del *reporting*. La declaración STARD es una iniciativa desarrollada para establecer directrices de cómo comunicar los estudios de exactitud de pruebas diagnósticas, pero no para la evaluación del riesgo de sesgo de estos estudios.

5. Si los resultados de los diferentes estudios han sido mezclados para obtener un resultado combinado, ¿era razonable hacer eso?

PISTA: considera si:

- *Los resultados de los estudios eran similares entre sí.*
- *Los resultados de todos los estudios incluidos están claramente presentados.*
- *Están discutidos los motivos de cualquier variación de los resultados.*

Sí | No sé | No ✓

El principal problema de la revisión es que los resultados de los diferentes estudios se han metaanalizado para obtener un resultado combinado cuando presentaban muchas diferencias. Esto se puede apreciar en los *forest plots* de la figura 2 del artículo. Aunque tomándolo con cautela como hemos visto en el capítulo, la cuantificación de la heterogeneidad medida con el I^2 de Higgins fue de un 89 y 97% para sensibilidad y especificidad para el test Mini-Cog, del 92 y 94% para sensibilidad y especificidad para el test MMSE y del 53 y 87% para sensibilidad y especificidad respectivamente en el test ACE-R (*forest plot* figura 2).

Las diferencias entre estudios o heterogeneidad podrían ser atribuibles a diferencias en la población en cuanto a edad, nivel educativo, socioeconómico, etc.; diferencias en cuanto al ámbito en que se ha realizado el estudio o por la prueba de referencia utilizada (v. tabla 2). Los autores detectaron diferencias en los puntos de corte de positividad utilizados por los estudios (por ejemplo, para el MMSE el punto de corte para demencia más común [44% estudios] fue de 23 a 24, mientras que un 20% utilizaron un punto de corte de 25 a 26).

Solo en el caso del MMSE el número de estudios incluidos permitió realizar un análisis de subgrupos para identificar como posibles fuentes de heterogeneidad (v. información suplementaria del artículo y comentario en pregunta 6) la región del estudio y el ámbito de reclutamiento de pacientes, sin encontrar diferencias significativas.

(Continúa)

B) ¿Cuáles son los resultados?

6. ¿Cuál es el resultado global de la revisión?

PISTA: considera:

- *Si tienes claros los resultados últimos de la revisión.*
- *¿Cuáles son? (Numéricamente, si es apropiado.)*
- *¿Cómo están expresados los resultados? (sensibilidad, especificidad, cocientes de probabilidad [CP], etc.).*

Los resultados principales de la revisión se expresan como sensibilidad y especificidad y CP promedio en la tabla 3 del artículo y en la figura 2. Aquí se presenta un resumen:

Test (tiempo de aplicación)	Número de estudios *(n)*	Sensibilidad (IC 95%)	Especificidad (IC 95%)	CP+	CP–
MMSE (≤ 10 min)	102 (36.080)	81% (78 a 84)	89% (87 a 91)	7,5	0,21
AMT (≤ 10 min)	13 (5.273)	88% (82 a 92)	85% (81 a 89)	5,9	0,15
MoCA (≤ 10 min)	20 (4.647)	91% (84 a 95)	81% (71 a 88)	4,8	0,12
3MS (≤ 20 min)	6 (4.271)	86% (83 a 89)	85% (74 a 92)	5,8	0,17
Mini-Cog (≤ 5 min)	9 (4.178)	91% (80 a 96)	86% (74 a 93)	6,6	0,10
IQCODE *long-form* (≤ 20 min)	15 (3.203)	84% (81 a 87)	82% (75 a 87)	4,7	0,19
IQCODE *short-form* (≤ 20 min)	7 (2.117)	89% (85 a 92)	82% (63 a 93)	4,9	0,14
Verbal fluency test (≤ 5 min)	7 (3.038)	80% (73 a 86)	82% (73 a 88)	4,4	0,24
MIS (≤ 5 min)	6 (2.635)	80% (68 a 86)	91% (84 a 96)	9,2	0,23
CDT Shulman (≤ 5 min)	9 (2.096)	83% (75 a 89)	84% (69 a 92)	5,0	0,20
CDT Sunderland (≤ 5 min)	9 (1.757)	76% (69 a 83)	85% (76 a 91)	5,1	0,28
ACE-R (≤ 20 min)	12 (2.070)	92% (90 a 94)	89% (84 a 93)	8,6	0,09
GPCOG (≤ 10 min)	5 (1.082)	92% (81 a 97)	87% (83 a 90)	6,8	0,10

En el suplemento del artículo se ofrecen resultados de:

- Un análisis de subgrupos *(etable 3)* considerando las variables región (América, Asia, Europa) y ámbito de reclutamiento de pacientes (comunitario, clínica, hospital, atención primaria, otros).
- Curva HSROC *(efigure 1)* del test MMSE en la detección de demencia, incluyendo resultado combinado de sensibilidad y especificidad y elipse de confianza.
- Plano ROC *(efigure 2)* con los estimadores combinados de sensibilidad y especificidad y sus correspondientes elipses de confianza para los test ACE-R, Mini-Cog Test y MMSE para la detección de demencia.

En este caso, el abordaje más adecuado de análisis, dada la presencia de diferentes umbrales de positividad entre los estudios, debería haberse orientado a estimar la curva ROC resumen que subyace entre los estudios incluidos (curva SROC) para cada uno de los test evaluados y evitar proporcionar un valor de sensibilidad y especificidad promedio, porque no podemos conocer a qué punto de corte de la prueba diagnóstica corresponden.

7. ¿Cuál es la precisión del resultado?

PISTA: busca los intervalos de confianza de las parejas de índices.

En la tabla 3 del artículo se muestran los intervalos de confianza tanto para sensibilidad y especificidad como para los cocientes de probabilidad. En las figuras 1 y 2 del suplemento también se describe la precisión con la representación gráfica de las elipses de confianza *(confidence regions)*. Pero insistimos en lo comentado en el punto 6 de que se debería haber evitado calcular una sensibilidad y especificidad promedio.

C) ¿Son los resultados aplicables al escenario?

Sí No sé No ✓

Los resultados de la revisión no son claramente aplicables al escenario que se presenta. La población incluida en la revisión es bastante heterogénea, incluyendo pacientes reclutados en distintos ámbitos, y nuestro interés se centra en el ámbito particular de la atención primaria. No define si eran participantes asintomáticos o bien presentaban algún tipo de déficit cognitivo, probablemente incluya ambos. Otra dificultad es la variabilidad en los puntos de corte de los test en los estudios incluidos, ¿qué punto de corte usó? Probablemente establecer un único punto de corte válido para todos los pacientes sea complejo, la valoración de la situación basal y su evolución con medidas repetidas en el tiempo serán casi siempre las que determinen el verdadero positivo o negativo del cribado.

Bibliografía

1. Tsoi KKF, Chan JYC, Hirai HW, Wong SYS, Kwok TCY. Cognitive Tests to Detect Dementia: A Systematic Review and Meta-analysis. JAMA Intern Med 2015;175(9):1450-8.
2. Roqué M, Martínez-García L, Solà I, Alonso-Coello P, Bonfill X, Zamora J. Toolkit of methodological resources to conduct systematic reviews. F1000Research 2020;9:82.
3. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;332(7549):1089-92.
4. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282(11):1061-6.
5. Rutjes AWS, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PMM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;174(4):469-76.
6. Leeflang MMG, Scholten RJPM, Rutjes AWS, Reitsma JB, Bossuyt PMM. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. J Clin Epidemiol 2006;59(3):234-40.
7. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, the PRISMA-DTA Group, 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):388-96.
8. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(8):529-36.
9. Whiting PF, Sterne JAC, Westwood ME, Bachmann LM, Harbord R, Egger M, et al. Graphical presentation of diagnostic information. BMC Med Res Methodol 2008;8:20.
10. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327(7414):557-60.
11. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PMM. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 2003;56(11):1129-35.
12. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005;58(10):982-90.
13. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 2001;20(19):2865-84.
14. Walter SD. The partial area under the summary ROC curve. Stat Med 2005;24(13):2025-40.
15. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol 2005;58(9):882-93.
16. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med 2002;137(7):598-602.

Cómo citar este capítulo:

Plana MN, Zamora J. Lectura crítica de revisiones sistemáticas de pruebas diagnósticas. En: Cabello Juan B, editor. Lectura crítica de la evidencia clínica, 2.ª ed. Barcelona: Elsevier; 2022. p. 167-181.