

UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE ESTUDIOS ESTADÍSTICOS



TESIS DOCTORAL

**Modelos estadísticos para el metanálisis de estudios de
exactitud diagnóstica. Aplicación a diferentes escenarios de
variabilidad entre estudios**

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

María Nieves Plana Farrás

Directores

**Javier Zamora Romero
Teresa Pérez Pérez**

Madrid

UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE ESTUDIOS ESTADÍSTICOS



TESIS DOCTORAL

Modelos estadísticos para el metanálisis de estudios de exactitud diagnóstica. Aplicación a diferentes escenarios de variabilidad entre estudios

MEMORIA PARA OPTAR AL GRADO DE DOCTORA

PRESENTADA POR

María Nieves Plana Farrás

DIRECTOR

Dr. Javier Zamora Romero
Dra. Teresa Pérez Pérez

UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE ESTUDIOS ESTADÍSTICOS



TESIS DOCTORAL

Modelos estadísticos para el metanálisis de estudios de exactitud diagnóstica. Aplicación a diferentes escenarios de variabilidad entre estudios

MEMORIA PARA OPTAR AL GRADO DE DOCTORA

PRESENTADA POR

María Nieves Plana Farrás

DIRECTOR

Dr. Javier Zamora Romero
Dra. Teresa Pérez Pérez

Las cosas que no se saben son las que convierten la vida en algo fascinante

Wisława Szymborska

*De eso se trata, de coincidir con gente que te haga ver cosas que tú no ves. Que
te enseñen a mirar con otros ojos*

José Saramago

Para ser alguien entre cielo y suelo

y salvarse del odio y sus resabios

nada como el amor y su consuelo

Mario Benedetti

A los tesoros que iluminan y sonríen mi vida, mi compañero y mis hijos

A mi padre, que me transmitió su pasión y compromiso por el trabajo

A mi madre y mi hermana, siempre queriéndome

A mis amigos, poetas, héroes y fieles

Agradecimientos

A la Unidad, por generar el espíritu de colaboración, generosidad y alegría que han sido imprescindibles para alcanzar este objetivo

A todos vosotros, coautores y colaboradores, gracias, a unos por la oportunidad y a otros, por el esfuerzo y por poner vuestro conocimiento a mi lado

A Teresa, por tu disponibilidad, aportaciones y correcciones

A Javier, cuando la gratitud es tan absoluta las palabras sobran, gracias por el enorme cariño que has puesto conmigo en esta tesis

ÍNDICE

PREÁMBULO	16
CAPÍTULO 1	18
1.1 MARCO DE EVALUACIÓN DE LAS PRUEBAS DIAGNÓSTICAS	22
1.2 EVALUACIÓN DE LA EXACTITUD DIAGNÓSTICA. INVESTIGACIÓN PRIMARIA	28
1.3 REVISIÓN SISTEMÁTICA. INVESTIGACIÓN SECUNDARIA.....	35
1.4 METANÁLISIS	39
1.5 HETEROGENEIDAD	40
1.6 OBJETIVOS	46
1.7 ESQUEMA DE LA EXPOSICIÓN.....	47
CAPÍTULO 2	51
2.1 MODELOS ESTADÍSTICOS DE METANÁLISIS	51
2.2 MODELOS DE SÍNTESIS DE RESULTADOS DE EXACTITUD DIAGNÓSTICA	52
2.3 ESTADÍSTICOS DE ANÁLISIS Y CUANTIFICACIÓN DE LA HETEROGENEIDAD	70
CAPÍTULO 3	77
3.1 OBJETIVOS	78
3.2 METODOLOGÍA.....	78
3.3 RESUMEN DE RESULTADOS	79
3.4 CONCLUSIONES	79
3.5 PUBLICACIÓN.....	80

CAPÍTULO 4	95
4.1 OBJETIVOS	96
4.2 SOFTWARE	97
4.3 METODOLOGÍA.....	109
4.4 VALIDACIÓN DE LA APLICACIÓN	110
4.5 PUBLICACIÓN.....	115
CAPÍTULO 5	118
5.1 OBJETIVOS	119
5.2 METODOLOGÍA.....	119
5.3 RESUMEN DE RESULTADOS	120
5.4 CONCLUSIONES	121
5.5 PUBLICACIÓN.....	121
CAPÍTULO 6	208
6.1 OBJETIVOS	209
6.2 METODOLOGÍA.....	209
6.3 RESUMEN DE RESULTADOS	210
6.4 CONCLUSIONES	211
6.5 PUBLICACIÓN.....	211
CAPÍTULO 7	226
7.1 OBJETIVOS	227
7.2 METODOLOGÍA.....	228
7.3 RESUMEN DE RESULTADOS	228

7.4 CONCLUSIONES	229
7.5 PUBLICACIÓN.....	229
7.6 REANÁLISIS DE LA HETEROGENEIDAD EN LAS REVISIONES SISTEMÁTICAS PUBLICADAS	240
CAPÍTULO 8	245
DISCUSIÓN.....	263
CONCLUSIONES	270
10.1 CONCLUSIONES METODOLÓGICAS.....	270
10.2 CONCLUSIONES CLÍNICAS	272
RESUMEN.....	274
SUMMARY	277
ANEXOS.....	280
ANEXO I.....	281
ANEXO II	285
ANEXO III.....	295
REFERENCIAS BIBLIOGRÁFICAS	297

ÍNDICE DE TABLAS

TABLA 1. CARACTERÍSTICAS DE LOS DIFERENTES MODELOS DE EVALUACIÓN	27
TABLA 2. TABLA CRUZADA RESULTADOS ESTUDIO PRIMARIO DE DIAGNÓSTICO	30
TABLA 3. MÉTODOS ESTADÍSTICOS DE METANÁLISIS.....	53
TABLA 4. DISTRIBUCIÓN DE LOS SUJETOS EN EL ESTUDIO <i>I</i> -ÉSIMO	58
TABLA 5. RESULTADOS METANÁLISIS UTILIZANDO METADISC 2.0.....	111
TABLA 6. RESULTADOS METANÁLISIS COMPARANDO METADISC 2.0 CON METANDI Y METADAS..	113
TABLA 7. COMPARACIÓN ANÁLISIS DE METARREGRESIÓN UTILIZANDO METADISC 2.0, METANDI Y METADAS	114

ÍNDICE DE FIGURAS

FIGURA 1. CURVA ROC.....	34
FIGURA 2. MODELO DE EFECTOS FIJOS	55
FIGURA 3. MODELO DE EFECTOS ALEATORIOS.....	57
FIGURA 4. CURVA SROC SIMÉTRICA Y ASIMÉTRICA	63
FIGURA 5. PORTADA PRESENTACIÓN METADiSc 2.0.....	98
FIGURA 6. MENÚ PRINCIPAL METADiSc 2.0	99
FIGURA 7. PANTALLA DE VISUALIZACIÓN DEL FICHERO DE DATOS ACTIVO	101
FIGURA 8. FOREST-PLOT DE SENSIBILIDAD DE LA PULSIOXIMETRÍA POR LA VARIABLE "TIMING"	103
FIGURA 9. PANTALLA METADiSc 2.0 CORRESPONDIENTE AL MODELO BIVARIANTE	104
FIGURA 10. PANTALLA DE METADiSc 2.0 CORRESPONDIENTE A LA PESTAÑA SROC	106
FIGURA 11. PANTALLA DE METADiSc 2.0 CORRESPONDIENTE AL ANÁLISIS DE SUBGRUPOS.....	107
FIGURA 12. PANTALLA DE METADiSc 2.0 CORRESPONDIENTE AL ANÁLISIS DE SENSIBILIDAD	108
FIGURA 13. PANTALLA DE METADiSc 2.0 CORRESPONDIENTE AL RESUMEN DE RESULTADOS	109
FIGURA 14. REPRESENTACIÓN GRÁFICA EN EL PLANO ROC DE LA SENSIBILIDAD Y ESPECIFICIDAD DE LAS REVISIONES INCLUIDAS EN LOS CAPÍTULO 5, 6 Y 7.	242

ABREVIATURAS

ACR: American College of Radiology

AUC: Area under the curve

CCC: Cardiopatías críticas congénitas

cfDNA: Cell free DNA

CIE: Clasificación Internacional de Enfermedades

CINAHL: Cumulative Index to Nursing and Allied Health Literature

CP: Cociente de probabilidad

DOR: Diagnostic odds ratio

FN: Falsos negativos

FP: Falsos positivos

GLM: General Linear Model

GLMM: Generalized linear Mixed Model

GLZ: Generalized Linear Model

HSRpROJ: Health Services Research Projects in Progress

IC: Intervalo de confianza

LR: Likelihood ratio

MeSH: Medical Subject Heading

OR: Odds ratio

ORD: Odds ratio diagnostic

PAPP-A: proteína plasmática asociada al embarazo

QUADAS: Quality Assessment of Diagnostic Accuracy Studies

ROC: Receiver Operating Characteristics

sROC: Summary Receiver Operating Characteristics

T21: Trisomía cromosoma 21

T18: Trisomía cromosoma 18

T13: Trisomía cromosoma 13

TFP: Tasa de falsos positivos

TNR: True negative rate

TPR: True positive rate

VN: Verdaderos negativos

VP: Verdaderos positivos

VPP: Valor predictivo positivo

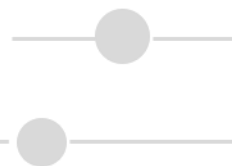
Preámbulo

El proyecto de investigación industrial realizado en esta tesis doctoral se enmarca en un convenio de colaboración entre el Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS) y la Universidad Complutense de Madrid (UCM).

El IRYCIS tiene como misión liderar una investigación biomédica de excelencia que genere conocimientos útiles para el diagnóstico y tratamiento de enfermedades actuales o potenciales desde una perspectiva traslacional dirigida a obtener un impacto en la sociedad. La investigación se ha realizado en el seno del grupo de Epidemiología y Bioestadística Clínica del área 4 del IRYCIS. Este grupo promueve el empleo de metodología de investigación óptima y técnicas bioestadísticas avanzadas para el desarrollo de investigación clínico-epidemiológica. La tesis doctoral se encuadra en la línea de investigación de síntesis, concretamente en la línea sobre metodología de revisiones sistemáticas y metanálisis en la que el grupo de Epidemiología y Bioestadística Clínica del IRYCIS tiene una amplia experiencia y proyección internacional.

La UCM, en la figura del Departamento de Estadística y Ciencia de los Datos, y en el marco del Programa de Doctorado “Análisis de Datos (*Data Science*)”, ha actuado como responsable de la dirección académica de este trabajo de investigación velando por la rigurosidad de los métodos estadísticos de análisis que se han implementado y por el cumplimiento de los objetivos propuestos en la memoria técnica justificativa. Estas sinergias entre el IRYCIS y la UCM han promovido la ejecución de la presente tesis doctoral por la que se opta a la Mención de Doctorado Industrial.

Capítulo 2



Capítulo 2

Tal y como se refiere en el Capítulo 1, el metanálisis integra los resultados obtenidos en estudios primarios. Esta combinación debe hacerse, teniendo en cuenta la contribución que cada estudio individual va a tener en la estimación global, a través de los denominados modelos meta-analíticos.

2.1 Modelos Estadísticos de Metanálisis

Los modelos meta-analíticos representan el avance y profundización metodológica en técnicas óptimas para la integración de resultados de investigación. Estas propuestas de modelización de la síntesis cuantitativa de la evidencia han ido evolucionando a modelos matemáticos complejos. De una manera genérica, los principales problemas metodológicos para establecer un método coherente que integre los resultados de investigación, son la existencia de distintas unidades de medida de los estudios a integrar y la presencia de variaciones en las características básicas de los distintos estudios.

En la investigación sobre diagnóstico se añaden otras particularidades que aumentan la complejidad: el parámetro de interés no es unidimensional sino bivalente, cada estudio se resume con una pareja de índices; estos índices frecuentemente están correlacionados entre sí y, la precisión de medida en ambos índices es distinta dependiendo del número de enfermos y no enfermos incluidos en cada estudio. Esto ha condicionado que los aspectos estadísticos relacionados con las revisiones sistemáticas de diagnóstico sigan generando retos y controversia en la comunidad científica.

El presente capítulo se organiza en dos bloques, en un primer bloque se exponen los diferentes modelos propuestos para la síntesis de resultados de la medida del efecto, en este caso la exactitud diagnóstica. En un segundo bloque se exponen los estadísticos disponibles para analizar y cuantificar la variabilidad entre estudios (heterogeneidad).

2.2 Modelos de síntesis de resultados de exactitud diagnóstica

En la Tabla 3 se enmarcan los diferentes modelos propuestos para la síntesis de evidencia en pruebas diagnósticas dentro de las familias de modelos estadísticos, incorporando además las características y denominación de cada uno de estos modelos.

Tabla 3. Métodos estadísticos de Metanálisis

	Características	Denominación
Modelos Lineales <i>(Linear Models):</i> Modelo de regresión lineal Análisis de la varianza Modelo Lineal General <i>(GLM, General Linear Model)</i>	Variable dependiente continua Efectos fijos Errores con distribución normal, independientes y homogéneos	Modelos Univariantes de Efectos Fijos y Aleatorios (Aproximación Normal) Modelo de Moses y Littenberg (Moses, Shapiro y Littenberg 1993)
Modelos Mixtos <i>(Mixed Models)</i>	Variable dependiente continua Efectos fijos y aleatorios Distribución normal de los errores Los errores pueden mostrar correlación o heterogeneidad No requiere función de enlace	Modelo Bivariante (Reitsma et al. 2005)

Tabla 3. Métodos estadísticos de Metanálisis

Tipo de modelo	Características	Denominación
Modelos para datos no normales:		
Modelo Lineal Generalizado (GLZ, <i>Generalized Linear Model</i>)	Variable dependiente puede ser discreta, ordinal, o categórica (binaria o multinomial) Efectos fijos Errores con distribución dentro de la familia exponencial, independientes y homogéneos Requiere una función de enlace no lineal	
Modelo Lineal Generalizado Mixto (GLMM, <i>Generalized Linear Mixed Model</i>)	Variable dependiente puede ser discreta, ordinal, o categórica (binaria o multinomial) Efectos fijos y aleatorios Errores con distribución dentro de la familia exponencial, pueden mostrar correlación y/o heterogeneidad Requiere una función de enlace no lineal	Modelo Univariante de Efectos Aleatorios (Método exacto) Modelo Bivariante (Chu y Cole 2006)
Modelo no Lineal Generalizado Mixto	Variable dependiente puede ser discreta, ordinal, o categórica (binaria o multinomial) Efectos fijos y aleatorios Errores con distribución dentro de la familia exponencial, pueden mostrar correlación y/o heterogeneidad Requiere una función de enlace no lineal	Modelo Jerárquico HSROC (Rutter y Gatsonis 2001)

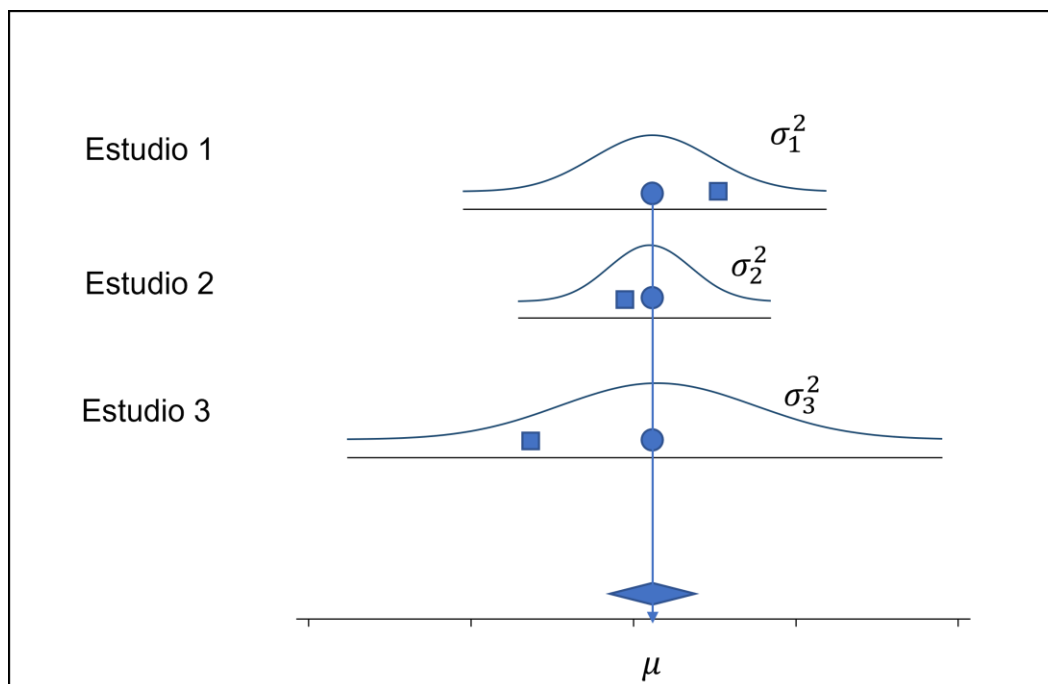
A continuación, se describen cada uno de estos métodos de metanálisis. Primero se exponen los modelos univariantes, después el modelo de regresión de Moses y Littenberg para ajustar la curva sROC y, por último, los modelos jerárquicos (bivariante y HSROC).

2.2.1 Modelos Univariantes. Aproximación Normal y Método Exacto

Los modelos univariantes de efectos fijos y aleatorios sintetizan la sensibilidad y especificidad de forma individual ignorando cualquier correlación entre ambos índices.

En el modelo de efectos fijos se asume que los estudios incluidos en el metanálisis están estimando un mismo y único efecto, por lo que la única variabilidad en los estudios individuales es la debida al error de muestreo aleatorio, que se cuantifica estimando la varianza intra-estudio para la sensibilidad o la especificidad, σ_i^2 (Figura 2).

Figura 2. Modelo de efectos fijos



En la Figura 2 se muestra un ejemplo con 3 estudios. El modelo asume que los estudios están estimando el mismo efecto fijo (representado en cada estudio por un círculo) y en cada uno de los estudios se obtiene un estimador de ese efecto fijo (representado por un cuadrado). El rombo representa el promedio ponderado de estos estudios (μ). La variabilidad en dicho efecto es sólo debida a las distintas varianzas intra-estudio (σ_1^2 , σ_2^2 , σ_3^2).

Así, el efecto obtenido en el estudio i -ésimo se pondera por un factor w_i , que viene determinado exclusivamente por la varianza intra-estudio:

$$w_i = \frac{1}{\sigma_i^2}; i = 1, 2, \dots, k \quad (6)$$

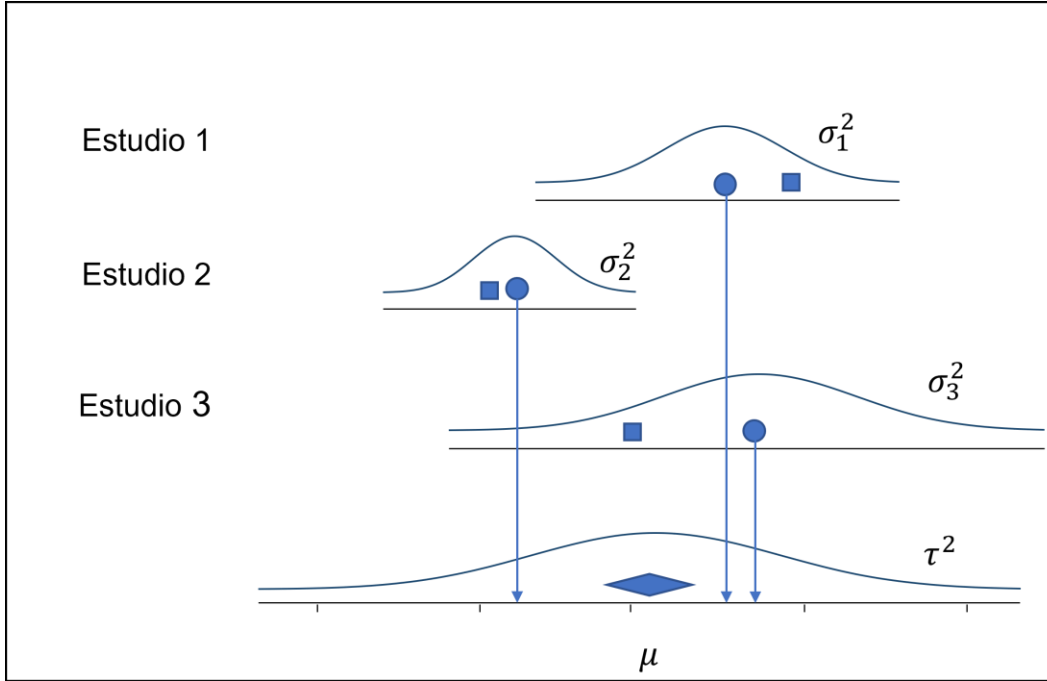
donde k denota el número de estudios incluidos en el metanálisis.

En el modelo de efectos aleatorios se asume que los diferentes estudios estiman efectos diferentes que siguen una distribución con media μ y varianza τ^2 (Figura 3). De nuevo, el rombo representa el promedio ponderado de estos estudios. El modelo de efectos aleatorios tiene en cuenta, por tanto, además de la varianza intra-estudio (σ_i^2), la varianza entre-estudios τ^2 .

El factor de ponderación, w_i , viene determinado por ambas varianzas, la varianza intra-estudio y la varianza entre estudios:

$$w_i = \frac{1}{\sigma_i^2 + \tau^2}; i = 1, 2, \dots, k \quad (7)$$

Figura 3. Modelo de efectos aleatorios



La varianza intra-estudio, σ_i^2 , la vamos a denotar por σ_{Ai}^2 , cuando nos referimos a la sensibilidad y por σ_{Bi}^2 cuando nos referimos a la especificidad.

El método de efectos aleatorios más usado y simple para sintetizar la evidencia (DerSimonian y Laird 1986), considera los logit de sensibilidad y especificidad para cada estudio:

$$\mu_{Ai} = \ln \left(\frac{\text{sensibilidad}_i}{(1 - \text{sensibilidad}_i)} \right)$$

$$\mu_{Bi} = \ln \left(\frac{\text{especificidad}_i}{(1 - \text{especificidad}_i)} \right); i = 1, 2, \dots, k \quad (8)$$

como variables aleatorias y asume que siguen una distribución normal:

$$\begin{aligned}\mu_{Ai} &\sim N(\mu_A, \sigma_A^2) \\ \mu_{Bi} &\sim N(\mu_B, \sigma_B^2); i = 1, 2, \dots, k\end{aligned}\tag{9}$$

La variabilidad intra-estudio se puede modelizar utilizando la Aproximación a la distribución Normal o el Método Exacto. El método de Aproximación a la distribución Normal considera como estimadores de estos logits, las expresiones:

$$\hat{\mu}_{Ai} = \ln\left(\frac{n_{A1i}}{n_{Ai} - n_{A1i}}\right); \quad \hat{\mu}_{Bi} = \ln\left(\frac{n_{B0i}}{n_{Bi} - n_{B0i}}\right); i = 1, 2, \dots, k\tag{10}$$

donde, tal y como aparece representado en la Tabla 4, n_{Ai} es el número total de enfermos, n_{A1i} el número de verdaderos positivos, n_{Bi} el número total de no enfermos y n_{B0i} el número de verdaderos negativos, todos ellos correspondientes al estudio i -ésimo:

Tabla 4. Distribución de los sujetos en el estudio i -ésimo

Prueba Referencia (D)		
Prueba en evaluación (Y)	Con Evento (A)	Sin evento (B)
Positivo (1)	n_{A1i}	n_{B1i}
Negativo (0)	n_{A0i}	n_{B0i}
Total	n_{Ai}	n_{Bi}

y asume que la distribución asintótica de estos estimadores es una distribución normal.

$$\hat{\mu}_{Ai} \sim N(\mu_{Ai}, \hat{\sigma}_{Ai}^2)\tag{11}$$

$$\hat{\mu}_{Bi} \sim N(\mu_{Bi}, \hat{\sigma}_{Bi}^2); i = 1, 2, \dots, k$$

donde:

$$\hat{\sigma}_{Ai}^2 = \frac{1}{n_{A1i}} + \frac{1}{n_{Ai} - n_{A1i}}$$

$$\hat{\sigma}_{Bi}^2 = \frac{1}{n_{B0i}} + \frac{1}{n_{Bi} - n_{B0i}}; i = 1, 2, \dots, k \quad (12)$$

Los parámetros del modelo (μ_A, τ_A^2) para sensibilidad y (μ_B, τ_B^2) para especificidad, se pueden estimar utilizando el método de máxima verosimilitud (ML), el de máxima verosimilitud restringido (REML) o el método de los momentos (DerSimonian y Laird 1986). La diferencia principal entre los tres métodos radica en la forma de estimar la varianza entre estudios. El método de máxima verosimilitud da resultados sesgados, infraestimando la varianza entre estudios, mientras que los resultados que se obtienen con el método REML y el de momentos son similares y por ese motivo son los que habitualmente se utilizan.

Este método tiene una importante limitación. Si alguna celda de la Tabla 4 es igual a cero, entonces no es posible calcular los estimadores (10) (12) y para solucionarlo es necesario aplicar una corrección de continuidad, típicamente sumar 0.5 a las celdas de todos los estudios, incluyendo los que no tienen valores cero. Está descrito que esta corrección puede sesgar los resultados (Dahabreh et al. 2012). Además, hay una alta

correlación entre $\hat{\mu}_{Ai}$ y $\hat{\sigma}_{Ai}^2$ (equivalentemente entre $\hat{\mu}_{Bi}$ y $\hat{\sigma}_{Bi}^2$). La correlación es positiva si μ_A es positiva y negativa en otro caso, motivando un sesgo hacia cero o hacia 0.5 en la estimación de la sensibilidad (equivalentemente para la especificidad). Este efecto se debe a que la media y la varianza de $\hat{\mu}_{Ai}$ están determinadas por el mismo parámetro.

Una alternativa más idónea al método de aproximación normal para estimar la variabilidad intra-estudio es el Método Exacto (Hamza, van Houwelingen y Stijnen 2008). Este método se basa, si nos fijamos en la sensibilidad, en que el número de verdaderos positivos observados en el estudio i -ésimo, n_{Ai} , sigue una distribución binomial

$$n_{Ai} \sim \text{Binomial}(n_{Ai}, p_{Ai}); i = 1, 2, \dots, k \quad (13)$$

donde n_{Ai} es el número total de enfermos del estudio i -ésimo y p_{Ai} la sensibilidad del estudio i -ésimo que se define como:

$$p_{Ai} = \frac{1}{1 + \exp(-\mu_{Ai})}; i = 1, 2, \dots, k \quad (14)$$

De forma análoga se define la especificidad del estudio i -ésimo como:

$$p_{Bi} = \frac{1}{1 + \exp(-\mu_{Bi})}; i = 1, 2, \dots, k \quad (15)$$

En este caso, no es necesario hacer una corrección de continuidad ante la presencia de valores cero, ni existe el problema de correlación anteriormente descrito.

El método se corresponde con un GLMM y los parámetros (μ_A, τ_A^2) para sensibilidad y (μ_B, τ_B^2) para especificidad, se pueden estimar utilizando el método de máxima verosimilitud.

2.2.2 Modelo de regresión de Moses-Littenberg

Este método fue uno de los primeros métodos propuestos para ajustar la curva sROC (Moses, Shapiro y Littenberg 1993). Es un modelo de efectos fijos que sólo tiene en cuenta la variabilidad relacionada con la presencia de efecto umbral y la varianza inter-estudio. Propone los siguientes pasos:

- i. Hacer una transformación logit de la sensibilidad y de la especificidad para cada estudio. Utilizando la notación dada en (8)

$$\mu_{Ai} = \ln\left(\frac{\text{sensibilidad}_i}{1 - \text{sensibilidad}_i}\right); \mu_{Bi} = \ln\left(\frac{\text{especificidad}_i}{1 - \text{especificidad}_i}\right); i = 1, 2, \dots, k$$

Si para algún estudio se observa que la Tabla 4 presenta algún cero, se suma 0.5 a sus cuatro celdas.

- ii. Calcular para cada estudio $D_i = \mu_{Ai} + \mu_{Bi}$ y $S_i = \mu_{Ai} - \mu_{Bi}$. Esta expresión de D es equivalente al logaritmo del ORD y S representa la proporción de resultados positivos del test y se puede interpretar como una medida del umbral.

- iii. Ajustar un modelo de regresión lineal simple donde D es la variable dependiente, S la independiente y ε es el error aleatorio:

$$D_i = \beta_0 + \beta_1 S_i + \varepsilon_i; \quad i = 1, 2, \dots, k \quad (16)$$

Los parámetros del modelo (16) se estiman utilizando el método de mínimos cuadrados ordinario, si se asume homogeneidad de la varianza entre estudios, o ponderado, en otro caso. Los pesos pueden ser la inversa de la varianza del logaritmo del ORD, o simplemente el tamaño muestral.

- iv. Contrastar el efecto umbral y deshacer la transformación logarítmica

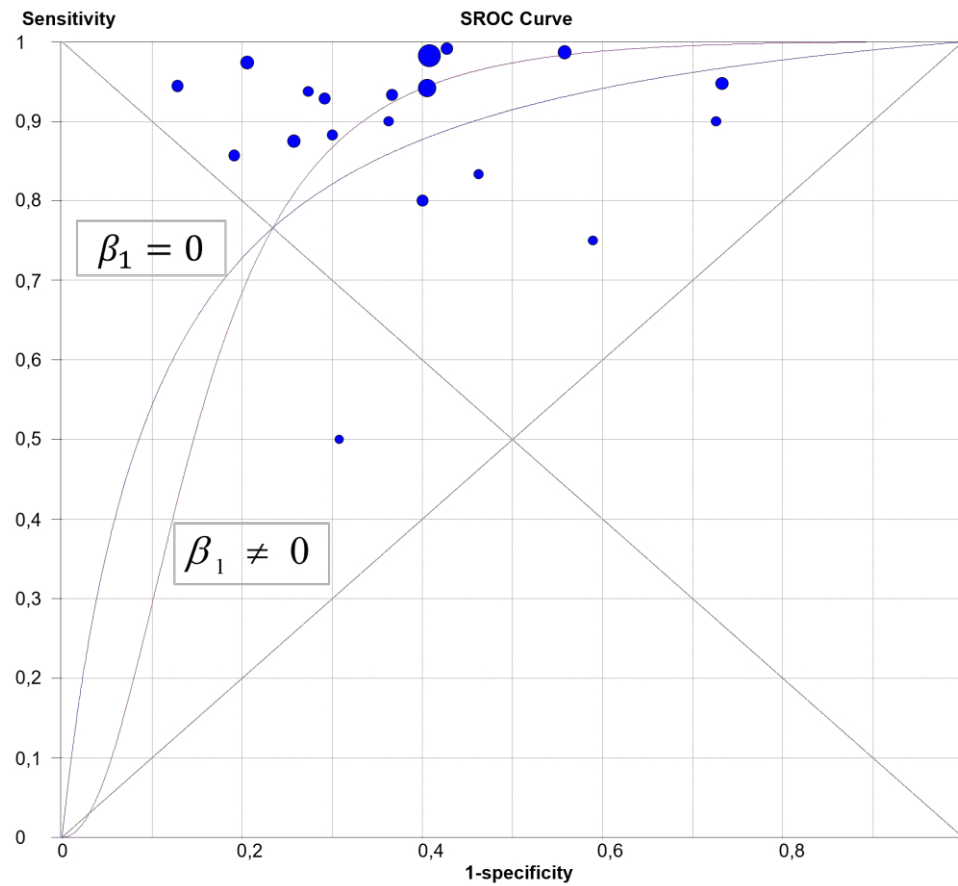
Contrastar si hay variación del rendimiento diagnóstico (medido por el ORD) debido al umbral de cada estudio, es equivalente a resolver un contraste acerca de la pendiente β_1 . En particular, no rechazar la hipótesis nula $H_0 : \beta_1 = 0$ es asumir que no hay variación y los estudios son homogéneos. En este caso el método da lugar a una curva ROC simétrica y $\exp(\beta_0)$ es la estimación del ORD global (Figura 4). La ecuación de la curva sería

$$Sensibilidad(t) = \frac{1}{1 + \frac{1}{\exp(\beta_0)} \times \left(\frac{1 - Especificidad(t)}{Especificidad(t)} \right)}; \quad t \in (0,1) \quad (17)$$

Sin embargo, rechazar $H_0 : \beta_1 = 0$ es asumir que sí existe variación entre estudios debido al umbral. En este caso la curva ROC es asimétrica (Figura 4) y está dada por la ecuación

$$Sensibilidad(t) = \frac{1}{1 + \frac{1}{\exp(\beta_0/(1-\beta_1))} \times \left(\frac{1 - Especificidad(t)}{Especificidad(t)} \right)^{\left(\frac{1+\beta_1}{1-\beta_1} \right)}}; t \in (0,1) \quad (18)$$

Figura 4. Curva sROC simétrica y asimétrica



En la Figura 4 se representa una curva sROC simétrica ($\beta_1 = 0$) y una curva sROC asimétrica ($\beta_1 \neq 0$).

El modelo de Moses y Littenberg (1993) tiene algunas limitaciones serias que aconsejan su uso sólo para estudios exploratorios (Arends et al. 2008; Ma et al. 2016). De una parte, el modelo no tiene en cuenta la correlación que existe entre sensibilidad y especificidad, ni que la precisión con la que se estimaron ambos índices en los estudios individuales difiere y, de otra parte, el método asume que la variable S no es aleatoria y está fijada por el investigador.

2.2.3 Modelos jerárquicos

Los modelos jerárquicos también denominados modelos mixtos o multinivel se caracterizan porque consideran tanto la correlación entre sensibilidad y especificidad como la variabilidad del rendimiento diagnóstico entre estudios, incluyendo un término de efectos aleatorios. Estos modelos son más robustos desde el punto de vista estadístico y se han propuesto como el estándar para el metanálisis de estudios de exactitud diagnóstica (Harbord et al. 2008; Macaskill et al. 2010).

Los dos principales abordajes, el modelo bivalente y el modelo HSROC, difieren en la parametrización, pero son matemáticamente equivalentes cuando no incluyen covariables para explicar fuentes de heterogeneidad (Harbord et al. 2007). La decisión sobre qué modelo utilizar depende de la presencia o no de efecto umbral y por tanto si el foco del análisis se puede poner en estimar una sensibilidad y especificidad promedio o por el contrario debe ajustarse una curva sROC.

2.2.3.1 Modelo bivalente basado en un modelo lineal mixto

Reitsma et al. (2005) proponen aplicar un modelo lineal mixto al metanálisis de estudios de pruebas diagnósticas, considerando un enfoque bivalente (Van Houwelingen, Zwinderman y Stijnen 1993).

Parten de nuevo de la sensibilidad y la especificidad p_{Ai} y p_{Bi} , obtenidas en el estudio i -ésimo en el que un total de n_{Ai} sujetos presentan la condición de interés y n_{Bi} sujetos no la presentan y sus correspondientes transformaciones logit μ_{Ai} y μ_{Bi} definidos en (8). Este modelo también asume que se cumplen las condiciones dadas en (9), es decir, que cada μ_{Ai} es una variable aleatoria que sigue una distribución normal de media μ_A y varianza común σ_A^2 . De forma análoga, se considera que cada μ_{Bi} sigue una distribución normal de media μ_B y varianza común σ_B^2 . Por tanto, el vector (μ_{Ai}, μ_{Bi}) sigue una distribución normal bivalente con vector de medias (μ_A, μ_B) y matriz de varianzas-covarianzas $\Sigma = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix}$ donde σ_{AB} representa la covarianza entre μ_{Ai} y μ_{Bi} .

$$\begin{pmatrix} \mu_{Ai} \\ \mu_{Bi} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix} \right) \quad (19)$$

El modelo se puede reescribir en términos de la correlación que suele ser más fácil de interpretar que la covarianza.

$$\begin{pmatrix} \mu_{Ai} \\ \mu_{Bi} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_A^2 & \rho_{AB} \sigma_A \sigma_B \\ \rho_{AB} \sigma_A \sigma_B & \sigma_B^2 \end{pmatrix} \right) \quad (20)$$

La extensión propuesta por Reitsma et al. (2005) asume que la precisión con la que se han medido los parámetros de sensibilidad y especificidad depende de cada estudio individual.

Así, cuando n_{Ai} y n_{Bi} son suficientemente grandes y $0 < p_{Ai} < 1$; $0 < p_{Bi} < 1$ asumen que

$$\begin{pmatrix} \hat{\mu}_{Ai} \\ \hat{\mu}_{Bi} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \begin{pmatrix} \sigma_A^2 & \rho_{AB}\sigma_A\sigma_B \\ \rho_{AB}\sigma_A\sigma_B & \sigma_B^2 \end{pmatrix} + \Sigma_i \right) \quad (21)$$

$$\text{con } \Sigma_i = \begin{pmatrix} \hat{\sigma}_{Ai}^2 & 0 \\ 0 & \hat{\sigma}_{Bi}^2 \end{pmatrix}$$

Definidos μ_{Ai} y μ_{Bi} en (8)

donde

$$\hat{\sigma}_{Ai}^2 = \frac{1}{n_{Ai} \times p_{Ai} \times (1 - p_{Ai})} \quad (22)$$

$$\hat{\sigma}_{Bi}^2 = \frac{1}{n_{Bi} \times p_{Bi} \times (1 - p_{Bi})} \quad (23)$$

Y el modelo, en ausencia de covariables, está compuesto por cinco parámetros μ_A , μ_B , σ_A^2 , σ_B^2 , ρ_{AB} , lo que exige disponer de un mínimo de estudios en el análisis para poder estimar todos los parámetros.

2.2.3.2 Modelo bivalente lineal generalizado mixto

Una limitación que presenta el modelo de Reitsma et al. (2005) es que, al igual que ocurría en el Método de Aproximación a la Normal, en algunos casos es necesario hacer una corrección por continuidad. Esto sucede, por ejemplo, cuando en uno de los estudios incluidos se observa que al menos una de las cuatro celdas de la Tabla 4 está vacía, es decir, o bien el número de verdaderos positivos, n_{Ai} , verdaderos negativos, n_{B0i} , falsos positivos, n_{B1i} , o falsos negativos, n_{A0i} , es igual a cero (Sweeting, Sutton y Lambert 2004) y tampoco es apropiado cuando los tamaños muestrales son pequeños.

Chu y Cole (2006) proponen un abordaje de análisis mediante un modelo bivalente lineal generalizado mixto. Para describir la intra-variabilidad de cada proporción (sensibilidad o especificidad), el modelo bivalente lineal generalizado mixto considera la distribución exacta del número de aciertos en cada uno de los grupos (con evento/sin evento).

$$(n_{Ai}/\varepsilon_{Ai}) \sim \text{Binomial}(n_{Ai}, \text{sensibilidad}) \text{ con}$$

$$\text{logit}(\text{sensibilidad}) = \mu_{Ai} = \alpha X_i + \varepsilon_{Ai}$$

$$(n_{B0i}/\varepsilon_{Bi}) \sim \text{Binomial}(n_{Bi}, \text{especificidad}) \text{ con}$$

(24)

$$\text{logit}(\text{especificidad}) = \mu_{Bi} = \beta Z_i + \varepsilon_{Bi}$$

$$y \begin{pmatrix} \varepsilon_{Ai} \\ \varepsilon_{Bi} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\varepsilon_{Ai}}^2 & \rho \sigma_{\varepsilon_{Ai}} \sigma_{\varepsilon_{Bi}} \\ \rho \sigma_{\varepsilon_{Ai}} \sigma_{\varepsilon_{Bi}} & \sigma_{\varepsilon_{Bi}}^2 \end{pmatrix} \right) \quad i = 1, 2, \dots, k$$

donde X_i y Z_i son vectores de covariables relacionadas con la sensibilidad y la especificidad respectivamente.

Esta alternativa no necesita asumir la condición de que los tamaños n_{Ai} y n_{Bi} sean suficientemente grandes ni tampoco que $0 < p_{Ai} < 1$; $0 < p_{Bi} < 1$, porque en este caso las varianzas pueden ser aproximadas como:

$$Var(\hat{\mu}_{Ai}) = \frac{1}{n_{Ai}} + \frac{1}{n_{A0i}} \quad (25)$$

$$Var(\hat{\mu}_{Bi}) = \frac{1}{n_{Bi}} + \frac{1}{n_{B0i}} \quad (26)$$

con $\hat{\mu}_{Ai}$ y $\hat{\mu}_{Bi}$ obtenidas utilizando el modelo (24).

2.2.3.3 Modelo jerárquico HSROC

El modelo SROC jerárquico (HSROC) (Rutter y Gatsonis 2001) es un modelo no lineal generalizado mixto y constituye una generalización de un modelo de regresión binomial.

El modelo se formula en términos de la probabilidad π_{ij} de que un paciente en un estudio i con la enfermedad en el estado j , tenga un resultado positivo en la prueba, siendo $j = 0$ el estado correspondiente a que el paciente no tiene la enfermedad y $j = 1$ el correspondiente a que sí tiene la enfermedad. La tasa de verdaderos positivos, o sensibilidad, ahora se denota por π_{i1} y la tasa de falsos positivos por π_{i0} , equivalente a 1-especificidad. Adoptando la notación del resto de las secciones de este capítulo:

$$\pi_{ij} = \begin{cases} \pi_{i1} = p_{Ai} & j = 1 \\ \pi_{i0} = 1 - p_{Bi} & j = 0 \end{cases}$$

El modelo HSROC se define por dos ecuaciones para la variabilidad intra-estudio (nivel 1) y variabilidad entre estudios (nivel 2). La formulación bayesiana original (Rutter y Gatsonis 2001) permite además especificar los parámetros a priori del modelo.

El nivel 1 modeliza la variabilidad intra-estudio. Para un estudio i , el logit de la probabilidad de resultado positivo toma la siguiente forma:

$$\text{logit}(\pi_{ij}) = (\theta_i + \alpha_i X_{ij}) \exp(-\beta X_{ij}); i = 1, 2, \dots, k; j = 1, 2$$

$$\text{con } X_{ij} = \begin{cases} 1/2 & j = 1 \\ -1/2 & j = 0 \end{cases} \quad (27)$$

Los parámetros θ_i y α_i son específicos de cada estudio. El parámetro θ_i definido como “umbral de positividad” modeliza la relación entre sensibilidad y especificidad, al incrementar θ_i se incrementa tanto la tasa de verdaderos positivos como la de falsos positivos. El parámetro α_i definido como “parámetro de exactitud” mide la diferencia entre la fracción de verdaderos positivos y falsos positivos. Cuando el coeficiente β es nulo, el ORD de cada estudio no depende del parámetro de umbral, θ_i , y en ese caso α_i es igual al logaritmo del ORD. Cuando $\beta \neq 0$ el ORD sí varía con el umbral, incluso si permanece fijo el parámetro α_i . Como la estimación del parámetro β requiere utilizar información de varios estudios, se considera el caso más sencillo, es decir, se asume que es constante para todos los estudios.

El nivel 2 modeliza la variación de los parámetros θ_i y α_i entre estudios. En el caso más sencillo se asume que ambos siguen una distribución normal, cuyas medias

pueden ser representadas por una función lineal de las covariables del estudio. Por ejemplo, para una única covariable Z que afecte tanto al umbral, como a los parámetros de exactitud las distribuciones se expresan como:

$$\theta_i \sim N(\Theta + \gamma Z_i, \sigma_\theta^2) \quad (28)$$

$$\alpha_i \sim N(\Lambda + \lambda Z_i, \sigma_\alpha^2) \quad (29)$$

donde los coeficientes γ y λ expresan el efecto de la covariable Z en los parámetros de umbral (θ_i) y exactitud (α_i) respectivamente, siendo Θ y σ_θ^2 , la media y varianza de la distribución del parámetro θ_i y siendo Λ y σ_α^2 , la media y varianza de la distribución del parámetro α_i .

En general estos modelos jerárquicos se pueden extender para incluir covariables a nivel de estudio que pudieran explicar la variabilidad entre estudios. Es lo que se conoce con el término de metarregresión.

2.3 Estadísticos de análisis y cuantificación de la heterogeneidad

Un aspecto crucial en el metanálisis es la evaluación rigurosa de la heterogeneidad. La interpretabilidad de los resultados de exactitud diagnóstica tiene que apoyarse en las conclusiones sobre la variabilidad presente en la revisión.

A continuación, se describen las distintas propuestas de análisis y cuantificación de la heterogeneidad en función de si los modelos de síntesis, descritos en el Apartado 2.2, son univariantes o bivariantes.

2.3.1 Métodos de cuantificación de la heterogeneidad de los modelos

univariantes

La primera evaluación de la heterogeneidad es la que se realiza gráficamente, bien utilizando los denominados *forest plots* o bien representando la sensibilidad y la especificidad de los estudios incluidos en la revisión en el plano ROC. En los modelos univariantes, uno de los métodos estadísticos para evaluar la presencia de heterogeneidad es el Q-test, que pertenece a la familia de los tests con distribución chi-cuadrado. Una limitación importante es que tiene poca potencia para detectar heterogeneidad cuando hay pocos estudios y excesiva cuando el número de estudios es grande (Hardy y Thompson 1998; Jackson 2006).

Sin embargo, dado que la presencia de variabilidad es inevitable, seamos o no capaces de detectarla con un test estadístico, el foco debe dirigirse no tanto en probar que existe la heterogeneidad sino en cuantificarla, tratar de identificar las fuentes que la originan y valorar su impacto en los resultados del metanálisis (Higgins et al. 2003). Para cuantificar la heterogeneidad en un modelo univariante de efectos aleatorios lo más inmediato sería utilizar la varianza entre-estudios, τ^2 , definida en el Apartado 2.2.1, pero tiene un problema de interpretabilidad dado que está en escala logit.

El estadístico más utilizado para cuantificar la heterogeneidad es el I^2 (Higgins y Thompson 2002). Calcula la razón entre el estimador de la varianza entre estudios, $\hat{\tau}^2$, y la suma del estimador de la varianza intra-estudio, $\hat{\sigma}^2$, y la varianza anterior. Se interpreta como la proporción de la variabilidad total que es debida a la variabilidad entre estudios.

$$I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2} \quad (30)$$

Tiene una ventaja frente a considerar solo la varianza entre estudios y es que se expresa en escala de porcentaje. Cuando la varianza entre estudios es cero, el estadístico I^2 es igual a 0%, en el otro extremo, cuando la varianza entre estudios es mucho mayor que la varianza intra-estudio, el estadístico I^2 se aproxima al 100%.

Los mismos autores hicieron otras propuestas para cuantificar la heterogeneidad, como son los estadísticos H^2 y R^2 (Higgins y Thompson 2002) y a partir de estos se derivaron otras alternativas de I^2 univariante (White 2011; Zhou y Dendukuri 2014). Tal y como se discute en el Capítulo 3, los test univariantes dependen del tamaño del estudio (Rücker et al. 2008), del número de enfermos y no enfermos. Su uso está desaconsejado y solo estaría justificada su utilización en aquellos casos en los que sí sea apropiado considerar métodos de síntesis univariante, por ejemplo, en ausencia de correlación o efecto umbral entre sensibilidad y especificidad.

2.3.2 Métodos de cuantificación de la heterogeneidad de los modelos bivariantes

2.3.2.1 Varianza de los logit de sensibilidad y especificidad

Las varianzas entre estudios de los logit de sensibilidad y especificidad que se obtienen a partir de los modelos jerárquicos (Reitsma et al. 2005; Chu y Cole 2006; Rutter y Gatsonis 2001) son las medidas de cuantificación recomendadas para medir la heterogeneidad en revisiones sistemáticas de diagnóstico (Macaskill et al. 2010; Rücker et al. 2008; Zhou y Dendukuri 2014). Presentan una importante ventaja y es que no

dependen del tamaño muestral de los estudios. Sin embargo, tienen la limitación de la dificultad de interpretación por estar en escala logit.

2.3.2.2 Elipse de predicción

La elipse de predicción es una representación bidimensional de la región de predicción del 95% alrededor del estimador resumen de los logit de sensibilidad y especificidad. Leeflang et al. (2012) propusieron cuantificar la heterogeneidad calculando el área de esta elipse. Es decir:

$$\text{área} = \pi ab \quad (31)$$

Siendo a y b los ejes menor y mayor de la elipse, respectivamente. La longitud de ambos ejes depende de las varianzas de los logit de sensibilidad y especificidad y de la correlación entre ambos. Esta área tiene la misma dificultad de interpretación que las varianzas porque igualmente está dada en escala logit. Además, no permite determinar el nivel de confianza de la estimación.

2.3.2.3 I^2 bivalente

Los I^2 bivalentes (I^2_{Biv}) o multivalentes (Jackson, White y Riley 2012; Zhou y Dendukuri 2014) representan una extensión del I^2 univariante, y permiten cuantificar la heterogeneidad, teniendo en cuenta la correlación entre índices. En concreto en el Capítulo 3 se utilizará el propuesto por Zhou y Dendukuri (2014) que se calcula como:

$$I_{Biv}^2 = \frac{|\hat{\Sigma}|^{1/2}}{|E(\hat{\Sigma}_D)|^{1/2} + |\hat{\Sigma}|^{1/2}} \quad (32)$$

Siendo $\hat{\Sigma} = \begin{pmatrix} \hat{\sigma}_A^2 & \hat{\rho}\hat{\sigma}_A\hat{\sigma}_B \\ \hat{\rho}\hat{\sigma}_A\hat{\sigma}_B & \hat{\sigma}_B^2 \end{pmatrix}$ el estimador de la matriz de varianzas-covarianzas entre estudios definida en (20), $\hat{\Sigma}_D$ el estimador de la matriz de varianzas-covarianzas intra-estudio

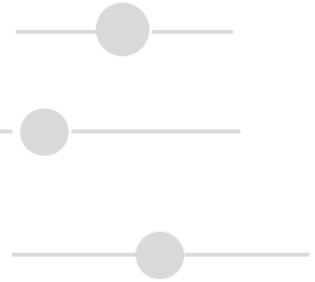
$$\hat{\Sigma}_D = \begin{pmatrix} \hat{\sigma}_{Ai}^2 & 0 \\ 0 & \hat{\sigma}_{Bi}^2 \end{pmatrix} \quad (33)$$

$|E(\hat{\Sigma}_D)|$ el determinante de la esperanza de la matriz $\hat{\Sigma}_D$ y $|\hat{\Sigma}|$ el determinante de la matriz $\hat{\Sigma}$.

De forma similar al I^2 univariante, el rango del I_{Biv}^2 varía de 0 a 1. Un valor elevado indica una alta heterogeneidad entre estudios comparada con la variabilidad intra-estudio esperada, pero también tiene la limitación de que depende del tamaño de los estudios incluidos.

En el Capítulo 3 se expone el estudio meta-epidemiológico que se ha realizado con el objetivo de investigar los métodos usados en la literatura científica para evaluar y cuantificar la heterogeneidad estadística en metanálisis de evaluación de exactitud diagnóstica y se presentan dos aportaciones metodológicas nuevas para complementar el estudio de la cuantificación en este tipo de revisiones.

Capítulo 3



Capítulo 3

En este capítulo, se presenta el estudio meta-epidemiológico realizado con el objetivo de describir los métodos usados para la cuantificación de la heterogeneidad en revisiones sistemáticas de evaluación de pruebas diagnósticas publicadas en la Cochrane Library⁴. Los métodos para evaluar y reportar la heterogeneidad entre estudios en revisiones sistemáticas que evalúan el rendimiento de pruebas diagnósticas no están completamente estandarizados. Se proponen medidas alternativas de cuantificación de la heterogeneidad para facilitar su interpretabilidad.

⁴ Página web Cochrane Library: www.cochranelibrary.com

3.1 Objetivos

Describir los métodos usados para cuantificar la heterogeneidad en revisiones sistemáticas de evaluación de pruebas diagnósticas publicadas por la Cochrane Library y proponer medidas alternativas de cuantificación de la heterogeneidad.

3.2 Metodología

En primer lugar, se realizó un estudio meta-epidemiológico que incluyó todas las revisiones sistemáticas de exactitud diagnóstica publicadas en la Cochrane Library hasta el 6 de octubre de 2019 utilizando el filtro de “*diagnosis*”. De cada revisión se extrajeron características relativas a autores, grupo editorial, año de publicación, tipo de prueba diagnóstica, métodos de metanálisis y de cuantificación de la heterogeneidad y, se extrajeron las tablas 2x2 del análisis principal de la revisión (número de verdaderos positivos, n_{Ai} , falsos positivos, n_{Bi} , falsos negativos, n_{AOi} , verdaderos negativos, n_{BOi}).

Una vez analizadas las características de las revisiones sistemáticas incluidas en este trabajo se proponen dos alternativas de cuantificación de la heterogeneidad, calcular: i. el área de la elipse de predicción en el plano ROC en referencia al área total del plano y ii. la mediana de OR que traslada la varianza de los logit de sensibilidad y especificidad a escala de OR.

En una segunda fase, se seleccionaron aquellas revisiones con un número de estudios mayor que 4 y que habían utilizado un modelo bivalente para la síntesis de resultados. Este grupo de revisiones se reanalizó utilizando modelos bivariantes de metanálisis y se caracterizó la heterogeneidad utilizando las varianzas de los logit de

sensibilidad y especificidad, el I^2 bivalente (Zhou y Dendukuri 2014), la odds ratio (OR) mediana para sensibilidad y especificidad y el área de la elipse de predicción con una confianza del 95%.

3.3 Resumen de resultados

Inicialmente, se seleccionaron las 124 revisiones sistemáticas que evaluaban pruebas diagnósticas publicadas en la Cochrane Library a la fecha de la búsqueda. De éstas, 91 (73%) incluían un metanálisis, pero solo 26 aportaban medidas de heterogeneidad, en 21 (23%) metanálisis se ilustraba con la elipse de predicción y únicamente 5 comunicaban las varianzas de los logit de sensibilidad y especificidad.

El reanálisis de 60 revisiones mostró que, en la mayoría de las revisiones, los autores describen la heterogeneidad de forma subjetiva como moderada o extrema (31 de 60, 52%) y esto no se corresponde con las áreas de las elipses de predicción y las medianas de los OR de sensibilidad y especificidad que muestran una gran variabilidad, desde áreas muy pequeñas en dichos escenarios a muy altas (5% a 97%) y medianas de los OR de sensibilidad de 1.13 a 10.7 y para especificidad de 1.18 a 19.68. La rutina en R para el cálculo del área de la elipse de predicción se incluye como Anexo I de esta memoria.

3.4 Conclusiones

En las revisiones de exactitud diagnóstica la descripción de la heterogeneidad es irregular e insuficiente y puede caracterizarse mejor introduciendo nuevas medidas de cuantificación. Proponemos utilizar, de forma complementaria a las varianzas de los logit

de sensibilidad y la especificidad, el área de la elipse de predicción al 95% y el OR mediano para sensibilidad y especificidad como medidas útiles para mejorar la comunicación y la interpretabilidad de la heterogeneidad presente en este tipo de revisiones.

3.5 Publicación

Se incluye como cuerpo del capítulo el texto completo del estudio publicado en la revista *Journal of Clinical Epidemiology* (Factor de impacto año 2020: 6.437; rango: Q1-primer decil, 5/108 de la categoría *Health Care Sciences and Services*):

Plana MN, Pérez T, Zamora J. New measures improved the reporting of heterogeneity in diagnostic test accuracy reviews: a meta-epidemiological study. *J Clin Epidemiol*. 2021 Mar;131:101-112.

ORIGINAL ARTICLE

New measures improved the reporting of heterogeneity in diagnostic test accuracy reviews: a metaepidemiological study

María Nieves Plana^{a,b,c,*}, Teresa Pérez^{d,e}, Javier Zamora^{b,f}^aDepartment of Preventive Medicine and Public Health, Príncipe de Asturias University Hospital, Alcalá de Henares, Madrid, Spain^bCIBER Epidemiology and Public Health, Clinical Biostatistics Unit, Ramón y Cajal University Hospital, IRYCIS, Madrid, Spain^cFaculty of Statistics Studies, Complutense University of Madrid, Madrid, Spain^dDepartment of Statistics and Data Science, Complutense University of Madrid, Madrid, Spain^eBarts Research Centre for Women's Health, WHO Collaborating Centre, Queen Mary University of London, London, UK^fWHO Collaborating Centre for Global Women's Health, Institute of Metabolism and Systems Research, Institute of Applied Health Research, University of Birmingham, Birmingham, UK

Accepted 13 November 2020; Published online 21 November 2020

Abstract

Objectives: To describe the methods used to quantify heterogeneity and to propose alternative measures to improve reporting of heterogeneity in Cochrane diagnostic test accuracy (DTA) reviews.

Study Design and Setting: Our metaepidemiological study included all DTA reviews in the Cochrane Library up to October 6th, 2019. We summarized reviews' characteristics focusing on heterogeneity analysis. We selected reviews with a bivariate model and ≥ 4 studies for reanalysis. In this group, we fitted bivariate random effects models and we quantified heterogeneity by means of logit variances of sensitivity and specificity, bivariate I^2 , median odds ratio (OR), and the area of the 95% prediction ellipse. We provided a narrative interpretation of these measures in different scenarios.

Results: There were 124 Cochrane DTA reviews of which 91 (73%) included meta-analysis. Only in 5 meta-analyses, variances of the logit sensitivity and specificity were reported, and in 21 meta-analyses (23%), the 95% prediction ellipse was reported without any calculation of its area. We selected 60 of these 91 reviews to explore the behavior of all measures of heterogeneity. We found that most reviews described the subjective heterogeneity as moderate or extreme ($n = 31/60$, 52%), whereas the area of the 95% prediction ellipse and the median OR for sensitivity and specificity showed high variability; the area ranged from 5% to 97%, the median OR of sensitivity ranged from 1.13 to 10.7, and the median OROR of specificity ranged from 1.18 to 19.68.

Conclusion: Cochrane DTA reviews show a poor reporting of between-study heterogeneity. Using median OR and the area of the 95% prediction ellipse will improve reporting and interpretation of this crucial aspect of DTA meta-analysis. © 2020 Elsevier Inc. All rights reserved.

Keywords: Heterogeneity; Diagnostic test accuracy; Meta-analysis; Systematic review

1. Background

Systematic reviews (SRs) of diagnostic test accuracy (DTA) face a particular challenge: dealing with greater heterogeneity than that expected by chance. The more unexplained variability there is in accuracy between

studies beyond chance, the more difficult it is to come to robust conclusions about the clinical implications of a DTA meta-analysis. Frequently, published SRs limit themselves to obtaining summary estimates of sensitivity and specificity without a deeper analysis and quantification of the amount of between-studies heterogeneity. Dissecting and quantifying this variability should be mandatory in any meta-analysis. Furthermore, clear reporting of the amount of heterogeneity offers valuable help to readers in interpreting the accuracy of a test in the appropriate context.

Methods to evaluate and report statistical heterogeneity in DTA reviews are not entirely standardized. Thus, the

Competing interests: María N Plana, Teresa Pérez, and Javier Zamora declare no competing interests.

* Corresponding author. Department of Preventive Medicine and Public Health, Príncipe de Asturias University Hospital, Alcalá de Henares, Madrid, Spain.

E-mail address: planafn@gmail.com (M.N. Plana).

What is new?

Key findings

- Reporting of heterogeneity in Cochrane Diagnostic Test Accuracy (DTA) reviews is poor.
- Author's subjective judgments concerning the heterogeneity are often inconsistent with the statistically estimated heterogeneity.

What this adds to what was known?

- We propose complementary statistics to improve reporting of heterogeneity in DTA systematic reviews.
- Reporting the area of the prediction ellipse allows interpreting heterogeneity as a proportion of the total area of the ROC space.
- Median Odds Ratio can facilitate interpretation of heterogeneity.

What is the implication and what should change now?

- DTA systematic reviewers should consider using these additional statistics to better report the quantification of heterogeneity in their reviews.

different methods that have penetrated into current reviews are very heterogeneous. Naaktgeboren et al. [1] proposed a structured way for evaluation of heterogeneity in DTA reviews in five steps: 1) visualizing total variability; 2) judging whether there is more variability in sensitivity and specificity than that expected by chance; 3) measuring the total between-study variability; 4) attributing the between-study variability due to threshold effect; and 5) exploring what study features might explain some of the between study variability. The evaluation of the approaches used in DTA reviews showed that most reviews still use unrecommended univariate methods to quantify heterogeneity [1], such as Cochran's Q statistic [2] or Higgins-Thompson I^2 statistic separately for sensitivity and specificity [3]. Univariate heterogeneity statistics had limitations related to their dependence on study sample size [4] and are difficult to interpret as all heterogeneity may be caused by threshold effects. Multivariate extensions of I^2 statistics have been proposed to account for the correlation between sensitivity and specificity [5,6], but they still show a dependence on sample size and are not commonly used in practice.

Heterogeneity can be evaluated by visually exploring forest plots and the scatter plot of sensitivities and specificities in the receiver operating characteristics (ROC) plane. Estimates of the random effect terms in the bivariate model [7–9] quantify the amount of heterogeneity in terms of

between-study variances (τ^2). Leeflang et al. [10] proposed calculating the area of the prediction ellipse in a logit scale. Despite these advances, quantification of heterogeneity is rarely reported in DTA reviews as noted by Naaktgeboren [1], and more importantly, it is rarely interpreted or discussed. Of 53 non-Cochrane DTA reviews evaluated, 7 (13%) presented between-study variances and only 3 of these showed the prediction ellipse without any quantification of its area [1]. There is a need for developing better ways to assess and report heterogeneity in DTA syntheses.

In our study, we systematically investigated the methods used to evaluate and quantify statistical heterogeneity in Cochrane DTA SRs. Readers interested in a more comprehensive discussion on other sources of heterogeneity (i.e., clinical, methodological, and the heterogeneity due to threshold effect) are invited to read the Cochrane handbook for DTA systematic reviews [11]. We inspected whether heterogeneity analysis influenced the decision to perform meta-analysis. Using data collected in Cochrane DTA reviews, we also empirically examined the behavior of some proposed alternative measures of variability in DTA reviews to improve the interpretability of heterogeneity.

2. Methods

2.1. Study design

This is a metaepidemiological study, reported as per the adapted preferred reporting items for systematic reviews and meta-analyses statement [12].

2.2. Search strategy

We searched Cochrane Library using the “diagnosis” filter to identify all Cochrane DTA reviews published from its inception to October 6, 2019.

2.3. Data extraction

We designed and piloted a data collection form to extract information on review characteristics (author, editorial group, year of publication, type of test, methods for meta-analysis, and evaluation of heterogeneity) and the measures of accuracy (number of true positives, false positives, false negatives, and true negatives) of the main analysis in the review. If multiple meta-analyses were reported in the same SR, we extracted data from the first meta-analysis reported in the abstract. Two reviewers independently conducted a pilot study on five SRs to agree on the interpretation of the form and to solve difficulties during data extraction.

2.4. Statistical analysis

We summarized DTA reviews' characteristics (type of diagnostic test, year of publication, and Cochrane editorial group), the statistical method for conducting the meta-

analysis (univariate, bivariate, and hierarchical summary ROC (HSROC)), and the degree of completeness in heterogeneity analysis. We also noted whether the authors mentioned in methods, results, or discussion if they deployed assessment of heterogeneity to guide their meta-analytical approach. From each review, we also abstracted in verbatim form, the text reflecting the authors' subjective assessment of the heterogeneity.

We selected reviews with a bivariate model and ≥ 4 studies for reanalysis. In this group, we fitted bivariate random effects models and we quantified heterogeneity by means of logit variances of sensitivity and specificity, bivariate I^2 , median odds ratio (OR), and the area of the 95% prediction ellipse. We computed the measures to quantify heterogeneity using data extracted from Cochrane DTA reviews.

To demonstrate the behavior of the proposed heterogeneity measures empirically under specific distributions of individual accuracy results, we selected four exemplar reviews: one with high variability in specificity but not in sensitivity; one with low variability in both sensitivity and specificity; one with variation in sensitivity but not in specificity, and one with high level of variability in both accuracy measures.

2.4.1. Area of the 95% prediction ellipse

The area of the 95% prediction ellipse in the ROC plane could be interpreted in reference to the total area of that plane. Unlike previous attempts made to quantify this area in the logit scale [10], we propose a more straightforward interpretation in the ROC space. The R code to perform this computation is provided in [Appendix 1](#).

2.4.2. Median odds ratio

The median OR is a measure commonly used in epidemiological research to analyze the variability between sampling units when these units are clusters of individuals e.g., cities, neighborhoods, clinics, etc. To properly account for the sampling design, analysis of such epidemiological studies is frequently based on multilevel random effects models. These models estimate the effect on outcomes (i.e., a binary response) of either participant level covariates or contextual factors evaluated at cluster level. In addition, with these models, researchers can explore the components of variance attributable to each level [13,14].

Median OR can be easily extended to the DTA meta-analysis framework [7,9]. Studies in a meta-analysis are the same as the clusters of individuals in epidemiological research. Median OR is a translation of between-study variance of accuracy on an odds ratio (OR) scale. We propose a median OR for sensitivity and another median OR for specificity. The median OR values range from 1 (no heterogeneity) to infinite (high heterogeneity). It can be calculated easily as a simple function of the between-study variance of logit sensitivity and specificity (τ^2):

$$\text{Median OR} = \exp\left(0.6745 \sqrt{2\tau^2}\right)$$

A theoretical derivation of this formula is provided by Larsen and Merlo [15].

The median OR for sensitivity can be interpreted as the increment in the odds of being “true positive” (in median) if a diseased individual would change from one study with lower sensitivity to another study with higher sensitivity. Similarly, median OR for specificity can be interpreted as the increment in the odds of being “true negative” (in median) if a nondiseased individual would change from one study with lower specificity to another study with higher specificity.

Our empirical comparison of heterogeneity measures was restricted to those SRs with at least one meta-analysis including more than 3 studies. We replicated the meta-analysis performed in each SR. A bivariate model was fitted [7–9], and the following parameters were estimated: summary estimates of sensitivity and specificity (logit (sensitivity) and logit (specificity)) with their variances and the correlation between them.

2.4.3. Bivariate I-squared

We also calculated the bivariate I^2 (I^2_{Biv}) statistic proposed by Zhou et al. [6] under a Bayesian framework using jargs R package [16]. The I^2_{Biv} takes into account simultaneously the between-study variances of logit sensitivity and specificity as well as the correlation among both indices, which would otherwise be ignored by univariate I^2 statistics. Alike other univariate I-squared statistics, the bivariate version is also affected by study sample size. We empirically evaluated the association of effective sample size with I^2_{Biv} using Spearman rho correlation coefficient. The interpretation of this statistic is similar to its univariate counterparts. It ranges between 0 and 1 and higher values indicate higher between-study variability across studies than the expected within-study variability [6].

3. Results

3.1. Characteristics of included DTA SRs

We evaluated all DTA SRs included in the Cochrane Library up to October 6, 2019 as it is shown in [Figure 1](#). [Appendix 2](#) shows the list of the 124 DTA Cochrane reviews included. Cochrane DTA reviews were more frequent in the following editorial groups: Dementia and Cognitive Improvement, 20 (16%); Skin, 10 (8%); Infectious Diseases, 10 (8%); and Hepatic–biliary, 10 (8%). In particular, 52 (42%) reviews evaluated imaging tests, 42 (34%) assessed laboratory tests, and 30 (24%) other types of tests. Of the total SRs, 90% were published between 2014 and 2019 ([Table 1](#)). One-quarter of the reviews ($n = 33$; 27%) provided a narrative synthesis. In 18 of these (55%), authors reported the presence of heterogeneity as

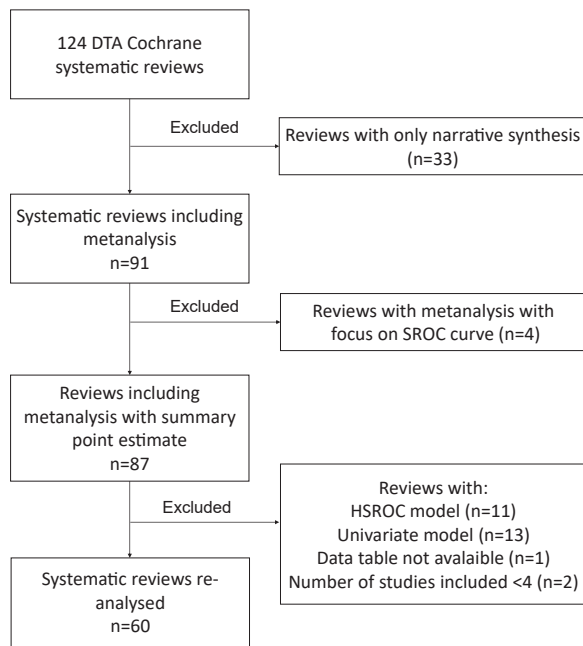


Fig. 1. Flow diagram representing the selection process of Cochrane diagnostic accuracy reviews that were included in the heterogeneity reanalysis.

the main reason for not performing a meta-analysis. In most of them ($n = 17$, 94%), the authors stated that they did not do so because there was too much clinical or methodological heterogeneity, only in one review, the decision is based on statistical assessment of heterogeneity.

Most of the reviews included a meta-analysis ($n = 91$; 73%) and around half of them provided a pooled accuracy estimate separately for sensitivity and specificity ($n = 53$; 58%) or only one index ($n = 4$; 4%), a summary ROC curve ($n = 4$; 4%) or both (sensitivity and specificity and ROC curve) ($n = 30$; 33%). Software used was SAS ($n = 38$, 42%), Stata ($n = 32$, 35%), or R ($n = 6$, 7%) exclusively; in 4 reviews (4%), it was not reported and in 11 (12%) used a combination. For those SRs obtaining summary measures (sensitivity, specificity, or both ($n = 87$)), the bivariate model was used ($n = 63$, 72%) more often than the HSROC ($n = 11$, 13%). In 13 SRs (15%), a univariate meta-analysis is performed (2 fixed effects and 11 random effects) to obtain a pooled estimate (Table 2).

3.2. Degree of completeness in heterogeneity analysis

Table 2 shows the results of the evaluation of the degree of completeness in heterogeneity analysis in Cochrane DTA reviews that included at least one meta-analysis ($n = 91$). All SRs graphically showed forest plots with point estimates and 95% confidence intervals (CI) for sensitivity and specificity for individual studies included in the meta-analysis. Of these, 81 (89%) included a scatter plot in the ROC

Table 1. Characteristics of the DTA Cochrane reviews

All reviews ($n = 124$)	N (%)
Type of diagnostic test	
Image test	52 (41.9)
Laboratory test	42 (33.9)
Clinical examination	12 (9.7)
Questionnaire scales	12 (9.7)
Other	6 (4.8)
Year of publication	
2008–2013	12 (9.7)
2014–2019	112 (90.3)
Cochrane editorial group	
Dementia and Cognitive Improvement	20 (16.1)
Skin	12 (9.7)
Infectious Diseases	10 (8.1)
Hepato–biliary	10 (8.1)
Gynecological, Neuro-oncology orphan Cancer	9 (7.3)
Upper GI and Pancreatic Diseases	9 (7.3)
Pregnancy and Childbirth	7 (5.7)
Gynecology and Fertility	5 (4.0)
Kidney and Transplant	4 (3.2)
Back and Neck	4 (3.2)
Vascular	4 (3.2)
Bone, Joint and Muscle Trauma	3 (2.4)
Eyes and Vision	3 (2.4)
Neonatal	3 (2.4)
Stroke	3 (2.4)
Schizophrenia	2 (1.6)
Oral Health	2 (1.6)
Acute Respiratory Infectious	2 (1.6)
Airways	2 (1.6)
Anesthesia, Critical, and Emergency Care Group	2 (1.6)
Injuries	2 (1.6)
Childhood Cancer	1 (0.8)
Colorectal Cancer	1 (0.8)
Developmental, Psychosocial and Learning Problems	1 (0.8)
Emergency and Critical Care	1 (0.8)
Lung Cancer	1 (0.8)
Urology	1 (0.8)

plane. In 35 (38%), different symbol sizes were used for each point estimate to represent its precision, and in 9 (10%) crosshairs, graphs were included with the same aim. In 10 (11%) of the SRs, heterogeneity was only assessed by visual inspection of forest plots. Among SRs that reported only one pooled estimate, (i.e., sensitivity), just one reported the value of the test for heterogeneity (Cochran's Q test). In 5 (5%) SRs, heterogeneity was quantified reporting the variances of the logit transformed of both

Table 2. Heterogeneity analysis in DTA reviews with meta-analysis

Degree of completeness in heterogeneity analysis, <i>n</i> = 91 (73%)	N (%)
Step 1. Visualize total variability	
Forest plot	10 (11)
Forest plot and ROC plane	81 (89)
The review shows the precision of the estimates in the ROC plane using	
Crosshairs graphs	9 (10)
Varying the size of the points	35 (38)
Step 2. Judge whether there is more variability in sensitivity and specificity than can be expected due to chance alone	
Graphical	76 (84)
Statistical (Cochran's Q test)	1 (1)
Step 3. Measure the total between-study variability	
Univariate fixed model (<i>n</i> = 2)	
Univariate random model (11)	
I^2	1 (1)
I^2 and H^2	1 (1)
τ^2	1 (1)
Bivariate model (<i>n</i> = 63)	
HSROC model (<i>n</i> = 11)	
Prediction ellipse	21 (23)
Logit variances (sen and spe)	5 (5)
Step 4. Attribute some of the between-study variability to the threshold effect	
Threshold effect	42 (46)
Step 5. Explore what study features might explain some of the variability	
Type of analysis proposed in methods section	
Metaregression	27 (30)
Subgroup analysis	6 (7)
Both	55 (60)
Heterogeneity results were used by the authors to guide their analytical approach (reported by the authors)	
To decide the analysis method (univariate vs. bivariate, fixed vs. random univariate model)	19 (21)
Variability of diagnostic thresholds across studies was used to select the type of analysis (reported by the authors)	
To estimate an SROC curve or a summary point	24 (26)

sensitivity and specificity. In 21 (23%), the 95% prediction ellipse was included but without calculating its area. In 42 SRs (46%), authors attributed part of the heterogeneity to the threshold effect without any specific analysis. All SRs stated that they planned to explore sources of heterogeneity, but in 34 (37%), it was not possible due to the small number of studies included.

3.3. Heterogeneity results after reanalyzing a subgroup of cochrane DTA SR

For reanalysis, 60 SRs had at least one meta-analysis with ≥ 4 studies (Figure 1). Tables 3 and 4 provide details of the bivariate analysis results along with the different measures for quantifying heterogeneity we proposed in methods section. Most SRs describe the observed heterogeneity as moderate or extreme ($n = 31$, 52%). The different heterogeneity measures showed high variability. The I^2_{Biv} ranged from 0.090 to 0.920, and the area of the 95% prediction ellipse ranged from 5% to 97%. The I^2_{Biv} was highly correlated with study effective sample size (Spearman $\rho = 0.61$; $P < 0.0001$). Variances of logit sensitivity and specificity ranged from 0.017 to 6.173 and from 0.029 to 9.756, respectively. The median OR of sensitivity ranged from 1.13 to 10.7, and the median OR of the specificity ranged from 1.18 to 19.68 (Table 5).

In those SRs that described heterogeneity as little, minimal, or some heterogeneity ($n = 5$, 8%), the I^2_{Biv} ranged from 0.01 to 0.77 and the area of the 95% prediction ellipse ranged from 1 to 38%. Variances of logit sensitivity and specificity ranged from 0 to 1.424 and from 0.409 to 1.915, respectively. The median OR of sensitivity ranged from 1 to 3.12, and the median OR of the specificity ranged from 1.84 to 3.74 (Table 5).

In 8 SRs (13%), the authors' subjective evaluation of the heterogeneity was focused only in the variability of sensitivity. In this subgroup, the I^2_{Biv} ranged from 0 to 0.91 and the area of the 95% prediction ellipse ranged from 0.8 to 67%. Variances of logit sensitivity and specificity ranged from 0.102 to 2.186 and from 0.001 to 6.038, respectively. The median OR of sensitivity ranged from 1.36 to 4.10, and the median OR of the specificity ranged from 1.03 to 10.42 (Table 5).

In those SRs where no subjective assessment of heterogeneity was reported ($n = 16$, 27%), the I^2_{Biv} ranged from 0 to 0.770 and the area of the 95% prediction ellipse ranged from 5% to 70%. Variances of logit sensitivity and specificity ranged from 0.090 to 2.803 and from 0.231 to 6.079, respectively. The median OR of the sensitivity ranged from 1.33 to 4.94, and the median OR of the specificity ranged from 1.58 to 10.51 (Table 5).

Figure 2 illustrates the behavior of the proposed measures for quantifying heterogeneity in the four exemplar reviews selected.

In the first exemplar SR (Nicholson BD, 2015 [17]), authors made a subjective assessment of "considerable heterogeneity." The variances of the logit sensitivity and specificity are 0.017 and 1.733, respectively. The odd of being "true positive" is 1.13 times more likely (in median) if a diseased individual is changed from one study with lower sensitivity to another study with higher sensitivity. Conversely, the odd of being "true negative" is 3.51 times more likely (in median) if a nondiseased individual is changed from one study with lower specificity to another

Table 3. Bivariate results including heterogeneity statistical estimates

Review data						Summary results		
	ID review	studies (n)	Sen (min; max)	Spe (min; max)	dis/nondis	Sen	Spe	rho
1	CD012233	6	0.448; 1	0.8; 0.985	220/16,040	0.84	0.96	−0.89
2	CD007394	27	0.00; 1	0.208; 1	394/3549	0.77	0.85	−0.29
3	CD007431	10	0.786; 0.98	0.1; 0.82	5,457/1,104	0.93	0.31	−0.15
4	CD008054	10	0.600; 0.959	0.396; 0.786	493/4,768	0.91	0.61	−0.53
5	CD008587	25	0.607; 1	0.638; 0.953	2432/135798	0.93	0.89	0.02
6	CD008759	10	0.560; 0.925	0.522; 0.976	723/1,331	0.71	0.80	−0.76
7	CD008760	15	0.645; 1	0.333; 1	640/296	0.85	0.84	−0.27
8	CD008803	51	0.375; 0.978	0.667; 1	3,755/3,438	0.77	0.90	−0.51
9	CD008874	24	0.000; 0.745	0.64; 1	1,220/20,959	0.22	0.94	−0.85
10	CD008892	14	0.563; 1	0.688; 0.96	627/4,258	0.79	0.88	−0.07
11	CD009020	6	0.912; 1	0.667; 1	263/84	0.98	0.80	1
12	CD009044	12	0.762; 0.984	0.2; 1	1,198/427	0.94	0.80	−0.77
13	CD009175	6	0.120; 0.435	0.549; 0.843	311/980	0.25	0.67	−1
14	CD009185	6	0.654; 1	0.25; 0.892	276/158	0.86	0.74	0.56
15	CD009372	8	0.833; 1	0.889; 1	180/346	0.94	0.99	−1
16	CD009519	18	0.467; 1	0.598; 1	819/2371	0.76	0.86	0.04
17	CD009551	33	0.222; 1	0.286; 1	628/4,361	0.79	0.79	−0.45
18	CD009579	74	0.155; 1	0.244; 1	25,195/77,252	0.75	0.87	−0.16
19	CD009591	8	0.828; 1	0.857; 1	198/567	0.93	0.95	−0.14
20	CD009593	86	0.429; 1	0.52; 1	11,904/30,187	0.85	0.98	−0.38
21	CD009647	4	0.000; 0.2	0.667; 1	66/64	0.17	0.91	NA
22	CD009944	50	0.118; 1	0; 1	2498/1,887	0.86	0.90	−0.40
23	CD010079	11	0.667; 0.909	0.617; 0.956	379/2265	0.80	0.85	−1
24	CD010276	15	0.561; 1	0; 0.933	628/710	0.84	0.70	−0.12
25	CD010296	16	0.333; 1	0.636; 1	305/830	0.84	0.92	0.14
26	CD010339	5	0.667; 0.941	0.923; 1	180/138	0.83	1.00	0.27
27	CD010360	38	0.636; 1	0.963; 1	3,200/7,981	0.90	1.00	0.29
28	CD010502	105	0.386; 1	0.541; 1	17,278/40,966	0.86	0.95	−0.17
29	CD010542	8	0.750; 1	0.8; 1	278/64	0.94	0.89	−1
30	CD010653	20	0.267; 0.778	0.388; 0.944	2657/2422	0.57	0.81	−0.55

^a Median.

study with higher specificity. The area of the 95% prediction ellipse is 0.10, and the value of the I^2_{Biv} is 0.52. Both statistics show moderate heterogeneity globally. Despite this, as per the median OR, the heterogeneity is restricted to specificity which is three times higher than the heterogeneity in sensitivity. This numerical assessment matches well what is seen graphically (Figure 2A).

In the second exemplar SR (Arbyn M, 2013 [18]), authors do not make any subjective statement about heterogeneity in the review. The variances of the logit sensitivity and specificity are quite similar (0.195 and 0.231). The median OR for sensitivity and median OR for specificity are both similar and close to 1. The area of the 95% prediction ellipse is 0.12, and the value of the I^2_{Biv} is 0.68. The apparent discordance between these measures could be a

consequence of the high sample size, especially of nondiseased individuals, that largely impacts on I^2_{Biv} .

In the third exemplar SR (Plana MN, 2018 [19]), authors made a subjective assessment of “greater variability in estimated sensitivity.” The variances of the logit sensitivity and specificity were 0.102 and 2, respectively. The odd of being “true positive” is 1.36 times more likely (in median) if a diseased individual is changed from one study with lower sensitivity to another study with higher sensitivity. Conversely, the median OR for specificity is 3.85, meaning that the odd of being “true negative” result is almost 4 times more likely (in median) if a nondiseased individual is changed from one study with lower specificity to another study with higher specificity. The area of the 95% prediction ellipse is 0.01, and the value of the I^2_{Biv} is 0.38. Both

Heterogeneity statistic estimates						
Var (logit sen)	Var (logit spe)	MOR sen	MOR spe	I^2_{Biv} ^a	Prediction ellipse area	Heterogeneity description (by authors)
2.186	0.868	4.1	2.43	0.64	0.091	“significant heterogeneity in sensitivity”
0.88	1.89	2.45	3.71	0.64	0.510	“very heterogeneous”
0.74	1.243	2.27	2.9	0.78	0.432	“large”
0.195	0.231	1.52	1.58	0.68	0.123	
1.236	0.306	2.89	1.69	0.92	0.143	“serious inconsistency”
0.203	0.669	1.54	2.18	0.69	0.211	
0.469	0.888	1.92	2.46	0.37	0.288	
0.617	0.409	2.12	1.84	0.75	0.134	“little variation”
1.933	2.175	3.77	4.08	0.75	0.401	
0.395	0.477	1.82	1.93	0.74	0.224	
0.586	0.029	2.08	1.18	0.24	0.146	“considerable”
0.982	0.851	2.57	2.41	0.66	0.151	“substantial variation”
0.255	0.278	1.62	1.65	0.79	0.147	“considerable heterogeneity”
0.695	0.847	2.21	2.41	0.63	0.680	“very high”
0.046	1.087	1.23	2.7	0.07	0.025	“minimal”
0.884	0.921	2.45	2.5	0.78	0.455	“substantial”
0.939	1.97	2.52	3.81	0.74	0.490	“large”
0.821	1.173	2.37	2.81	0.91	0.395	“greater heterogeneity for sensitivity”
0.242	0.292	1.6	1.67	0.33	0.057	“high”
0.69	1.619	2.21	3.37	0.64	0.102	“some variability particularly in sensitivity”
0	1.915	1	3.74	0.08	NA	“some heterogeneity”
1.41	0.707	3.1	2.23	0.54	0.216	“large”
0.09	0.495	1.33	1.96	0.56	0.050	“substantial”
1.039	0.61	2.64	2.11	0.63	0.486	“considerable variation”
2.205	0.847	4.12	2.41	0.52	0.441	
0.214	6.038	1.55	10.42	0.05	0.674	“heterogeneity mainly in sensitivity”
0.358	0.757	1.77	2.29	0.36	0.010	“homogeneous enough to combine results”
0.655	0.753	2.16	2.29	0.78	0.102	“substantial”
0.576	0.001	2.06	1.03	0.1	0.121	“no evidence of heterogeneity in sensitivity”
0.287	0.832	1.67	2.39	0.8	0.305	“large”

heterogeneity measures show relatively low heterogeneity globally. Because specificity is close to 1 in all studies, small variations in specificity are accompanied by large variations in its logit. Thus, the variance of logit specificity is high in contrast to what could be expected after visual assessment of Figure 2C. Similar situations are observed in the SR of Ochodo EA, 2015 [20] (CD009579), Horne DJ, 2019 [21] (CD009593), and Gurusamy KS, 2015 [22] (CD010339) (data shown in Tables 3 and 4).

Finally, in the fourth exemplar SR of Leeftang MMG, 2015 [23], authors classify the scenario as “very heterogeneous.” The variance of the logit sensitivity (0.88) is about half the variance of the logit specificity (1.89). The odd of being “true positive” is 2.45 times more likely (in median) if a diseased individual is changed from one study with

lower sensitivity to another study with higher sensitivity. Conversely, the median OR for specificity is 3.71. The area of the 95% prediction ellipse exceeds 50% (0.510), and the value of the I^2_{Biv} is 0.64. Both statistics show high heterogeneity globally. This can be appreciated also in the large median OR for both indices that are higher than 2. This is concordant with the scatter plots in Figure 2D.

4. Discussion

We have presented a thorough exploration of how Cochrane DTA reviews deal with heterogeneity analysis. We focused on Cochrane DTA reviews as we expect them to adhere to the highest methodological standards. However,

Table 4. Bivariate results including heterogeneity statistical estimates

Review data						Summary results		
	ID review	studies (n)	Sen (min max)	Spe (min max)	dis/nondis	Sen	Spe	rho
31	CD010657	20	0.122; 0.909	0.146; 1	1016/2710	0.44	0.78	−0.84
32	CD010705	19	0.571; 1	0.773; 1	869/1,354	0.86	0.99	0.37
33	CD010772	13	0.708; 1	0.39; 0.88	1,403/1,342	0.91	0.66	−0.23
34	CD011121	7	0.733; 0.992	0.112; 0.928	1,016/2643	0.92	0.64	−0.54
35	CD011134	7	0.650; 0.909	0.339; 0.978	444/1,071	0.82	0.80	−1
36	CD011145	15	0.362; 1	0.298; 0.991	1,133/9,835	0.85	0.90	−0.54
37	CD011420	8	0.227; 0.682	0.752; 1	1,277/2172	0.42	0.92	−0.64
38	CD011427	5	0.825; 0.957	0.25; 0.93	877/715	0.88	0.66	0.43
39	CD011431	11	0.667; 0.9	0.89; 1	958/5,921	0.78	0.99	−0.42
40	CD011548	5	0.320; 1	0.769; 0.974	162/361	0.73	0.91	−0.48
41	CD011549	13	0.750; 1	0.851; 1	686/851	0.95	0.97	0.67
42	CD011551	7	0.333; 1	0; 1	83/32	0.83	0.48	−0.66
43	CD011871	15	0.134; 1	0; 1	660/1,224	0.70	0.95	−0.37
44	CD011901	8	0.200; 0.9	0.286; 0.998	1,586/5,431	0.67	0.89	−0.76
45	CD011902	26	0.531; 1	0.284; 0.997	1,664/21,505	0.87	0.90	−0.08
46	CD011912	19	0.000; 1	0.988; 1	345/436413	0.76	1.00	−0.75
47	CD011926	35	0.375; 1	0.318; 1	738/6,601	0.90	0.93	0.05
48	CD012165	8	0.191; 1	0.5; 1	241/188	0.96	0.84	0.03
49	CD012179	4	0.561; 0.874	0.571; 0.925	435/324	0.81	0.82	−0.61
50	CD012399	12	0.522; 1	0.404; 0.943	229/894	0.85	0.83	−0.62
51	CD012546	13	0.496; 1	0.737; 1	788/857	0.88	0.90	−0.12
52	CD012663	12	0.516; 1	0.088; 0.657	1,133/1,958	0.91	0.37	−1
53	CD012669	34	0.262; 1	0.592; 1	2237/6,398	0.74	0.96	0.28
54	CD012768	33	0.000; 1	0.846; 1	433/3,387	0.72	0.98	0.40
55	CD012777	5	0.350; 0.816	0.204; 0.92	368/511	0.63	0.69	−0.84
56	CD012806	11	0.000; 1	0.735; 1	542/2062	0.35	0.94	0.04
57	CD013186	22	0.167; 1	0.199; 0.978	1,063/7,929	0.90	0.74	0.01
58	CD013187	7	0.909; 0.996	0.545; 1	1,045/219	0.98	0.90	−0.17
59	CD013190	9	0.625; 1	0.575; 0.953	370/1,082	0.90	0.82	−0.39
60	CD013193	4	0.931; 1	0.25; 0.964	452/265	0.95	0.84	−1

^a Median.

we found that the description and quantification of heterogeneity in these reviews was highly variable. It was far from comprehensive. It was not standardized and not properly discussed. Although graphical description of this heterogeneity and exploration of its sources was frequent and satisfactorily performed, most reviews did not provide any numerical quantification of heterogeneity. The use of bivariate and hierarchical models was common within Cochrane DTA reviews, but this was only followed up with subjective assessments of heterogeneity rather than thorough descriptions and quantifications. Furthermore, heterogeneity analysis results rarely drove subsequent meta-analyses or their interpretation. We found inconsistencies in some reviews when comparing authors' subjective judgments about variability and the statistical measures of

heterogeneity provided, or even worse, there are misinterpretations and mismatches between interpretation of graphical exploration of heterogeneity (based on forest plots and the ROC plane) and corresponding heterogeneity statistics.

Using only one measure to summarize heterogeneity does not provide a complete view of the variability in DTA reviews [6]. To better characterize heterogeneity, graphical description has to be complemented with additional heterogeneity statistics. We have introduced two alternative approaches to quantify heterogeneity aiming to promote a better understanding of variability and more transparent reporting. These are the area of the 95% prediction ellipse and median OR. These two measures should be complementary to the variances of the logit accuracy indices already proposed. Although variances of logit

Heterogeneity statistic estimates						Heterogeneity description (by authors)
Var (logit sen)	Var (logit spe)	MOR sen	MOR spe	I ² Biv ^a	Prediction ellipse area	
0.742	1.346	2.27	3.02	0.78	0.380	“high”
0.556	2.325	2.04	4.28	0.25	0.224	
0.599	0.541	2.09	2.02	0.81	0.282	“significant”
0.641	1.709	2.15	3.48	0.76	0.419	“significant”
0.017	1.733	1.13	3.51	0.52	0.101	“considerable”
1.424	1.856	3.12	3.67	0.77	0.382	“some differences”
0.387	0.538	1.81	2.01	0.62	0.274	“moderate to high”
0.061	1.536	1.27	3.26	0.53	0.277	“extreme”
0.09	1.175	1.33	2.81	0.56	0.051	
1.354	0.232	3.03	1.58	0.56	0.417	
0.383	0.86	1.8	2.42	0.22	0.062	
0.641	6.079	2.15	10.51	0.04	0.695	
6.173	9.756	10.7	19.68	0.17	0.973	“substantial”
0.873	5.183	2.44	8.77	0.56	0.602	“highly heterogeneous”
0.952	1.901	2.54	3.73	0.82	0.401	“substantial”
0.102	2	1.36	3.85	0.38	0.008	“greater variability in estimated sensitivity”
2.803	1.665	4.94	3.42	0.55	0.494	
5.111	0.855	8.64	2.42	0.36	0.755	“substantial”
0.433	0.598	1.87	2.09	0.62	0.741	“high”
1.236	0.681	2.89	2.2	0.64	0.316	“sensitivity was more heterogeneous than specificity”
2.289	0.695	4.23	2.21	0.54	0.397	“great heterogeneity”
1.062	0.333	2.67	1.73	0.65	0.078	“considerable”
1.064	1.937	2.67	3.77	0.66	0.387	“considerable”
0.953	1.058	2.54	2.66	0.22	0.127	
0.82	1.489	2.37	3.2	0.77	0.677	“very heterogeneous”
2.282	1.389	4.22	3.08	0.59	0.643	
1.14	1.342	2.77	3.02	0.77	0.489	“considerable, very variable specificity”
0.765	0.465	2.3	1.92	0.38	0.138	
0.449	0.898	1.89	2.47	0.69	0.271	
0.032	2.815	1.19	4.96	0.21	0.274	“considerable”

indices are a direct output of the bivariate model, however, the interpretation is not straightforward because they are on the logit scale. Among the 91 Cochrane reviews that included a meta-analysis, only 5 provide variances of logit sensitivity and specificity despite being recommended by Cochrane DTA methods group [11]. It is likely that this lack of interpretability is one of the causes of its infrequent reporting.

Our's was an empirical evaluation that presented variations in outputs related to differences in statistical approaches to heterogeneity evaluation in DTA syntheses. It was not our intention to produce a rank order of the heterogeneity measures. This may be seen as a limitation. Instead of establishing a single preferred measure of heterogeneity, it is, in our view, far more preferable to promote the use of

several complementary assessments of heterogeneity. This approach would allow a richer exploration of heterogeneity assessments in DTA reviews. Our findings highlight the complementarity that exists between the measures. It is generally recommended by DTA Cochrane methods group and other experts [4,6,11] that between-study variance should be used as the key statistic to quantify heterogeneity. Tau-squared captures between-study variability and underpins the two different metrics for heterogeneity assessment that we propose: the area of the confidence ellipse and the median ORs. Using these metrics along with a visual inspection of ROC space and forest plots would provide a better insight into interpretation of the accuracy data with respect to heterogeneity. We have also incorporated into the list of heterogeneity statistics the (bivariate) I-squared

Table 5. Summary of heterogeneity measures as per authors' subjective evaluation of the heterogeneity

Heterogeneity description (by authors)	Var (logit sen)	Var (logit spe)	MOR sen	MOR spe	I ² Biv ^a	Prediction ellipse area
			range (min-max)			
Not reported (<i>n</i> = 16)	0.090–2.803	0.231–6.079	1.33–4.94	1.58–10.51	0–0.77	0.051–0.695
Little, minimal, or some (<i>n</i> = 5)	0–1.424	0.409–1.915	1–3.12	1.84–3.74	0.01–0.77	0.010–0.382
Considerable to extreme (<i>n</i> = 31)	0.017–6.173	0.029–9.756	1.13–10.7	1.18–19.68	0.09–0.92	0.050–0.973
Only subjective evaluation of variability in one index (sensitivity) (<i>n</i> = 8)	0.102–2.186	0.001–6.038	1.36–4.10	1.03–10.42	0–0.91	0.008–0.674

^a Median.

as a descriptive measure of heterogeneity. We did so because reviewers tend to be familiar with its interpretation as a proportion of variability in study accuracy estimates that is due to heterogeneity rather than due to within-study variance. Although the bivariate form of the I-squared solves one of the criticisms of its univariate counterparts (i.e., takes into account the correlation between accuracy indices), we acknowledge that its dependence on sample size is a weakness that makes it difficult to compare I-squared values among meta-analysis with different number of studies and different number of diseased and nondiseased participants. For all the above reasons, we prefer

to avoid recommending a single measure of heterogeneity as the most valid measure superior to other measures. A more balanced approach would be one where addressing the issue of heterogeneity assessment involves various measures simultaneously.

We propose using the area of the 95% prediction ellipse and median OR as measures to further describe heterogeneity. These measures can be easily obtained and interpreted. The area of the 95% prediction ellipse in the ROC plane allows interpreting heterogeneity as a proportion of the total area of the ROC space. We found that 24% of the Cochrane DTA reviews, already provided the 95% prediction ellipses.

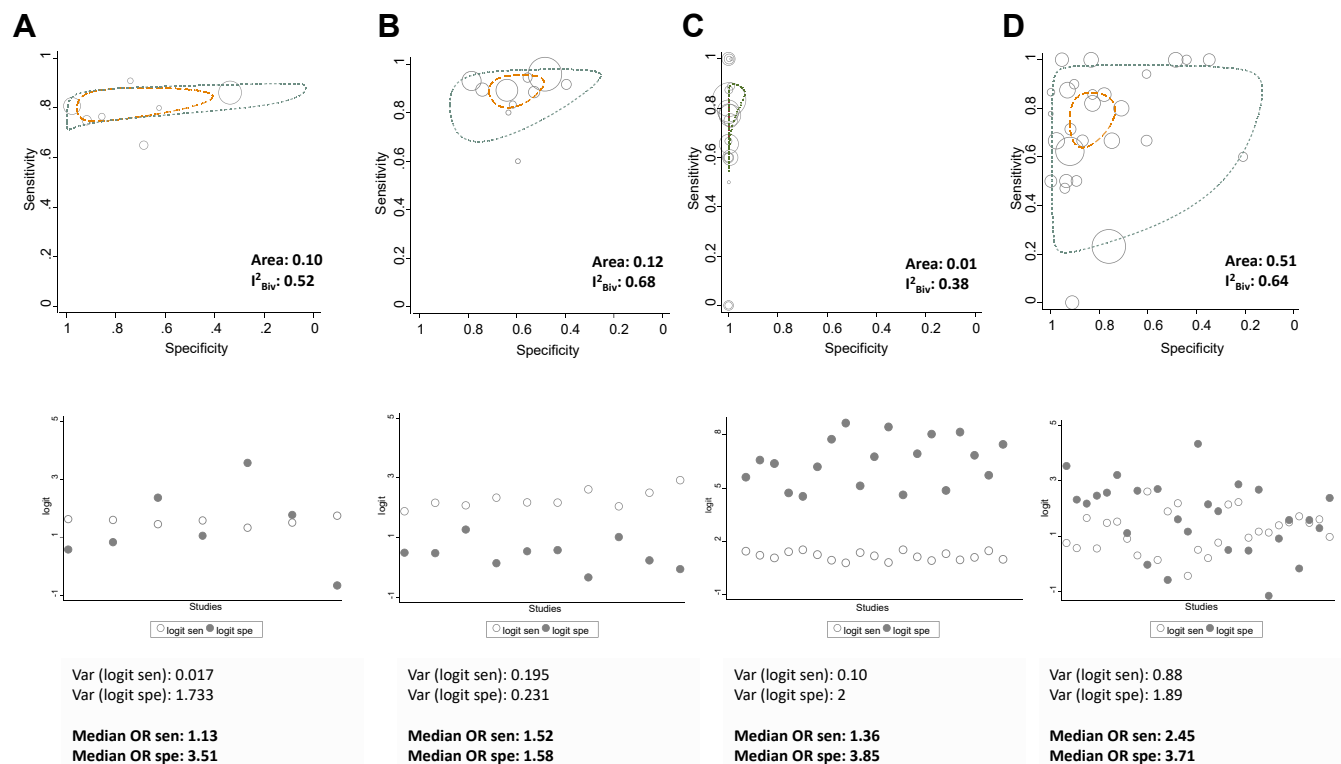


Fig. 2. Scatter plots of four exemplar reviews selected to illustrate the heterogeneity measures (see text for details) A: Nicholson BD, 2015 (CD011134) [17]. B: Arbyn M, 2013 (CD008054) [18]. C: Plana MN, 2018 (CD011912) [19]. D: Leeflang MMG, 2015 (CD007394) [23]. Upper panel: ROC space with accuracy estimates of individual studies included in the review (gray circles). Orange dashed line represents the 95% confidence ellipse around the summary estimate (not drawn). Green dotted line represents the 95% prediction ellipse. Area corresponds to the area of the 95% prediction ellipse. I² Biv, bivariate I-squared; OR, odds ratio; Var, variances; Sen and spe, sensitivity and specificity, respectively. Bottom panel: Scatter plots representing the logits of sensitivity and specificity (y-axis) for each study (x-axis) included in the review.

Calculating the corresponding areas of these prediction regions could quantify the amplitude of the region where average sensitivity and specificity of new studies will lay as an assessment of variability. The median OR is a measure that can be obtained for both accuracy indices separately (i.e., sensitivity and specificity) and are derived from their (logit) variances. Its interpretation is closer to clinicians' understanding of variability. It can be interpreted in relative terms, i.e., it represents the median impact in the odds of being correctly classified as diseased or nondiseased if an individual participant is shifted between two studies. Extending the use of this more intuitive measure will promote more discussion in DTA SRs about how to interpret heterogeneity. This, in turn, will provide more tools for authors to make better informed decisions on whether to pool or not to pool individual studies estimates into an overall accuracy summary.

5. Conclusion

In this study, we found that in DTA reviews current reporting of heterogeneity is patchy and it is possible to augment the assessment of heterogeneity with new measures. We propose to include complementary to variances of the logit sensitivity and specificity, the area of the 95% prediction ellipse, and the median OR as tools to aid in better interpretation of heterogeneity. Adopting these measures may improve reporting of this important aspect of DTA meta-analysis.

CRedit authorship contribution statement

María Nieves Plana: Conceptualization, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Teresa Pérez:** Conceptualization, Methodology, Writing - review & editing, Supervision. **Javier Zamora:** Conceptualization, Methodology, Validation, Writing - review & editing, Resources, Supervision.

Acknowledgments

This work has been partially supported by research grants PI19/00481, from the Instituto de Salud Carlos III, Ministerio de Ciencia e Innovación, Spain, FEDER “Una manera de hacer Europa” and Grant PID2019-104681RB-I00, Ministerio de Ciencia e Innovación, Spain.

Supplementary data

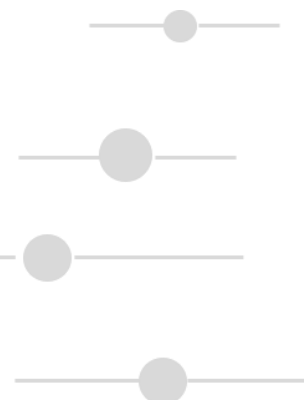
Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jclinepi.2020.11.011>.

References

- [1] Naaktgeboren CA, Ochodo EA, Van Enst WA, de Groot JAH, Hooft L, Leeftang MMG, et al. Assessing variability in results in systematic reviews of diagnostic studies. *BMC Med Res Methodol* 2016.
- [2] DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7:177–88.
- [3] Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;21:1539–58.
- [4] Rucker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. *BMC Med Res Methodol* 2008;8:79.
- [5] Jackson D, White IR, Riley RD. Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. *Stat Med* 2012;31:3805–20.
- [6] Zhou Y, Dendukuri N. Statistics for quantifying heterogeneity in univariate and bivariate meta-analyses of binary data: the case of meta-analyses of diagnostic accuracy. *Stat Med* 2014;33:2701–17.
- [7] 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;58:982–90.
- [8] Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2001;20:2865–84.
- [9] Chu H, Guo H, Zhou Y. Bivariate random effects meta-analysis of diagnostic studies using generalized linear mixed models. *Med Decis Making* 2010;30:499–508.
- [10] Leeftang MMG, Deeks JJ, Rutjes AWS, Reitsma JB, Bossuyt PM. Bivariate meta-analysis of predictive values of diagnostic tests can be an alternative to bivariate meta-analysis of sensitivity and specificity. *J Clin Epidemiol* 2012;65:1088–97.
- [11] 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. Available at: <http://srdta.cochrane.org/>. Accessed September 30, 2020.
- [12] Murad MH, Wang Z. Guidelines for reporting meta-epidemiological methodology research. *Evid Based Med* 2017;22(4):139–42.
- [13] Austin PC, Merlo J. Intermediate and advanced topics in multilevel logistic regression analysis. *Stat Med* 2017;36:3257–77.
- [14] Merlo J, Chaix B, Ohlsson H, Beckman A, Johnell K, Hjerpe P, et al. A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *J Epidemiol Community Health* 2006;60:290–7.
- [15] Larsen K, Merlo J. Appropriate assessment of neighborhood effects on individual health: integrating random and fixed effects in multilevel logistic regression. *Am J Epidemiol* 2005;161:81–8.
- [16] Plummer M. JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling. Vienna Austria: 3rd Int Workshop Distrib Stat Comput DSC 2003; 2003:124.
- [17] Nicholson BD, Shinkins B, Pathiraja I, Roberts NW, James TJ, Mallett S, et al. Blood CEA levels for detecting recurrent colorectal cancer. In: The Cochrane Collaboration, editor. *Cochrane Database of Systematic Reviews*. Chichester, UK: John Wiley & Sons, Ltd; 2015.
- [18] Arbyn M, Roelens J, Simoons C, Buntinx F, Paraskevaidis E, Martin-Hirsch PP, et al. The Cochrane Collaboration. Human papillomavirus testing versus repeat cytology for triage of minor cytological cervical lesions. In: *Cochrane Database of Systematic Reviews [Internet]*. Chichester, UK: John Wiley & Sons, Ltd; 2013.
- [19] Plana MN, Zamora J, Suresh G, Fernandez-Pineda L, Thangaratinam S, Ewer AK. Pulse oximetry screening for critical congenital heart defects. *Cochrane Database Syst Rev* 2018;3:CD011912.

- [20] Ochodo EA, Gopalakrishna G, Spek B, Reitsma JB, van Lieshout L, Polman K, et al, The Cochrane Collaboration. Circulating antigen tests and urine reagent strips for diagnosis of active schistosomiasis in endemic areas. In: Cochrane Database of Systematic Reviews [Internet]. Chichester, UK: John Wiley & Sons, Ltd; 2015.
- [21] Horne DJ, Kohli M, Zifodya JS, Schiller I, Dendukuri N, Tollefson D, et al. Xpert MTB/RIF and Xpert MTB/RIF Ultra for pulmonary tuberculosis and rifampicin resistance in adults. *Cochrane Database Syst Rev* 2019;6:CD009593.
- [22] Gurusamy KS, Giljaca V, Takwoingi Y, Higgie D, Poropat G, Štimac D, et al, The Cochrane Collaboration. Endoscopic retrograde cholangiopancreatography versus intraoperative cholangiography for diagnosis of common bile duct stones. In: Cochrane Database of Systematic Reviews [Internet]. Chichester, UK: John Wiley & Sons, Ltd; 2015.
- [23] Leeflang MM, Debets-Ossenkopp YJ, Wang J, Visser CE, Scholten RJ, Hooft L, et al, The Cochrane Collaboration. Galactomannan detection for invasive aspergillosis in immunocompromised patients. In: Cochrane Database of Systematic Reviews [Internet]. Chichester, UK: John Wiley & Sons, Ltd; 2015.

Capítulo 4



Capítulo 4

En este capítulo se presenta una actualización del software Meta-DiSc® en el que se pone a disposición de los usuarios los últimos avances metodológicos y estadísticos para el metanálisis de estudios de exactitud diagnóstica.

Originariamente, MetaDiSc® fue un programa desarrollado por la Unidad de Bioestadística Clínica del Hospital Ramón y Cajal de Madrid en el contexto de los proyectos de investigación FIS PI02/0954 y PI04/1055 y con financiación parcial de la red temática de investigación cooperativa G03/090 de "Desarrollo de metodologías para la aplicación y gestión del conocimiento en la práctica clínica".

Este programa constituye uno de los programas estadísticos más utilizados para la realización de metanálisis de estudios de exactitud diagnóstica por la comunidad científica. La publicación en la que se detalla su desarrollo en la revista *BMC Medical Research Methodology* (Zamora et al. 2006) ha sido ampliamente citada. La consulta,

realizada con fecha 15/06/2021, arroja un total de 1341 citas en artículos indexados en bases de datos internacionales según la plataforma SCOPUS y 1711 según Google Scholar. El amplio uso de este software responde a su carácter amigable y de fácil uso en ámbitos donde la utilización de los paquetes estadísticos en STATA, SAS o R es difícil si no se cuenta con soporte estadístico adecuado.

Desde el año 2019 el programa dejó de estar disponible para su descarga porque los métodos estadísticos implementados en su momento como el modelo lineal propuesto por Moses y Littenberg (1993) o el uso del índice I^2 univariante (Higgins et al. 2003) para cuantificar la heterogeneidad solo son aplicables a escenarios muy concretos de análisis. Los estándares de análisis actualmente recomendados, en particular en las revisiones Cochrane, son los modelos jerárquicos desarrollados en el Capítulo 2 de la presente memoria. Incorporar en MetaDiSc los métodos jerárquicos y los estadísticos de heterogeneidad propuestos en el Capítulo 3, es uno de los objetivos de esta investigación.

4.1 Objetivos

Desarrollar una herramienta de transferencia del conocimiento, en forma de software de análisis estadístico que incorpore los métodos de metanálisis de pruebas diagnósticas avanzados y los métodos de cuantificación de la heterogeneidad propuestos.

4.2 Software

El producto de transferencia consiste en una aplicación web de acceso libre que se ha registrado como MetaDiSc 2.0. El desarrollo de la aplicación se ha realizado utilizando el software estadístico R (<https://www.r-project.org>). Shiny es un paquete de R que permite la creación de aplicaciones web y gráficos interactivos desde el entorno de desarrollo integrado Rstudio, donde el paquete Shiny viene instalado por defecto. La primera versión de este paquete se presentó en 2012 dentro de las herramientas de Rstudio y ha ido evolucionado en los últimos años (<http://shiny.rstudio.com/>).

4.2.1 MetaDiSc 2.0 paso a paso

Interfaz de usuario

La interfaz de usuario se ha diseñado para facilitar un uso amigable e intuitivo. Toda la información está en inglés. Se puede acceder a MetaDiSc 2.0 en el siguiente enlace: www.metadisc.es. La primera pantalla que nos muestra es la portada de presentación de la aplicación (Figura 5).

Figura 5. Portada presentación MetaDiSc 2.0

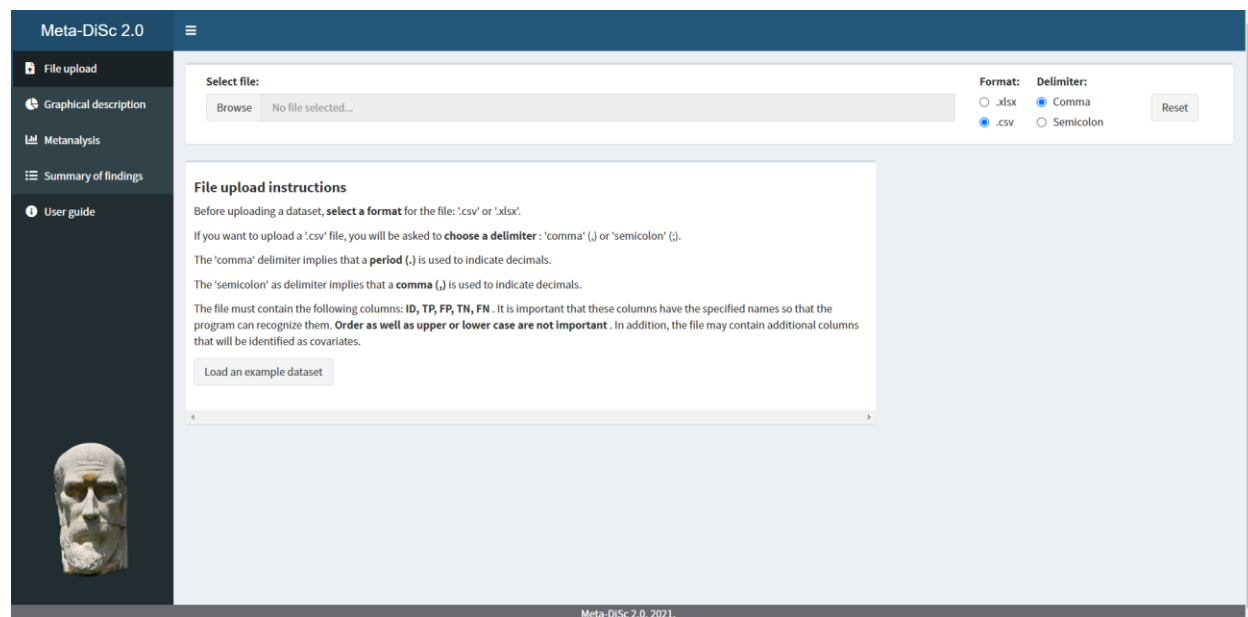


Una vez que entramos en la aplicación se visualiza en la esquina superior izquierda el título, un panel lateral izquierdo y un panel superior. Desde el panel lateral izquierdo se muestran las opciones para:

- i. Carga de datos (*File upload*)
- ii. Descripción gráfica (*Graphical description*)
- iii. Metanálisis (*Metanalysis*)
- iv. Resumen de resultados (*Summary of findings*)
- v. Guía de usuario (*User Guide*)

Hasta que no se cargan los datos, las entradas en el menú lateral izquierdo permanecen inactivas salvo la de Guía de usuario (Figura 6).

Figura 6. Menú principal MetaDiSc 2.0



i. Carga de datos (*File upload*)

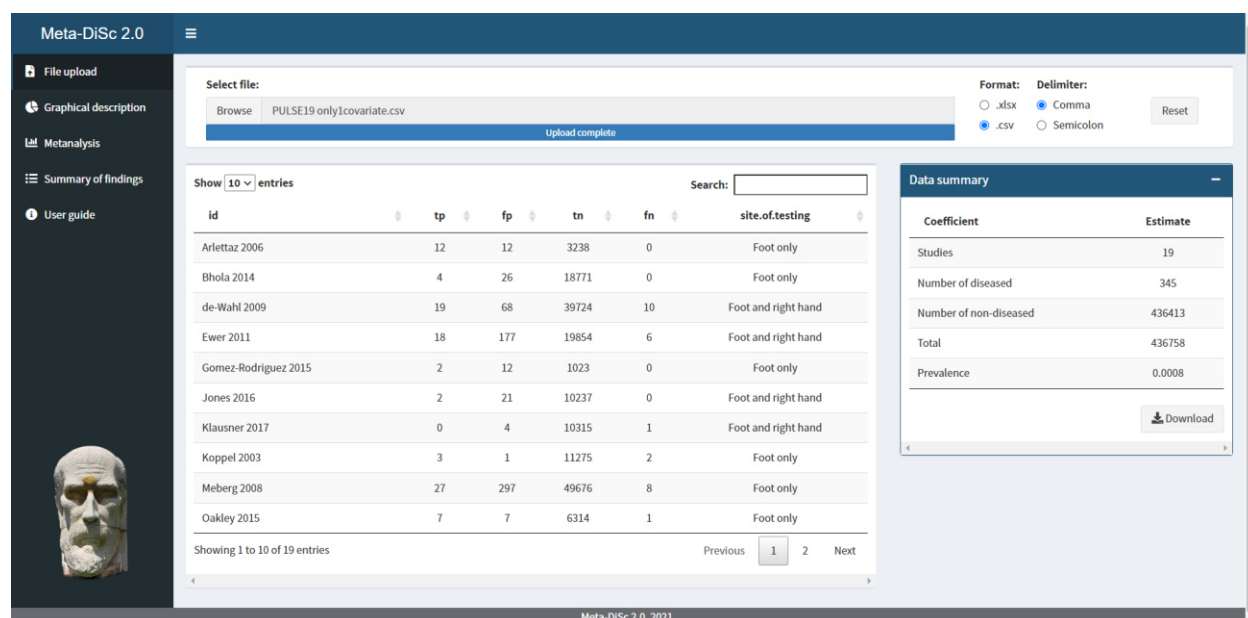
Para cargar los datos debe clicarse en el botón *Select file*. La aplicación permite cargar datos en formato de archivo delimitados por comas (extensión csv) y formato Excel (extensión xlsx). En caso de utilizar un archivo .csv el usuario debe indicar si las columnas están separadas por un carácter coma (*comma*) o punto y coma (*semicolon*). Los archivos deben obligatoriamente contener las columnas: identificador del estudio (variable ID), verdaderos positivos (*true positives*, variable TP), falsos positivos (*false positives*, variable FP), falsos negativos (*false negatives*, variable FN) y verdaderos negativos (*true negatives*, variable TN). No importa el orden de estas columnas en el archivo de datos ni si el nombre de la variable está en mayúsculas o minúsculas. El archivo de datos puede contener opcionalmente otras columnas con covariables que

servirán para realizar análisis de subgrupos o análisis de sensibilidad. Si el archivo no tiene el formato correcto aparece un mensaje de error en pantalla.

Para ilustrar la interfaz de usuario y los cálculos que realiza la aplicación, se ha tomado el conjunto de datos de la revisión sistemática que se incluye en el Capítulo 5 de esta tesis. El objetivo era evaluar la exactitud diagnóstica de la pulsioximetría como prueba de cribado neonatal de cardiopatías críticas congénitas en recién nacidos asintomáticos. Una vez cargados los datos de este estudio se pueden visualizar en pantalla las columnas correspondientes al ID, las correspondientes a la tabla 2x2 (TP, FP, FN, TN) y las columnas que se hayan incluido como covariables (Figura 7). El conjunto de datos se puede ordenar alfabéticamente por la columna ID, que generalmente corresponde al autor del estudio, o de menor a mayor a partir de las columnas TP, FP, FN, TN. También se puede ordenar el conjunto de datos por las categorías de las covariables. La caja de búsqueda (*search*) se puede emplear para buscar estudios concretos.

En el lateral derecho de la Figura 7 aparece una tabla descriptiva (*Data summary*) incluyendo el número de estudios de la revisión, el número de enfermos (TP + FN), el número de sanos (TN + FP), el total de sujetos de estudio y la prevalencia de la condición clínica que se esté tratando de identificar en la revisión. La aplicación permite descargar las tablas y los gráficos en formatos .png y .svg. Desde esta ventana también se puede descargar esta tabla con el resumen de datos de la revisión (*Data summary*) en formato delimitado por comas (csv).

Figura 7. Pantalla de visualización del fichero de datos activo



ii. Descripción gráfica (*Graphical description*)

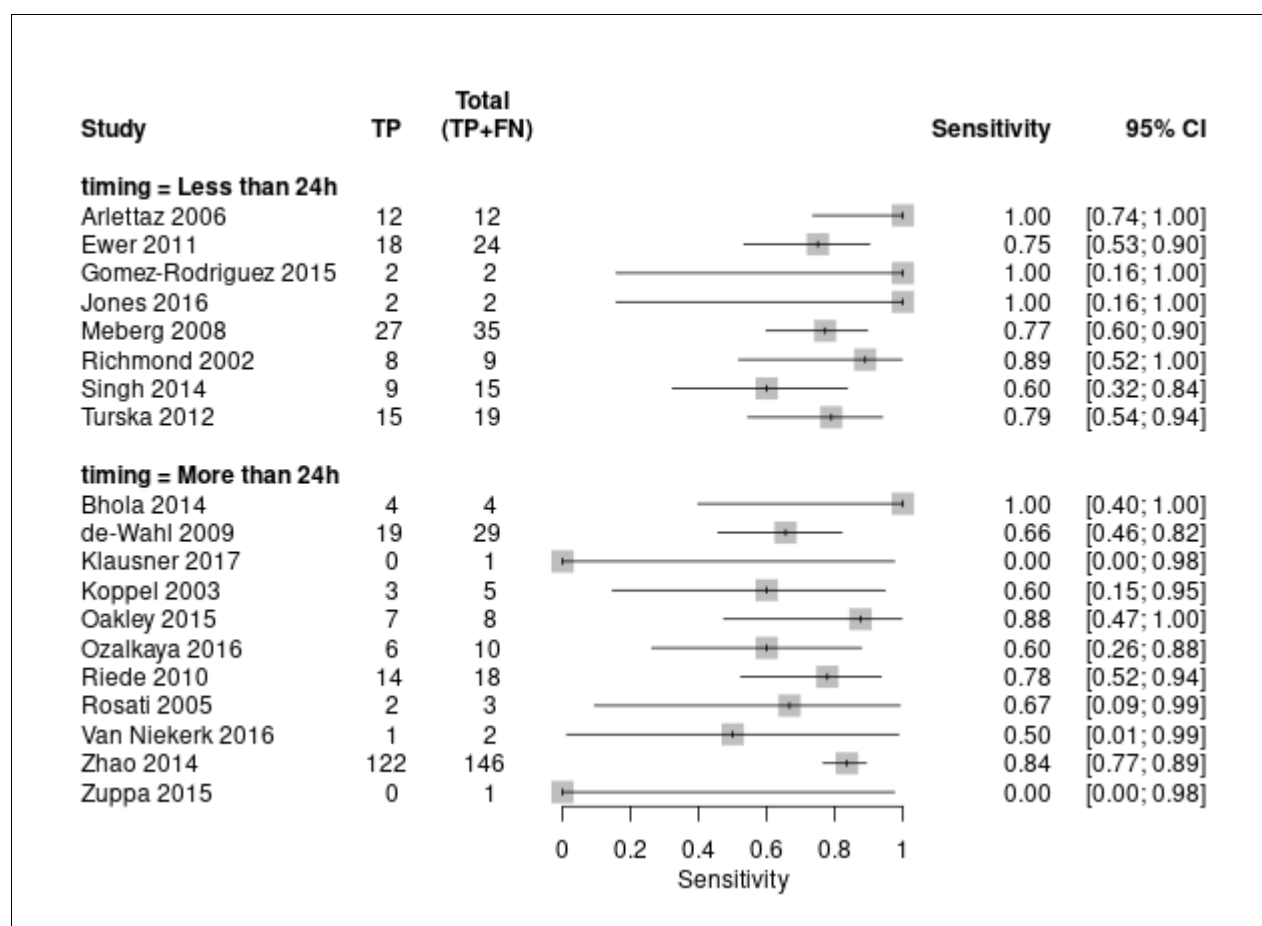
La segunda entrada en el menú lateral izquierdo es la que permite hacer una Descripción gráfica (*Graphical description*) de los estudios incluidos en la revisión. Se incluyen dos pestañas: en una se visualizan los *forest-plot* en los que se representa la sensibilidad y especificidad para cada estudio junto a sus intervalos de confianza al 95% y en otra se representan estos índices en un plano ROC. Los *forest-plot* son descargables en formatos .png y .svg. El plano ROC permite no sólo representar los índices de sensibilidad y especificidad de los estudios individuales incluidos en la revisión, sino también se pueden mostrar unas líneas verticales y horizontales representando los intervalos de confianza al 95% de sensibilidad y especificidad respectivamente. Este gráfico también se puede descargar en los mismos formatos mencionados .png y .svg.

Este paso de descripción gráfica permite reflexionar sobre el escenario de variabilidad o heterogeneidad que presenta nuestra revisión. En el ejemplo que se está utilizando correspondiente a la revisión sistemática presentada en el Capítulo 5, la especificidad de la prueba es prácticamente del 100% en todos los estudios mientras que la sensibilidad varía con un rango entre 0 y 100%.

En esta entrada del menú lateral también tenemos la opción de repetir la descripción gráfica en función de las categorías de alguna de las covariables que incluya el conjunto de datos. La

Figura 8 corresponde al *forest-plot* de sensibilidad de acuerdo a los niveles de una de las covariables del estudio. En este ejemplo se trata del momento de realización de la pulsioximetría al neonato (*test timing*) que pudo ser dentro de las primeras 24 horas de vida o pasadas estas 24 horas (*less than 24h vs more than 24 h*).

Figura 8. Forest-plot de sensibilidad de la Pulsioximetría por la variable "timing"

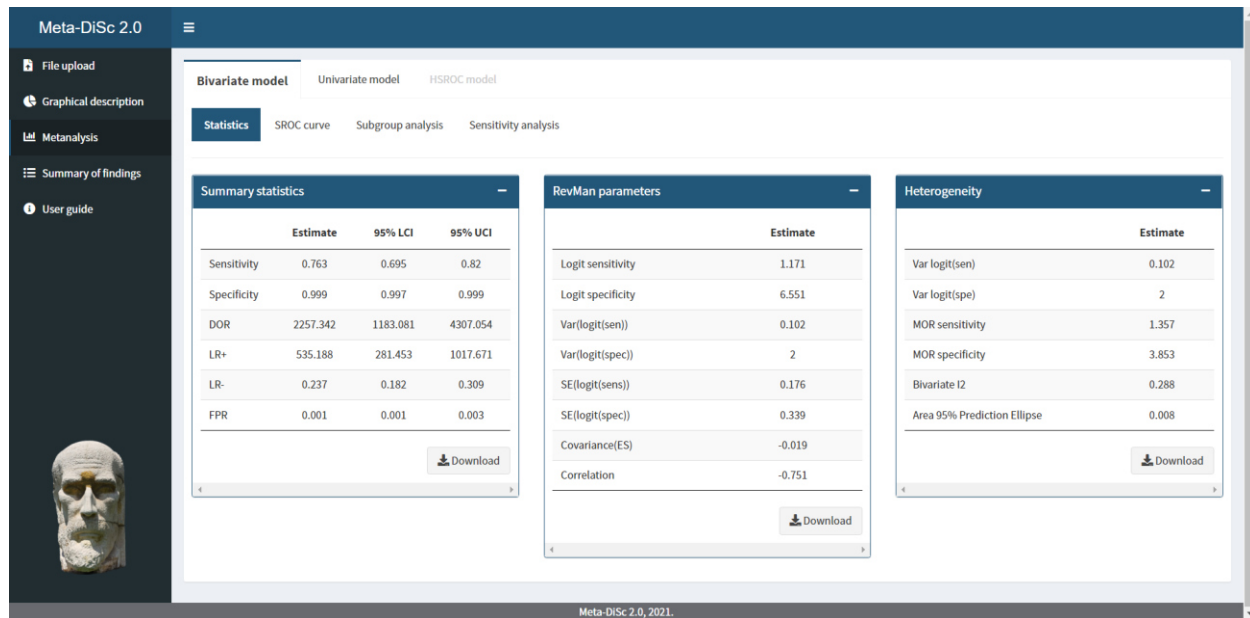


iii. Metanálisis (*Metanalysis*)

La tercera entrada en el menú lateral izquierdo es la relativa al Metanálisis (*Metanalysis*). En esta versión del software sólo está activa la opción de utilizar el modelo univariante de efectos aleatorios y el modelo bivalente (Chu y Cole 2006). Está previsto que en una futura actualización de la aplicación se incorpore el modelo jerárquico HSROC (Rutter y Gatsonis 2001). La pestaña de *bivariate model* (Figura 9) contiene las salidas de resultados estadísticos (*Statistics*), la representación gráfica *sROC* y sendas

opciones para hacer tanto análisis de subgrupos (*Subgroup analysis*) como análisis de sensibilidad (*Sensitivity analysis*).

Figura 9. Pantalla MetaDiSc 2.0 correspondiente al modelo bivariante



En la pestaña de *Statistics* se muestran tres tablas de resultados:

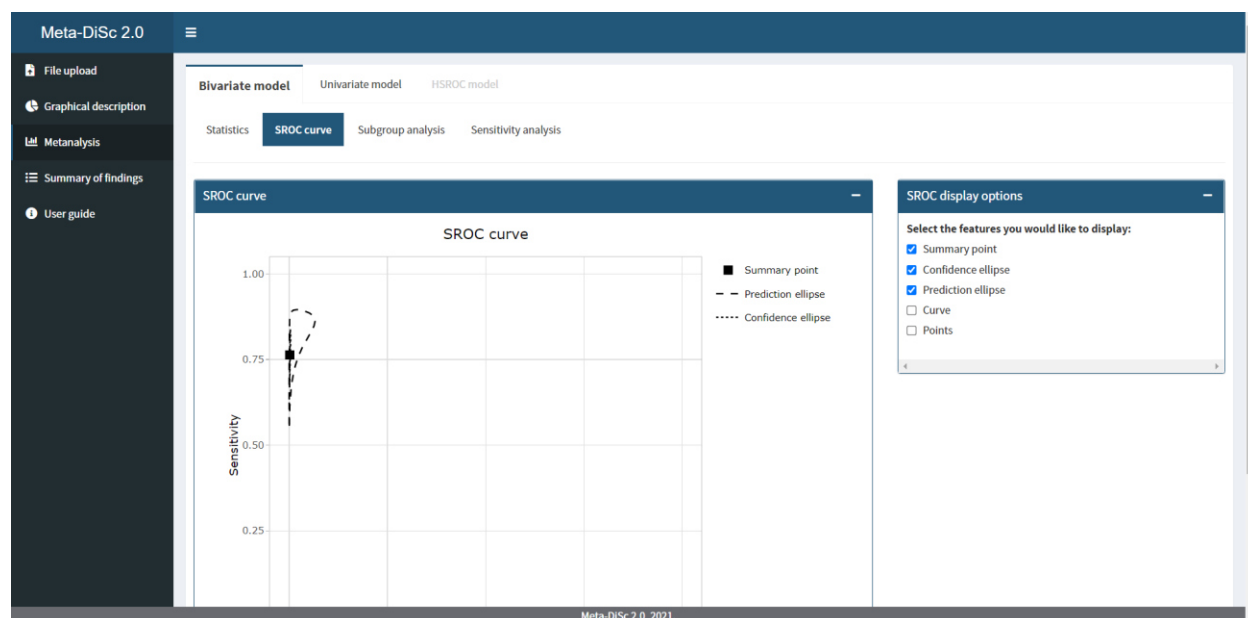
- a. *Summary Statistics*, que nos muestra el promedio ponderado de la sensibilidad (*Sensitivity*) y especificidad (*Specificity*) junto a sus IC al 95%, (*LCI* y *UCI*), de los cocientes de probabilidad positivo y negativo (*LR+* y *LR-*), el ORD (*DOR*) y la tasa de falsos positivos (*FPR*);
- b. *RevMan parameters*, con los parámetros del modelo en escala logit, (*Logit sensitivity* y *Logit specificity*), las varianzas de los logit (*Var(logit sen)* y *Var(logit spe)*), sus errores estándar (*SE(logit sen)* y *SE(logit spe)*), la covarianza y la correlación entre ellos. Estos parámetros se pueden incorporar en el software de Cochrane RevMan para obtener la salida gráfica de la curva sROC e incluirla en revisiones sistemáticas Cochrane. Esta tabla también puede descargarse.
- c. *Heterogeneity*, con las medidas de heterogeneidad que se han propuesto en el Capítulo 3 de esta tesis.

En la pestaña de *sROC curve* se muestra la curva *sROC* que permite una serie de opciones, como decidir incluir los estudios en la gráfica (*Studies*), la propia curva SROC (*Curve*), el promedio ponderado de sensibilidad y especificidad (*Summary point*) y las elipses de confianza (*Confidence ellipse*) y predicción (*Prediction ellipse*) (Figura 10).

Cuando no se puedan estimar todos los parámetros del modelo bivalente, o cuando este ajuste del modelo presente problemas de convergencia, el software está programado para que se estimen los distintos resultados del metanálisis utilizando dos modelos univariantes de efectos aleatorios (para sensibilidad y especificidad

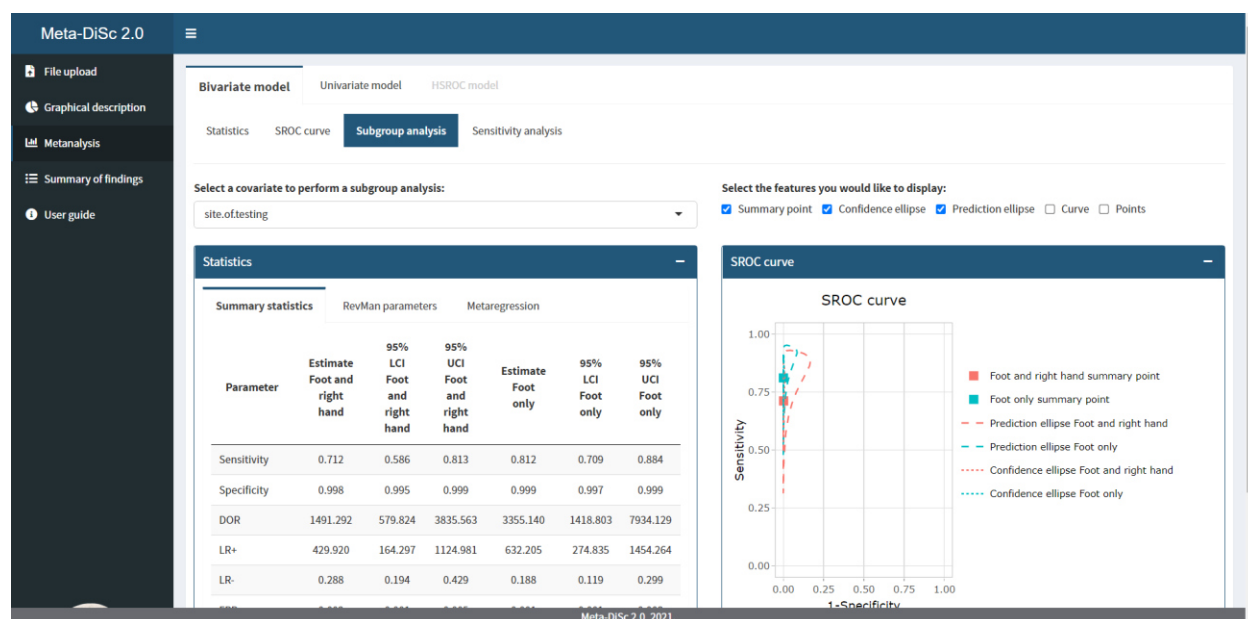
respectivamente) ofreciendo al usuario un mensaje de advertencia. También es posible utilizar directamente un modelo univariante de efectos aleatorios si lo considera oportuno, desde la pestaña de *univariate model*.

Figura 10. Pantalla de MetaDiSc 2.0 correspondiente a la pestaña sROC



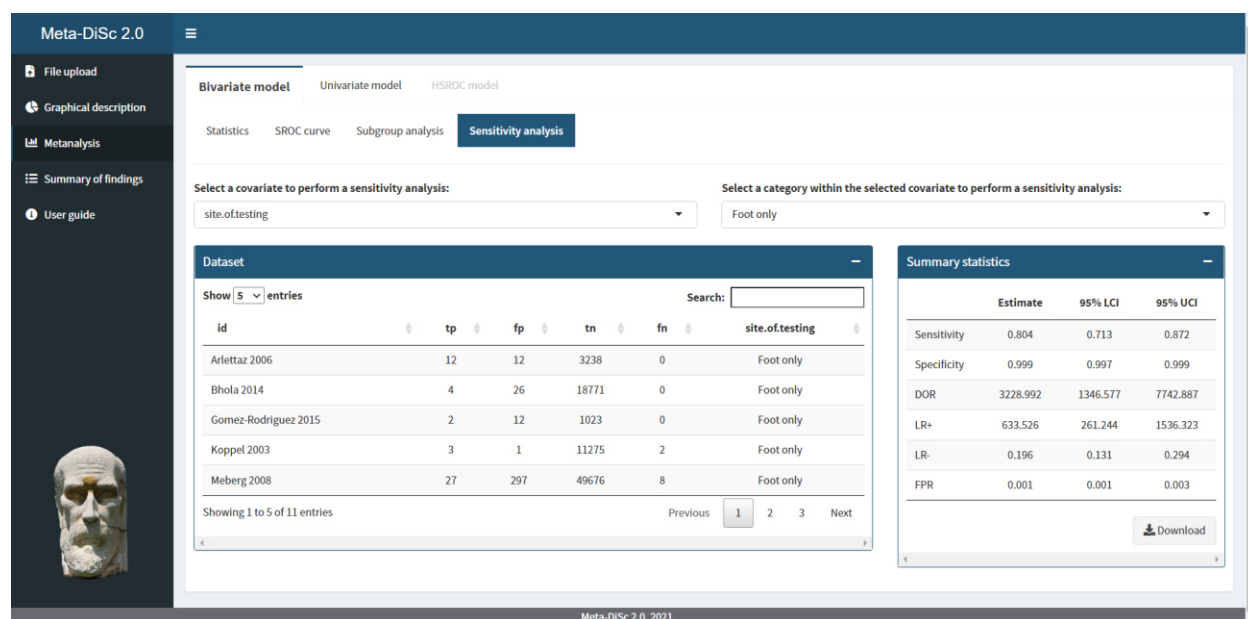
La tercera pestaña de la Figura 9, corresponde al análisis de subgrupos (*Subgroup analysis*). Se pueden realizar los mismos análisis seleccionando la covariable que formará los subgrupos de interés. Para cada una de las categorías de la covariable, se muestran las tablas de resultados (*Summary Statistics*, *RevMan parameters* y *Heterogeneity*), las curvas sROC correspondientes y los resultados del análisis de metarregresión (Figura 11), que permiten contrastar las posibles diferencias entre los distintos subgrupos.

Figura 11. Pantalla de MetaDiSc 2.0 correspondiente al análisis de subgrupos



Finalmente, la cuarta pestaña de la Figura 9, permite hacer análisis de sensibilidad. Por ejemplo, podemos estar interesados en evaluar si los resultados obtenidos difieren si sólo incluimos estudios evaluados como de bajo riesgo de sesgo. Si nuestro conjunto de datos contiene la variable riesgo de sesgo, la podemos seleccionar en el desplegable y elegir la categoría de bajo riesgo de sesgo para realizar el metanálisis incluyendo sólo dichos estudios. En este caso, la salida en pantalla será por un lado el subconjunto seleccionado con información relativa a ID, TP, FP, FN, TN y covariables, y por otro lado, nos muestra la tabla de *Summary Statistics* con los resultados de sensibilidad, especificidad, ORD, cocientes de probabilidad positivo y negativo y tasa de falsos positivos con sus intervalos de confianza del 95% para ese subconjunto de estudios (Figura 12).

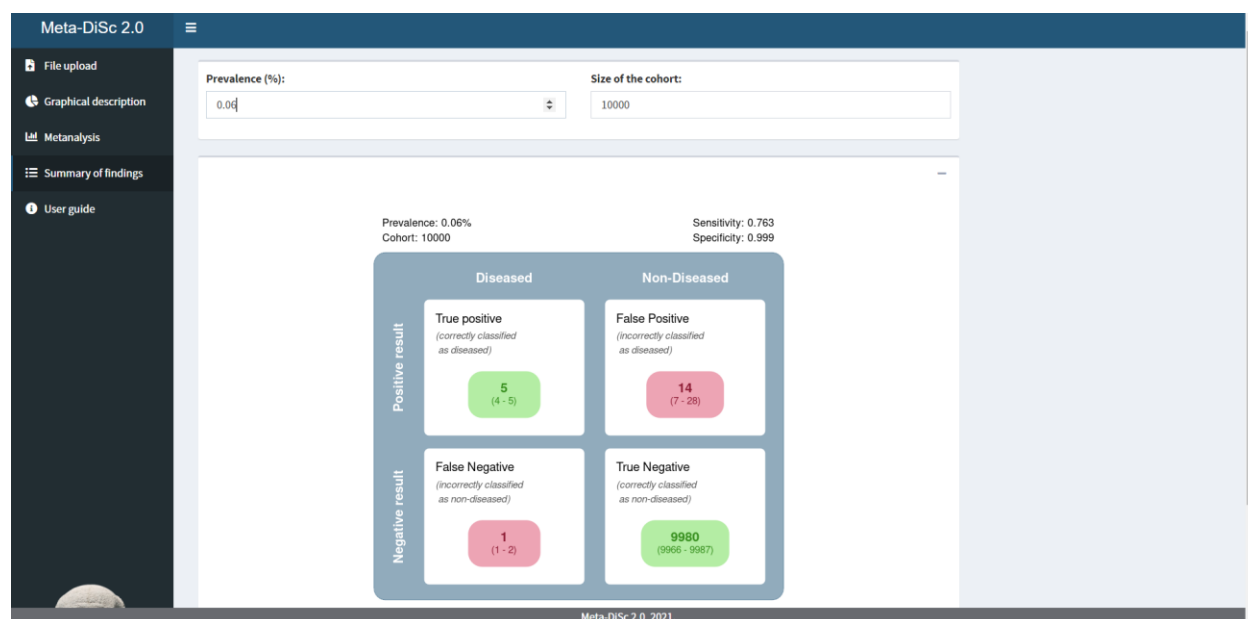
Figura 12. Pantalla de MetaDiSc 2.0 correspondiente al análisis de sensibilidad



iv. Resumen de resultados (*Summary of findings*)

La cuarta entrada en el menú lateral izquierdo se ha definido como de Resumen de Resultados (*Summary of findings*). En esta pantalla, definiendo el tamaño de una hipotética cohorte y la prevalencia de la condición de interés se calcula el número de verdaderos positivos, falsos positivos, falsos negativos y verdaderos negativos que se obtendrían tras aplicar la prueba diagnóstica (Figura 13). Estos resultados se suelen incluir en las tablas de resumen de resultados (*summary of findings*) de las revisiones sistemáticas Cochrane para ofrecer una idea del impacto de la prueba en obtener un resultado incorrecto, en el caso del ejemplo que hemos tratado de la pulsi-oximetría, cuántos recién nacidos serían clasificados erróneamente como portadores de una cardiopatía congénita crónica (FP) y cuántos se clasificarían erróneamente como no portadores cuando sí la tienen (FN).

Figura 13. Pantalla de MetaDiSc 2.0 correspondiente al resumen de resultados



v. Menú de ayuda (*User guide*)

La quinta entrada en el menú lateral izquierdo corresponde al menú de ayuda (*user guide*) con unas instrucciones de uso.

4.3 Metodología

Los parámetros del modelo bivalente lineal generalizado mixto (Chu y Cole 2006) descrito en el Apartado 2.2.3.2, se estiman utilizando la función `glmer` del paquete `lme4` (Douglas et al. 2015). Los cocientes de probabilidad positivo y negativo y el ORD, definidos en las ecuaciones (3), (4) y (5) se calculan directamente a partir de los parámetros estimados en el modelo (logits de sensibilidad y especificidad). El cálculo de los errores estándar de estos cocientes de probabilidad y del ORD se han realizado mediante el método delta implementado en el paquete `msm` (Jackson 2011) de R. La curva sROC se representa a partir de los parámetros HSROC que estima el modelo

bivariante. Para los análisis de metarregresión se ha utilizado el paquete *lmtest* (Zeileis 2002) y el cálculo del área de la elipse de predicción del 95% se ha hecho a través de la función *polyarea* del paquete *pracma* (Borchers 2021) de R.

Como ya se ha comentado en el Apartado 4.2.1, la aplicación está disponible de forma gratuita en el servidor de Shiny (<https://ciberisciii.shinyapps.io/MetaDiSc2/>) o desde www.metadisc.es. El certificado de registro de propiedad intelectual de la aplicación se incluye como Anexo III.

4.4 Validación de la aplicación

Esta versión de la aplicación MetaDiSc 2.0 se ha pilotado, en primer lugar, sobre las revisiones sistemáticas incluidas en el cuerpo de esta memoria en el Capítulo 5 (Plana et al. 2018), Capítulo 6 (Xiao et al. 2018) y Capítulo 7 (Gil et al. 2019). En la Tabla 5 se muestran los resultados de los metanálisis utilizando el software estadístico empleado en cada una de las publicaciones y para su comparación, se muestran también los resultados obtenidos con MetaDiSc 2.0.

Tabla 5. Resultados metanálisis utilizando MetaDiSc 2.0

Primer autor, fecha	N	Resultados (publicados)		Resultados MetaDiSc 2.0	
		Sen/Spe (IC 95%)	Heterogeneidad	Sen/Spe (IC 95%)	Heterogeneidad
Plana et al. 2018	19	76.3 (69.5; 82.0) 99.9 (99.7; 99.9)	Var logit (sen): 0.1 Var logit (spe): 2	76.3 (69.5; 82.0) 99.9 (99.7; 99.9)	Var logit (sen): 0.1 Var logit (spe): 2 MOR sen: 1.4 MOR spe: 3.9 I ² biv: 0.3 AEP 95%: 0.008
Xiao et al. 2018*	6	85.0 (59.3; 95.6) 96.1 (65.2; 99.7)	Var logit (sen): 2.6 Var logit (spe): 9.9	85.0 (59.5; 95.6) 96.1 (65.7; 99.7)	Var logit (sen): 2.6 Var logit (spe): 9.8 MOR sen: 4.7 MOR spe: 19.9 I ² biv: 0.7 AEP 95%: 0.815
Gil et al. 2019 ^{&}	8	98.2 (83.2; 99.8) FPR: 0.05 (0.01; 0.26)	Var logit (sen): 0.02 Var logit (spe): 0.01	No converge-univariante 98.2 (88.4; 99.7) 99.9 (99.8; 100)	Var logit (sen): 0 Var logit (spe): 0 MOR sen: 1 MOR spe: 1

*Se han seleccionado los análisis corresponden a pancreatitis aguda; [&] Se han seleccionado los análisis corresponden a Trisomía 21; N= N° de estudios; Sen: sensibilidad; Spe: especificidad; biv: bivalente; AEP: área de la elipse de predicción 95%. Los análisis en Plana et al 2018 se realizaron utilizando la macro METADAS de SAS (nlmixed model). Los análisis en Xiao et al 2018 se realizaron con la macro Metandi de STATA (xtmlogit model). Los análisis en Gil et al. 2019 se realizaron con MIDAS de STATA (xtmlogit model)

La aplicación también se ha validado utilizando los datos de varias revisiones sistemáticas que se han utilizado frecuentemente en la literatura para ilustrar estos métodos (Fahey, Irwig y Macaskill 1995; Scheidler et al. 1997; Honest et al. 2002; Verde 2010). En la Tabla 6 se muestran los resultados para las parejas de índices de sensibilidad y especificidad, cocientes de probabilidad positivo y negativo y ORD obtenidos utilizando la aplicación MetaDiSc 2.0 con los obtenidos utilizando la macro Metandi de STATA (Harbord y Whiting 2009) y METADAS de SAS (Takwoingi y Deeks 2010). En la Tabla 7 se comparan también los resultados de los análisis de metarregresión utilizando MetaDiSc 2.0 y METADAS sobre tres conjuntos de datos (Verde 2010; Schuetz et al. 2010; Plana et al. 2018). En este caso la macro Metandi de STATA no permite hacer análisis de metarregresión por lo que se muestran solo los resultados obtenidos con SAS mediante la macro METADAS.

Tabla 6. Resultados metanálisis comparando MetaDiSc 2.0 con Metandi y METADAS

Primer autor, fecha	Software	Sensibilidad (IC 95%)	Especificidad (IC 95%)	CP+ (IC 95%)	CP- (IC 95%)	ORD (IC 95%)
Fahey et al. 1995	MetaDiSc 2.0	0.66 (0.58; 0.73)	0.75 (0.68; 0.80)	2.58 (2.14; 3.11)	0.46 (0.39; 0.55)	5.59 (4.34; 7.19)
	Metandi	0.66 (0.58; 0.72)	0.75 (0.68; 0.80)	2.58 (2.14; 3.12)	0.46 (0.39; 0.55)	5.58 (4.33; 7.20)
	METADAS	0.66 (0.58; 0.72)	0.75 (0.68; 0.80)	2.58 (2.14; 3.12)	0.46 (0.39; 0.54)	5.59 (4.34; 7.20)
Scheidler et al. 1997	MetaDiSc 2.0	0.67 (0.61; 0.74)	0.84 (0.76; 0.89)	4.14 (2.70; 6.36)	0.39 (0.31; 0.49)	10.64 (5.83; 19.40)
	Metandi	0.67 (0.60; 0.74)	0.84 (0.76; 0.89)	4.15 (2.69; 6.40)	0.39 (0.31; 0.49)	10.65 (5.81; 19.54)
	METADAS	0.67 (0.60; 0.74)	0.84 (0.76; 0.89)	4.14 (2.69; 6.38)	0.39 (0.31; 0.49)	10.64 (5.82; 19.46)
Honest et al. 2002	MetaDiSc 2.0	0.58 (0.45; 0.71)	0.84 (0.78; 0.88)	3.57 (2.75; 4.63)	0.50 (0.38; 0.67)	7.14 (4.59; 11.13)
	Metandi	0.58 (0.44; 0.71)	0.84 (0.78; 0.88)	3.57 (2.75; 4.64)	0.50 (0.37; 0.67)	7.15 (4.57; 11.19)
	METADAS	0.57 (0.43; 0.70)	0.84 (0.77; 0.88)	3.49 (2.67; 4.56)	0.51 (0.38; 0.68)	6.82 (4.34; 10.71)
Verde 2010	MetaDiSc 2.0	0.96 (0.95; 0.97)	0.95 (0.94; 0.97)	20.61 (14.65; 29.0)	0.05 (0.04; 0.06)	452.28 (296.93; 688.90)
	Metandi	0.96 (0.95; 0.97)	0.95 (0.93; 0.97)	20.59 (14.59; 29.06)	0.05 (0.04; 0.06)	451.76 (295.52; 690.60)
	METADAS	0.96 (0.95; 0.97)	0.95 (0.94; 0.97)	20.61 (14.64; 29.01)	0.05 (0.04; 0.06)	452.28 (296.78; 689.26)

Tabla 7. Comparación análisis de metarregresión utilizando MetaDiSc 2.0, Metandi y METADAS

Primer autor, Fecha Covariable	Software	Sensibilidad relativa (IC 95%)	P valor (comparación sensibilidad)	Especificidad relativa (IC 95%)	P valor (comparación especificidad)
Verde 2010 (retrospective vs prospective)	MetaDiSc 2.0	1.008 (0.988; 1.028)	0.452	0.986 (0.953; 1.020)	0.405
	METADAS	1.008 (0.988; 1.028)	0.450	0.986 (0.953; 1.020)	0.419
	Metandi	-	-	-	-
Schuetz et al. 2010 (MRI vs CT)	MetaDiSc 2.0	0.918 (0.867; 0.972)	0.000	0.805 (0.691; 0.938)	0.000
	METADAS	0.918 (0.867; 0.972)	0.003	0.805 (0.691; 0.938)	0.005
	Metandi	-	-	-	-
Plana et al. 2018 (more than 24 h vs less than 24 h)	MetaDiSc 2.0	0.926 (0.775; 1.105)	0.366	1.004 (1.001; 1.007)	0.001
	METADAS	0.926 (0.775; 1.105)	0.393	1.004 (1.000; 1.007)	0.027
	Metandi	-	-	-	-

El desarrollo de la aplicación se ha integrado también dentro de un proyecto intramural financiado por la Fundación para la Investigación Biomédica del Hospital Ramón y Cajal que lleva por título “Revisiones rápidas (*rapid reviews*) de diagnóstico para la toma de decisiones en salud: análisis de puntos críticos y desarrollo de software”. Por último, este desarrollo cuenta con el soporte del Consorcio de Investigación Biomédica en Red (CIBER) de Epidemiología y Salud Pública, institución que financia la suscripción a la plataforma shinyapps.io donde se aloja la aplicación desarrollada.

4.5 Publicación

En el Anexo II se presenta el borrador del artículo sobre la aplicación MetaDiSc 2.0 que se está preparando para su envío como publicación a la revista *BMC Medical Research Methodology*.

Capítulo 5



Capítulo 5

En este capítulo se presenta el estudio de revisión sistemática realizado con el objetivo de evaluar la exactitud diagnóstica de la pulsioximetría como prueba de cribado neonatal de cardiopatías críticas congénitas en recién nacidos asintomáticos.

El diagnóstico precoz de cardiopatías congénitas críticas (CCC) en los recién nacidos antes del colapso circulatorio permite mejorar su pronóstico. El diagnóstico por ecocardiografía fetal y la exploración física del recién nacido no detecta un porcentaje significativo de estos defectos. La pulsioximetría se propone como una prueba de cribado que podría mejorar la detección de CCC.

5.1 Objetivos

Determinar la exactitud diagnóstica de la pulsioximetría como método de cribado para la detección de cardiopatías críticas congénitas en recién nacidos asintomáticos.

Determinar posibles fuentes de heterogeneidad, incluyendo las siguientes variables: características de la población (inclusión o exclusión de defectos congénitos detectados de forma antenatal); tiempo de medida (menos de 24 horas vs mayor o igual a 24 horas tras el nacimiento); localización (pre-ductal vs post-ductal); saturación de oxígeno (funcional vs fraccional); diseño del estudio (retrospectivo vs prospectivo) y en función del riesgo de sesgo en el dominio “*flow and timing*” de la herramienta QUADAS-2 (Whiting et al. 2011).

5.2 Metodología

Las bases de datos utilizadas en la búsqueda fueron el registro CENTRAL de ensayos clínicos de la Cochrane Library, MEDLINE, Embase, CINAHL (*Cumulative Index to Nursing and Allied Health Literature*) y HSRpROJ (*Health Services Research Projects in Progress*). Se revisaron las referencias de los estudios incluidos y revisiones sistemáticas relevantes sobre la pregunta para identificar estudios adicionales. No se aplicaron restricciones por idioma de publicación. Se incluyeron estudios tanto con diseño transversal como de cohortes que evaluaran la exactitud diagnóstica de la pulsioximetría frente a una prueba de referencia de diagnóstico ecocardiográfico y seguimiento clínico, incluyendo hallazgos postmortem, datos de mortalidad o registros de anomalías congénitas.

Se extrajeron los datos de exactitud para el umbral utilizado en los estudios primarios. La variabilidad entre estudios y la correlación entre índices se exploró visualmente mediante las gráficas de *forest-plot* y plano ROC. Se evaluó el riesgo de sesgo en los estudios incluidos mediante la herramienta QUADAS-2 (Whiting et al. 2011). Se utilizó el modelo bivalente para la síntesis de resultados, las varianzas de los logit de sensibilidad y especificidad como medidas de heterogeneidad y se analizaron posibles fuentes de heterogeneidad mediante análisis de subgrupos y metarregresión.

5.3 Resumen de resultados

Se incluyeron 21 estudios correspondientes a 457,202 participantes. El análisis principal, considerando un umbral de saturación de oxígeno $< 95\%$ o $\leq 95\%$ se realizó sobre 19 estudios (N=436,758). La sensibilidad de la pulsioximetría para la detección de CCC fue del 76.3% (IC 95% 69.5; 82.0) (baja calidad de evidencia). La especificidad fue del 99.9% (IC 95% 99.7; 99.9), con una tasa de falsos positivos de 0.14% (IC 95% 0.07; 0.22) (alta calidad de evidencia). Los cocientes de probabilidad positivo y negativos fueron de 535.6 (IC 95% 280.3; 1023.4) y 0.24 (IC 95% 0.18; 0.31), respectivamente. Sobre una hipotética población de 10000 recién nacidos asintomáticos, seis presentarían una cardiopatía congénita crítica, basándonos en la mediana de prevalencia que presentan los estudios incluidos en la revisión. El cribado con pulsioximetría detectaría 5 de los 6 casos patológicos y no detectaría 1. A su vez, el cribado con esta prueba identificará de forma errónea a otros 14 recién nacidos de los 10,000 como portadores de una CCC sin tenerla.

La tasa de falsos positivos (TFP) para la detección de CCC fue inferior cuando la pulsioximetría se había realizado pasadas 24 horas desde el nacimiento comparada con cuando se realizó dentro las primeras 24 horas (TFP=0.06%, IC 95% 0.03; 0.13; versus 0.42%, IC 95% 0.20; 0.89; $P=0.027$ respectivamente). La evaluación de la heterogeneidad entre estudios se realizó tanto de forma gráfica a través de los forest-plots y el plano sROC, como estadística mediante la estimación de las varianzas de los logit de sensibilidad y especificidad. Gráficamente se observó una mayor heterogeneidad en sensibilidad que en la especificidad. Se exploraron distintas fuentes de heterogeneidad mediante análisis de subgrupos y metarregresión, pero no se encontró una explicación a la variabilidad observada en la sensibilidad entre los estudios.

5.4 Conclusiones

La pulsioximetría es una prueba diagnóstica muy específica y moderadamente sensible para la detección de CCC con una baja tasa de falsos positivos. La evidencia disponible apoya la introducción del cribado con pulsioximetría para la detección de CCC de rutina en los recién nacidos asintomáticos de forma previa al alta.

5.5 Publicación

Se incluye como cuerpo del capítulo el texto completo del estudio de revisión sistemática publicado en la revista Cochrane Database of Systematic Reviews (Factor de impacto año 2018: 7.755; rango: Q1-primer decil, 11/160 de la categoría Medicine General and Internal):

Plana MN, Zamora J, Suresh G, Fernandez-Pineda L, Thangaratinam S, Ewer AK. Pulse oximetry screening for critical congenital heart defects. Cochrane Database Syst Rev. 2018 Mar 1;3(3):CD011912.

El número de citas de esta publicación (07/07/2021) es de 5 citas en la plataforma Scopus y 113 en Google Scholar.



Cochrane
Library

Cochrane Database of Systematic Reviews

Pulse oximetry screening for critical congenital heart defects (Review)

Plana MN, Zamora J, Suresh G, Fernandez-Pineda L, Thangaratinam S, Ewer AK

Plana MN, Zamora J, Suresh G, Fernandez-Pineda L, Thangaratinam S, Ewer AK.

Pulse oximetry screening for critical congenital heart defects.

Cochrane Database of Systematic Reviews 2018, Issue 3. Art. No.: CD011912.

DOI: 10.1002/14651858.CD011912.pub2.

www.cochranelibrary.com

TABLE OF CONTENTS

HEADER	1
ABSTRACT	1
PLAIN LANGUAGE SUMMARY	2
BACKGROUND	3
OBJECTIVES	5
METHODS	5
RESULTS	7
Figure 1.	8
Figure 2.	10
Figure 3.	11
Figure 4.	12
Figure 5.	13
DISCUSSION	18
AUTHORS' CONCLUSIONS	19
ACKNOWLEDGEMENTS	20
REFERENCES	20
CHARACTERISTICS OF STUDIES	26
DATA	71
Test 1. All studies.	71
Test 2. Primary analysis (threshold < 95% or $\leq$ 95%).	72
ADDITIONAL TABLES	73
APPENDICES	76
CONTRIBUTIONS OF AUTHORS	80
DECLARATIONS OF INTEREST	80
SOURCES OF SUPPORT	80
DIFFERENCES BETWEEN PROTOCOL AND REVIEW	81

Pulse oximetry screening for critical congenital heart defects

Maria N Plana¹, Javier Zamora², Gautham Suresh³, Luis Fernandez-Pineda⁴, Shakila Thangaratinam⁵, Andrew K Ewer⁶

¹Clinical Biostatistics Unit, Ramón y Cajal Hospital (IRYCIS), CIBER Epidemiology and Public Health (CIBERESP), Madrid, Spain. ²Clinical Biostatistics Unit, Ramon y Cajal Institute for Health Research (IRYCIS), CIBER Epidemiology and Public Health (CIBERESP), Madrid (Spain) and Women's Health Research Unit, Centre for Primary Care and Public Health, Queen Mary University of London, London, UK. ³Section of Neonatology, Department of Pediatrics, Baylor College of Medicine, Houston, Texas, USA. ⁴Pediatric Cardiology Unit, Hospital Ramon y Cajal, Madrid, Spain. ⁵Centre for Primary Care and Public Health, Women's Health Research Unit, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK. ⁶Institute of Metabolism and Systems Research, University of Birmingham, Birmingham, UK

Contact address: Andrew K Ewer, Institute of Metabolism and Systems Research, University of Birmingham, Birmingham Women's Hospital, Edgbaston, Birmingham, B15 2TT, UK. a.k.ewer@bham.ac.uk.

Editorial group: Cochrane Neonatal Group.

Publication status and date: New, published in Issue 3, 2018.

Citation: Plana MN, Zamora J, Suresh G, Fernandez-Pineda L, Thangaratinam S, Ewer AK. Pulse oximetry screening for critical congenital heart defects. *Cochrane Database of Systematic Reviews* 2018, Issue 3. Art. No.: CD011912. DOI: 10.1002/14651858.CD011912.pub2.

Copyright © 2018 The Cochrane Collaboration. Published by John Wiley & Sons, Ltd.

ABSTRACT

Background

Health outcomes are improved when newborn babies with critical congenital heart defects (CCHDs) are detected before acute cardiovascular collapse. The main screening tests used to identify these babies include prenatal ultrasonography and postnatal clinical examination; however, even though both of these methods are available, a significant proportion of babies are still missed. Routine pulse oximetry has been reported as an additional screening test that can potentially improve detection of CCHD.

Objectives

- To determine the diagnostic accuracy of pulse oximetry as a screening method for detection of CCHD in asymptomatic newborn infants
- To assess potential sources of heterogeneity, including:
 - characteristics of the population: inclusion or exclusion of antenatally detected congenital heart defects;
 - timing of testing: < 24 hours versus ≥ 24 hours after birth;
 - site of testing: right hand and foot (pre-ductal and post-ductal) versus foot only (post-ductal);
 - oxygen saturation: functional versus fractional;
 - study design: retrospective versus prospective design, consecutive versus non-consecutive series; and
 - risk of bias for the “flow and timing” domain of QUADAS-2.

Search methods

We searched the Cochrane Central Register of Controlled Trials (CENTRAL; 2017, Issue 2) in the Cochrane Library and the following databases: MEDLINE, Embase, the Cumulative Index to Nursing and Allied Health Literature (CINAHL), and Health Services Research Projects in Progress (HSRProj), up to March 2017. We searched the reference lists of all included articles and relevant systematic reviews to identify additional studies not found through the electronic search. We applied no language restrictions.

Selection criteria

We selected studies that met predefined criteria for design, population, tests, and outcomes. We included cross-sectional and cohort studies assessing the diagnostic accuracy of pulse oximetry screening for diagnosis of CCHD in term and late preterm asymptomatic newborn infants. We considered all protocols of pulse oximetry screening (eg, different saturation thresholds to define abnormality, post-ductal only or pre-ductal and post-ductal measurements, test timing less than or greater than 24 hours). Reference standards were diagnostic echocardiography (echocardiogram) and clinical follow-up, including postmortem findings, mortality, and congenital anomaly databases.

Data collection and analysis

We extracted accuracy data for the threshold used in primary studies. We explored between-study variability and correlation between indices visually through use of forest and receiver operating characteristic (ROC) plots. We assessed risk of bias in included studies using the QUADAS-2 tool. We used the bivariate model to calculate random-effects pooled sensitivity and specificity values. We investigated sources of heterogeneity using subgroup analyses and meta-regression.

Main results

Twenty-one studies met our inclusion criteria (N = 457,202 participants). Nineteen studies provided data for the primary analysis (oxygen saturation threshold < 95% or $\leq$ 95%; N = 436,758 participants). The overall sensitivity of pulse oximetry for detection of CCHD was 76.3% (95% confidence interval [CI] 69.5 to 82.0) (*low certainty of the evidence*). Specificity was 99.9% (95% CI 99.7 to 99.9), with a false-positive rate of 0.14% (95% CI 0.07 to 0.22) (*high certainty of the evidence*). Summary positive and negative likelihood ratios were 535.6 (95% CI 280.3 to 1023.4) and 0.24 (95% CI 0.18 to 0.31), respectively. These results showed that out of 10,000 apparently healthy late preterm or full-term newborn infants, six will have CCHD (median prevalence in our review). Screening by pulse oximetry will detect five of these infants as having CCHD and will miss one case. In addition, screening by pulse oximetry will falsely identify another 14 infants out of the 10,000 as having suspected CCHD when they do not have it.

The false-positive rate for detection of CCHD was lower when newborn pulse oximetry was performed longer than 24 hours after birth than when it was performed within 24 hours (0.06%, 95% CI 0.03 to 0.13, vs 0.42%, 95% CI 0.20 to 0.89; P = 0.027).

Forest and ROC plots showed greater variability in estimated sensitivity than specificity across studies. We explored heterogeneity by conducting subgroup analyses and meta-regression of inclusion or exclusion of antenatally detected congenital heart defects, timing of testing, and risk of bias for the “flow and timing” domain of QUADAS-2, and we did not find an explanation for the heterogeneity in sensitivity.

Authors' conclusions

Pulse oximetry is a highly specific and moderately sensitive test for detection of CCHD with very low false-positive rates. Current evidence supports the introduction of routine screening for CCHD in asymptomatic newborns before discharge from the well-baby nursery.

PLAIN LANGUAGE SUMMARY

Pulse oximetry for diagnosis of critical congenital heart defects

Review question

We reviewed evidence on the accuracy of pulse oximetry for detection of critical congenital heart defects (CCHDs) in asymptomatic newborn infants.

Background

CCHDs occur in around two in 1000 newborn infants and are a leading cause of infant death. Timely diagnosis is crucial for best outcomes for these babies, but current screening methods may miss up to 50% of affected newborn infants before birth, and those sent home before diagnosis frequently die or endure major morbidity. However, babies with CCHD often have low blood oxygen levels, which can be detected quickly and non-invasively by pulse oximetry, using a sensor placed on the newborn infant's hand or foot. A pulse oximeter is a machine that can measure, non-invasively, the amount of oxygen carried around the body by red blood cells. Oxygen from the lungs is bound to hemoglobin in red blood cells, forming oxyhemoglobin. If oxygen is not bound, de-oxyhemoglobin is formed. In

health, almost all hemoglobin is oxyhemoglobin, and so oxygen saturation (ie, the percentage of hemoglobin that has bound oxygen) is close to 100%. The pulse oximeter measures this by passing light through peripheral blood vessels (eg, a fingertip in an adult, in a hand or foot in a baby). Oxyhemoglobin and de-oxyhemoglobin absorb this light in different ways, and the proportion of light absorbed can be analyzed by software within the oximeter, which then calculates the percentage of hemoglobin saturated with oxygen.

Study characteristics

We searched until March 2017 for evidence on use of pulse oximetry to detect CCHD in newborn infants and found 21 studies. These studies used different thresholds to define a pulse oximetry test as positive. We combined all studies using a threshold around 95% (19 studies with 436,758 newborn infants).

Key results

This review found that for every 10,000 apparently healthy newborn infants screened, around six of them will have CCHD. The pulse oximetry test will correctly identify five of these newborn infants with CCHD (but will miss one case). Newborn infants who are missed could die or experience major morbidity.

For every 10,000 apparently healthy newborn infants screened, 9994 will not have CCHD. The pulse oximetry test will correctly identify 9980 of them (but 14 newborn infants will be investigated for suspected CCHD). Some of these infants may be exposed to unnecessary additional tests and a prolonged hospital stay, but a proportion will have a potentially serious non-cardiac illness.

The number of newborn infants incorrectly investigated for CCHD decreases when pulse oximetry is performed longer than 24 hours after birth.

Certainty of evidence

We judged the included studies to be mainly at low or unclear risk of bias for several of the certainty domains assessed. Some studies used less robust methods to verify negative results. We considered the overall certainty of the evidence as moderate.

BACKGROUND

Congenital heart defects (CHDs) constitute the most common group of congenital malformations, with an incidence of 4 to 10 per 1000 live births ([Botto 2001](#); [Lloyd-Jones 2009](#); [Mahle 2009](#); [Wren 2008](#)); they account for more deaths than any other congenital malformation ([Heron 2007](#); [Mahle 2009](#); [Office of National Statistics, 2015](#)), and up to 10% of all infant deaths are attributed to them ([Abu-Harb 1994](#); [Boneva 2001](#); [Knowles 2005](#); [Lloyd-Jones 2009](#); [Wren 2008](#)). Life-threatening critical CHDs (CCHDs) account for approximately 15% to 25% of all CHDs ([Mahle 2009](#); [Wren 2008](#)). Most CCHDs are amenable to treatment, but poor clinical condition at the time of surgery increases mortality and has been shown to result in worse outcomes for conditions such as hypoplastic left heart ([Brown 2001](#); [Brown 2006](#)), coarctation of the aorta ([Franklin 2002](#)), and transposition of the great arteries ([Tworetzky 2001](#)). Early detection of these conditions can reduce the risk of acute cardiovascular collapse and death ([Abu-Harb 1994](#); [Mahle 2009](#)).

Most newborns with a CCHD are asymptomatic at birth ([Wren 2008](#)); detection before the onset of symptoms usually involves

routine screening by antenatal ultrasound scan, as described by [Allan 1986](#) and [Bull 1999](#), and by postnatal clinical examination of the cardiovascular system, as reported by [Hall 1999](#). Unfortunately, both methods have a variable, and often low, detection rate ([Abu-Harb 1994a](#); [Carvalho 2002](#); [Chew 2007](#); [Garne 2001](#); [Tegnander 2006](#); [Westin 2006](#); [Wren 1999](#)), and up to 30% of infants born with CCHDs are discharged home before the diagnosis has been established ([Abu-Harb 1994](#); [Brown 2006](#); [Mellander 2006](#); [Wren 2008](#)), with reported mortality rates as high as 50% ([Chang 2008](#)).

Although antenatal detection rates following screening ultrasonography are improving, average detection of isolated CCHD remains less than 50% ([Abu-Harb 1994a](#); [Carvalho 2002](#); [Chew 2007](#); [Garne 2001](#); [Tegnander 2006](#); [Westin 2006](#); [Wren 1999](#)). Clinical examination abnormalities such as murmur and weak pulse are often absent in early postnatal life, and the more common finding of cyanosis (bluish discoloration of the skin due to reduced oxygen in the blood) is frequently clinically undetectable ([Mahle 2008](#); [O'Donnell 2007](#)). The fact that most infants with CCHD will have such mild cyanosis has led to the exploration of pulse

oximetry assessment as a possible screening test to identify affected infants (Ewer 2012a; Knowles 2005; Lloyd-Jones 2009).

Following publication of several large test accuracy studies, several countries adopted pulse oximetry screening as routine practice, and many more are considering its introduction (de-Wahl Granelli 2014; Ewer 2014; Kuelling 2009; Mahle 2012; Manzoni 2017).

In addition to test accuracy, studies have demonstrated that pulse oximetry screening is cost-effective (Knowles 2005; Peterson 2013; Roberts 2012), and that it is acceptable to both parents and clinical staff (Narayan 2017; Powell 2013).

The vast majority of babies studied have been screened in a hospital setting - specifically, the well-baby nursery - at low altitude. However, screening has been reported recently in other settings including neonatal units (Iyengar 2014; Suresh 2013), as well as out of hospital settings such as home births - reported by Cawsey 2016; Lhost 2014; and Narayan 2016a - and births at moderate altitude (Han 2013; Wright 2014).

This review does not include settings outside the well-baby nursery.

Target condition being diagnosed

The definition of CCHD is not consistent, and the literature reveals many interpretations (Ewer 2012a). One of the difficulties arises because some conditions (such as coarctation of the aorta and pulmonary stenosis) may or may not predispose to acute collapse, depending on relative severity. For the purposes of this review, we have used a previously described definition of CCHD, that is, "any potentially life-threatening duct-dependent heart lesion from which infants either die or require invasive procedures (surgery or cardiac catheterization) in the first 28 days of life" (Ewer 2012a; Wren 2008). The definition includes all infants with hypoplastic left heart syndrome, pulmonary atresia with intact ventricular septum, simple transposition of the great arteries, or interruption of the aortic arch. In addition, all infants dying or needing surgery or catheter in the first 28 days of life with coarctation of the aorta, aortic valve stenosis, pulmonary valve stenosis, tetralogy of Fallot, pulmonary atresia with ventricular septal defect, or total anomalous pulmonary venous connection are classified as having critical congenital heart defects. This definition offers the advantages that it allows a degree of assessment of severity of certain lesions based on early death or intervention, it is relatively easy to categorize, and it has been used in several test accuracy studies and previous systematic reviews (Ewer 2011; Ewer 2012a; Thangaratinam 2012; Zhao 2014).

Index test(s)

Pulse oximetry measurement of oxygen saturations in an asymptomatic newborn infant can be used to identify CCHD before

discharge from hospital. Pulse oximetry is an accurate and well-established test used to quantify hypoxemia (low oxygen levels in the blood) that is rapid, painless, and easy to perform in all patient groups including newborn infants (Ewer 2012a; Ewer 2013; Knowles 2005; Lloyd-Jones 2009; Mahle 2009; Narayan 2016b). Any trained individual can perform pulse oximetry screening, and results can be obtained in approximately five minutes. For infants enrolled in pulse oximetry screening studies, pulse oximetry probes (to measure oxygen saturations) are placed on the foot only (post-ductal) or on the right hand and foot (pre-ductal and post-ductal) (Ewer 2012b; Ewer 2013; Ewer 2016). The index test allows screening to reduce the number of infants discharged from hospital before diagnosis of CCHD, and it can be performed at any point before discharge before or after the clinical examination.

Clinical pathway

Standard screening for CCHD usually includes midtrimester ultrasonography of pregnant women, which includes assessment of fetal cardiac anatomy. If a cardiac defect is suspected when this examination is performed, a detailed fetal echocardiogram may confirm the diagnosis. Most newborn infants also undergo one or more clinical examinations before discharge from the hospital, which include assessment of the cardiovascular system (auscultation of heart sounds, palpation of peripheral pulses). If a cardiac defect is suspected upon completion of either of these screening tests, then a postnatal diagnostic echocardiogram is usually obtained. As described previously, these screening tests have variable, and often low, detection rates.

The population included in this review may or may not have had antenatal screening. All were asymptomatic at the time of pulse oximetry screening.

Alternative test(s)

In addition to the screening tests already described, alternatives such as routine screening fetal echocardiography and postnatal echocardiography have been proposed but are unlikely to be cost-effective (Knowles 2005). This review did not assess the accuracy of existing screening tests (ie, antenatal ultrasonography and physical examination).

Rationale

Hypoxemia, or suboptimal arterial oxygen saturation, is present in most infants with CCHD (Ewer 2012a; Lloyd-Jones 2009; Mahle 2009). Some may have overt cyanosis, but in many, the degree of hypoxaemia may be difficult to discern on clinical examination. Pulse oximetry is a quick, painless, non-invasive, and reliable method used to determine arterial oxygen saturation levels; it has been widely used in many areas of clinical medicine

for over 30 years. The concept of using oxygen saturations as a screen for critical heart defects was first reported more than 15 years ago. Pulse oximetry screening may allow detection of infants who have been missed by other screening methods before they are discharged from hospital, allowing urgent cardiac intervention before the onset of life-threatening cardiorespiratory collapse. Systematic reviews of pulse oximetry screening studies have been published (Thangaratinam 2007; Thangaratinam 2012), and indeed this screening technique is now common practice in the United States and in some European countries. However, this is the only review that includes recent large test accuracy studies, including one reported from a middle-income country - China.

We performed a systematic review of studies assessing the diagnostic accuracy of screening with pulse oximetry (index test) in relation to echocardiography or clinical follow-up (reference standard) for detection of CCHD in asymptomatic newborn infants. It is important to note that we wanted to determine how well a negative pulse oximetry test result rules out a CCHD diagnosis. Several previous reviews have explored this topic (Ewer 2012a; Ewer 2013; Knapp 2010; Knowles 2005; Lloyd-Jones 2009; Narayan 2016b; Thangaratinam 2007).

OBJECTIVES

- To determine the diagnostic accuracy of pulse oximetry as a screening method for detection of CCHD in asymptomatic newborn infants

Secondary objectives

- To assess potential sources of heterogeneity, including:
 - characteristics of the population: inclusion or exclusion of antenatally detected congenital heart defects;
 - timing of testing: < 24 hours versus $\geq$ 24 hours after birth;
 - site of testing: right hand and foot (pre-ductal and post-ductal) versus foot only (post-ductal);
 - oxygen saturation: functional versus fractional;
 - study design: retrospective versus prospective design, consecutive versus non-consecutive series; and
 - risk of bias for the “flow and timing” domain of QUADAS-2.

METHODS

Criteria for considering studies for this review

Types of studies

We considered inclusion of prospective or retrospective cohort and cross-sectional studies evaluating the diagnostic accuracy of pulse oximetry as a screening method for detection of critical congenital heart defects in asymptomatic newborn infants. A study of diagnostic accuracy should provide sufficient data for construction of the two-by-two table showing the cross-classification of disease status (CCHD) and test outcome (pulse oximetry). We excluded studies if we could not extract true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) values after contacting corresponding authors of primary studies when necessary. We excluded case reports and studies of case-control design.

Participants

We included studies that recruited asymptomatic (with no signs of respiratory or cardiac illness) term or near-term newborns before discharge from hospital.

Index tests

The test under evaluation was pulse oximetry screening to identify low oxygen saturation. We included all protocols of screening (eg, post-ductal [foot] only vs pre-ductal and post-ductal [right hand and foot], different saturation thresholds to define abnormality, different numbers of repeat tests). Criteria for defining a screen as positive or negative in this review were those used by the authors of respective publications.

Target conditions

Critical congenital heart defects as defined above.

Reference standards

Reference standards were diagnostic echocardiography (echocardiogram) and clinical follow-up in the first 28 days of life, including postmortem findings and information from mortality and congenital anomaly databases, to identify patients with false-negative findings.

Search methods for identification of studies

Electronic searches

Information Specialists of the Cochrane Neonatal Review Group performed the searches. Using the strategy described in [Appendix 1](#), they searched the following databases.

- Cochrane Central Register of Controlled Trials (CENTRAL; 2017, Issue 2) in the Cochrane Library.
- MEDLINE via PubMed (1966 to March 2017).

- Embase via Ovid (1980 to March 2017).
- Cumulative Index to Nursing and Allied Health Literature (CINAHL) (1982 to March 2017).

The MEDLINE search strategy included medical subject headings (MeSH) and free text words (see [Appendix 1](#)). We adjusted this strategy for use with the other electronic databases. We considered a combination of medical subject headings and text terms to generate three subsets of citations: one subset indexing the index test (pulse oximetry), a second subset indexing the target population (infant-newborn), and a third subset indexing the clinical condition (congenital heart disease). We combined these subsets to generate a set of citations relevant to our research question. We considered both published and unpublished reports for inclusion and excluded studies published in abstract form only. We applied no language restriction to the electronic searches.

Searching other resources

We used the Science Citation Index, accessed via the Institute for Scientific Information (ISI) Web of Science, to retrieve reports citing the studies included in this review. We searched for similar systematic reviews in the Database of Abstracts of Reviews of Effects (DARE) to March 2017, to cross-reference results. We also searched the Health Services Research Projects in Progress (HSRProj) database (<http://www.nlm.nih.gov/hsrproj/>) (searched on March 20, 2017). We handsearched the reference lists of all relevant primary studies on the topic of our interest to identify cited articles not captured by our electronic searches (up to March 15, 2017). We applied no language restrictions.

Data collection and analysis

Selection of studies

Two review authors (MNP and JZ) independently screened titles and abstracts identified through electronic literature searches to identify potentially eligible studies. First, we excluded those records classified by both review authors as “excluded.” Second, we independently assessed the full text of reports classified as “unsure” or “potentially eligible” by applying the selection criteria outlined above in the [Criteria for considering studies for this review](#) section. We resolved disagreements through discussion. If finally we reached no consensus, we consulted a third review author (AKE).

Data extraction and management

We used a standardized data extraction form to aid extraction of relevant information and data from each included study. Three review authors (MNP, LFP, and AKE) separately participated in data extraction. MNP and LFP extracted data corresponding to study

design, participant details, method of testing, threshold saturation level, and type of oxygen saturation measured, as well as timing of the test and inclusion or exclusion of infants with suspected congenital heart defects after antenatal ultrasound screening in pregnancy, reference tests, and funding. AKE and MNP extracted the following data to reconstruct the two-by-two table: true-positive, false-positive, true-negative, and false-negative values or, if not available, relevant parameters (sensitivity, specificity, or positive and negative predictive values). Two review authors (MNP and JZ) incorporated data and study characteristics into Review Manager 5.3 ([RevMan 2014](#)).

Dealing with duplicate publications

We included only once those studies that have been published in duplicate, ensuring that we extracted all relevant data from all publications.

Inconclusive results

Although we did not anticipate uninterpretable results, when we detected these cases, we excluded them from analysis and adequately reported their frequency in tables.

Assessment of methodological quality

Two review authors (MNP and LFP) independently appraised the methodological quality of each included study using the QUADAS-2 tool ([Whiting 2011](#)). QUADAS-2 consists of four domains, each requiring a risk of bias categorization of low, high, or unclear risk. The first three domains are also assessed in terms of concerns about applicability (applicability concerns ratings). Each domain comprises a set of signaling questions that should be marked as “yes,” “no,” or “unclear.” We tailored QUADAS-2 for our specific review question by modifying signaling questions accordingly and providing guidance on how to assess risk of bias and applicability concerns ratings ([Appendix 2](#)). We resolved disagreements between risk of bias and applicability concern ratings through discussion or by consultation with a third review author (AKE). We summarized our results in the text and in tables and corresponding figures. We decided post hoc to assess the certainty of evidence by using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach ([GRADEpro GDT](#); [Hultcrantz 2017](#); [Schunemann 2008](#)).

Statistical analysis and data synthesis

We performed analyses using methods described in Chapter 10 of the *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* ([Macaskill 2010](#)).

We considered pulse oximetry screening as positive if the oxygen saturation level was below the threshold defined in the primary

study, and negative if it was above that threshold. Cross-classification of these test results with those of the reference standard(s) produced the numbers of true positives, false positives, true negatives, and false negatives for each study, based on the ability of pulse oximetry to detect CCHD.

We used data from the two-by-two tables to calculate sensitivity and specificity for individual studies. We present individual study results by plotting the estimates of sensitivity and specificity (and their 95% confidence intervals) in both forest plots and receiver operating characteristic (ROC) scatter plots. We extracted accuracy data for the threshold used in primary studies.

We performed meta-analyses using a bivariate model (Chu 2006; Reitsma 2005). This model accounts for intra-study accuracy variability and inter-study variations in test performance with inclusion of random effects. We analyzed studies sharing the same threshold and obtained summary accuracy estimates (when the number of studies was enough). We present these estimates with a 95% confidence ellipse in the ROC space. We used pooled estimates of sensitivity and specificity to derive positive and negative likelihood ratios that can be used to update the prior probability of having CCHD to a post-test probability of having CCHD after a positive or negative pulse oximetry result. The greater the positive likelihood ratio and the lower the negative likelihood ratio, the more important the effect of the test on changing pretest into post-test probabilities. We did not calculate positive and negative predictive values because these indices depend on the prevalence of the target condition (ie, CCHD).

For analyses, we used a METADAS SAS macro that estimates parameters for the model with SAS Proc NLMIXED (SAS Institute Inc. 2004; Takwoingi 2010). We entered parameter estimates from the bivariate model into RevMan to produce the summary operating point with a 95% confidence region and a 95% prediction region (Chu 2006; Reitsma 2005; RevMan 2014).

Investigations of heterogeneity

We explored between-study variability and correlation between indices visually through forest and ROC plots. We measured total between-study variability in sensitivity and in specificity through variances of the random effects for logit(sensitivity), logit(specificity), and their covariance of the bivariate model. We also provided confidence and prediction ellipses. We further investigated heterogeneity by exploring effects of several study-level factors through subgroup and meta-regression analyses including covariate terms to the bivariate model (Chu 2006; Reitsma 2005). When available, we examined the following covariates.

- Inclusion or exclusion of antenatally detected congenital heart defects.
- Screening test method (the screening test may be performed at different times after birth, oxygen saturations may be measured at pre-ductal and post-ductal sites or at post-ductal

sites only, and, finally, the oxygen saturation measured could be expressed as “functional” [which refers to the proportion of oxygenated hemoglobin that is *capable* of binding oxygen] or “fractional” [which refers to the percentage of total hemoglobin that is oxygenated]). In most cases, differences between the two values are very small, and most modern pulse oximeters measure functional saturations only.

- Study design (included studies may be prospective or retrospective, and may enroll consecutive patients or not). We expect that retrospective studies are more prone to information and selection biases. In this review, it is more likely that medical records of infants with a positive index test result include more information as compared with medical records of infants with a negative test result (information bias). In a similar way, infants with any CCHD are more likely than infants without CCHD to be detected and included in the study after a retrospective medical records review (selection bias).

- Risk of bias of the “flow and timing” domain of the QUADAS-2 questionnaire (unclear/high vs low risk of bias). It is expected that studies used different reference standards to confirm index test results (echocardiogram, clinical follow-up, registries in mortality, and congenital anomaly databases).

Sensitivity analyses

We examined the robustness of meta-analyses by conducting sensitivity analyses. We checked the impact of excluding studies from analysis according to domains of the QUADAS-2 assessment. Additionally, we decided to perform ad hoc sensitivity analyses to explore how sensitivity and specificity vary by including or excluding studies with different thresholds.

Assessment of reporting bias

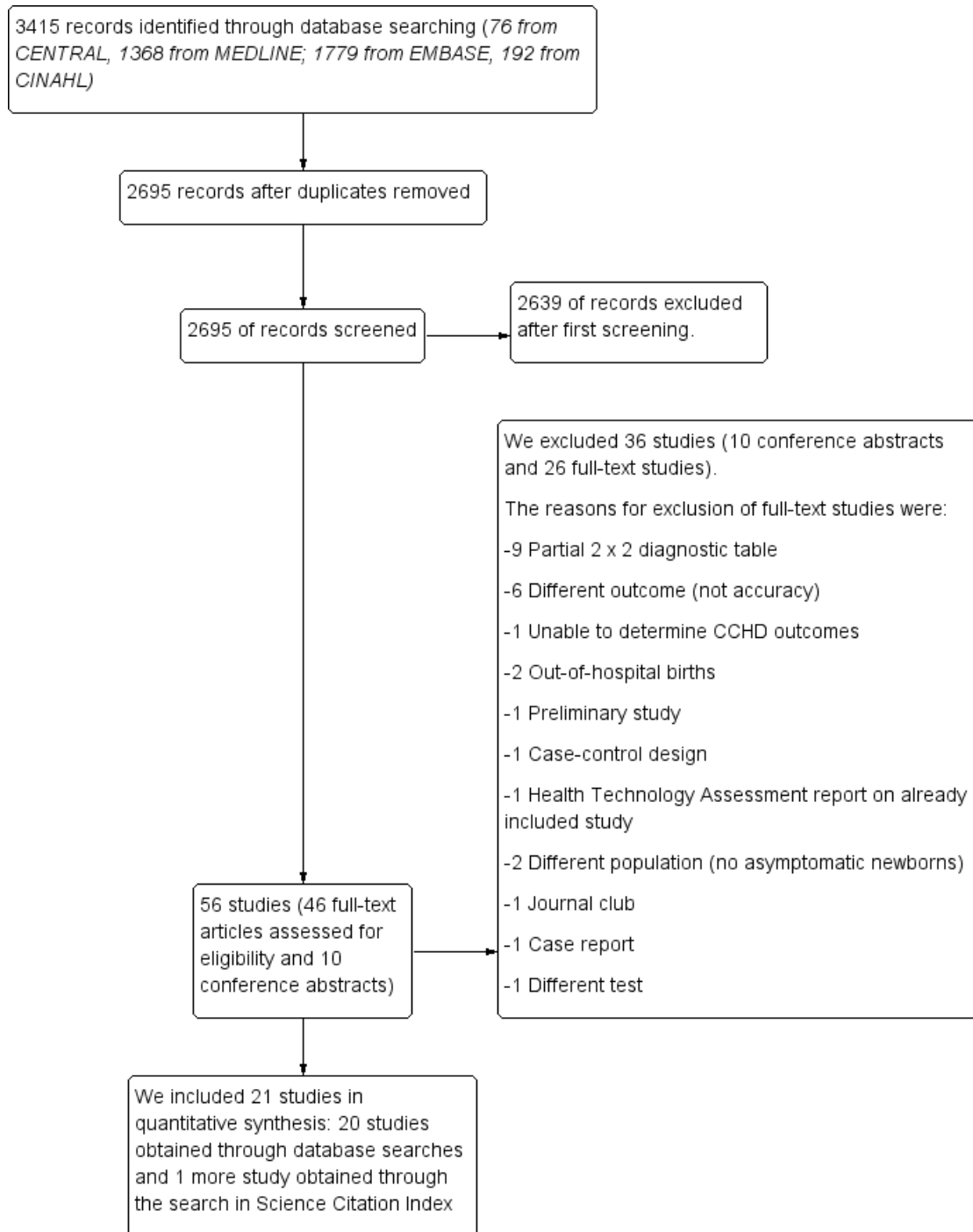
We did not investigate reporting bias, given the limited power of available tests and uncertainty about interpreting statistical evidence of funnel plot asymmetry as necessarily implying publication bias (Leeftang 2008).

RESULTS

Results of the search

Figure 1 shows details of the search and selection process. Electronic database searches yielded a total of 3415 references from CENTRAL, MEDLINE, Embase, and CINAHL. Searches for primary studies through other resources did not reveal additional potentially eligible studies.

Figure 1. Flow of studies through the screening process. CCHD: critical congenital heart defect.



After de-duplication, two review authors (MNP and JZ) independently assessed 2695 references against the inclusion criteria. During initial screening of titles and abstracts, we identified 56 studies (46 full-text papers and 10 conference abstracts). We excluded 2639 references because they did not meet the inclusion criteria. We also excluded those published in abstract form only ($n = 10$). Of 46 full-text studies, nine studies provided a partial two-by-two diagnostic table, and we excluded them. We excluded 17 other studies for the following reasons.

- Different outcomes (not accuracy) ($n = 6$).
- Inability to determine CCHD outcomes ($n = 1$).
- Out-of-hospital births ($n = 2$).
- Preliminary studies ($n = 1$).
- Different population ($n = 2$).
- Health technology assessment report on already included study ($n = 1$).
- Case-control study ($n = 1$).
- Journal club ($n = 1$).
- Case report ($n = 1$).
- Different index test ($n = 1$) (see [Characteristics of excluded studies](#)).

We obtained one additional study by searching Science Citation Index ([Gomez-Rodriguez 2015](#)). We included 21 studies in a quantitative synthesis ([Arlettaz 2006](#); [Bakr 2005](#); [Bhola 2014](#); [de-Wahl Granelli 2009](#); [Ewer 2011](#); [Gomez-Rodriguez 2015](#); [Jones 2016](#); [Klausner 2017](#); [Koppel 2003](#); [Meberg 2008](#); [Oakley 2015](#); [Ozalkaya 2016](#); [Richmond 2002](#); [Riede 2010](#); [Rosati 2005](#); [Sendelbach 2008](#); [Singh 2014](#); [Turska 2012](#); [Van Niekerk 2016](#); [Zhao 2014](#); [Zuppa 2015](#)).

Characteristics of studies

We provide in the [Characteristics of included studies](#) table details on the design, setting, population, index test, target condition, and reference standard of all included studies. We prepared an additional table ([Table 1](#)) to summarize the main characteristics. Of 3415 references, we identified 21 primary studies that were eligible for inclusion and provided data for 457,202 newborn infants ([Figure 1](#)). Studies were published between 2002 and 2017. Countries included were United Kingdom ([Ewer 2011](#); [Jones 2016](#); [Oakley 2015](#); [Richmond 2002](#); [Singh 2014](#)), Italy ([Rosati 2005](#); [Zuppa 2015](#)), USA ([Klausner 2017](#); [Koppel 2003](#); [Sendelbach 2008](#)), Australia ([Bhola 2014](#)), China ([Zhao 2014](#)), Germany ([Riede 2010](#)), Mexico ([Gomez-Rodriguez 2015](#)), Norway ([Meberg 2008](#)), Poland ([Turska 2012](#)), Saudi Arabia ([Bakr 2005](#)), South Africa ([Van Niekerk 2016](#)), Sweden ([de-Wahl Granelli 2009](#)), Switzerland ([Arlettaz 2006](#)), and Turkey ([Ozalkaya 2016](#)). Sixteen studies included prospective cohorts ([Arlettaz 2006](#); [Bakr 2005](#); [de-Wahl Granelli 2009](#); [Ewer 2011](#); [Gomez-Rodriguez](#)

[2015](#); [Koppel 2003](#); [Meberg 2008](#); [Oakley 2015](#); [Richmond 2002](#); [Riede 2010](#); [Rosati 2005](#); [Sendelbach 2008](#); [Turska 2012](#); [Van Niekerk 2016](#); [Zhao 2014](#); [Zuppa 2015](#)), as well as five retrospective cohorts ([Bhola 2014](#); [Jones 2016](#); [Klausner 2017](#); [Ozalkaya 2016](#); [Singh 2014](#)). Seventeen studies excluded newborns who were suspected to have congenital heart disease after antenatal ultrasound screening during pregnancy ([Bakr 2005](#); [de-Wahl Granelli 2009](#); [Gomez-Rodriguez 2015](#); [Jones 2016](#); [Klausner 2017](#); [Koppel 2003](#); [Meberg 2008](#); [Oakley 2015](#); [Ozalkaya 2016](#); [Riede 2010](#); [Rosati 2005](#); [Sendelbach 2008](#); [Singh 2014](#); [Turska 2012](#); [Van Niekerk 2016](#); [Zhao 2014](#); [Zuppa 2015](#)) ([Table 1](#)).

Nine studies performed pulse oximetry within 24 hours after birth ([Arlettaz 2006](#); [Ewer 2011](#); [Gomez-Rodriguez 2015](#); [Jones 2016](#); [Meberg 2008](#); [Richmond 2002](#); [Sendelbach 2008](#); [Singh 2014](#); [Turska 2012](#)) ([Table 1](#)). Twelve studies used the foot alone (post-ductal) to measure oxygen saturation, and the remainder used both right hand and foot (pre-ductal and post-ductal) ([Table 1](#)). Investigators used several different pulse oximeter models (see description in [Table 1](#)). Two studies measured fractional saturations ([Bakr 2005](#); [Richmond 2002](#)) ([Table 1](#)). Eight studies used a post-ductal saturation threshold of less than 95% ([Arlettaz 2006](#); [Bhola 2014](#); [Gomez-Rodriguez 2015](#); [Meberg 2008](#); [Oakley 2015](#); [Richmond 2002](#); [Turska 2012](#); [Zuppa 2015](#)), three studies used a post-ductal saturation threshold $\leq 95\%$ ([Koppel 2003](#); [Riede 2010](#); [Rosati 2005](#)), and six studies used both pre-ductal and post-ductal saturations less than 95% ([de-Wahl Granelli 2009](#); [Ewer 2011](#); [Klausner 2017](#); [Singh 2014](#); [Van Niekerk 2016](#); [Zhao 2014](#)). Two studies reported different positive thresholds ([Bakr 2005](#) reported both pre-ductal and post-ductal fractional saturation $\leq 94\%$, and [Sendelbach 2008](#) reported post-ductal saturation $< 96\%$) ([Table 1](#)). In summary, the most common threshold was less than 95% ($n = 14$); five studies reported a threshold lower than or equal to 95%, and two studies reported thresholds $\leq 94\%$ and $< 96\%$, respectively. No study reported results for more than one threshold. Studies used different methods to verify test results: Investigators verified positive test results by echocardiography and negative results by interrogation of congenital anomaly registers, mortality data, or clinical follow-up ([Table 1](#)).

Methodological quality of included studies

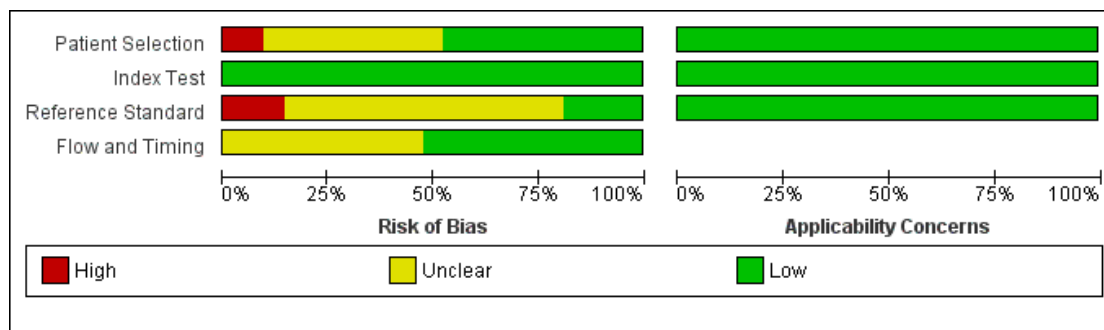
We appraised the quality of primary diagnostic accuracy studies using the QUADAS-2 tool. We present quality assessment results for individual studies in the [Characteristics of included studies](#) table and in [Figure 2](#). We have summarized the overall risk of bias and applicability concerns of studies in [Figure 3](#).

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

	Risk of Bias				Applicability Concerns		
	Patient Selection	Index Test	Reference Standard	Flow and Timing	Patient Selection	Index Test	Reference Standard
Arlettaz 2006	+	+	+	?	+	+	+
Bakr 2005	+	+	?	?	+	+	+
Bhola 2014	?	+	?	?	+	+	+
de-Wahl Granelli 2009	+	+	?	+	+	+	+
Ewer 2011	+	+	+	+	+	+	+
Gomez-Rodriguez 2015	?	+	?	+	+	+	+
Jones 2016	+	+	?	?	+	+	+
Klausner 2017	+	+	?	?	+	+	+
Koppel 2003	+	+	+	+	+	+	+
Meberg 2008	?	+	?	+	+	+	+
Oakley 2015	+	+	?	+	+	+	+
Ozalkaya 2016	+	+	+	+	+	+	+
Richmond 2002	+	+	?	+	+	+	+
Riede 2010	?	+	?	+	+	+	+
Rosati 2005	?	+	?	?	+	+	+
Sendelbach 2008	?	+	?	?	+	+	+
Singh 2014	?	+	?	+	+	+	+
Turska 2012	?	+	+	?	+	+	+
Van Niekerk 2016	+	+	+	?	+	+	+
Zhao 2014	?	+	?	+	+	+	+
Zuppa 2015	+	+	+	?	+	+	+

+ High
 ? Unclear
 + Low

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



We judged the risk that patient selection (QUADAS-2, domain 1) had introduced bias as low in 10 studies (Arlettaz 2006; Bakr 2005; de-Wahl Granelli 2009; Ewer 2011; Jones 2016; Klausner 2017; Koppel 2003; Ozalkaya 2016; Richmond 2002; Zuppa 2015), high in two because investigators did not avoid inappropriate exclusions (Oakley 2015; Van Niekerk 2016), and unclear in the remaining nine studies (Bhola 2014; Gomez-Rodriguez 2015; Meberg 2008; Riede 2010; Rosati 2005; Sendelbach 2008; Singh 2014; Turska 2012; Zhao 2014). Applicability was of low concern for all studies in the patient selection domain.

For the index test assessment (QUADAS-2, domain 2), we considered all studies to be at low risk of bias and low concern regarding applicability.

We judged the risk that conduct or interpretation of reference standard(s) (QUADAS-2, domain 3) had introduced bias as low in four studies because investigators used echocardiography to confirm both positive and negative pulse oximetry cases (Ozalkaya 2016), or because they used echocardiography to confirm pulse oximetry positives and clinical follow-up in the first 28 days of life, which included postmortem findings and mortality and congenital anomaly databases to identify false-negative screening cases (Ewer 2011; Koppel 2003; Turska 2012). This comprehensive combination of clinical follow-up and review of registries and databases was considered as having low risk of bias. We considered that three studies reporting only echocardiography as the reference standard for positive pulse oximetry results were at high risk of bias (Arlettaz 2006; Van Niekerk 2016; Zuppa 2015). We considered risk for the remaining 14 studies as unclear because they used an incomplete reference standard to identify false-negative cases (Bakr 2005; Bhola 2014; de-Wahl Granelli 2009; Gomez-Rodriguez 2015; Jones 2016; Klausner 2017; Meberg 2008; Oakley 2015;

Richmond 2002; Riede 2010; Rosati 2005; Sendelbach 2008; Singh 2014; Zhao 2014); six studies used echocardiography and follow-up (Gomez-Rodriguez 2015; Klausner 2017; Meberg 2009; Rosati 2005; Sendelbach 2008; Zhao 2014), and eight studies used echocardiography and different mortality and malformations registries (Bakr 2005; Bhola 2014; de-Wahl Granelli 2009; Jones 2016; Oakley 2015; Richmond 2002; Riede 2010; Singh 2014). It is noteworthy that only one study used echocardiography for positive and negative pulse oximetry results (Ozalkaya 2016). Applicability was of low concern for all studies in the reference standard(s) domain.

For flow and timing assessment (QUADAS-2, domain 4), 11 studies were at low risk of bias (de-Wahl Granelli 2009; Ewer 2011; Gomez-Rodriguez 2015; Koppel 2003; Meberg 2008; Oakley 2015; Ozalkaya 2016; Richmond 2002; Riede 2010; Singh 2014; Zhao 2014), and the remaining studies were at unclear risk because information reported was insufficient to permit judgment (Arlettaz 2006; Bakr 2005; Bhola 2014; Jones 2016; Klausner 2017; Rosati 2005; Sendelbach 2008; Turska 2012; Van Niekerk 2016; Zuppa 2015).

Findings

Results of meta-analysis

We considered for primary analysis all studies with thresholds around 95% (< 95% and ≤ 95%). As expected, this was the most common threshold among included studies (n = 19 studies; 436,758 participants) (Arlettaz 2006; Bhola 2014; de-Wahl Granelli 2009; Ewer 2011; Gomez-Rodriguez 2015; Jones 2016; Klausner 2017; Koppel 2003; Meberg 2008; Oakley 2015; Ozalkaya 2016; Richmond 2002; Riede 2010; Rosati 2005; Singh

2014; Turska 2012; Van Niekerk 2016; Zhao 2014; Zuppa 2015). The overall sensitivity of pulse oximetry for detection of critical congenital heart defects was 76.3% (95% confidence interval [CI] 69.5 to 82.0). Specificity was 99.9% (95% CI 99.7 to 99.9) with a false-positive rate of 0.14% (95% CI 0.07 to 0.22) (Summary of findings). Summary positive and negative likelihood ratios were 535.6 (95% CI 280.3 to 1023.4) and 0.24 (95% CI 0.18 to 0.31), respectively.

Fourteen out of 19 studies used a threshold lower than 95% (Arlettaz 2006; Bhola 2014; de-Wahl Granelli 2009; Ewer 2011; Gomez-Rodriguez 2015; Klausner 2017; Meberg 2008; Oakley 2015; Richmond 2002; Singh 2014; Turska 2012; Van Niekerk 2016; Zhao 2014; Zuppa 2015), and five studies used a threshold lower than or equal to 95% (Jones 2016; Koppel 2003; Ozalkaya 2016; Riede 2010; Rosati 2005).

Two additional studies used different thresholds: One used a threshold lower than or equal to 94% with sensitivity and specificity of 100% (95% CI 29 to 100) and 100% (95% CI 100 to 100), respectively (Bakr 2005); the other used a threshold of less

than 96% with sensitivity and specificity of 100% (95% CI 3 to 100) and 100% (95% CI 100 to 100), respectively (Sendelbach 2008).

Overall, we have included in this review 349 cases of CCHD. The median prevalence of CCHD was 0.6 per 1000 live births (range 0.1 to 3.7; interquartile range 0.4 to 1.2).

Investigations of heterogeneity

To visualize total variability in sensitivity and specificity, we present the data in forest and ROC scatter plots (Figure 4; Figure 5). Forest plots show studies in increasing order of specificity (Figure 4). Sensitivity of the 21 studies ranged from 0% to 100%, and specificity from 99% to 100%. Forest and ROC plots show greater variability in estimated sensitivity than specificity across studies. Given results from investigations of heterogeneity, we used the bivariate model to estimate summary sensitivity and specificity (summary points) instead of the hierarchical summary ROC model to estimate summary ROC curves.

Figure 4. Forest plot of sensitivity and specificity. The figure shows the estimated sensitivity and specificity of the study (blue square) and its 95% confidence interval (black horizontal line). Studies are ordered by ascending specificity.

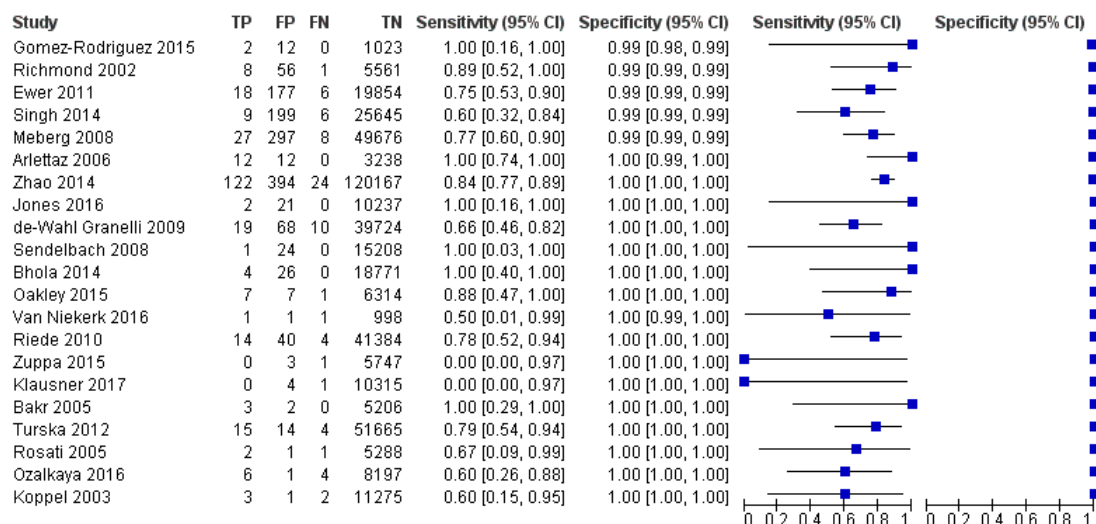
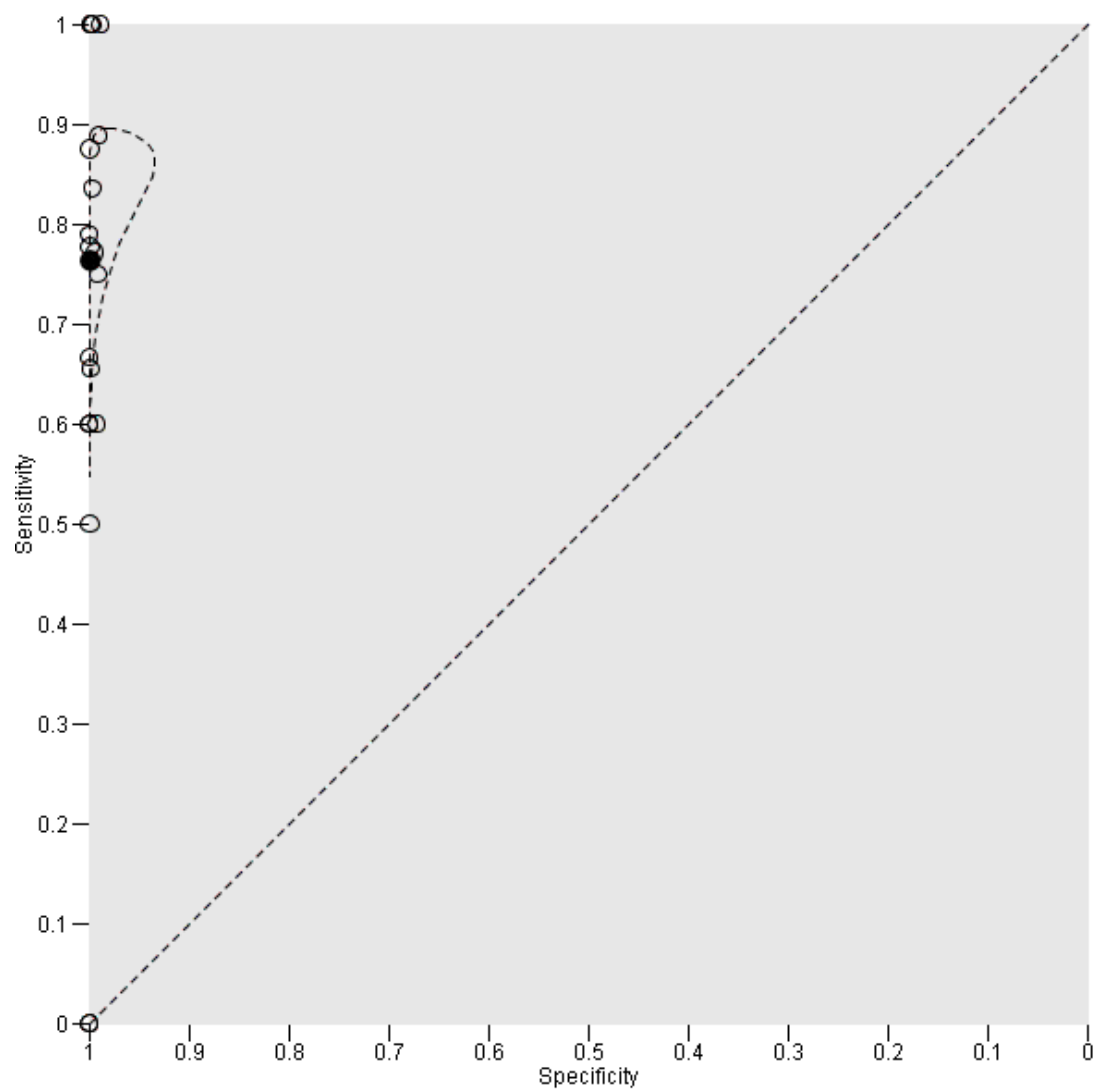


Figure 5. Summary ROC plot for pulse oximetry using a threshold lower than or lower than or equal to 95% (n = 19 studies). The solid circle corresponds to the summary estimate of sensitivity and specificity, and is shown with a 95% prediction region (dashed line).



For the primary analysis, we measured total between-study variability in sensitivity and in specificity through variances of the random effects for $\text{logit}(\text{sensitivity})$, $\text{logit}(\text{specificity})$, and their covariance, which were 0.102, 2.001, and -0.340, respectively. We represented the summary operating point with a 95% prediction region (Figure 5) and explored heterogeneity by differentiating studies on the basis of antenatal screening for CHD, timing of testing, type of oxygen saturation, study design, and risk of bias for the “flow and timing” domain of QUADAS-2. We plotted subgroups of studies in the ROC space.

Subgroup analysis and meta-regression

Table 2 summarizes results of the subgroup analysis including sensitivity and false-positive rates.

Antenatal diagnosis

Four studies included newborn infants with antenatal screening (Arlettaz 2006; Bhola 2014; Ewer 2011; Richmond 2002), and 15 studies did not (de-Wahl Granelli 2009; Gomez-Rodriguez 2015; Jones 2016; Klausner 2017; Koppel 2003; Meberg 2008; Oakley 2015; Ozalkaya 2016; Riede 2010; Rosati 2005; Singh 2014; Turska 2012; Van Niekerk 2016; Zhao 2014; Zuppa 2015). Summary estimates of sensitivity were 86.3% (95% CI 71.8 to 94.0) for studies that included antenatal screening, and 74.1% (95% CI 65.7 to 81.1) for studies that did not include antenatal screening. Summary estimates of specificity were 99.5% (95% CI 98.4 to 99.9) with a false-positive rate of 0.46% (95% CI 0.13 to 1.59) for studies with antenatal screening, and 99.9% (95% CI 99.8 to 100) with a false positive rate of 0.10% (95% CI 0.05 to 0.21) for studies that did not include antenatal screening. Sensitivity ($P = 0.071$) and specificity ($P = 0.231$) did not change significantly when newborn infants with antenatal suspicion of congenital heart defects were included compared with when they were excluded.

Test timing

Eleven studies performed pulse oximetry screening after 24 hours from birth (Bhola 2014; de-Wahl Granelli 2009; Klausner 2017; Koppel 2003; Oakley 2015; Ozalkaya 2016; Riede 2010; Rosati 2005; Van Niekerk 2016; Zhao 2014; Zuppa 2015), and the other eight studies performed pulse oximetry within 24 hours of birth (Arlettaz 2006; Ewer 2011; Gomez-Rodriguez 2015; Jones 2016; Meberg 2008; Richmond 2002; Singh 2014; Turska 2012). Summary estimates of sensitivity and specificity of studies that performed screening after 24 hours were 73.6% (95% CI 62.8 to 82.1) and 99.9% (95% CI 99.9 to 100). For studies that performed screening within 24 hours, summary estimates of sensitivity and specificity were 79.5% (95% CI 70.0 to 86.6) and 99.6%

(95% CI 99.1 to 99.8). Test timing to perform pulse oximetry had no significant effect on sensitivity ($P = 0.393$), but the false-positive rate for detection of CCHD was lower when newborn pulse oximetry was done after 24 hours from birth than when it was done within 24 hours (0.06% [95% CI 0.03 to 0.13] vs 0.42% [95% CI 0.20 to 0.89]; $P = 0.027$).

Limbs

Eleven studies used the foot alone (post-ductal) to measure oxygen saturation (Arlettaz 2006; Bhola 2014; Gomez-Rodriguez 2015; Koppel 2003; Meberg 2008; Oakley 2015; Richmond 2002; Riede 2010; Rosati 2005; Turska 2012; Zuppa 2015); summary estimates of sensitivity and specificity were 81.2% (95% CI 70.9 to 88.4) and 99.9% (95% CI 99.7 to 100), respectively, with a false-positive rate of 0.13% (95% CI 0.05 to 0.31). Eight studies used both right hand and foot (pre-ductal and post-ductal) (de-Wahl Granelli 2009; Ewer 2011; Jones 2016; Klausner 2017; Ozalkaya 2016; Singh 2014; Van Niekerk 2016; Zhao 2014); summary estimates of sensitivity and specificity for this group of studies were 71.2% (95% CI 58.5 to 81.3) and 99.8% (95% CI 99.5 to 99.9), respectively, with a false-positive rate of 0.17% (95% CI 0.06 to 0.46). We noted no significant differences in sensitivity ($P = 0.197$) nor in specificity ($P = 0.718$) for pulse oximetry when measures were obtained in the foot alone rather than in both the foot and the right hand.

Risk of bias

We judged nine studies as having unclear risk of bias for the “flow and timing” domain of QUADAS-2 (Arlettaz 2006; Bhola 2014; Klausner 2017; Koppel 2003; Riede 2010; Rosati 2005; Turska 2012; Van Niekerk 2016; Zuppa 2015). Summary estimates of sensitivity and specificity were 77.8% (95% CI 64.1 to 87.3) and 100% (95% CI 99.9 to 100), respectively, with a false-positive rate of 0.05% (95% CI 0.02 to 0.12). We judged the remaining 10 studies as having low risk of bias (de-Wahl Granelli 2009; Ewer 2011; Gomez-Rodriguez 2015; Jones 2016; Meberg 2008; Oakley 2015; Ozalkaya 2016; Richmond 2002; Singh 2014; Zhao 2014); summary estimates of sensitivity and specificity were 77.3% (95% CI 68.8 to 84.0) and 99.7% (95% CI 99.3 to 99.8), respectively, with a false-positive rate of 0.34% (95% CI 0.17 to 0.66). Risk of bias for this domain had no significant effect on sensitivity ($P = 0.937$), but studies judged as having unclear risk of bias for the “flow and timing” domain had higher specificity ($P = 0.016$).

Sensitivity analysis

We performed a sensitivity analysis while excluding from the primary analysis studies with a threshold $\leq 95\%$ (Jones 2016; Koppel

2003; Ozalkaya 2016; Riede 2010; Rosati 2005). For this analysis, sensitivity and specificity were 78.1% (95% CI 71.2 to 83.7) and 99.8% (95% CI 99.6 to 99.9) with a false-positive rate of 0.23% (95% CI 0.12 to 0.44). Exclusion of these studies increased the sensitivity and false-positive rate of pulse oximetry screening.

We also performed a sensitivity analysis for which we added to the primary analysis studies with a threshold $\leq 94\%$ and $< 96\%$ (Bakr 2005; Sendelbach 2008). For this analysis, sensitivity and specificity were 77% (95% CI 70 to 82) and 100% (95% CI 100

to 100), respectively. Inclusion of these studies produced a slight improvement in the sensitivity of the test.

Furthermore, we investigated the effects of potential sources of bias by removing the four studies judged as having high risk of bias in one of the QUADAS-2 domains (Arlettaz 2006; Oakley 2015; Van Niekerk 2016; Zuppa 2015). For this analysis, sensitivity and specificity were similar to those in the main analysis, at 75.5% (95% CI 68.2 to 81.6) and 99.79% (95% CI 99.7 to 99.9), respectively.

Summary of findings

Should pulse oximetry be used to diagnose CCHD in asymptomatic newborns?								
Patient or population: asymptomatic newborns at the time of pulse oximetry screening								
Setting: hospital births								
Index test: pulse oximetry								
Reference test: Reference standards were both diagnostic echocardiography (echocardiogram) and clinical follow-up in the first 28 days of life, including postmortem findings and mortality and congenital anomaly databases to identify false-negative patients								
Studies: We included prospective or retrospective cohorts and cross-sectional studies. We excluded case reports and studies of case-control design								
Threshold	Summary racy (95% CI)	accu- racy	Number of participants (diseased / non-diseased) Number of studies	Prevalence dian (range)	Implications (in a cohort of 10,000 newborns tested [95% CI])			Certainty of the evidence (GRADE)
					Prevalence 0.6 per 1000	Prevalence 0.1 per 1000	Prevalence 3.7 per 1000	
95% (less than or less than or equal to)	Sensitivity 76.3% (69.5 to 82.0) Specificity 99.9% (99.7 to 99.9)		436,758 (345/ 436,413) 19 studies	0.6 per 1000 (0.1 to 3.7)	True positives	5	1	28
					(newborns with CCHD)	(4 to 5)	(1 to 1)	(26 to 30)
					False negatives	1	0	9
					(newborns in- correctly classified as not having CCHD)	(1 to 2)	(0 to 0)	(7 to 11)
					True negatives	9980	9985	9949
					(newborns without CCHD)	(9966 to 9987)	(9971 to 9992)	(9935 to 9956)
								LOW* ⊕⊕○○
								HIGH ⊕⊕⊕⊕

	False positives (newborns in- correctly classified as having CCHD)	14 (7 to 28)	14 (7 to 28)	14 (7 to 28)
CCHD: critical congenital heart defect; CI: confidence interval				
Sensitivity: <i>*We have downgraded certainty of the evidence from high to low because the low number of CCHD cases included in the review (serious imprecision) and secondly, there was a serious risk of differential verification bias (ie, diagnosis was established by echocardiography in test positive cases however test negatives were usually confirmed by clinical follow-up or by accessing congenital malformation registries and mortality databases)."</i>				
Certainty of the evidence (Balshem 2011) <i>High certainty:</i> We are very confident that the true effect lies close to that of the estimate of the effect. <i>Moderate certainty:</i> We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different. <i>Low certainty:</i> Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect. <i>Very low certainty:</i> We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect				

DISCUSSION

Summary of main results

For this review, we have identified and summarized the results of all available cohort studies reporting the test accuracy of pulse oximetry screening for detection of critical congenital heart defects (CCHDs) in asymptomatic late preterm and full-term infants in postnatal wards or well-baby nurseries. We have presented the main results in [Summary of findings](#). We analyzed data on 457,202 participants from 21 included studies. We restricted the primary analysis to studies with thresholds around 95% ($< 95\%$ and $\leq 95\%$). Analysis, including 436,758 participants from 19 studies, showed that pulse oximetry screening is a highly specific screening test with moderate sensitivity and a low overall false-positive rate. Overall sensitivity was 76.3%, specificity was 99.9%, and the false-positive rate was 0.14%. Summary positive and negative likelihood ratios were 535.6 and 0.24, respectively. Inclusion of studies that used different saturation thresholds from those in the primary analysis slightly improved the sensitivity of the test. Exclusion of studies at high risk of bias did not significantly alter overall sensitivity or specificity. Between-study heterogeneity was higher in sensitivity than in specificity estimates.

Most studies were conducted in high-income countries (USA, Europe); however, we also included studies from middle-income countries, which increases the generalizability of review findings. We noted methodological variation between studies with respect to inclusion or exclusion of babies with a suspected antenatal diagnosis, timing of testing (before or after 24 hours of age), site of testing (post-ductal only or pre-ductal and post-ductal), functional or fractional saturation measurement, and study design (prospective or retrospective). Subgroup analysis showed no effect on sensitivity or specificity among these variables, although later screening was associated with a lower false-positive rate than was reported with earlier screening.

The definition of CCHD provided in the published literature is highly variable. We attempted to address this by applying a strict definition (see above) to categorize CCHD in a standardized manner, thus reducing the risk of an incorrect diagnosis.

Strengths and weaknesses of the review

Strengths of this review include a comprehensive literature search performed to identify all relevant studies, rigorous assessment of risk of bias of included studies using the QUADAS-2 tool, duplicate data extraction, and performance of subgroup and sensitivity analyses to investigate differences in estimates of accuracy of pulse oximetry among studies with high, low, or unclear risk of bias. However, only one study included more than 100 CCHD cases, and 12 studies included fewer than 10 cases. The relatively low

number of CCHD cases included in this review indicates that the precision of sensitivity is still low.

Our review has explored and quantified the heterogeneity, and review authors have tried to identify possible sources of heterogeneity. Exploration of sources of heterogeneity has produced different results for sensitivity and specificity. Sensitivity has not been affected by any of the a priori selected sources of heterogeneity. We cannot rule out the presence of unexplained heterogeneity in this accuracy index, although it is highly likely that some of the variability observed in sensitivities of individual studies could be explained by the paucity of CCHD cases. Use of different strategies for confirming pulse oximetry negative cases (ie, passively with mortality or registry data rather than active clinical follow-up) could well have introduced some degree of heterogeneity into sensitivity results. However, this post hoc exploration was not performed, given the scarcity of data. This means that sensitivity estimates are somewhat unstable with wide confidence intervals. At the same time, this scarcity made analysis of heterogeneity underpowered. Conversely, specificity was affected by the timing of the test and by the risk of bias due to the flow and timing domain of the QUADAS-2 tool. Statistical significance achieved by the specificity analysis is a direct consequence of the large number of healthy newborns included in the review. On the other hand, the magnitude of differences between subgroup analyses was small. False-positive rates were 0.06% and 0.42% for newborns screened after and before 24 hours of birth, respectively. The absolute difference was 0.36% with more false-positives in the earlier screening group as compared with the late screening group. This means, in relative terms, seven times more false positives are seen in the earlier screening group than in the late screening group. Similarly, false-positive rates varied between studies judged as having unclear or low risk of bias for the “flow and timing” domain of QUADAS-2 (ie, 0.05% vs 0.34% for unclear and low risk groups of studies, respectively). The absolute difference a 0.29% reduction in false positives in the unclear risk group, which equates almost seven times fewer false positives in relative terms.

Agreements and disagreements with other studies or reviews

This review includes more studies and a larger body of data from a significantly greater number of infants than were included in similar previous systematic reviews of the test accuracy of pulse oximetry screening to detect CCHD ([Mahle 2009](#); [Thangaratinam 2007](#); [Thangaratinam 2012](#)), which reported identical statistical methods and meta-analyses. Compared with the largest prior review ([Thangaratinam 2012](#)), authors of this review screened a significantly larger number of references (2695 vs 552) and included data from over 220,000 more babies, allowing greater precision of the estimates of test accuracy, and providing the most complete meta-analysis available so far.

Overall sensitivity is similar (76.3% vs 76.5%) to that described by [Thangaratinam 2012](#) and is similar to the overall false-positive

rate (0.14% vs 0.14%). The statistically significant lower false-positive rate between early and later screening persists (0.06% vs 0.42% and 0.05% vs 0.5%).

Applicability of findings to the review question

This review includes a large number of relevant studies that met our inclusion criteria, and review authors had few concerns about the relevance of their findings to our review questions. We mainly judged included studies to be at low or unclear risk of bias in QUADAS-2 domains. Most studies had a prospective design with consecutive enrollment of participants and included an adequate description of the index test. Some studies reported the exclusion criteria poorly. Data were complete and were available for all included studies.

Risk of differential verification bias was unavoidable as diagnosis was established by echocardiography in test-positive cases; however, test-negative cases were usually confirmed by clinical follow-up or by examination of congenital malformation registries and mortality databases; risk of bias in the conduct or interpretation of reference standard(s) was unclear in most studies that used incomplete reference standards. This of course raises the possibility that some of the false negatives may be misclassified as true negatives. This misclassification overestimates sensitivity and specificity. Differential verification may have had an impact on the sensitivity estimate. For this reason, along with the potential for imprecision, given the small number of CCHD cases, we have downgraded the GRADE certainty of evidence for sensitivity to “low.” In our review, studies judged as having unclear risk of bias for the “flow and timing” domain showed higher specificity.

AUTHORS' CONCLUSIONS

Implications for practice

This review provides further compelling evidence for the use of pulse oximetry as a routine screening test for early identification of CCHD in asymptomatic babies in the well-baby nursery. The test has high specificity and moderate sensitivity and meets the criteria for universal screening.

Current evidence supports the introduction of routine screening for CCHD in asymptomatic newborns before discharge from the well-baby nursery. The test appears feasible in various middle-income countries and shows consistent test accuracy.

Some important elements regarding specific screening algorithms need further consideration. Data show no difference in sensitivity based on the site of testing (pre-ductal or pre-ductal and post-ductal). However, only two studies using pre-ductal and post-ductal saturations reported absolute saturation values rather than just

test results (de-Wahl Granelli 2009; Ewer 2011). As has been reported previously by Ewer 2016 and Thangaratinam 2012, several CCHD cases that were detected by pre-ductal and post-ductal testing would have been missed by post-ductal testing alone, but the numbers are too small to affect sensitivity analysis results.

In addition, the finding of a lower false-positive rate with screening after 24 hours needs to be balanced against the fact that many countries discharge babies within 24 hours and - as is important to note - most reported studies did not take into account the risk that a baby with CCHD or other serious illness may present before screening takes place (de-Wahl Granelli 2014; Ewer 2011; Ewer 2016; Riede 2010; Thangaratinam 2012).

The prevalence of CCHD is quite low, and most test-positive infants do not have the target condition. The false-positive rate is variable and depends largely on the timing of the screening (earlier screening - within 24 hours of age - has a higher false-positive rate than screening after 24 hours). This raises concerns that a false-positive test may unnecessarily increase parental anxiety and may lead to avoidable investigations and delayed discharge. Investigators in the UK PulseOx study assessed the acceptability of pulse oximetry screening and reported on anxiety created by the test - particularly among mothers of false-positive (FP) babies (Ewer 2012a; Powell 2013). Investigators quantified satisfaction with, and perceptions of, the test and anxiety and depression following screening by using validated questionnaires on samples of mothers whose babies were true positive, false positive, and true negative. All participants were predominantly satisfied with screening, and it is important to note that mothers given false-positive results after screening were no more anxious than those given true-negative results. Many studies report identification of alternative non-cardiac conditions by pulse oximetry screening. Although these conditions - such as congenital pneumonia and early-onset sepsis - are technically false positives, their identification may be seen as a positive additional benefit of screening; they are more likely to be detected within the first 24 hours, allowing early treatment of individuals with these potentially serious conditions. Healthcare providers must consider the potential for overdiagnosis of these conditions following screening and must apply rigorous criteria to classify these conditions.

Implications for research

The large sample size of this review along with precise estimates of sensitivity and specificity suggests that further research into the accuracy of this screening method is unnecessary. In addition, several countries, including the USA, have already implemented screening. However, given concerns related to differential verification, we propose that monitoring of screening outcomes (including possible reduction in early mortality) and management of false positives should be performed in a rigorous manner.

Further evidence regarding the routine screening of babies outside

the well-baby nursery (including non-intensive care unit [NICU] stays and out-of-hospital births) is required. Additional raw saturation data and further analysis are required to further elucidate the relative sensitivities of post-ductal versus pre-ductal and post-ductal saturation testing.

The ability of pulse oximetry to detect non-cardiac illness such as respiratory and infectious conditions has been well described, but test accuracy remains unclear.

ACKNOWLEDGEMENTS

We would like to thank Colleen Ovelman (Managing Editor and Assistant Information Specialist), Yolanda Brosseau (Webmaster), and Jennifer Spano (Information Specialist) of the Neonatal Group for their kind and efficient support during preparation of this review.

REFERENCES

References to studies included in this review

Arlettaz 2006 *{published data only}*

Arlettaz R, Bauschatz AS, Monkhoff M, Essers B, Bauersfeld U. The contribution of pulse oximetry to the early detection of congenital heart disease in newborns. *European Journal of Pediatrics* 2006;**165**(2):94–8. PUBMED: 10.1007/s00431-005-0006-y; PUBMED: 16211399]

Bakr 2005 *{published data only}*

Bakr AF, Habib HS. Combining pulse oximetry and clinical examination in screening for congenital heart disease. *Pediatric Cardiology* 2005;**26**(6):832–5. [DOI: 10.1007/s00246-005-0981-9; PUBMED: 16088415]

Bhola 2014 *{published data only}*

Bhola K, Kluckow M, Evans N. Post-implementation review of pulse oximetry screening of well newborns in an Australian tertiary maternity hospital. *Journal of Paediatrics and Child Health* 2014;**50**(11):920–5. [DOI: 10.1111/jpc.12651; PUBMED: 24923996]

de-Wahl Granelli 2009 *{published data only}*

de-Wahl Granelli A, Wennergren M, Sandberg K, Mellander M, Bejlum C, Inganas L, et al. Impact of pulse oximetry screening on the detection of duct dependent congenital heart disease: a Swedish prospective screening study in 39, 821 newborns. *BMJ (Clinical Research Ed.)* 2009;**338**:a3037. [DOI: 10.1136/bmj.a3037; PUBMED: 19131383]

Ewer 2011 *{published data only}*

Ewer AK, Middleton LJ, Furmston AT, Bhojar A, Daniels JP, Thangaratinam S, et al. PulseOx Study Group. Pulse oximetry screening for congenital heart defects in newborn infants (PulseOx): a test accuracy study. *Lancet* 2011;**378**(9793):785–94. [DOI: 10.1016/S0140-6736(11)60753-8; PUBMED: 21820732]

Gomez-Rodriguez 2015 *{published data only}*

Gomez-Rodriguez G, Quezada-Herrera A, Amador-Licona N, Carballo-Magdalena D, Rodriguez-Mejia EJ, Guizar-Mendoza JM. Pulse oximetry as a screening test for critical congenital heart disease in term newborns. *Revista de Investigacion Clinica* 2015;**67**(2):130–4. PUBMED: 25938847]

Jones 2016 *{published data only}*

Jones AJ, Howarth C, Nicholl R, Mat-Ali E, Knowles R. The impact and efficacy of routine pulse oximetry screening

for CHD in a local hospital. *Cardiology in the Young* 2016;**26**(7):1397–405. [DOI: 10.1017/S1047951115002784; PUBMED: 26905447]

Klausner 2017 *{published data only}*

Klausner R, Shapiro ED, Elder RW, Colson E, Loyal J. Evaluation of a screening program to detect critical congenital heart defects in newborns. *Hospital Pediatrics* 2017;**7**(4):214–8. [DOI: 10.1542/hpeds.2016-0176; PUBMED: 28250095]

Koppel 2003 *{published data only}*

Koppel RI, Druschel CM, Carter T, Goldberg BE, Mehta PN, Talwar R, et al. Effectiveness of pulse oximetry screening for congenital heart disease in asymptomatic newborns. *Pediatrics* 2003;**111**(3):451–5. PUBMED: 12612220]

Meberg 2008 *{published data only}*

Meberg A, Brugmann-Pieper S, Due R Jr, Eskedal L, Fagerli I, Farstad T, et al. First day of life pulse oximetry screening to detect congenital heart defects. *Journal of Pediatrics* 2008;**152**(6):761–5. [DOI: 10.1016/j.jpeds.2007.12.043; PUBMED: 18492511]

Oakley 2015 *{published data only}*

Oakley JL, Soni NB, Wilson D, Sen S. Effectiveness of pulse-oximetry in addition to routine neonatal examination in detection of congenital heart disease in asymptomatic newborns. *Journal of Maternal-Fetal & Neonatal Medicine* 2015;**28**(14):1736–9. [DOI: 10.3109/14767058.2014.967674; PUBMED: 25241768]

Ozalkaya 2016 *{published data only}*

Ozalkaya E, Akdag A, Sen I, Comert E, Melek Yaren H. Early screening for critical congenital heart defects in asymptomatic newborns in Bursa province. *Journal of Maternal-Fetal & Neonatal Medicine* 2016;**29**(7):1105–7. [DOI: 10.3109/14767058.2015.1035642; PUBMED: 25902399]

Richmond 2002 *{published data only}*

Richmond S, Reay G, Abu Harb M. Routine pulse oximetry in the asymptomatic newborn. *Archives of Disease in Childhood. Fetal and Neonatal Edition* 2002;**87**(2):F83–8. PUBMED: 12193511]

Riede 2010 *{published data only}*

Riede FT, Worner C, Dahnert I, Mockel A, Kostelka M, Schneider P. Effectiveness of neonatal pulse oximetry

- screening for detection of critical congenital heart disease in daily clinical routine - results from a prospective multicenter study. *European Journal of Pediatrics* 2010; **169**(8):975–81. [DOI: 10.1007/s00431-010-1160-4; PUBMED: 20195633]
- Rosati 2005** *{published data only}*
Rosati E, Chitano G, Dipaola L, De Felice C, Latini G. Indications and limitations for a neonatal pulse oximetry screening of critical congenital heart disease. *Journal of Perinatal Medicine* 2005;**33**(5):455–7. [DOI: 10.1515/JPM.2005.080; PUBMED: 16238542]
- Sendelbach 2008** *{published data only}*
Sendelbach DM, Jackson GL, Lai SS, Fixler DE, Stehel EK, Engle WD. Pulse oximetry screening at 4 hours of age to detect critical congenital heart defects. *Pediatrics* 2008;**122**(4):e815–20. [DOI: 10.1542/peds.2008-0781; PUBMED: 18762486]
- Singh 2014** *{published data only}*
Singh A, Rasiyah SV, Ewer AK. The impact of routine pre-discharge pulse oximetry screening in a regional neonatal unit. *Archives of Disease in Childhood. Fetal and Neonatal Edition* 2014;**99**(4):F297–302. [DOI: 10.1136/archdischild-2013-305657; PUBMED: 24646619]
- Turska 2012** *{published data only}*
Turska Kmiec A, Borszewska Kornacka MK, Blaz W, Kawalec W, Zuk M. Early screening for critical congenital heart defects in asymptomatic newborns in Mazovia province: experience of the POLKARD pulse oximetry programme 2006–2008 in Poland. *Kardiologia Polska* 2012; **70**(4):370–6. 22528711]
- Van Niekerk 2016** *{published data only}*
Van Niekerk AM, Cullis RM, Linley LL, Zuhlke L. Feasibility of pulse oximetry pre-discharge screening implementation for detecting critical congenital heart lesions in newborns in a secondary level maternity hospital in the Western Cape, South Africa: The 'POPSICLE' study. *South African Medical Journal* 2016;**106**(8):817–21. PUBMED: 27499412]
- Zhao 2014** *{published data only}*
Zhao QM, Ma XJ, Ge XL, Liu F, Yan WL, Wu L, et al. Neonatal Congenital Heart Disease Screening Group. Pulse oximetry with clinical assessment to screen for congenital heart disease in neonates in China: a prospective study. *Lancet* 2014;**384**(9945):747–54. [DOI: 10.1016/S0140-6736(14)60198-7; PUBMED: 24768155]
- Zuppa 2015** *{published data only}*
Zuppa AA, Riccardi R, Catenazzi P, D'Andrea V, Cavani M, D'Antuono A, et al. Clinical examination and pulse oximetry as screening for congenital heart disease in low-risk newborn. *Journal of Maternal-Fetal & Neonatal Medicine* 2015;**28**(1):7–11. [DOI: 10.3109/14767058.2014.899573; PUBMED: 24588079]
- Andrews 2014** *{published data only}*
Andrews JP, Ross AS, Salazar MA, Tracy NA, Burke BL Jr. Smooth implementation of critical congenital heart defect screening in a newborn nursery. *Clinical Pediatrics* 2014;**53**(2):173–6. [DOI: 10.1177/0009922813502850; PUBMED: 24037922]
- Ewer 2012a** *{published data only}*
Ewer AK, Furmston AT, Middleton LJ, Deeks JJ, Daniels JP, Pattison HM, et al. Pulse oximetry as a screening test for congenital heart defects in newborn infants: a test accuracy study with evaluation of acceptability and cost-effectiveness. *Health Technology Assessment (Winchester, England)* 2012;**16**(2):v-xiii, 1–184. [DOI: 10.3310/hta16020; PUBMED: 22284744]
- Hoke 2002** *{published data only}*
Hoke TR, Donohue PK, Bawa PK, Mitchell RD, Pathak A, Rowe PC, et al. Oxygen saturation as a screening test for critical congenital heart disease: a preliminary study. *Pediatric Cardiology* 2002;**23**(4):403–9. [DOI: 10.1007/s00246-002-1482-8; PUBMED: 12170356]
- John 2016** *{published data only}*
John C, Phillips J, Hamilton C, Lastliger A. Implementing universal pulse oximetry screening in West Virginia: findings from year one. *West Virginia Medical Journal* 2016; **112**(4):42–6. PUBMED: 27491102]
- Kardasevic 2016** *{published data only}*
Kardasevic M, Jovanovic I, Samardzic JP. Modern strategy for identification of congenital heart defects in the neonatal period. *Medical Archives* 2016;**70**(5):384–8. [DOI: 10.5455/medarh.2016.70.384-388; PUBMED: 27994302]
- Kochilas 2013** *{published data only}*
Kochilas LK, Lohr JL, Bruhn E, Borman-Shoap E, Gams BL, Pylipow M, et al. Implementation of critical congenital heart disease screening in Minnesota. *Pediatrics* 2013;**132**(3):e587–94. [DOI: 10.1542/peds.2013-0803; PUBMED: 23958775]
- Lhost 2014** *{published data only}*
Lhost JJ, Goetz EM, Belling JD, van Roojen WM, Spicer G, Hokanson JS. Pulse oximetry screening for critical congenital heart disease in planned out-of-hospital births. *Journal of Pediatrics* 2014;**165**(3):485–9. [DOI: 10.1016/j.jpeds.2014.05.011; PUBMED: 24948344]
- Meberg 2009** *{published data only}*
Meberg A, Andreassen A, Brunvand L, Markestad T, Moster D, Nietsch L, et al. Pulse oximetry screening as a complementary strategy to detect critical congenital heart defects. *Acta Paediatrica* 2009;**98**(4):682–6. [DOI: 10.1111/j.1651-2227.2008.01199.x; PUBMED: 19154526]
- Movahedian 2016** *{published data only}*
Movahedian AH, Mosayebi Z, Sagheb S. Evaluation of pulse oximetry in the early detection of cyanotic congenital heart disease in newborns. *Journal of Tehran Heart Center* 2016;**11**(2):73–8. PUBMED: 27928258]

References to studies excluded from this review

Narayan 2016a {published data only}

Narayan IC, Blom NA, Bourgonje MS, Haak MC, Smit M, Posthumus F, et al. Pulse oximetry screening for critical congenital heart disease after home birth and early discharge. *Journal of Pediatrics* 2016;**170**:188–92.e1. PUBMED: 10.1016/j.jpeds.2015.12.004]

Prudhoe 2013 {published data only}

Prudhoe S, Abu-Harb M, Richmond S, Wren C. Neonatal screening for critical cardiovascular anomalies using pulse oximetry. *Archives of Disease in Childhood. Fetal and Neonatal Edition* 2013;**98**(4):F346–50. [DOI: 10.1136/archdischild-2012-302045; PUBMED: 23341250]

Reich 2003 {published data only}

Reich JD, Miller S, Brogdon B, Casatelli J, Gompf TC, Huhta JC, et al. The use of pulse oximetry to detect congenital heart disease. *Journal of Pediatrics* 2003;**142**(3):268–72. [DOI: 10.1067/mpd.2003.87; PUBMED: 12640374]

Reich 2008 {published data only}

Reich JD, Connolly B, Bradley G, Littman S, Koeppl W, Lewycky P, et al. The reliability of a single pulse oximetry reading as a screening test for congenital heart disease in otherwise asymptomatic newborn infants. *Pediatric Cardiology* 2008;**29**(5):885–9. [DOI: 10.1007/s00246-008-9214-3; PUBMED: 18347842]

Reich 2008a {published data only}

Reich JD, Connolly B, Bradley G, Littman S, Koeppl W, Lewycky P, et al. Reliability of a single pulse oximetry reading as a screening test for congenital heart disease in otherwise asymptomatic newborn infants: the importance of human factors. *Pediatric Cardiology* 2008;**29**(2):371–6. [DOI: 10.1007/s00246-007-9105-z; PUBMED: 17932712]

Riede 2009 {published data only}

Riede FT, Dahnert I, Schneider P, Mockel A. Pulse oximetry screening at 4 hours of age to detect critical congenital heart defects. *Pediatrics* 2009; Vol. 123, issue 3:e542. [DOI: 10.1542/peds.2008-3598; PUBMED: 19254989]

Ruangritnamchai 2007 {published data only}

Ruangritnamchai C, Bunjapamai W, Pongpanich B. Pulse oximetry screening for clinically unrecognized critical congenital heart disease in the newborns. *Images in Paediatric Cardiology* 2007;**9**(1):10–5. PUBMED: 22368668]

Saha 2014 {published data only}

Saha A, Mathew JL, Chawla D, Kumar D. How useful is pulse oximetry for screening of congenital heart disease in newborns?. *Indian Pediatrics* 2014;**51**(11):913–5. PUBMED: 25432224]

Saxena 2015 {published data only}

Saxena A, Mehta A, Ramakrishnan S, Sharma M, Salhan S, Kalaivani M, et al. Pulse oximetry as a screening tool for detecting major congenital heart defects in Indian newborns. *Archives of Disease in Childhood. Fetal and Neonatal Edition* 2015;**100**(5):F416–21. [DOI: 10.1136/archdischild-2014-307485; PUBMED: 26038347]

Schena 2017 {published data only}

Schena F, Picciolli I, Agosti M, Zuppa AA, Zuccotti G, Parola L, et al. Neonatal Cardiology Study Group of the Italian Society of Neonatology. Perfusion index and pulse oximetry screening for congenital heart defects. *Journal of Pediatrics* 2017;**183**:74–9.e1. [DOI: 10.1016/j.jpeds.2016.12.076; PUBMED: 28153478]

Studer 2014 {published data only}

Studer MA, Smith AE, Lustik MB, Carr MR. Newborn pulse oximetry screening to detect critical congenital heart disease. *Journal of Pediatrics* 2014;**164**(3):505–9.e1-2. [DOI: 10.1016/j.jpeds.2013.10.065; PUBMED: 24315501]

Taksande 2013 {published data only}

Taksande AM, Lakhkar B, Gadekar A, Suwarnakar K, Japzape T. Accuracy of pulse oximetry screening for detecting critical congenital heart disease in the newborns in rural hospital of Central India. *Images in Paediatric Cardiology* 2013;**15**(4):5–10. PUBMED: 26236364]

Tautz 2010 {published data only}

Tautz J, Merkel C, Loersch F, Egen O, Hagele F, Thon HM, et al. Implication of pulse oximetry screening for detection of congenital heart defects. *Klinische Padiatrie* 2010;**222**(5):291–5. [DOI: 10.1055/s-0030-1253391; PUBMED: 20458668]

Tsao 2016 {published data only}

Tsao PC, Shiau YS, Chiang SH, Ho HC, Liu YL, Chung YF, et al. Development of a newborn screening program for critical congenital heart disease (CCHD) in Taipei. *PLoS One* 2016;**11**(4):e0153407. [DOI: 10.1371/journal.pone.0153407; PUBMED: 27073996]

Vaidyanathan 2011 {published data only}

Vaidyanathan B, Sathish G, Mohanan ST, Sundaram KR, Warriar KK, Kumar RK. Clinical screening for congenital heart disease at birth: a prospective study in a community hospital in Kerala. *Indian Pediatrics* 2011;**48**(1):25–30. PUBMED: 20972295]

Valmari 2006 {published data only}

Valmari P, Jauhola R, Leskinen M, Heikinheimo M. [Oxygen saturation to detect asymptomatic congenital heart disease of newborns] [Vastasyntyneen kavalat sydanviat—aika aloittaa happikylläisyyssuulontaa?]. *Duodecim; Laaketieteellinen Aikakauskirja* 2006;**122**(2):173–6. PUBMED: 16509060]

Walsh 2011 {published data only}

Walsh W. Evaluation of pulse oximetry screening in Middle Tennessee: cases for consideration before universal screening. *Journal of Perinatology* 2011;**31**(2):125–9. [DOI: 10.1038/jp.2010.70; PUBMED: 20508595]

Additional references**Abu-Harb 1994**

Abu-Harb M, Hey E, Wren C. Death in infancy from unrecognized congenital heart disease. *Archives of Disease in Childhood* 1994;**71**(1):3–7. [PUBMED: 8067789]

Abu-Harb 1994a

Abu-Harb M, Wyllie J, Hey E, Richmond S, Wren C. Presentation of obstructive left heart malformations in infancy. *Archives of Disease in Childhood. Fetal and Neonatal Edition* 1994;**71**(3):F179–83. [PUBMED: 7820713]

Allan 1986

Allan LD, Crawford DC, Chita SK, Tynan MJ. Prenatal screening for congenital heart disease. *British Medical Journal (Clinical Research Ed.)* 1986;**292**(6537):1717–9. [PUBMED: 3089369]

Balslem 2011

Balslem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. *Journal of Clinical Epidemiology* 2011;**64**(4):401–6. [PUBMED: 21208779]

Boneva 2001

Boneva RS, Botto LD, Moore CA, Yang Q, Correa A, Erickson JD. Mortality associated with congenital heart defects in the United States: trends and racial disparities, 1979–1997. *Circulation* 2001;**103**(19):2376–81. [PUBMED: 11352887]

Botto 2001

Botto LD, Correa A, Erickson JD. Racial and temporal variations in the prevalence of heart defects. *Pediatrics* 2001;**107**(3):E32. [PUBMED: 11230613]

Brown 2001

Brown JW, Park HJ, Turrentine MW. Arterial switch operation: factors impacting survival in the current era. *Annals of Thoracic Surgery* 2001;**71**(6):1978–84. [PUBMED: 11426778]

Brown 2006

Brown KL, Ridout DA, Hoskote A, Verhulst L, Ricci M, Bull C. Delayed diagnosis of congenital heart disease worsens preoperative condition and outcome of surgery in neonates. *Heart (British Cardiac Society)* 2006;**92**(9):1298–302. [DOI: 10.1136/hrt.2005.078097; PUBMED: 16449514]

Bull 1999

Bull C. Current and potential impact of fetal diagnosis on prevalence and spectrum of serious congenital heart disease at term in the UK. British Paediatric Cardiac Association. *Lancet* 1999;**354**(9186):1242–7. [PUBMED: 10520632]

Carvalho 2002

Carvalho JS, Mavrides E, Shinebourne EA, Campbell S, Thilaganathan B. Improving the effectiveness of routine prenatal screening for major congenital heart defects. *Heart (British Cardiac Society)* 2002;**88**(4):387–91. [PUBMED: 12231598]

Cawsey 2016

Cawsey MJ, Noble S, Cross-Sudworth F, Ewer AK. Feasibility of pulse oximetry screening for critical congenital heart defects in homebirths. *Archives of Disease in Childhood. Fetal and Neonatal Edition* 2016;**101**(4):F349–51. [PUBMED: 26915671]

Chang 2008

Chang RK, Gurvitz M, Rodriguez S. Missed diagnosis of critical congenital heart disease. *Archives of Pediatrics & Adolescent Medicine* 2008;**162**(10):969–74. [DOI: 10.1001/archpedi.162.10.969; PUBMED: 18838650]

Chew 2007

Chew C, Halliday JL, Riley MM, Penny DJ. Population-based study of antenatal detection of congenital heart disease by ultrasound examination. *Ultrasound in Obstetrics & Gynecology* 2007;**29**(6):619–24. [DOI: 10.1002/uog.4023; PUBMED: 17523161]

Chu 2006

Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology* 2006; Vol. 59, issue 12:1331–2; author reply 1332–3.

de-Wahl Granelli 2014

de-Wahl Granelli A, Meberg A, Ojala T, Steensberg J, Oskarsson G, Mellander M. Nordic pulse oximetry screening - implementation status and proposal for uniform guidelines. *Acta Paediatrica* 2014;**103**(11):1136–42. [DOI: 10.1111/apa.12758; PUBMED: 25060211]

Ewer 2012b

Ewer AK. How to develop a business case to establish a neonatal pulse oximetry programme for screening of congenital heart defects. *Early Human Development* 2012;**88**(12):915–9. [DOI: 10.1016/j.earlhumdev.2012.09.013; PUBMED: 23089530]

Ewer 2013

Ewer AK. Review of pulse oximetry screening for critical congenital heart defects in newborn infants. *Current Opinion in Cardiology* 2013;**28**(2):92–6. [DOI: 10.1097/HCO.0b013e32835d7e42; PUBMED: 23381095]

Ewer 2014

Ewer AK. Pulse oximetry screening for critical congenital heart defects in newborn infants: should it be routine?. *Archives of Disease in Childhood. Fetal and Neonatal Edition* 2014;**99**(1):F93–5. [DOI: 10.1136/archdischild-2013-303968; PUBMED: 23934364]

Ewer 2016

Ewer AK, Martin GR. Newborn pulse oximetry screening: which algorithm is best?. *Pediatrics* 2016;**138**(5):e20161206. [DOI: 10.1542/peds.2016-1206; PUBMED: 27940777]

Franklin 2002

Franklin O, Burch M, Manning N, Sleeman K, Gould S, Archer N. Prenatal diagnosis of coarctation of the aorta improves survival and reduces morbidity. *Heart* 2002;**87**(1):67–9. [PUBMED: 11751670]

Garne 2001

Garne E, Stoll C, Clementi M, Euroscan Group. Evaluation of prenatal diagnosis of congenital heart diseases by ultrasound: experience from 20 European registries. *Ultrasound in Obstetrics & Gynecology* 2001;**17**(5):386–91. [DOI: 10.1046/j.1469-0705.2001.00385.x; PUBMED: 11380961]

GRADEpro GDT [Computer program]

GRADEproGDT [Computer program]. Version accessed March 2017. GRADE Handbook. Hamilton (ON): McMaster University (developed by Evidence Prime), 2015. Available from gradepr.org.

Hall 1999

Hall DM. The role of the routine neonatal examination. *BMJ (Clinical Research Ed.)* 1999; Vol. 318, issue 7184: 619–20. [PUBMED: 10066183]

Han 2013

Han LM, Klewer SE, Blank KM, Seckeler MD, Barber BJ. Feasibility of pulse oximetry screening for critical congenital heart disease at 2643-foot elevation. *Pediatric Cardiology* 2013;**34**(8):1803–7. [DOI: 10.1007/s00246-013-0716-2; PUBMED: 23677390]

Heron 2007

Heron MP, Smith BL. Deaths: leading causes for 2003. *National Vital Statistics Reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System* 2007;**55**(10):1–92. [17408087]

Hultcrantz 2017

Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, et al. The GRADE Working Group clarifies the construct of certainty of evidence. *Journal of Clinical Epidemiology* 2017;**87**:4–13.

Iyengar 2014

Iyengar H, Kumar P, Kumar P. Pulse-oximetry screening to detect critical congenital heart disease in the neonatal intensive care unit. *Pediatric Cardiology* 2014;**35**(3): 406–10. [DOI: 10.1007/s00246-013-0793-2; PUBMED: 24026810]

Knapp 2010

Knapp AA, Metterville DR, Kemper AR, Prosser L, Perrin JM. Evidence review: critical congenital cyanotic heart disease. Prepared for: Maternal And Child Health Bureau. Massachusetts General Hospital, Center for Child and Adolescent Health Policy and Health Resources and Services Administration (HRSA), U.S. Department of Health and Human Services (DHHS) 2010. [<https://www.hrsa.gov/advisorycommittees/mchbadvisory/heritabledisorders/nominatecondition/reviews/cyanoticheart.pdf>]

Knowles 2005

Knowles R, Griebsch I, Dezateux C, Brown J, Bull C, Wren C. Newborn screening for congenital heart defects: a systematic review and cost-effectiveness analysis. *Health Technology Assessment* 2005;**9**(44):1-152, iii-iv. [PUBMED: 16297355]

Kuelling 2009

Kuelling B, Arlettaz Mieth R, Bauersfeld U, Balmer C. Pulse oximetry screening for congenital heart defects in Switzerland: most but not all maternity units screen their neonates. *Swiss Medical Weekly* 2009;**139**(47-48):699–704. [DOI: smw-12880; PUBMED: 20047132]

Leefflang 2008

Leefflang MM, Deeks JJ, Gatsonis C, Bossuyt PM, Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. *Annals of Internal Medicine* 2008;**149**(12):889–97. [PUBMED: 19075208]

Lloyd-Jones 2009

Lloyd-Jones D, Adams R, Carnethon M, De Simone G, Ferguson TB, Flegal K, et al. American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics - 2009 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. *Circulation* 2009;**119**(3):480–6. [DOI: 10.1161/CIRCULATIONAHA.108.191259; PUBMED: 19171871]

Macaskill 2010

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. Available from sdta.cochrane.org.

Mahle 2008

Mahle WT. Physical examination and pulse oximetry in newborn infants: out with the old, in with the new?. *Journal of Pediatrics* 2008; Vol. 152, issue 6:747–8. [DOI: 10.1016/j.jpeds.2008.01.043; PUBMED: 18492505]

Mahle 2009

Mahle WT, Newburger JW, Matherne GP, Smith FC, Hoke TR, Koppel R, et al. American Heart Association Congenital Heart Defects Committee of the Council on Cardiovascular Disease in the Young, Council on Cardiovascular Nursing, and Interdisciplinary Council on Quality of Care and Outcomes Research, American Academy of Pediatrics Section on Cardiology, Cardiac Surgery. Committee on Fetus and Newborn. Role of pulse oximetry in examining newborns for congenital heart disease: a scientific statement from the AHA and AAP. *Pediatrics* 2009;**124**(2):823–36. [DOI: 10.1542/peds.2009-1397; PUBMED: 19581259]

Mahle 2012

Mahle WT, Martin GR, Beekman RH 3rd, Morrow WR, Section on Cardiology and Cardiac Surgery Executive Committee. Endorsement of Health and Human Services recommendation for pulse oximetry screening for critical congenital heart disease. *Pediatrics* 2012;**129**(1):190–2. [DOI: 10.1542/peds.2011-3211; PUBMED: 22201143]

Manzoni 2017

Manzoni P, Martin GR, Sanchez Luna M, Mestrovic J, Simeoni U, Zimmermann LJI, et al. Pulse oximetry screening for critical congenital heart defects: a European consensus statement. *Lancet Child and Adolescent Health* 2017;**1**:88–90. [DOI: 10.1016/S2352-4642(17)30066-4]

Mellander 2006

Mellander M, Sunnegårdh J. Failure to diagnose critical heart malformations in newborns before discharge - an

- increasing problem?. *Acta Paediatrica* 2006;**95**(4):407–13. [DOI: 10.1080/08035250500541910; PUBMED: 16720486]
- Narayan 2016b**
Narayan IC, Blom NA, Ewer AK, Vento M, Manzoni P, te Pas AB. Aspects of pulse oximetry screening for critical congenital heart defects: when, how and why. *Archives of Disease in Childhood. Fetal and Neonatal Edition* 2016;**101**: F162–7.
- Narayan 2017**
Narayan IC, Kaptein AA, Hogewoning JA, Blom NA, Te Pas AB. Maternal acceptability of pulse oximetry screening at home after home birth or very early discharge. *European Journal of Pediatrics* 2017;**176**(5):669–72. [DOI: 10.1007/s00431-017-2883-2; PUBMED: 28281093]
- O'Donnell 2007**
O'Donnell CP, Kamlin CO, Davis PG, Carlin JB, Morley CJ. Clinical assessment of infant colour at delivery. *Archives of Disease in Childhood. Fetal and Neonatal Edition* 2007;**92**(6):F465–7. [DOI: 10.1136/adc.2007.120634; PUBMED: 17613535]
- Office of National Statistics, 2015**
Office of National Statistics (ONS). Statistical bulletin: deaths registered in England and Wales: 2014. Deaths, stillbirths and infant mortality including death rates, causes, age, and area of residence. <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsregistrationsummarytables/previousReleases> 2015.
- Peterson 2013**
Peterson C, Grosse SD, Oster ME, Olney RS, Cassell CH. Cost-effectiveness of routine screening for critical congenital heart disease in US newborns. *Pediatrics* 2013;**132**(3): e595–603. [DOI: 10.1542/peds.2013-0332; PUBMED: 23918890]
- Powell 2013**
Powell R, Pattison HM, Bhoyar A, Furmston AT, Middleton LJ, Daniels JP, et al. Pulse oximetry screening for congenital heart defects in newborn infants: an evaluation of acceptability to mothers. *Archives of Disease in Childhood. Fetal and Neonatal Edition* 2013;**98**(1):F59–63. [DOI: 10.1136/fetalneonatal-2011-301225; PUBMED: 22611113]
- 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. [DOI: 10.1016/j.jclinepi.2005.02.022; PUBMED: 16168343]
- RevMan 2014 [Computer program]**
Nordic Cochrane Centre, The Cochrane Collaboration. Review Manager 5 (RevMan 5). Version 5.3. Copenhagen: Nordic Cochrane Centre, The Cochrane Collaboration, 2014.
- Roberts 2012**
Roberts TE, Barton PM, Auguste PE, Middleton LJ, Furmston AT, Ewer AK. Pulse oximetry as a screening test for congenital heart defects in newborn infants: a cost-effectiveness analysis. *Archives of Disease in Childhood* 2012;**97**(3):221–6. [DOI: 10.1136/archdischild-2011-300564; PUBMED: 22247242]
- SAS Institute Inc. 2004**
SAS Institute Inc. What's New in SAS® 9.0, 9.1, 9.1.2, and 9.1.3. SAS OnlineDoc® 9.1.3. Cary, NC 2004. [<https://support.sas.com/documentation/onlinedoc/91pdf/sasdoc/913/whatsnew/10878.pdf>]
- Schunemann 2008**
Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. GRADE Working Group. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ (Clinical Research Ed.)* 2008;**336**(7653):1106–10. [DOI: 10.1136/bmj.39500.677199.AE; PUBMED: 18483053]
- Suresh 2013**
Suresh GK. Pulse oximetry screening for critical congenital heart disease in neonatal intensive care units. *Journal of Perinatology* 2013;**33**(8):586–8. [DOI: 10.1038/jp.2012.161; PUBMED: 23897309]
- Takwoingi 2010**
Takwoingi Y, Deeks JJ. MetaDAS: A SAS macro for meta-analysis of diagnostic accuracy studies. Quick reference and worked example. Available from <http://srdta.cochrane.org/> Version 1.3. Accessed July 2010.
- Tegnander 2006**
Tegnander E, Williams W, Johansen OJ, Blaas HG, Eik-Nes SH. Prenatal detection of heart defects in a non-selected population of 30,149 fetuses - detection rates and outcome. *Ultrasound in Obstetrics & Gynecology* 2006;**27**(3):252–65. [DOI: 10.1002/uog.2710; PUBMED: 16456842]
- Thangaratinam 2007**
Thangaratinam S, Daniels J, Ewer AK, Zamora J, Khan KS. Accuracy of pulse oximetry in screening for congenital heart disease in asymptomatic newborns: a systematic review. *Archives of Disease in Childhood. Fetal and Neonatal Edition* 2007;**92**(3):F176–80. [DOI: 10.1136/adc.2006.107656; PUBMED: 17344253]
- Thangaratinam 2012**
Thangaratinam S, Brown K, Zamora J, Khan KS, Ewer AK. Pulse oximetry screening for critical congenital heart defects in asymptomatic newborn babies: a systematic review and meta-analysis. *Lancet* 2012;**379**(9835):2459–64. [DOI: 10.1016/S0140-6736(12)60107-X; PUBMED: 22554860]
- Tworetzky 2001**
Tworetzky W, McElhinney DB, Reddy VM, Brook MM, Hanley FL, Silverman NH. Improved surgical outcome after fetal diagnosis of hypoplastic left heart syndrome. *Circulation* 2001;**103**(9):1269–73. [PUBMED: 11238272]

Westin 2006

Westin M, Saltvedt S, Bergman G, Kublickas M, Almstrom H, Grunewald C, et al. Routine ultrasound examination at 12 or 18 gestational weeks for prenatal detection of major congenital heart malformations? A randomised controlled trial comprising 36,299 fetuses. *BJOG* 2006;**113**(6):675–82. [DOI: 10.1111/j.1471-0528.2006.00951.x; PUBMED: 16709210]

Whiting 2011

Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. 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. [DOI: 10.7326/0003-4819-155-8-201110180-00009; PUBMED: 22007046]

Wren 1999

Wren C, Richmond S, Donaldson L. Presentation of

congenital heart disease in infancy: implications for routine examination. *Archives of Disease in Childhood. Fetal and Neonatal Edition* 1999;**80**(1):F49–53. [PUBMED: 10325813]

Wren 2008

Wren C, Reinhardt Z, Khawaja K. Twenty-year trends in diagnosis of life-threatening neonatal cardiovascular malformations. *Archives of Disease in Childhood. Fetal and Neonatal Edition* 2008;**93**(1):F33–5. [DOI: 10.1136/adc.2007.119032; PUBMED: 17556383]

Wright 2014

Wright J, Kohn M, Niermeyer S, Rausch CM. Feasibility of critical congenital heart disease newborn screening at moderate altitude. *Pediatrics* 2014;**133**(3):e561–9. [DOI: 10.1542/peds.2013-3284; PUBMED: 24567022]

* Indicates the major publication for the study

CHARACTERISTICS OF STUDIES

Characteristics of included studies [ordered by study ID]

Arlettaz 2006

Study characteristics	
Patient sampling	Prospective multicenter study with consecutive enrollment of participants
Patient characteristics and setting	Country: Switzerland Setting: 4 hospitals in Zurich: 3 maternity hospitals and the Division of Cardiology of the University Children's Hospital Study period: 1-year period (from May 13, 2003, to May 12, 2004) Inclusion criteria: all newborn infants from 35 weeks' gestation Exclusion criteria: Premature infants below 35 weeks' gestation Infants with a respiratory disorder Live birth cohort, n = 3663 (401 infants excluded according to exclusion criteria) N screened: 3262 (89%) (1764 at the University Hospital, 1011 at the Zollikerberg Hospital, 487 at the Triemli Hospital) Gestational age: median: 39 weeks (range 35 to 42) Prevalence of CCHD: 3.7 per 1000 live births
Index tests	Pulse oximetry was performed with the Nellcor NPB-40 handheld pulse oximeter and the Nellcor Max-N Oximax adhesive sensors Screening protocol: Site of testing: right or left foot Test timing: within 24 hours (<i>in 48 cases [1%], pulse oximetry was performed too early in part because of immediate postnatal transfer to the cardiology unit or because patients were discharged before 6 hours of age; in 255 cases (8%), pulse oximetry was performed after 12 hours; 2959 measurements [91%] were performed at between 6 and 12 hours</i>) Oxygen saturation: functional Threshold: < 95% Measurement did not exceed 2 minutes. If saturation was below 95%, a senior house officer performed a full clinical examination of the newborn. If the infant had saturation below 90% or any signs suggestive of a CHD, echocardiography was performed immediately. In the case of an asymptomatic newborn with borderline values (90% to 94%), a second measurement was performed 4 to 6 hours later
Target condition and reference standard(s)	Target condition: Congenital heart disease was defined as the presence of a gross structural abnormality of the heart or intrathoracic great vessels that is actually or potentially of functional significance Reference standard(s): Reference standard used for positive pulse oximetry results: echocardiography (<i>complete M-mode, 2-dimensional, and Doppler echocardiograms were performed either at the University Hospital with an Acuson 128XP/10 [Siemens, Erlangen, Germany] with a 7.5-mHz transducer, or at the University Children's Hospital with a Sonos 5500 [Philips, Amsterdam, Netherlands], both equipped with all Doppler modalities</i>) Reference standard used for negative pulse oximetry results: not stated

Flow and timing	Duration of follow-up: not stated Loss to follow-up: none (n analysed: 3262)		
Comparative			
Notes			
Methodological quality			
Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
		Low	Low
DOMAIN 2: Index Test Pulse oximetry			
If a threshold was used, was it pre-specified?	Yes		
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Unclear		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
		High	Low
DOMAIN 4: Flow and Timing			

Were all patients included in the analysis?	Yes		
Was there at least 28 days of appropriate follow up?	Unclear		
Did all patients receive a reference standard?	Unclear		
		Unclear	

Bakr 2005

Study characteristics	
Patient sampling	Prospective study with consecutive enrollment of participants
Patient characteristics and setting	Country: Saudi Arabia Setting: neonatology department of King Abdel-Aziz Specialist Hospital Study period: 6-month period (January 2004 to July 2004) Inclusion criteria: asymptomatic newborns Exclusion criteria: Those admitted to the neonatal intensive care unit at birth N screened: 5211 Prevalence of CCHD: 3.7 per 1000 live births
Index tests	Pulse oximetry was performed with a Digioxi PO 920 pulse oximeter (Digicare Biomedical Technology, West Palm Beach, FL, USA) Screening protocol: Site of testing: right upper and lower limbs Test timing: longer than 24 hours. Average age at screening was 31.7 hours Oxygen saturation: fractional Threshold: $\leq 94\%$ <i>"Any infant who had an oxygen saturation < 90% from either limb was examined by echocardiography. Saturations between 90% and 94% were verified by three readings; if they persisted in this range, echocardiography was also done."</i>
Target condition and reference standard(s)	Target condition: congenital heart disease Reference standard(s): Reference standard used for positive pulse oximetry results: echocardiography Reference standard used for negative pulse oximetry results: cardiology service of the only pediatric hospital in the region to identify patients who had received a diagnosis of CHD after discharge from the well-baby nursery
Flow and timing	Duration of follow-up: not stated Loss to follow-up: none (n = 5211)
Comparative	

Notes			
Methodological quality			
Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
		Low	Low
DOMAIN 2: Index Test Pulse oximetry			
If a threshold was used, was it pre-specified?	Yes		
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	No		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
		Unclear	Low
DOMAIN 4: Flow and Timing			
Were all patients included in the analysis?	Yes		
Was there at least 28 days of appropriate follow up?	Unclear		

Did all patients receive a reference standard?	Unclear		
		Unclear	

Bhola 2014**Study characteristics**

Patient sampling	Retrospective observational study
Patient characteristics and setting	Country: Australia Setting: Royal Prince Alfred Hospital (tertiary maternity hospital delivering over 5000 newborns a year) Study period: 42-month period (from April 2008 to December 2011) Inclusion criteria: all newborns (routine neonatal examination) Exclusion criteria: not stated Live birth cohort, n = 19,765 N screened: 18,801 (95.1%) (648 had been admitted to the nursery and did not qualify for screening, 316 missed) Prevalence of CCHD: 0.2 per 1000 live births
Index tests	Pulse oximetry was performed with a Masimo Radical 5 portable oximeter (Masimo Corporation, Irvine, CA, USA) with a reusable probe with disposable Coban tape (1-inch self-adherent wrap, manufactured by 3M, Australia) Screening protocol: Site of testing: post-ductal (foot) Test timing: longer than 24 hours (between 24 and 72 hours of life) Oxygen saturation: functional Threshold: < 95% <i>“If the post-ductal saturation was 95% or more, the result was assigned as a pass. Readings between 90% and 95% led to a repeat saturation measurement in the next 1-2 hours. If the post-ductal saturation remained below 95% on repeat testing, the newborn was referred for review and examination by a senior neonatal paediatrician. If the saturation was less than 90%, at any time, the newborn was referred for review by a senior neonatal paediatrician without waiting for a repeat test.”</i>
Target condition and reference standard(s)	Target condition: congenital heart disease Reference standard(s): Reference standard used for positive pulse oximetry results: echocardiography Reference standard used for negative pulse oximetry results: database of the Heart Centre for Children at Children's Hospital at Westmead for any newborns undergoing cardiac surgery or catheter intervention in the first year of life
Flow and timing	Duration of follow-up: not stated Loss to follow-up: 316 missed (not performed by the resident, performed but not recorded, screening not completed before early discharge or owing to compliance issues when the new protocol was originally introduced)

Comparative			
Notes			
Methodological quality			
Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Unclear		
Did the study avoid inappropriate exclusions?	Unclear		
		Unclear	Low
DOMAIN 2: Index Test Pulse oximetry			
If a threshold was used, was it pre-specified?	Yes		
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	No		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
		Unclear	Low
DOMAIN 4: Flow and Timing			
Were all patients included in the analysis?	Unclear		

Was there at least 28 days of appropriate follow up?	Unclear		
Did all patients receive a reference standard?	Yes		
		Unclear	

de-Wahl Granelli 2009

Study characteristics	
Patient sampling	Prospective study with consecutive enrollment of participants
Patient characteristics and setting	Country: Sweden Setting: 5 maternity units in West Götaland Study period: July 2004 to March 31, 2007 Inclusion criteria: all newborn infants Exclusion criteria: Admitted to neonatal special care units Live birth cohort, n = 46,963 (7064 excluded owing to rolling start of study or admission to neonatal intensive care) Eligible, n = 39,899 (Östran n = 13,455, Mölndal n = 8953, Trollhättan n = 7019, Borås n = 5382, Skövde n = 5090) Excluded, n = 1470 - refusal (19), oximeter failure (18), staff shortage (2), incomplete record of screening results (39) or of physical examination (1392) N screened: 38,429 (<i>flowchart page 4</i>) Prevalence of CCHD: 0.7 per 1000 live births
Index tests	Pulse oximetry was performed with a pulse oximeter Radical SET, version 4 (average time set on 8 seconds) with multisite LNOP YI sensors, Masimo, Irvine, CA, USA Screening protocol: Site of testing: pre-ductal (palm of right hand) and post-ductal (either foot) Test timing: longer than 24 hours Oxygen saturation: functional Threshold: < 95% <i>“When both pre-ductal and post-ductal oxygen saturation was < 95% or the difference between the two measurements was > 3% (≥ 2 standard deviations of interobserver measurement variability) the baby was provisionally considered to be screening positive, but a repeat measurement was performed. Babies with three repeated positive measurements were supposed to have an echocardiogram performed the same day according to the study protocol, but with some babies scheduled for early discharge only two pulse oximetry screenings were managed before the discharge examination was performed. Babies were considered screening positive until a measurement not fulfilling screening positive criteria was obtained. If saturation $\leq 90\%$ the newborn was referred for an echocardiogram the same day.”</i>
Target condition and reference standard(s)	Target condition: congenital heart disease Reference standard(s): Reference standard used for positive pulse oximetry results: echocardiography

	Reference standard used for negative pulse oximetry results: mortality data of the National Board of Forensic Medicine (information on all deaths due to undiagnosed cardiovascular malformations in children younger than 1 year in Sweden born during the study)		
Flow and timing	Duration of follow-up: not stated Inconclusive results: 73 (results for only 1 site [34], oxygen saturation < 90% but not optimal [39]) N analyzed: 39,821		
Comparative			
Notes			
Methodological quality			
Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
		Low	Low
DOMAIN 2: Index Test Pulse oximetry			
If a threshold was used, was it pre-specified?	Yes		
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	No		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		

		Unclear	Low
DOMAIN 4: Flow and Timing			
Were all patients included in the analysis?	No		
Was there at least 28 days of appropriate follow up?	Yes		
Did all patients receive a reference standard?	Yes		
		Low	

Ewer 2011

Study characteristics	
Patient sampling	Prospective multicenter study with consecutive enrollment of participants
Patient characteristics and setting	Country: United Kingdom Setting: 6 obstetrical units in the West Midlands Study period: February 2008 to January 2009 Inclusion criteria: asymptomatic newborns (gestation > 34 weeks) Study includes newborns in whom congenital heart defects were suspected antenatally after midtrimester ultrasonography Exclusion criteria: Newborns with symptoms suggestive of cardiac disease that were detected before screening Livebirth cohort, n = 26,513 (3768 missed, 2005 declined, 685 ineligible) N screened: 20,055 (75.6%) Prevalence of CCHD: 1.2 per 1000 live births
Index tests	Pulse oximetry was performed with the Radical-7 pulse oximeter with reusable probe LNOP Y1 (Masimo, Irvine, CA, USA) Screening protocol: Site of testing: pre-ductal (right hand) and post-ductal (either foot in non-specified order) Test timing: within 24 hours (median age at testing of 12.4 hours for the full cohort) Oxygen saturation: functional Threshold: < 95% <i>“A saturation of less than 95% in either limb or a difference of more than 2% between the limb saturation readings (if both were ≥ 95%) was judged to be abnormal. Clinical examination was expedited if an abnormal test result was obtained. If this examination was unremarkable, oximetry was repeated 1 to 2 hours later. If abnormalities of the cardiovascular system were detected with expedited examination, or saturations remained abnormal during a second test, the newborn were classified as test positive.”</i>

Target condition and reference standard(s)	Target condition: critical congenital heart defects (ie, death or requiring invasive intervention before 28 days) All infants with hypoplastic left heart, pulmonary atresia with intact ventricular septum, simple transposition of the great arteries, or interruption of the aortic arch All infants dying or requiring surgery within the first 28 days of life with coarctation of the aorta, aortic valve stenosis, pulmonary valve stenosis, tetralogy of Fallot, pulmonary atresia with ventricular septal defect, or total anomalous pulmonary venous connection Reference standard(s): Reference standard used for positive pulse oximetry results: echocardiography Reference standard used for negative pulse oximetry results: clinical follow-up, use of cardiology databases and congenital anomaly registries
Flow and timing	Duration of follow-up: up to 12 months Loss to follow-up: none (n analyzed: 20,055)
Comparative	
Notes	Funding: National Institute for Health Research Health Technology Assessment (NIHR HTA) program (project number 06/06/03)

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
		Low	Low
DOMAIN 2: Index Test Pulse oximetry			
If a threshold was used, was it pre-specified?	Yes		
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			

Ewer 2011 (Continued)

Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
		Low	Low
DOMAIN 4: Flow and Timing			
Were all patients included in the analysis?	Yes		
Was there at least 28 days of appropriate follow up?	Yes		
Did all patients receive a reference standard?	Yes		
		Low	

Gomez-Rodriguez 2015

Study characteristics	
Patient sampling	Cross-sectional prospective study
Patient characteristics and setting	Country: Mexico Setting: Department of Neonatology, UMAE 48-Instituto Mexicano del Seguro Social (IMSS), León, Gto Study period: July 2010 to April 2011 Inclusion criteria: newborns > 6 hours of age in whom no CHD was suspected; only tested consecutive newborns who were available during the working hours of investigators Exclusion criteria: - newborns with lung disease - no informed consent N screened: 1037 Gestational age: mean (SD): 38.9 (1.1) weeks Prevalence of CCHD: 1.9 per 1000 live births
Index tests	Pulse oximetry was performed with a Rad-5 handheld pulse oximeter with multisite sensor Screening protocol: - Site of testing: left lower extremity (post-ductal) - Test timing: within 24 hours (mean age at pulse oximetry screening 12 hours - range 6 to 48 hours) - Oxygen saturation: functional

	Threshold: < 95% Measurement was taking during 2 minutes until the reading remained the same in 2 determinations		
Target condition and reference standard(s)	Target condition: critical congenital heart disease (no definition included) Reference standard(s): Reference standard used for positive pulse oximetry results: echocardiography Reference standard used for negative pulse oximetry results: clinical records of all follow-up at 6 months		
Flow and timing	Duration of follow-up: 6 months Loss to follow-up: none (n analyzed: 1037)		
Comparative			
Notes			
Methodological quality			
Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	No		
Did the study avoid inappropriate exclusions?	Yes		
		Unclear	Low
DOMAIN 2: Index Test Pulse oximetry			
If a threshold was used, was it pre-specified?	Yes		
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	No		

Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
		Unclear	Low
DOMAIN 4: Flow and Timing			
Were all patients included in the analysis?	Yes		
Was there at least 28 days of appropriate follow up?	Yes		
Did all patients receive a reference standard?	Yes		
		Low	

Jones 2016

Study characteristics	
Patient sampling	Retrospective observational study
Patient characteristics and setting	Country: United Kingdom Setting: neonatal intensive care unit, Northwick Park Hospital, Harrow, Middlesex (level-2 neonatal unit without on-site access to pediatric echocardiography) Study period: September 1, 2011, to August 31, 2013 Inclusion criteria: all newborns admitted to the neonatal unit during the study period Exclusion criteria: Antenatal diagnosis of CCHD Admitted to neonatal intensive care unit after birth Live birth cohort, n = 11,233 (973 neonatal unit admissions) N screened: 10,260 Gestational age: not stated Prevalence of CCHD: 0.2 per 1000 live births
Index tests	Type of pulse oximeter not stated Screening protocol: Site of testing: both pre-ductal and post-ductal Test timing: within 24 hours Oxygen saturation: not stated Threshold: $\leq 95\%$ (or pre-ductal and post-ductal difference $> 3\%$)

Target condition and reference standard(s)	Target condition: critical congenital heart disease defined as CHD resulting in death or requiring surgical intervention or therapeutic catheterization within the first 28 days of life Reference standard(s): Reference standard used for positive pulse oximetry results: echocardiography Reference standard used for negative pulse oximetry results: National Congenital Heart Disease Audit		
Flow and timing	Duration of follow-up: not stated Loss to follow-up: none (n analyzed: 10,260)		
Comparative			
Notes			
Methodological quality			
Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
		Low	Low
DOMAIN 2: Index Test Pulse oximetry			
If a threshold was used, was it pre-specified?	Yes		
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	No		
Were the reference standard results interpreted without knowledge	Unclear		

of the results of the index tests?			
		Unclear	Low
DOMAIN 4: Flow and Timing			
Were all patients included in the analysis?	Yes		
Was there at least 28 days of appropriate follow up?	Unclear		
Did all patients receive a reference standard?	Yes		
		Unclear	

Klausner 2017

Study characteristics	
Patient sampling	Retrospective observational study
Patient characteristics and setting	Country: USA Setting: 4 Yale-New Haven Health System hospitals in Connecticut Study period: January 1 and December 31, 2014 Inclusion criteria: all newborns delivered during the study period Exclusion criteria: Live-born infants who died before CCHD screening Antenatal screening Live birth cohort, n = 10,589 (171 [1.6%] underwent an echocardiogram before screening, and 98 [0.9%] were not screened; 96 were missed in error and parents refused in 2 instances) N screened: 10,320 Gestational age: 9584 (90.5%) were term (> 37 weeks) Prevalence of CCHD: 0 per 1000 live births
Index tests	Type of pulse oximeter not stated Screening protocol: Site of testing: both pre-ductal and post-ductal Test timing: longer than 24 hours Oxygen saturation: not stated Threshold: < 95%
Target condition and reference standard(s)	Target condition: critical congenital heart disease defined as structural defect associated with hypoxemia in the newborn period that requires surgical intervention before 1 year and, without intervention, can lead to significant morbidity and mortality Reference standard(s): Reference standard used for positive pulse oximetry results: echocardiogram Reference standard used for negative pulse oximetry results: follow-up

Flow and timing	Of 10,316 infants with negative pulse oximetry at the time of birth, possible to review postdischarge records of only 52.1% (n = 5367)		
Comparative			
Notes	Study was supported in part by a National Heart, Lung, and Blood Institute Medical Student Research Fellowship, National Institutes of Health (award T35HL007649; to Ms Klausner), and by grant UL1 TR001863 from the National Center for Advancing Translational Science at the National Institutes of Health and the NIH Roadmap for Medical Research. Funded by the National Institutes of Health (NIH)		
Methodological quality			
Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
		Low	Low
DOMAIN 2: Index Test Pulse oximetry			
If a threshold was used, was it pre-specified?	Yes		
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	No		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		

		Unclear	Low
DOMAIN 4: Flow and Timing			
Were all patients included in the analysis?	Yes		
Was there at least 28 days of appropriate follow up?	Unclear		
Did all patients receive a reference standard?	No		
		Unclear	

Koppel 2003

Study characteristics	
Patient sampling	Prospective study with consecutive enrollment of participants
Patient characteristics and setting	Country: USA Setting: well infant nurseries at 2 hospitals (New York) Study period: from May 1998 to November 1999 Inclusion criteria: all asymptomatic newborns Exclusion criteria: Admitted to neonatal special care units (infants who did manifest any of these clinical findings: cyanosis, tachypnea [respiratory rate: 60/min], grunting, flaring, retraction, murmur, active precordium, or diminished pulses) N screened: 11,281 (8642 at hospital A, 2639 at hospital B) Prevalence of CCHD: 0.4 per 1000 live births
Index tests	Pulse oximetry was performed with an Ohmeda Medical pulse oximeter Screening protocol: Site of testing: post-ductal Test timing: longer than 24 hours. Timing of oximetry determination was linked to state-mandated metabolic screening (24 hours of age) at hospital A. At hospital B, screening was performed immediately before discharge as part of a series of discharge procedures (average length of stay for vaginal delivery: 56.9 hours; for cesarean section: 103.2 hours) Oxygen saturation: functional Threshold: $\leq 95\%$ <i>"single determination of post-ductal saturation"</i>
Target condition and reference standard(s)	Target condition: critical congenital cardiovascular malformation (CCVM) defined as a lesion that would likely require surgical correction during the first month of life Reference standard(s): Reference standard used for positive pulse oximetry results: echocardiography Reference standard used for negative pulse oximetry results: clinical follow-up and New York State

Koppel 2003 (Continued)

	Congenital Malformations Registry (CMR)		
Flow and timing	Duration of follow-up: not stated Loss to follow-up: none (n analyzed = 11,281)		
Comparative			
Notes	Study was supported by a cooperative agreement from the Centers for Disease Control and Prevention Oximeters were provided by Ohmeda Medical.		
Methodological quality			
Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
		Low	Low
DOMAIN 2: Index Test Pulse oximetry			
If a threshold was used, was it pre-specified?	Yes		
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
		Low	Low

DOMAIN 4: Flow and Timing			
Were all patients included in the analysis?	Yes		
Was there at least 28 days of appropriate follow up?	Unclear		
Did all patients receive a reference standard?	Yes		
		Low	

Meberg 2008

Study characteristics	
Patient sampling	Prospective multicenter study with consecutive enrollment of participants
Patient characteristics and setting	Country: Norway Setting: 14 hospitals with obstetrical departments and pediatric services and neonatal special or intensive care units (50% of all deliveries in Norway) Study period: 1-year period (2005 to 2006) Inclusion criteria: healthy newborns Exclusion criteria: Prenatal diagnosis Live birth cohort, n = 57,959 (not screened: 7951 [14%]; 224 of 7951 newborns [3%] had CHD) N screened: 50,008 (86%) Prevalence of CCHD: 0.7 per 1000 live births
Index tests	Pulse oximetry was performed with a pulse oximeter type RAD-5v (Masimo Corporation, Irvine, CA) with a multisite reusable sensor (LNOP YI) Screening protocol: Site of testing: post-ductal (foot) Test timing: within 24 hours (first day) Oxygen saturation: functional Threshold: < 95% <i>"The probe was attached for at least 2 minutes, until a stable value was obtained. Retest if the result of pulse-ox was < 95%."</i>
Target condition and reference standard(s)	Target condition: congenital heart defects Reference standard(s): Reference standard used for positive pulse oximetry results: echocardiography Reference standard used for negative pulse oximetry results: clinical follow-up
Flow and timing	Duration of follow-up: 6 months after the last infants were born Loss to follow-up: none (n analyzed: 50,008)

Comparative			
Notes			
Methodological quality			
Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	No		
		Unclear	Low
DOMAIN 2: Index Test Pulse oximetry			
If a threshold was used, was it pre-specified?	Yes		
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	No		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
		Unclear	Low
DOMAIN 4: Flow and Timing			
Were all patients included in the analysis?	Yes		

Was there at least 28 days of appropriate follow up?	Yes		
Did all patients receive a reference standard?	Yes		
		Low	

Oakley 2015

Study characteristics	
Patient sampling	Prospective observational study
Patient characteristics and setting	Country: United Kingdom Setting: Tertiary Neonatal Unit from the Royal Gwent Hospital, Newport (3700 deliveries a year) Study period: 2 years (from January 2007 to December 2009) Inclusion criteria: all newborns at 35 weeks' gestation and above who were admitted to the postnatal ward Exclusion criteria: - Admitted to the neonatal intensive care unit - Antenatal diagnosis of CHD - Births on weekends and holidays - Live birth cohort, n = 9613 N screened: 6329 (65.8%) Gestational age of newborn infants included, range 35 to 42 weeks Prevalence of CCHD: 1.3 per 1000 live births
Index tests	Pulse oximetry was performed with a Nellcor NPB 40 pulse-oximeter (Pleasanton, CA) and a reusable OXI-A/N saturation probe Screening protocol: Site of testing: post-ductal (foot) Test timing: longer than 24 hours (all newborns were greater than 6 hours of age at the time of examination) Oxygen saturation: functional Threshold: < 95% <i>"Newborns with saturation readings < 95% had a repeat reading taken on the other leg after thirty minutes and if still < 95%, a further repeat reading after one hour. If the reading remained < 95%, it was considered abnormal."</i>
Target condition and reference standard(s)	Target condition: CCHD Reference standard(s): Reference standard used for positive pulse oximetry results: echocardiography Reference standard used for negative pulse oximetry results: regional pediatric cardiology database and local death records
Flow and timing	Duration of follow-up: not stated Loss to follow-up: none (n analyzed: 6369)

Comparative			
Notes			
Methodological quality			
Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	No		
Did the study avoid inappropriate exclusions?	No		
		High	Low
DOMAIN 2: Index Test Pulse oximetry			
If a threshold was used, was it pre-specified?	Yes		
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	No		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
		Unclear	Low
DOMAIN 4: Flow and Timing			
Were all patients included in the analysis?	Yes		

Oakley 2015 (Continued)

Was there at least 28 days of appropriate follow up?	Unclear		
Did all patients receive a reference standard?	Yes		
		Low	

Ozalkaya 2016

Study characteristics	
Patient sampling	Retrospective observational study
Patient characteristics and setting	Country: Turkey Setting: Bursa Sevkett Yilmaz Training and Research Hospital Study period: between January 2014 and December 2014 Inclusion criteria: asymptomatic newborns Exclusion criteria: Referred within first 24 hours of life or admitted to neonatal intensive care unit Perinatal CCHD Live birth cohort, n = 10,200 (excluded: hospitalized = 1100, referred = 890, perinatal CCHD = 2) N screened: 8208 Gestational age: not stated Prevalence of CCHD: 1 per 1000 live births
Index tests	Pulse oximetry was performed with a Nellcor pulse oximeter. Screening protocol: Site of testing: both pre-ductal and post-ductal Test timing: longer than 24 hours Oxygen saturation: functional Threshold: Screening test was considered positive in newborns whose saturation with pulse oximetry was less than or equal to 95% and/or who had a difference < 3% between right lower and right extremities
Target condition and reference standard(s)	Target condition: CCHD defined as congenital heart disease requiring catheter-based or surgical intervention within the first month of life, or causing high mortality and morbidity in the first weeks of life Reference standard(s): Reference standard used for positive pulse oximetry results: echocardiography Reference standard used for negative pulse oximetry results: echocardiography
Flow and timing	Duration of follow-up: not stated Loss to follow-up: none (n analyzed: 8208)
Comparative	
Notes	

Methodological quality			
Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
		Low	Low
DOMAIN 2: Index Test Pulse oximetry			
If a threshold was used, was it pre-specified?	Yes		
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
		Low	Low
DOMAIN 4: Flow and Timing			
Were all patients included in the analysis?	Yes		
Was there at least 28 days of appropriate follow up?	Unclear		
Did all patients receive a reference standard?	Yes		

		Low	
--	--	-----	--

Richmond 2002

Study characteristics			
Patient sampling	Prospective study with consecutive enrollment of participants		
Patient characteristics and setting	Country: United Kingdom Setting: Sunderland Royal Hospital Study period: from April 1, 1999, to March 31, 2001 Inclusion criteria: asymptomatic newborn without signs of respiratory or cardiac illness Exclusion criteria: Admitted to neonatal care units Live birth cohort, n = 6166 (540 excluded: 447 neonatal unit, 5 no consent, 88 newborns missed) N screened: 5626 (91%) Prevalence of CCHD: 1.6 per 1000 live births		
Index tests	Pulse oximetry was performed with a radiometer Oxi machine. Screening protocol: Site of testing: post-ductal (foot) Test timing: within 24 hours (after the age of 2 hours and before discharge) Oxygen saturation: fractional Threshold: < 95% “Any baby who did not achieve a post-ductal fractional saturation of at least 95% was clinically examined by the midwife. If no suspicions were raised by the examination, a second saturation measurement was performed an hour or two later. If either the examination or the repeat saturation measurement were not satisfactory, an echocardiogram was performed.”		
Target condition and reference standard(s)	Target condition: congenital cardiac malformation Reference standard(s): Reference standard used for positive pulse oximetry results: echocardiography Reference standard used for negative pulse oximetry results: Regional Perinatal Mortality Survey and Northern Congenital Abnormality Survey & Diagnostic Database at Freeman Hospital (referral hospital)		
Flow and timing	Duration of follow-up: not stated Loss to follow-up: none (n analyzed: 5626)		
Comparative			
Notes			
Methodological quality			
Item	Authors’ judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			

Richmond 2002 (Continued)

Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
		Low	Low
DOMAIN 2: Index Test Pulse oximetry			
If a threshold was used, was it pre-specified?	Yes		
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	No		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
		Unclear	Low
DOMAIN 4: Flow and Timing			
Were all patients included in the analysis?	Yes		
Was there at least 28 days of appropriate follow up?	Unclear		
Did all patients receive a reference standard?	Yes		
		Low	

Study characteristics			
Patient sampling	Prospective multicenter study with consecutive enrollment of participants		
Patient characteristics and setting	Country: Germany Setting: primary, secondary, and tertiary care (34 neonatal/obstetrical departments in Saxony) Study period: 2-year period (from July 2006 to June 2008) Inclusion criteria: full-term and post-term neonates (gestational age ≥ 37 weeks) Normal routine clinical examination Informed parental consent Exclusion criteria: Antenatal diagnosis/suspicion of congenital heart disease Livebirth cohort, n = 48,348 (excluded: 6108 newborns [72 clinical or prenatal diagnosis of CCHD; 6036 other]) N eligible for pulse oximetry screening: 42,240 (n = 727 [91%] did not receive pulse oximetry screening, mainly because of early discharge after birth) N screened: 41,445 (85.7%) Prevalence of CCHD: 0.4 per 1000 live births		
Index tests	Pulse oximetry was performed with a great variety of devices (no further information) Screening protocol: Site of testing: post-ductal (foot) Test timing: longer than 24 hours Oxygen saturation: functional Threshold: $\leq 95\%$ <i>“The study protocol included repeated SpO₂ measurements after 1 hour if the initial value was $< 96\%$.”</i>		
Target condition and reference standard(s)	Target condition: CCHD Reference standard(s): Reference standard used for positive pulse oximetry results: echocardiography Reference standard used for negative pulse oximetry results: Saxonian perinatal and neonatal registries		
Flow and timing	Duration of follow-up: not stated Loss to follow-up: 3 for violation of study protocol (n analyzed: 41,442)		
Comparative			
Notes			
Methodological quality			
Item	Authors’ judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		

Did the study avoid inappropriate exclusions?	No		
		Unclear	Low
DOMAIN 2: Index Test Pulse oximetry			
If a threshold was used, was it pre-specified?	Yes		
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	No		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
		Unclear	Low
DOMAIN 4: Flow and Timing			
Were all patients included in the analysis?	No		
Was there at least 28 days of appropriate follow up?	Unclear		
Did all patients receive a reference standard?	Yes		
		Low	

Rosati 2005

Study characteristics			
Patient sampling	Prospective study with consecutive enrollment of participants		
Patient characteristics and setting	Country: Italy Setting: Perrino Hospital (referral center of the area) Study period: from May 1, 2000, to November 30, 2004 Inclusion criteria: term newborns with uncomplicated neonatal courses Exclusion criteria: Infants who were symptomatic (ie, heart murmur, severe cyanosis) Prenatal diagnosis of critical congenital cardiovascular malformation N screened: 5292 Prevalence of CCHD: 0.6 per 1000 live births		
Index tests	Type of pulse oximeter not stated Screening protocol: Site of testing: post-ductal (foot) Test timing: longer than 24 hours Oxygen saturation: functional Threshold: $\leq 95\%$ <i>“Post-ductal saturation (SpO_2) and the monitoring of oxymetry values were evaluated for two minutes in each newborn.”</i>		
Target condition and reference standard(s)	Target condition: CCHD defined as lesions requiring surgical correction or interventional procedures during the first month of life Reference standard(s): Reference standard used for positive pulse oximetry results: echocardiography Reference standard used for negative pulse oximetry results: clinical follow-up		
Flow and timing	Duration of follow-up: not stated Loss to follow-up: none (n analyzed: 5292)		
Comparative			
Notes			
Methodological quality			
Item	Authors’ judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	No		
		Unclear	Low

DOMAIN 2: Index Test Pulse oximetry			
If a threshold was used, was it pre-specified?	Yes		
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	No		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
		Unclear	Low
DOMAIN 4: Flow and Timing			
Were all patients included in the analysis?	Yes		
Was there at least 28 days of appropriate follow up?	Unclear		
Did all patients receive a reference standard?	Unclear		
		Unclear	

Sendelbach 2008

Study characteristics	
Patient sampling	Prospective study with consecutive enrollment of participants
Patient characteristics and setting	Country: USA Setting: large public hospital (Parkland Health and Hospital System [PHHS]) in Dallas, TX, which serves a primarily indigent Hispanic population Study period: from March 1, 2006, to February 28, 2007

	Inclusion criteria: Term and late preterm neonates who did not have major malformations Gestational age criteria: ≥ 35 weeks Birth weight: ≥ 2100 grams Exclusion criteria: Admitted to neonatal intensive care unit (NICU) Respiratory distress and/or cyanosis before 4 hours of age Live birth cohort, n = 16,432 (excluded: 66 [0.4%]; 11 had CHD) N screened: 15,233 (99.6%) Prevalence of CCHD: 0.1 per 1000 live births		
Index tests	Pulse oximetry was performed with a Nellcor N-395 (Boulder, CO) pulse oximeter Screening protocol: Site of testing: post-ductal (foot) Test timing: within 24 hours (4 hours after delivery) Oxygen saturation: functional Threshold: < 96% “On the day of discharge, the 4-hour pulse oximetry result was made available to the provider. A pulse oximetry result of 96% was considered normal and was not repeated. For neonates who failed to achieve 96% on the 4-hour screen, a follow-up pulse oximetry reading was performed by either the nursing staff or the medical provider by using the procedure described above. When the discharge pulse oximetry reading was < 96%, echocardiography was performed.”		
Target condition and reference standard(s)	Target condition: CCHD including cyanotic defects such as tetralogy of Fallot, pulmonary atresia, truncus arteriosus, transposition of the great vessels, total anomalous pulmonary venous return, and tricuspid atresia, as well as left-sided obstructive lesions, including coarctation of the aorta, critical aortic stenosis, interrupted aortic arch, and hypoplastic left heart syndrome Reference standard(s): Reference standard used for positive pulse oximetry results: echocardiography Reference standard used for negative pulse oximetry results: clinical follow-up		
Flow and timing	Duration of follow-up: not stated Follow-up information not available for 19 (0.1%)		
Comparative			
Notes			
Methodological quality			
Item	Authors’ judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	Unclear		

Sendelbach 2008 (Continued)

		Unclear	Low
DOMAIN 2: Index Test Pulse oximetry			
If a threshold was used, was it pre-specified?	Yes		
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	No		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
		Unclear	Low
DOMAIN 4: Flow and Timing			
Were all patients included in the analysis?	No		
Was there at least 28 days of appropriate follow up?	Unclear		
Did all patients receive a reference standard?	No		
		Unclear	

Singh 2014

Study characteristics	
Patient sampling	Retrospective observational study

Patient characteristics and setting	Country: United Kingdom Setting: level 3 Neonatal Unit of Birmingham Women’s Hospital Study period: from April 1, 2010, to July 31, 2013 Inclusion criteria: screening is part of routine practice Exclusion criteria: Antenatal diagnosis of CCHD N screened: 25,859 Prevalence of CCHD: 0.6 per 1000 live births		
Index tests	Pulse oximetry was performed with a handheld oximeter with a reusable probe Screening protocol: Site of testing: post-ductal (foot) and pre-ductal (right hand) Test timing: within 24 hours Oxygen saturation: functional Threshold: < 95% “A saturation result of < 95% in either limb or a difference of > 2% between the readings (if both were ≥ 95%) was considered abnormal. Following an abnormal first test, an initial assessment was performed. If this was unremarkable, oximetry was repeated 1 to 2 hours later. If the saturations remained abnormal on second testing, or if there were concerns following the initial assessment, newborns were classified as test positive and were admitted to the neonatal unit for further assessment.”		
Target condition and reference standard(s)	Target condition: CCHD Reference standard(s): Reference standard used for positive pulse oximetry results: echocardiography Reference standard used for negative pulse oximetry results: the Regional Cardiac Centre database at Birmingham Children’s Hospital, the Regional Congenital Anomaly Register, and the local mortality database		
Flow and timing	Duration of follow-up: not stated Loss to follow-up: none (n analyzed: 25,859)		
Comparative			
Notes			
Methodological quality			
Item	Authors’ judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Unclear		
Did the study avoid inappropriate exclusions?	No		
		Unclear	Low

DOMAIN 2: Index Test Pulse oximetry			
If a threshold was used, was it pre-specified?	Yes		
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	No		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
		Unclear	Low
DOMAIN 4: Flow and Timing			
Were all patients included in the analysis?	Yes		
Was there at least 28 days of appropriate follow up?	Unclear		
Did all patients receive a reference standard?	Yes		
		Low	

Turska 2012

Study characteristics	
Patient sampling	Prospective multicenter study with consecutive enrollment of participants
Patient characteristics and setting	Country: Poland Setting: 51 neonatal units in the Mazovian province of Poland as part of the POLKARD 2006 to 2008 program Study period: 1 year (from January 16, 2007, to January 31, 2008)

	Inclusion criteria: Protocol B: asymptomatic newborns at ≥ 34 weeks' gestation Exclusion criteria: Circulatory symptoms or coexisting diseases Prenatal diagnosis Live birth cohort, n = 55,944 (in 2611 newborns, the test could not be performed owing to technical problems [equipment failure, absence of trained staff due to holiday], 340 no consent, 1295 newborns with symptoms) N screened: 51,698 (92.4%) Prevalence of CCHD: 0.4 per 1000 live births		
Index tests	Pulse oximetry was performed with Novametrix, Nellcor, and Masimo pulse oximeters Screening protocol: Site of testing: post-ductal (foot) Test timing: within 24 hours (between the 2nd and 24th hours of life) Oxygen saturation: functional Threshold: < 95% “The measurement was carried out by specially trained nurses for 2 to 3 min on the infant's lower extremity between the 2nd and 24th hour of life after normalisation of the plethysmographic curve of the pulse oximeter.”		
Target condition and reference standard(s)	Target condition: CCHD defined as requiring an interventional procedure or cardiac surgery in the first month of life Reference standard(s): Reference standard used for positive pulse oximetry results: echocardiography Reference standard used for negative pulse oximetry results: clinical follow-up or based on data from the Mazovian Centre of Public Health		
Flow and timing	Duration of follow-up: not stated Loss to follow-up: none (n analyzed: 51,698)		
Comparative			
Notes	Funding: Ministry of Health in Poland		
Methodological quality			
Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	No		
		Unclear	Low

DOMAIN 2: Index Test Pulse oximetry			
If a threshold was used, was it pre-specified?	Yes		
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
		Low	Low
DOMAIN 4: Flow and Timing			
Were all patients included in the analysis?	Yes		
Was there at least 28 days of appropriate follow up?	Unclear		
Did all patients receive a reference standard?	Unclear		
		Unclear	

Van Niekerk 2016

Study characteristics	
Patient sampling	Prospective observational study
Patient characteristics and setting	Country: South Africa Setting: Mowbray Maternity Hospital (MMH), a busy level-2 maternity hospital in the Western Cape Province, SA Study period: May 19 to September 19, 2014

	Inclusion criteria: All neonates > 6 hours old with no clinical signs of cardiovascular disease were eligible Exclusion criteria: “unwell” infants, those < 6 hours old, those born to mothers < 14 years of age or unable to give informed verbal consent (owing to illness, illiteracy, or language barriers); all infants with a prenatal diagnosis of CHD or any signs of CHD, including a heart murmur ($\geq 3/6$) or significant dysmorphic features Livebirth cohort, n = 2256 (1220 mothers not approached) N screened: 1001 (44%) Prevalence of CCHD: 1 per 1000 live births		
Index tests	Pulse oximetry was performed with Nellcor pulse oximeters. Screening protocol: Site of testing: right hand and any foot Test timing: longer than 24 hours Oxygen saturation: functional Threshold: < 95%		
Target condition and reference standard(s)	Target condition: CCHD, which leads to death or needs surgical intervention before 28 days Reference standard(s): Reference standard used for positive pulse oximetry results: echocardiography Reference standard used for negative pulse oximetry results: not stated		
Flow and timing	Duration of follow-up: no physical follow-up Loss to follow-up: none (n analyzed: 1001)		
Comparative			
Notes	Study was funded in part by the School of Child and Adolescent Health Research Committee, Department of Paediatrics, Red Cross War Memorial Children’s Hospital and University of Cape Town		
Methodological quality			
Item	Authors’ judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	No		
Did the study avoid inappropriate exclusions?	Unclear		
		High	Low
DOMAIN 2: Index Test Pulse oximetry			
If a threshold was used, was it pre-specified?	Yes		

Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	No		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
		High	Low
DOMAIN 4: Flow and Timing			
Were all patients included in the analysis?	Yes		
Was there at least 28 days of appropriate follow up?	No		
Did all patients receive a reference standard?	No		
		Unclear	

Zhao 2014

Study characteristics	
Patient sampling	Prospective multicenter study with consecutive enrollment of participants
Patient characteristics and setting	Country: China Setting: 18 hospitals Study period: from August 1, 2011, to November 30, 2012 Inclusion criteria: all consecutive newborns (irrespective of gestational age or neonatal intensive care unit status) Exclusion criteria: Prenatally diagnosed major CHD Livebirth cohort, n = 130,282 (not screened: 9575 [3571 incomplete screening data, 1450 lack of consent, 2496 transfer to superior hospital, 27 prenatally diagnosed major CHD, 2031 symptomatic

	newborns]) N screened: 120,707 (92.7%) Prevalence of CCHD: 1.2 per 1000 live births		
Index tests	Pulse oximetry was performed with a RAD-5V / Multisite reusable sensor (LNOP YI, Masimo) Screening protocol: Site of testing: pre-ductal (right hand) and post-ductal (foot) Test timing: longer than 24 hours Oxygen saturation: functional Threshold: < 95% “The clinician repeated pulse oximetry testing 4 hours later if the first pulse oximeter oxygen saturation measurement was between 90% and 95%. Screening was deemed positive if an SpO ₂ of less than 95% was obtained both on the right hand and on either foot on two measures, separated by 4 hours; a difference between the two extremities was more than 3% on two measures, separated by 4 hours; or any measure was less than 90%.”		
Target condition and reference standard(s)	Target condition: CCHD Reference standard(s): Reference standard used for positive pulse oximetry results: echocardiography Reference standard used for negative pulse oximetry results: clinical follow-up and parents’ feedback		
Flow and timing	Duration of follow-up: clinical examination at 6 weeks of age at the hospital “The Children’s Hospital of Fudan University provided help with further confirmation of diagnosis for all affected babies from the participating hospitals. All cases of congenital heart disease were followed up by telephone review at least 1 year of age.” Loss to follow-up: none (n analyzed: 120,707)		
Comparative			
Notes	Funding: Key Clinical Research Project sponsored by Ministry of Health, Shanghai Public Health Three-Year Action Plan, sponsored by Shanghai Municipal Government, and National Basic Research Project of China		
Methodological quality			
Item	Authors’ judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	No		
		Unclear	Low
DOMAIN 2: Index Test Pulse oximetry			

Zhao 2014 (Continued)

If a threshold was used, was it pre-specified?	Yes		
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	No		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
		Unclear	Low
DOMAIN 4: Flow and Timing			
Were all patients included in the analysis?	Yes		
Was there at least 28 days of appropriate follow up?	Yes		
Did all patients receive a reference standard?	Yes		
		Low	

Zuppa 2015

Study characteristics	
Patient sampling	Prospective study with consecutive enrollment of participants
Patient characteristics and setting	Country: Italy Setting: Agostino Gemelli General Hospital Study period: 2 years (from 2009 to 2010) Inclusion criteria: all newborns admitted to the nursery. These newborns by definition were considered healthy or were under observation for maternal disease, mild prematurity, or low birth weight Exclusion criteria:

	Newborns with syndrome Total number of newborn infants included: N screened: 5750 Prevalence of CCHD: 0.2 per 1000 births		
Index tests	Pulse oximetry was performed with an Ohmeda 3900 pulse oximeter Screening protocol: Site of testing: post-ductal (foot) Test timing: longer than 24 hours Oxygen saturation: functional Threshold: < 95% “The measurement was performed by a professional nurse in all newborns admitted to the nursery, between the 48th and 72nd hours of life, before discharge. The probe detector was placed on one of the two legs, making sure that the newborn was quiet and with warm ends. The measurement was performed in presence of stable, continuous and free of artefacts pulse wave, for at least 3 minutes. In case of positive screening, a second check was carried out by medical staff after 15 to 30 min.”		
Target condition and reference standard(s)	Target condition: CCHD defined as severe cardiac alterations that require cardiac surgery during the first year of life Reference standard(s): Reference standard used for positive pulse oximetry results: electrocardiographic and echocardiography (echocardiograph “HP Sonos 4500, Agilent Technologies” [Andover, MA], a multifrequency probe [5 to 12 MHz], suitable for study of the neonatal heart. Evaluation was performed by 2-dimensional analysis [2-D], analysis of M-mode, and Doppler ultrasound) Reference standard used for negative pulse oximetry results: not stated		
Flow and timing	Duration of follow-up: not stated Loss to follow-up: none (n analyzed: 5751)		
Comparative			
Notes			
Methodological quality			
Item	Authors’ judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient Selection			
Was a consecutive or random sample of patients enrolled?	Yes		
Did the study avoid inappropriate exclusions?	Yes		
		Low	Low
DOMAIN 2: Index Test Pulse oximetry			

If a threshold was used, was it pre-specified?	Yes		
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
		Low	Low
DOMAIN 3: Reference Standard			
Is the reference standards likely to correctly classify the target condition?	Unclear		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
		High	Low
DOMAIN 4: Flow and Timing			
Were all patients included in the analysis?	Yes		
Was there at least 28 days of appropriate follow up?	Unclear		
Did all patients receive a reference standard?	Unclear		
		Unclear	

CCHD: critical congenital heart defect.

CHD: congenital heart defect.

NICU: neonatal intensive care unit.

Characteristics of excluded studies *[ordered by study ID]*

Study	Reason for exclusion
Andrews 2014	Study includes only information about positive results. Investigators did not follow infants who passed the screen once they left the hospital
Ewer 2012a	Health Technology Assessment report on already included study
Hoke 2002	Case-control design
John 2016	Study includes only information about positive results.
Kardasevic 2016	Different population (not asymptomatic newborns)
Kochilas 2013	Different outcome (not accuracy)
Lhost 2014	Out-of-hospital births
Meberg 2009	Different outcome (not accuracy)
Movahedian 2016	Study includes only information about positive results.
Narayan 2016a	Out-of-hospital births
Prudhoe 2013	Study did not include enough information for construction of a 2 × 2 table. Study includes data contained in Richmond 2002 (study included)
Reich 2003	Study provides a partial 2 × 2 diagnostic table from which estimation of sensitivity was not possible
Reich 2008	Different outcome (not accuracy)
Reich 2008a	Different outcome (not accuracy)
Riede 2009	Preliminary study
Ruangritnamchai 2007	Study includes only information about positive results.
Saha 2014	Journal club
Saxena 2015	Different population (not asymptomatic newborns)
Schena 2017	Different index test (combined pulse oximetry and perfusion index)
Studer 2014	Different outcome (not accuracy)
Taksande 2013	Study includes only information about positive results. Definition of test positive was not given

(Continued)

Tautz 2010	No ability to determine CCHD outcomes
Tsao 2016	Study includes only information about positive results.
Vaidyanathan 2011	Study provides a partial 2 × 2 diagnostic table. No ability to determine CCHD outcomes
Valmari 2006	Case report
Walsh 2011	Different outcome (not accuracy)

CCHD: critical congenital heart defect.

DATA

Presented below are all the data for all of the tests entered into the review.

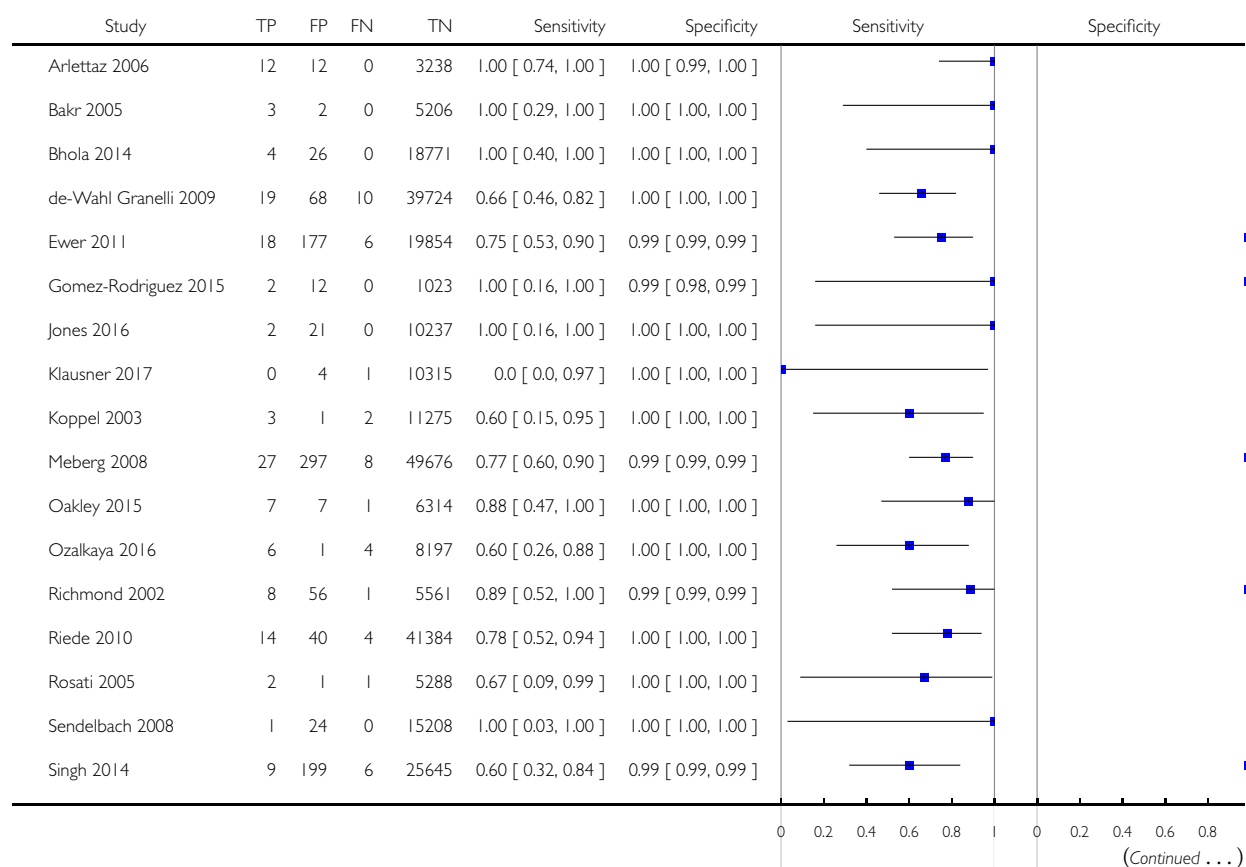
Tests. Data tables by test

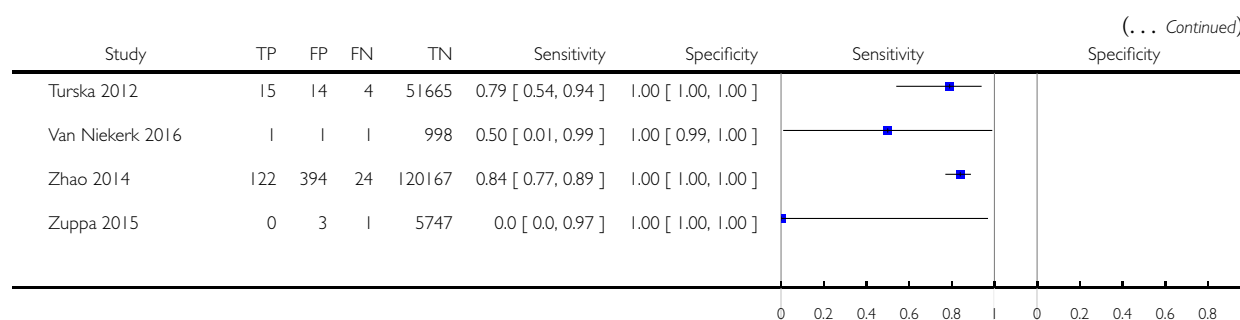
Test	No. of studies	No. of participants
1 All studies	21	457202
2 Primary analysis (threshold < 95% or ≤ 95%)	19	436758

Test 1. All studies.

Review: Pulse oximetry screening for critical congenital heart defects

Test: 1 All studies

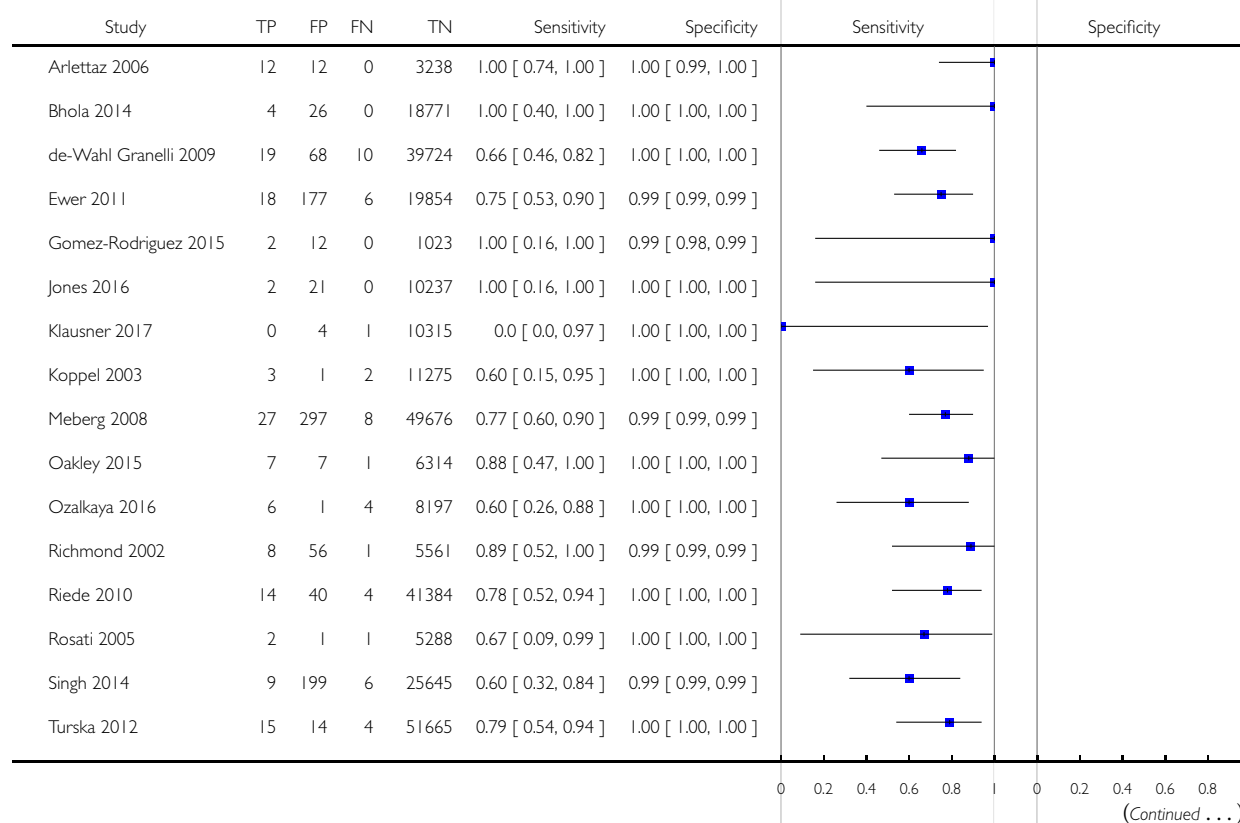


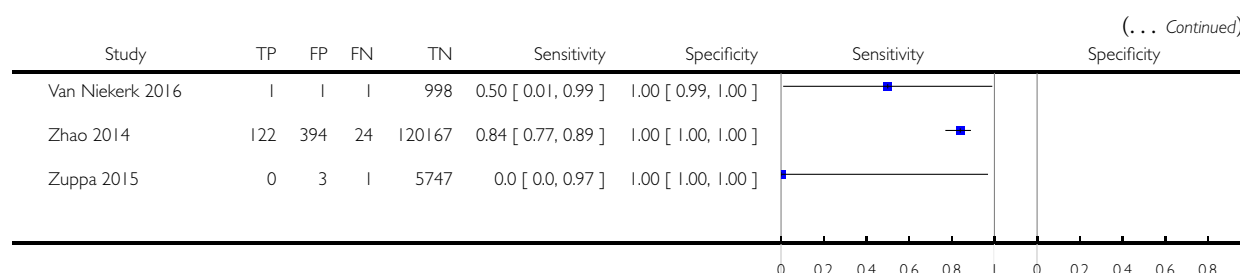


Test 2. Primary analysis (threshold < 95% or ≤ 95%).

Review: Pulse oximetry screening for critical congenital heart defects

Test: 2 Primary analysis (threshold < 95% or ≤ 95%)





ADDITIONAL TABLES

Table 1. Main studies characteristics

Study	Population	Index test					Reference standard(s)		
		<i>Antenatal diagnosis of CHD</i>	<i>Pulse oximeter</i>	<i>Limb</i>	<i>Test timing</i>	<i>Oxygen saturation</i>	<i>Threshold</i>	<i>Positive pulse oximetry</i>	<i>Negative pulse oximetry</i>
Arlettaz 2006	included		Nellcor NPB-40	post-ductal	within 24 hours	functional	< 95%	echocardiography	NA
Bakr 2005	excluded		Digiox PO 920	pre-ductal and post-ductal	longer than 24 hours	fractional	≤ 94%	echocardiography	cardiology database
Bhola 2014	included		Masimo Radical 5	post-ductal	longer than 24 hours	functional	< 95%	echocardiography	cardiology database
De-Wahl 2009	excluded		Radical SET v4	pre-ductal and post-ductal	longer than 24 hours	functional	< 95%	echocardiography	mortality data
Ewer 2011	included		Radical-7	pre-ductal and post-ductal	within 24 hours	functional	< 95%	echocardiography	clinical follow-up, cardiology database & congenital registry
Gomez-Rodriguez 2015	excluded		Radical-5	post-ductal	within 24 hours	functional	< 95%	echocardiography	clinical follow-up

Table 1. Main studies characteristics (Continued)

Jones 2016	excluded	NA	pre-ductal and post-ductal	within 24 hours	NA	≤ 95%	echocardiography	National Congenital Heart Disease Audit
Klausner 2017	excluded	NA	pre-ductal and post-ductal	longer than 24 hours	NA	< 95%	echocardiography	clinical follow-up
Koppel 2003	excluded	Ohmeda Medical	post-ductal	longer than 24 hours	functional	≤ 95%	echocardiography	clinical follow-up & congenital registry
Meberg 2008	excluded	RAD-5v	post-ductal	within 24 hours	functional	< 95%	echocardiography	clinical follow-up
Oakley 2015	excluded	Nellcor NPB 40	post-ductal	longer than 24 hours	functional	< 95%	echocardiography	cardiology database & mortality data
Ozalkaya 2016	excluded	Nellcor	pre-ductal and post-ductal	longer than 24 hours	functional	≤ 95%	echocardiography	echocardiography
Richmond 2002	included	Oxi machine	post-ductal	within 24 hours	fractional	< 95%	echocardiography	mortality data & congenital registry
Riede 2010	excluded	NA	post-ductal	longer than 24 hours	functional	≤ 95%	echocardiography	congenital registry
Rosati 2005	excluded	NA	post-ductal	longer than 24 hours	functional	≤ 95%	echocardiography	clinical follow-up
Sendelbach 2008	excluded	Nellcor N-395	post-ductal	within 24 hours	functional	< 96%	echocardiography	clinical follow-up
Singh 2014	excluded	NA	pre-ductal and post-ductal	within 24 hours	functional	< 95%	echocardiography	mortality data & congenital registry & cardiology database

Table 1. Main studies characteristics (Continued)

Turska 2012	excluded	Novamatrix, Nellcor & Masimo	post-ductal	within 24 hours	functional	< 95%	echocardiography	clinical follow-up and Public Health registries
Van Niekerk 2016	excluded	Nellcor	pre-ductal and post-ductal	longer than 24 hours	functional	< 95%	echocardiography	NA
Zhao 2014	excluded	RAD-5V	pre-ductal and post-ductal	longer than 24 hours	functional	< 95%	echocardiography	clinical follow-up
Zuppa 2015	excluded	Ohmeda 3900	post-ductal	longer than 24 hours	functional	< 95%	echocardiography	NA
NA: not available								

Table 2. Subgroup analysis

	N	Sensitivity (95% CI)	Relative sensitivity P value	False-positive rate (FPR) (95% CI)	Relative FPR P value
Antenatal diagnosis					
Included	4	86.3% (71.8 to 94.0)	0.071	0.46% (0.13 to 1.59)	0.231
Excluded	15	74.1% (65.7 to 81.1)		0.10% (0.05 to 0.21)	
Test timing					
Longer than 24 hours	11	73.6% (62.8 to 82.1)	0.393	0.06% (0.03 to 0.13)	0.027
Within 24 hours	8	79.5% (70.0 to 86.6)		0.42% (0.20 to 0.89)	
Limb					
Foot only	11	81.2% (70.9 to 88.4)	0.197	0.13% (0.05 to 0.31)	0.718
Foot and right hand	8	71.2% (58.5 to 81.3)		0.17% (0.06 to 0.46)	
Risk of bias ("flow and timing")					

Table 2. Subgroup analysis (Continued)

Unclear risk of bias	9	77.8% (64.1 to 87.3)	0.937	0.05% (0.02 to 0.12)	0.016
Low risk of bias	10	77.3% (68.8 to 84.0)		0.34% (0.17 to 0.66)	

APPENDICES

Appendix I. Searches performed

Date: March 2017	Search strategy	Hits retrieved
Cochrane Central Register of Controlled Trials (CENTRAL; 2017, Issue 2) in the Cochrane Library	(infant or newborn or neonate or neonatal or premature or preterm or very low birth weight or low birth weight or VLBW or LBW) AND (Congenital Heart Defects OR Heart Valve Diseases OR tetralogy near fallot* OR cyanotic near heart OR congenital near heart OR congenital near cardiac OR aortic near coarctation OR valve near diseases OR hypoplastic near syndrome OR pulmonary near atresia OR interruption of the aortic arch OR valve near stenosis OR pulmonary near atresia) AND (oximetry OR pulse near oximetr* OR oxygen near saturation OR O ₂ near saturation)	76
MEDLINE via PubMed (1966 to current)	(Infant, Newborn[MeSH] OR neonate* OR infant* OR newborn*) AND (Heart Defects, Congenital[MeSH] OR Heart Valve Diseases[MeSH] OR tetralogy fallot* OR cyanotic heart OR congenital heart OR congenital cardiac OR aortic coarctation OR valve diseases OR hypoplastic syndrome OR pulmonary atresia OR interruption of the aortic arch OR valve stenosis OR pulmonary atresia) AND (oximetry[MeSH] OR oximetry OR pulse oximetr* OR oxygen saturation OR O ₂ saturation)	1368

(Continued)

Embase via Ovid (1980 to current)	(exp Infant OR exp Newborn OR neonat*.mp OR infant*.mp OR newborn*.mp) AND ((exp congenital heart malformation/) OR (exp valvular heart disease/) OR (tetralogy adj3 fallot*).mp OR (cyanotic adj3 heart).mp OR (congenital adj3 heart).mp OR (congenital adj3 cardiac).mp OR (aortic adj3 coarctation).mp OR (valve adj3 diseases).mp OR (hypoplastic adj3 syndrome).mp OR (pulmonary adj3 atresia).mp OR (interruption of the aortic arch).mp OR (valve adj3 stenosis).mp) AND (exp oximetry OR (pulse adj3 oximetr*).mp OR (oxygen adj3 saturation).mp OR (O ₂ adj3 saturation).mp)	1779
CINAHL (1982 to current)	TX (Infant, Newborn OR neonate* OR infant* OR newborn*) AND TX (Congenital Heart Defects OR Heart Valve Diseases OR tetralogy fallot* OR cyanotic heart OR congenital heart OR congenital cardiac OR aortic coarctation OR valve diseases OR hypoplastic syndrome OR pulmonary atresia OR interruption of the aortic arch OR valve stenosis OR pulmonary atresia) AND TX (oximetry OR pulse oximetr* OR oxygen saturation OR O ₂ saturation)	192
TOTAL before de-duplication		3415
TOTAL after de-duplication		2695

Appendix 2. QUADAS 2

Item	Criteria for assessment
Domain 1: Patient selection	
Describe methods of patient selection (prior testing, presentation, intended use of index test and setting)	
A. Risk of bias	
Was a consecutive or random sample of patients enrolled?	“Yes” if described enrolling a consecutive or random sample of newborns before discharge from hospital “No” if criteria for “yes” not achieved

(Continued)

	“Unclear” if the study did not describe the method of enrollment
Did the study avoid inappropriate exclusions?	“Yes” if exclusions were detailed and review authors reached consensus on the appropriateness of any exclusion “No” if inappropriate exclusions were reported (eg, if cases with antenatally diagnosed congenital heart disease were excluded) “Unclear” if insufficient information was provided
Could selection of patients have introduced bias?	A judgment of low, high, or unclear risk of bias was based on a balanced assessment of responses to the above signaling questions
B. Concerns about applicability	
Is there concern that the included patients do not match the review question?	A judgment of low, high, or unclear concern about applicability was made on the basis of how closely the sample matches an asymptomatic newborn population screened for CCHD
Domain 2: Index test	
Describe the index test and how it was conducted and interpreted	
A. Risk of bias	
Were the index test results interpreted without knowledge of results of the reference standard?	“Yes” if pulse oximetry was conducted and interpreted before the echocardiogram or clear temporal pattern to the order of testing that precludes the need for formal blinding (eg, echocardiogram, clinical follow-up, and inclusion in congenital anomaly registries are always posterior to index test) “No” if reference standard results were available to those who conducted or interpreted the pulse oximetry “Unclear” if insufficient information was provided
If a threshold was used, was it prespecified?	“Yes” if a threshold was prespecified “No” if trial authors selected a cutoff value based on analysis of collected data “Unclear” if insufficient information was provided
Could the conduct or interpretation of the index test have introduced bias?	A judgment of low, high, or unclear risk of bias was based on a balanced assessment of responses to the above signaling questions
B. Concerns about applicability	
Is there concern that the index test, its conduct, or its interpretation differ from the review question?	A judgment of low, high, or unclear concern about applicability was based on a balanced assessment of information detailed under “index test” description
Domain 3: Reference standard	

(Continued)

Describe the reference standard(s) and how they were conducted and interpreted	
A. Risk of bias	
Is the reference standard likely to correctly classify the target condition?	“Yes” if the study used an appropriate reference standard (diagnostic echocardiography and clinical follow-up in the first 28 days of life, including postmortem findings and mortality and congenital anomaly databases to identify false-negative patients) “No” if the study did not use an appropriate reference standard “Unclear” if the reference standard used was not clearly specified
Were the reference standard results interpreted without knowledge of results of the index test?	“Yes” if the person undertaking the reference test did not know the results of the pulse oximetry “No” if pulse oximetry results were available to those who conducted or interpreted the echocardiogram “Unclear” if insufficient information was provided
Could the reference standard, its conduct, or its interpretation have introduced bias?	A judgment of low, high, or unclear risk of bias was based on a balanced assessment of responses to the above signaling questions
B. Concerns about applicability	
Is there concern that the target condition as defined by the reference standard does not match the question?	A judgment of low, high, or unclear concern about applicability was based on the possibility of reference standards mixing both critical and non-critical congenital heart disease
Domain 4: Flow and timing	
Describe any patients who did not receive the index test and/or reference standard(s) or who were excluded from the two-by-two table (refer to flow diagram), and describe the time interval and any interventions between index test and reference standard(s)	
A. Risk of bias	
Was at least 28 days of appropriate follow-up provided?	“Yes” if follow-up was at least 28 days “No” if follow-up was less than 28 days “Unclear” if insufficient information was provided
Did all patients receive a reference standard?	“Yes” if the study specifically stated that all patients received echocardiogram, clinical follow-up, or confirmation by mortality and congenital anomaly databases (for both positive and negative pulse oximetry results) “No” if some negative pulse oximetry participants were lost to follow-up without any confirmation “Unclear” if insufficient information was provided
Were all patients included in the analysis?	“Yes” if the study had no withdrawals or withdrawals were clearly described “No” if the number of patients contributing to the two-by-two

(Continued)

	tables did not match the number of patients recruited and no reasons for exclusions were described “Unclear” if information was not enough to establish the flow of participants
Could the patient flow have introduced bias?	A judgment of low, high, or unclear risk of bias was based on a balanced assessment of responses to the above signaling questions

CONTRIBUTIONS OF AUTHORS

Protocol development: Shakila Thangaratinam, Andrew K. Ewer, Maria Nieves Plana, Javier Zamora, Gautham Suresh.

Selection of studies and data extraction: Maria Nieves Plana, Andrew K. Ewer.

Quality assessment: Maria Nieves Plana, Luis Fernandez-Pineda, Andrew K. Ewer.

Statistical analysis and interpretation of data: Javier Zamora, Maria Nieves Plana, Shakila Thangaratinam, Andrew K. Ewer.

Drafting of the manuscript: Maria Nieves Plana, Javier Zamora, Shakila Thangaratinam, Andrew K. Ewer.

Critical revision: Shakila Thangaratinam, Andrew K. Ewer, Maria Nieves Plana, Javier Zamora, Gautham Suresh, Luis Fernandez-Pineda.

DECLARATIONS OF INTEREST

Maria Nieves Plana: none to declare.

Javier Zamora: none to declare.

Gautham Suresh: none to declare.

Luis Fernandez-Pineda: none to declare.

Shakila Thangaratinam: an author of one of the primary studies - Ewer 2011.

Andrew K. Ewer: an author of one of the primary studies - Ewer 2011.

SOURCES OF SUPPORT

Internal sources

- IRYCIS, Spain.
- Instituto Ramón y Cajal de Investigaciones Sanitarias, Madrid
- Universidad Rey Juan Carlos, Madrid, Spain.

External sources

- CIBERESP, Spain.

CIBER Epidemiology and Public Health

- Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Department of Health and Human Services, USA.

Editorial support of the Cochrane Neonatal Review Group has been funded with Federal funds from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Department of Health and Human Services, USA, under Contract No. HHSN275201600005C

DIFFERENCES BETWEEN PROTOCOL AND REVIEW

Searching other resources

The protocol describes our plan to search the MEDION database (www.mediondatabase.nl), but this resource is not longer available.

Statistical analysis

We planned to use a 95% saturation level as the primary threshold for the analysis and to perform separate analyses for other thresholds categorized as < 95% and > 95% saturation levels. However, at editorial phase, it was suggested to group thresholds. As many studies used a lower than or lower than or equal to 95% threshold, we decided to group all these studies for the main analysis.

The protocol describes our plan to switch the modeling strategy to fit two univariate random-effects logistic regression models by assuming no correlation between sensitivity and specificity if the number of studies was small (fewer than four), or if the proposed modeling strategy led to problems in achieving convergence. We identified sufficient studies to fit a bivariate model and had no problem achieving model convergence.

Certainty of the evidence

We decided post hoc to assess the certainty of evidence using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach ([Schunemann 2008](#)).

Subgroup analysis

We did not perform subgroup analyses by oxygen saturation or study design, given the low power of these subgroup analyses.

Sensitivity analysis

Additionally, we decided to perform ad hoc sensitivity analyses to describe how sensitivity and specificity vary by including or excluding studies with different thresholds.

Capítulo 6



Capítulo 6

En este capítulo se presenta el estudio de revisión sistemática realizado con el objetivo de evaluar la exactitud diagnóstica de los códigos de la Clasificación Internacional de Enfermedades (CIE) en la identificación de pacientes con pancreatitis aguda y crónica.

El diagnóstico certero de la pancreatitis continúa siendo complejo. No existen criterios diagnósticos aceptados para la pancreatitis crónica y los criterios de Atlanta utilizados para el diagnóstico de pancreatitis aguda requieren de la presencia de 2 de los 3 criterios (clínicos: dolor abdominal intenso de inicio súbito localizado en el epigastrio y

a menudo irradiado a la espalda; de laboratorio: amilasa (o lipasa) en sangre 3 veces por encima del valor normal; hallazgos característicos de PA en pruebas de imagen).

Cada pareja de criterios puede tener diferente sensibilidad y especificidad y hay una gran variabilidad en el valor predictivo positivo de los códigos CIE en el diagnóstico de pancreatitis aguda.

6.1 Objetivos

Determinar la exactitud diagnóstica de los códigos de la CIE en la identificación de pacientes con pancreatitis aguda y crónica e investigar el efecto de covariables en el rendimiento diagnóstico.

6.2 Metodología

Se realizó una búsqueda en las bases de datos MEDLINE, EMBASE y SCOPUS. Se incluyeron estudios que evaluaban la exactitud de los códigos según la CIE-8 o la CIE-9, o la CIE-10 (o una combinación de las anteriores) frente a una prueba de referencia consistente en el diagnóstico experto. Los códigos de la CIE evaluados fueron todos los subtipos de K85 (pancreatitis aguda), K86.0 y K86.1 (pancreatitis crónica) de la CIE 10 y 577.0 (pancreatitis aguda), 577.1 (pancreatitis crónica) de la CIE 8 y CIE 9.

Para los estudios con información de las cuatro celdas de la tabla de clasificación cruzada (VP, FP, FN, TN) se estimó la sensibilidad y la especificidad con su intervalo de confianza al 95%. Se ajustó una curva SROC utilizando un modelo bivariante lineal generalizado mixto y se calculó el área bajo la curva SROC. Se calcularon para cada estudio el valor predictivo positivo. El VPP promedio se obtuvo mediante un modelo de

regresión logística de efectos aleatorios. La heterogeneidad entre estudios se cuantificó con la varianza de los logit de sensibilidad y especificidad y el estadístico I^2 . El riesgo de sesgo de los estudios incluidos en la revisión se evaluó mediante la herramienta QUADAS (Whiting 2003). Los análisis estadísticos se realizaron utilizando la macro METANDI (Harbord y Whiting 2009) y Metaprop (Nyaga 2014) de STATA.

6.3 Resumen de resultados

Se incluyeron 24 estudios correspondientes a 18,106 participantes. En 21 estudios se evaluaba la exactitud de los códigos diagnósticos de la CIE en la identificación de pacientes con pancreatitis aguda y en 7 estudios en la identificación de pacientes con pancreatitis crónica.

El valor predictivo positivo promedio de los códigos CIE para la identificación de pacientes con pancreatitis aguda fue de 0.71 (IC 95% 0.61; 0.79; $p < 0.0001$; $I^2 = 98.5\%$)⁵. El promedio ponderado de sensibilidad y especificidad fue de 0.85 (IC 95% 0.59; 0.96) y 0.96 (IC 95% 0.65; 1.00) respectivamente, calculada sobre 6 estudios (N=10,018). Los cocientes de probabilidad positivo y negativos fueron de 21.6 (IC 95% 2.1; 223.7) y de 0.2 (IC 95% 0.1; 0.5) respectivamente. El área bajo la curva ROC fue de 0.95 (95% CI 0.56; 1.00). El análisis de subgrupos por tipo de CIE se hizo a partir de 14 estudios que utilizaron sólo la CIE-9 (N=10,809 participantes) y 3 estudios (N=2,855) que utilizaron sólo la CIE-10. El VPP de la codificación utilizando la CIE-9 fue de 0.69 (IC 95% 0.55;

⁵ Por consistencia con la publicación se muestran aquí los resultados como proporción.

0.81) con una Tau^2 de 0.271 y un I^2 de 98.3%, mientras que el VPP de la codificación utilizando la CIE-10 fue de 0.79 (IC 95% 0.69; 0.88) con una Tau^2 de 0.042 y un I^2 de 95.7% ($P=0.189$). El análisis de subgrupos de estudios en adultos ($N=14,938$) vs pediátricos ($N=694$), mostró un VPP para el subgrupo de adultos de 0.71 (IC 95% 0.61; 0.80) y de 0.68 (IC 95% 0.44; 0.88) ($P=0.826$). El análisis de sensibilidad con sólo cinco estudios ($N=1,718$) para la identificación de casos incidentes de pancreatitis aguda mostró un VPP de 0.78 (IC 95% 0.70; 0.85).

El valor predictivo positivo promedio de los códigos CIE para la identificación de pacientes con pancreatitis crónica fue de 0.71 (IC 95% 0.54; 0.85; $p < 0.0001$; $I^2 = 98.3\%$). La sensibilidad y especificidad fue de 0.75 (IC 95% 0.71; 0.80) y 0.94 (IC 95% 0.93; 0.96) respectivamente, calculada en base a dos estudios (1,123 participantes).

6.4 Conclusiones

La exactitud diagnóstica de los códigos CIE para la identificación de pacientes con pancreatitis no es óptima. El rendimiento podría ser mayor cuando los códigos se aplican a episodios incidentes de pancreatitis aguda y en adultos y también cuando se utiliza la CIE-10. Es necesario desarrollar nuevos criterios diagnósticos para los pacientes con pancreatitis aguda y crónica recurrente.

6.5 Publicación

Se incluye como cuerpo del capítulo la publicación del estudio de revisión sistemática en la revista Clinical and Translational Gastroenterology (Factor de impacto año 2018: 4.803; rango: Q1, 20/84 de la categoría Gastroenterology and Hepatology):

Xiao AY, Tan ML, Plana MN, Yadav D, Zamora J, Petrov MS. The Use of International Classification of Diseases Codes to Identify Patients with Pancreatitis: A Systematic Review and Meta-analysis of Diagnostic Accuracy Studies. Clin Transl Gastroenterol. 2018 Oct 4;9(10):191. Erratum in: Clin Transl Gastroenterol. 2018 Nov 21;9(11):209.

El número de citas de esta publicación (07/07/2021) es de 18 citas en la plataforma Scopus y 22 en Google Scholar.

REVIEW ARTICLE

Open Access

The Use of International Classification of Diseases Codes to Identify Patients with Pancreatitis: A Systematic Review and Meta-analysis of Diagnostic Accuracy Studies

Amy Y. Xiao, MBChB¹, Marianne L. Tan, MBChB¹, Maria N. Plana, MD², Dhiraj Yadav, MD, MPH³, Javier Zamora, MSc, PhD^{2,4} and Maxim S. Petrov, MD, MPH, PhD¹

Abstract

Background: Hospital discharge codes are increasingly used in gastroenterology research, but their accuracy in the setting of acute pancreatitis (AP) and chronic pancreatitis (CP), one of the most frequent digestive diseases, has never been assessed systematically. The aim was to conduct a systematic literature review and determine accuracy of diagnostic codes for AP and CP, as well as the effect of covariates.

Methods: Three databases (Pubmed, EMBASE and Scopus) were searched by two independent reviewers for relevant studies that used International Classification of Disease (ICD) codes. Summary estimates of sensitivity, specificity and positive predictive value were obtained from bivariate random-effects regression models. Sensitivity and subgroup analyses according to recurrence of AP and age of the study population were performed.

Results: A total of 24 cohorts encompassing 18,106 patients were included. The pooled estimates of sensitivity and specificity of ICD codes for AP were 0.85 and 0.96, respectively. The pooled estimates of sensitivity and specificity of ICD codes for CP were 0.75 and 0.94, respectively. The positive predictive value of ICD codes was 0.71 for either AP or CP. It increased to 0.78 when applied to incident episode of AP only. The positive predictive value decreased to 0.68 when the ICD codes were applied to paediatric patients.

Conclusion: Nearly three out of ten patients are misidentified as having either AP or CP with the indiscriminate use of ICD codes. Limiting the use of ICD codes to adult patients with incident episode of AP may improve identification of patients with pancreatitis in administrative databases.

Introduction

Advancements in information technology have revolutionised the way individual patient data are collected and

processed¹. Increasingly, more simultaneous documentation and execution has allowed large amounts of data to be amassed in a short time²—a phenomenon that has been penned 'big data'. 'Big data' is defined by characteristics of large variety of sources, volume and velocity³. In the health industry, these sources can vary from regional databases of electronic health records and cancer registries to individual smartphone monitoring of sleep and

Correspondence: Maxim S. Petrov (max.petrov@gmail.com)

¹Department of Surgery, University of Auckland, Auckland, New Zealand

²CIBER Epidemiology and Public Health (CIBERESP), Universidad Francisco de Vitoria, Madrid, Spain

Full list of author information is available at the end of the article.

© The Author(s) 2018



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

diet³. Digitalisation has enabled practical and low-cost accessibility of 'big data', and one example of it is the use of administrative diagnostic codes. Diagnostic coding is now used ubiquitously, including application for the purpose of research¹. Increasingly, larger cohorts are required to produce more generalisable results and distil out trends from background error⁴. Diagnostic codes are a practical method to achieve these goals¹ and, therefore they have become engrained in medical research in general and gastroenterology research in particular⁴.

Pancreatitis poses a significant burden to health systems⁵, at least in part because there are still obstacles to accurate diagnosis of pancreatitis. Chronic pancreatitis (CP) has no universally accepted diagnostic criteria⁶. The Atlanta criteria to diagnose acute pancreatitis (AP)⁷ offer a composite definition that is based on the presence of two out of the three domains (clinical, laboratory and radiological). Each pair of domains can have different diagnostic accuracy, and it is conceivable that individual doctors may favour one combination over another. Further, there is high variability in the reported positive predictive value of diagnostic coding in AP^{8,9}. This not only has implications for the studies that rely on diagnostic coding, but also suggests possible overdiagnosing of AP. Further, inflated estimates of burden of AP may lead to excessive cost allocation, unnecessary procedures and may deflate estimates of mortality¹⁰.

The aim of this study was to conduct a systematic literature review of cohort studies to assess the accuracy of diagnostic codes for AP and CP and investigate the effect of covariates.

Methods

Search strategy

Three electronic databases (Pubmed, EMBASE and Scopus) were used to search for articles from the earliest available date until February 1, 2016. The Pubmed and EMBASE search strategy contained three sets of terms and the Scopus search strategy contained four sets. The Boolean operator 'AND' was used between the sets whereas the operator 'OR' was used within each set. For Pubmed, the first set contained "Drug prescriptions", "Insurance, Health", "Databases as topic", "Clinical coding", "Registries", "Hospitalisation", "International Classification of Disease" and "ICD". The second set contained "Validation Studies as topic", "Epidemiologic Research Design", "Algorithm" and "Pancreatitis/epidemiology". The third set contained "Pancreatitis". These were all MeSH terms, except for "ICD". For EMBASE, the terms were searched by subject heading and exploded where possible. The first set contained the exploded terms of "Health Services Research", "Medical Records", "International Classification of Disease", "Prescriptions", "Hospital

Discharge", "Billing and Claims" and "Coding" and the terms searched by keyword "Health Information", "Surveillance", "Administrative Data", "Code\$" and "ICD\$". The second set contained the exploded terms of "Validity", "Validation Study" and "Algorithm" and the terms searched by keyword "Case Definition", "Sensitivity", "Specificity", "Positive Predictive Value" and "Negative Predictive Value". The third set contained the exploded term "Pancreatitis". For Scopus, the first set contained "Prescription", "Medical Records", "Insurance Claim", "Registries", "Database" and "Hospital Discharge". The second set contained "International Classification of Disease", "ICD*", "Coding" and "Code*". The third set contained "Case Definition", "Sensitivity", "Specificity", "Positive Predictive Value" and "Negative Predictive Value". The fourth set contained "Acute Pancreatitis". The search was limited to articles in English.

Inclusion criteria

Included studies required to have reported at least one measure of diagnostic accuracy (such as sensitivity, specificity, positive predictive value and negative predictive value) in the setting of AP and/or CP. The accuracy of codes according to either ICD-8 or ICD-9, or ICD-10 (or a combination of the above) had to be compared with an independent reference standard formulated by experts in the field. The ICD codes explored in this study were all subtypes of K85 and K86.0, 86.1 from ICD 10 CM and 577.0, 577.1 from ICD 8.9. Two independent reviewers (A.Y.X.) and (M.L.T.) screened for eligible studies and any discrepancies were discussed with the senior author (M.S.P.).

Exclusion criteria

Studies were excluded if there was inadequate information on the coding provided or no independent reference standard used. Cases of post-ERCP pancreatitis or postpartum pancreatitis were excluded. Studies with a sample size of less than 25 were also excluded, as well as studies focused on a particular aetiology of AP or CP.

Data extraction

Extraction was performed on the following variables: type of administrative code, coding position, number of cases identified by the administrative code, reference standard used, number of cases verified by reference code, positive predictive value, negative predictive value, sensitivity and specificity. Positive predictive value (PPV), negative predictive value, sensitivity and specificity were calculated if not reported in the primary article and required data were available. Positive and negative likelihood ratios, as well as diagnostic odds ratios, were calculated for each study if adequate information was available. Paediatric

and first episode of acute pancreatitis cases were also recorded.

Quality assessment

The QUADAS (Quality Assessment of Diagnostic Accuracy Studies) tool¹¹ was used to assess the methodological quality of the included studies based on a total of 14 items.

Statistical methods

For studies in which it was possible to extract information on all four cells of the 2×2 table, sensitivity and specificity were estimated with 95% confidence intervals (CI). A bivariate random-effects regression model was fitted to obtain a summary receiver operating characteristic (SROC) curve and the corresponding area under the curve in order to take the potential trade-off between sensitivity and specificity explicitly into consideration and incorporate this negative correlation into the analysis¹². Positive predictive values were calculated for all studies included. Mean PPV was obtained using a random-effects logistic regression. Sensitivity, specificity and PPV were represented graphically using the corresponding forest plots to investigate heterogeneity. Heterogeneity among studies was quantified with the variance of the logit of accuracy indices as estimated by the bivariate model, τ^2 and I^2 statistics. The minimum number of studies required to calculate heterogeneity was two. We selected a priori the following factors as potential sources of heterogeneity: ICD version, coding position, reference standard, recurrence of acute pancreatitis and age group of the patients. If the number of studies was sufficient, we investigated heterogeneity by adding covariate terms to the bivariate model to assess the effect of a covariate on accuracy. Statistical analyses were conducted using the Metandi and Metaprop_one programs for the STATA software¹³.

Results

Characteristics of the included studies

A total of 24 studies were included in the final analysis (Fig. 1). Baseline characteristics of all the included cohorts are shown in Table 1. A total of 21 cohorts investigated AP^{8,14–33} and seven cohorts—CP^{15,18,20,21,34–36}. In AP, two cohorts used ICD-8, 15—ICD-9 and five—ICD-10. In CP, two cohorts used ICD-8, five—ICD-9 and two—ICD-10. The total number of individuals in the source population was 18,106 (6858 with AP; 1927 with CP; 8537 with diseases other than AP and 784 with diseases other than CP). The total number of validated cases was 7464 (5668 with AP and 1796 with CP). The median study period was 3 years with an interquartile range of 2 to 10 years. Methodological quality of the included studies is presented in Tables 2 and 3.

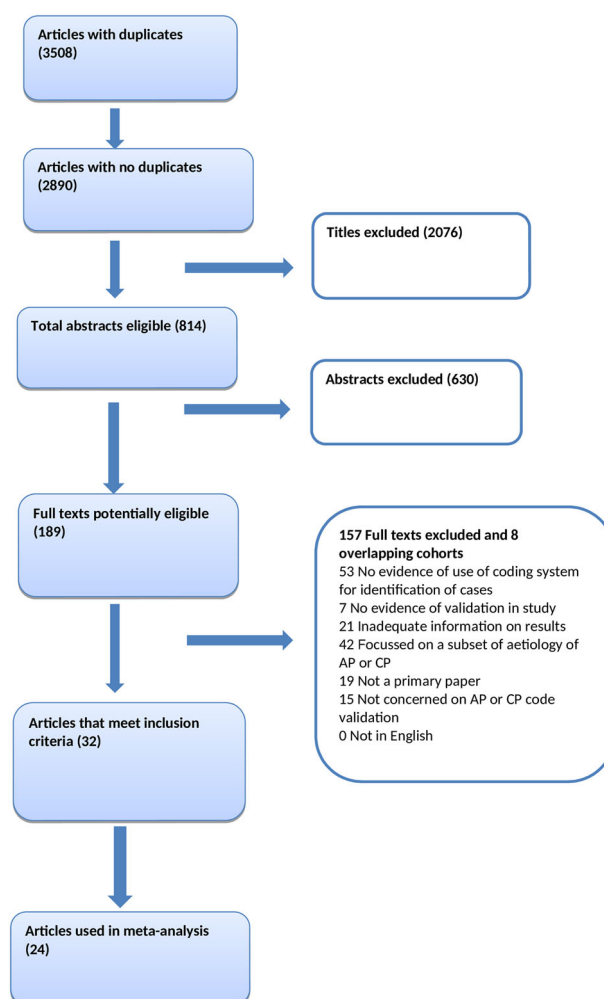


Fig. 1 Flow chart of the study selection process

Studies in acute pancreatitis

A total of 21 cohorts reported on the PPV of ICD codes for AP. The crude pooled PPV was 0.71 (95% CI 0.61–0.79; $p < 0.0001$; $I^2 = 98.5\%$) (Fig. 2). Six cohorts (10,018 participants) reported on sensitivity, specificity, positive likelihood ratio, negative likelihood ratio and diagnostic odds ratio. The crude pooled sensitivity and specificity were 0.85 (95% CI 0.59–0.96) and 0.96 (95% CI 0.65–1.00), respectively (Fig. 3). The crude pooled positive likelihood ratio, negative likelihood ratio and diagnostic odds ratio were 21.6 (95% CI 2.1–223.7), 0.2 (95% CI 0.1–0.5) and 137.8 (95% CI 19.0–1001.4), respectively. The SROC curve produced an area under the curve of 0.95 (95% CI 0.56–1.00) (Fig. 4).

The subgroup analysis according to the versions of ICD included 10,809 participants from 14 cohorts that used ICD-9 alone, as well as 2855 participants from three cohorts that used ICD-10 alone. The PPV for ICD-9 codes

Table 1 Characteristics of the included cohorts

Author	Country	Database	Study period	Patient population	Coding system	Coding position	Reference standard	Total (n)	Cases coded(n)	PPV	NPV	Sensitivity	Specificity	Likelihood ratios positive/negative
<i>Acute pancreatitis</i>														
Eland et al. ¹⁴	Netherlands	National Information System on Hospital Care	1985, 1990, 1995	Adult	ICD-9	Primary	Non-Atlanta criteria	101	101	0.82	–	–	–	–
Porta et al. ¹⁵	Spain	PANKRAS II Study	1992–1995	Adult	ICD-9	Primary	Not reported	600	37	0.54	0.96	0.50	0.97	17.81/0.47
Chwistek et al. ¹⁶	United Kingdom	Bridgeport Hospital	1994–1996	Adult	–	Primary or secondary	Non-Atlanta criteria	145	145	0.85	–	–	–	–
Floyd et al. ¹⁷	Denmark	Hospital Discharge Registry	1981–2000	Adult	ICD-8, ICD-10	Primary or secondary	Not reported	99	99	0.82	–	–	–	–
Quraishi et al. ⁸	United States	Henry Ford Health System	1998–2003	Adult	ICD-9	Primary or secondary	Non-Atlanta criteria	1393	128	0.22	1.00	1.00	0.93	13.65/0
Yadav et al. ¹⁸	United States	Veterans Outpatient Detoxification Programme	2002–2003	Adult	ICD-9	Primary or secondary	Non-Atlanta criteria	50	50	0.32	–	–	–	–
Kandula et al. ¹⁹	United States	Children's Hospital of Pittsburgh	1995–2004	Paediatric	ICD-9	Primary or secondary	Non-Atlanta criteria	109	109	0.80	–	–	–	–
Spanier et al. ²⁰	Netherlands	Academic Medical Centre	2002–2003	Adult	ICD-9	Primary or secondary	Non-Atlanta criteria	523	112	0.78	0.89	0.65	0.94	10.2/0.37
Nojgaard et al. ²¹	Denmark	Hvidovre Hospital Admissions	1983, 1994, 2005	Adult	ICD-8, ICD-10	Primary or secondary	Non-Atlanta criteria	165	165	0.64	–	–	–	–
Dore et al. ²²	United States	Normative Health Information Database	2005–2007	Adult	ICD-9	Primary or secondary	Atlanta criteria	585	585	0.50	–	–	–	–
Omdal et al. ²³	Norway	Haukeland University	1996–2006	Adult	ICD-9, ICD-10	Primary or secondary	Non-Atlanta criteria	724	724	0.78	–	–	–	–

Table 1 continued

Author	Country	Database	Study period	Patient population	Coding system	Coding position	Reference standard	Total (n)	Cases coded(n)	PPV	NPV	Sensitivity	Specificity	Likelihood ratios positive/negative
Razavi et al. ²⁴	Sweden	Swedish National Patient Registry	1998–2007	Adult	ICD-10	Primary or secondary	Atlanta criteria	603	530	0.83	0.69	0.95	0.36	1.49/0.14
Ma et al. ²⁵	United States	Yale New Haven Children's Hospital	1994– 2007	Paediatric	ICD-9	Primary or secondary	Non-Atlanta criteria	548	548	0.50	–	–	–	–
Edwards et al. ²⁶	United Kingdom	Derriford Hospital Emergency	2009– 2010	Adult	Not reported	Primary or secondary	Non-Atlanta criteria	231	231	0.29	–	–	–	–
Shen et al. ²⁷	Taiwan	National Health Insurance Research Database	2006–2008	Adult	ICD-9	Primary or secondary	Atlanta criteria	50	50	0.90	–	–	–	–
Saligram et al. ²⁸	United States	University of Pittsburgh Medical Centre	2000, 2002, 2005	Adult	ICD-9	Primary	Atlanta criteria	803	401	0.77	0.97	0.97	0.81	5.11/0.04
Podugu et al. ²⁹	United States	Cleveland Clinic	2010–2011	Adult and paediatric	ICD-9	Primary	Atlanta criteria	480	480	0.68	–	–	–	–
Wu et al. ³⁰	United States	Kaiser Permanente Southern California	2006–2012	Adult	ICD-9	Primary	Not reported	100	100	0.55	–	–	–	–
Shinagare et al. ³¹	United States	Brigham and Women's Hospital	2012– 2013	Adult	ICD-9	Primary or secondary	Atlanta criteria	115	115	0.89	–	–	–	–
Bertilsson et al. ³²	Sweden	Skane University Hospital	2003–2012	Adult	ICD-10	Primary or secondary	Atlanta criteria	2112	2112	0.87	–	–	–	–
Yang et al. ³³	United States	Mayo Clinic	2011–2013	Adult	ICD-9	Primary	Not reported	6096	273	1.00	0.94	0.44	1.00	Not calculable/0.56
<i>Chronic pancreatitis</i>														
Porta et al. ¹⁵	Spain	PANKRAS II Study	1992–1995	Adult	ICD-9	Primary	Non-standard	600	89	0.87	0.93	0.68	0.98	32.61/0.33

Table 1 continued

Author	Country	Database	Study period	Patient population	Coding system	Coding position	Reference standard	Total (n)	Cases coded(n)	PPV	NPV	Sensitivity	Specificity	Likelihood ratios positive/negative
Bagul et al. ³⁴	United Kingdom	Manchester Royal Infirmary	1993	Adult	ICD-9	Primary or secondary	Non-standard	45	45	0.91	–	–	–	–
Yadav et al. ¹⁸	United States	Veterans Outpatient Detoxification Programme	2002–2003	Adult	ICD-9	Primary or secondary	Ammann's criteria	15	15	0.07	–	–	–	–
Spanier et al. ²⁰	Netherlands	Academic Medical Centre	2002–2003	Adult	ICD-9	Primary or secondary	Non-standard	523	250	0.84	0.79	0.79	0.84	4.93/0.25
Joergensen et al. ³⁵	Denmark	Danish National Registry	1977– 2004	Adult	ICD-8, ICD-10	Primary or secondary	Mayo Clinic diagnostic scoring system	719	719	0.81	–	–	–	–
Nojgaard et al. ²¹	Denmark	Hvidovre Hospital Admissions	1983, 1994, 2005	Adult	ICD-8, ICD-10	Primary or secondary	Mayo Clinic diagnostic scoring system	185	185	0.72	–	–	–	–
Reddy et al. ³⁶	United States	University of Michigan Health Service Database	2005– 2008	Adult	ICD-9	Primary or secondary	Mayo Clinic diagnostic scoring system ; Ammann's criteria; Japanese Pancreas Society criteria	1343	1343	0.49	–	–	–	–

Table 2 QUADAS analysis of the acute pancreatitis cohorts

Study ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Eland et al. ¹⁴	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Porta et al. ¹⁵	Y	Y	U	Y	Y	Y	U	Y	N	Y	Y	Y	Y	Y
Chwistek et al. ¹⁶	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	N	Y	Y	Y
Floyd et al. ¹⁷	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Quraishi et al. ⁸	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Yadav et al. ¹⁸	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Kandula et al. ¹⁹	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Spanier et al. ²⁰	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Norjgaard et al. ²¹	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Dore et al. ²²	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Omdal et al. ²³	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Razavi et al. ²⁴	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Ma et al. ²⁵	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Edwards et al. ²⁶	Y	Y	N	Y	Y	Y	Y	N	Y	Y	N	Y	Y	Y
Shen et al. ²⁷	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Saligram et al. ²⁸	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Podugu et al. ²⁹	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Wu et al. ³⁰	Y	Y	U	Y	Y	Y	U	Y	N	Y	N	Y	Y	Y
Shinagare et al. ³¹	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Bertilsson et al. ³²	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Yang et al. ³³	Y	Y	U	Y	Y	Y	U	Y	N	Y	N	Y	Y	Y

Table 3 QUADAS analysis of the chronic pancreatitis cohorts

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Porta et al. ¹⁵	Y	Y	N	N	Y	Y	U	Y	Y	Y	Y	N	Y	Y
Bagul et al. ³⁴	Y	Y	U	N	Y	Y	U	Y	N	Y	N	N	Y	Y
Yadav et al. ¹⁸	N	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	N	Y	Y
Spanier et al. ²⁰	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	N	N	Y	Y
Joergensen et al. ³⁵	N	Y	Y	N	Y	Y	Y	Y	Y	Y	N	N	Y	Y
Norjgaard et al. ²¹	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	N	N	Y	Y
Reddy et al. ³⁶	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	N	N	Y	Y

was 0.69 (95% CI 0.55–0.81; $\text{Tau}^2 = 0.271$; $I^2 = 98.3\%$), whereas the one for ICD-10 was 0.79 (95% CI 0.69–0.88; $\text{Tau}^2 = 0.042$; $p = 0.189$; $I^2 = 95.7\%$), $p = 0.189$. The subgroup of adult patients only included 14,938 participants from 19 cohorts and yielded a PPV of 0.71 (95% CI

0.61–0.80; $\text{Tau}^2 = 0.207$; $I^2 = 98.4\%$), whereas the subgroup of paediatric patients only included 694 participants from three cohorts and yielded a PPV of 0.68 (95% CI 0.44–0.88; $\text{Tau}^2 = 0.173$; $I^2 = 95.5\%$), $p = 0.826$. The subgroup analysis according to definitions of AP showed that studies that used the Atlanta definition as the reference standard (4163 participants from seven cohorts) yielded a PPV of 0.79 (95% CI 0.67–0.88; $\text{Tau}^2 = 0.123$; $I^2 = 98.4\%$). The remaining 14 cohorts (10,725 participants) used a reference standard other than the Atlanta definition and yielded a PPV of 0.66 (95% CI 0.51–0.80; $\text{Tau}^2 = 0.337$; $I^2 = 98.5\%$). The subgroup analysis according to coding position included 9881 participants from seven cohorts with primary coding position and yielded the PPV of 0.75 (95% CI 0.59–0.88; $\text{Tau}^2 = 0.207$; $p = 0.596$; $I^2 = 98.3\%$). The PPV for primary or secondary coding position was 0.81 (95% CI 0.773–0.837; $\text{Tau}^2 = 0.003$; $p = 0.596$; $I^2 = 48.4\%$). The sensitivity analysis constrained to incident episode of AP only was based on five cohorts (1718 patients) and yielded a PPV of 0.78 (95% CI 0.70–0.85; $\text{Tau}^2 = 0.033$; $p = 0.209$; $I^2 = 87.1\%$).

Studies in chronic pancreatitis

A total of seven cohorts reported on the PPV of ICD codes for CP. The crude PPV was 0.71 (95% CI 0.54–0.85; $p < 0.0001$; $I^2 = 98.3\%$) (Fig. 5). The sensitivity analysis constrained to ICD-9 version only included five cohorts (2526 participants) and yielded a PPV of 0.67 (95% CI 0.42–0.89; $\text{Tau}^2 = 0.320$; $p = 0.301$; $I^2 = 98.2\%$). Only two cohorts (1123 participants) reported on sensitivity and specificity, which yielded pooled values of 0.75 (95% CI 0.71–0.80) and 0.94 (95% CI 0.93–0.96), respectively. There was an insufficient number of cohorts to perform other pre-specified analyses.

Discussion

This is the first systematic literature review and meta-analysis to report on pooled estimates of accuracy of the ICD codes for identifying patients with AP and CP. The pooled PPV for AP in the present study was 0.71. Systematic literature reviews on accuracy of ICD codes in other acute conditions found pooled estimates of PPV to be 0.82 in ischaemic stroke³⁷, 0.92 in myocardial infarction³⁸ and 0.93 in subarachnoid haemorrhage³⁷. Similarly, the pooled PPV for CP in the present study was 0.71. Systematic literature reviews on accuracy of ICD codes in other chronic conditions found pooled estimates of PPV to be 0.87 in heart failure³⁹ and 0.89 in depression⁴⁰. Taken together, the above findings suggest that accuracy of ICD codes in identifying patients with AP and CP is, in general, inferior to other acute and chronic conditions.

A series of pre-specified analyses showed that higher PPV of ICD codes for AP is reached when ICD-10, as opposed to ICD-9, is used; when the codes are applied to

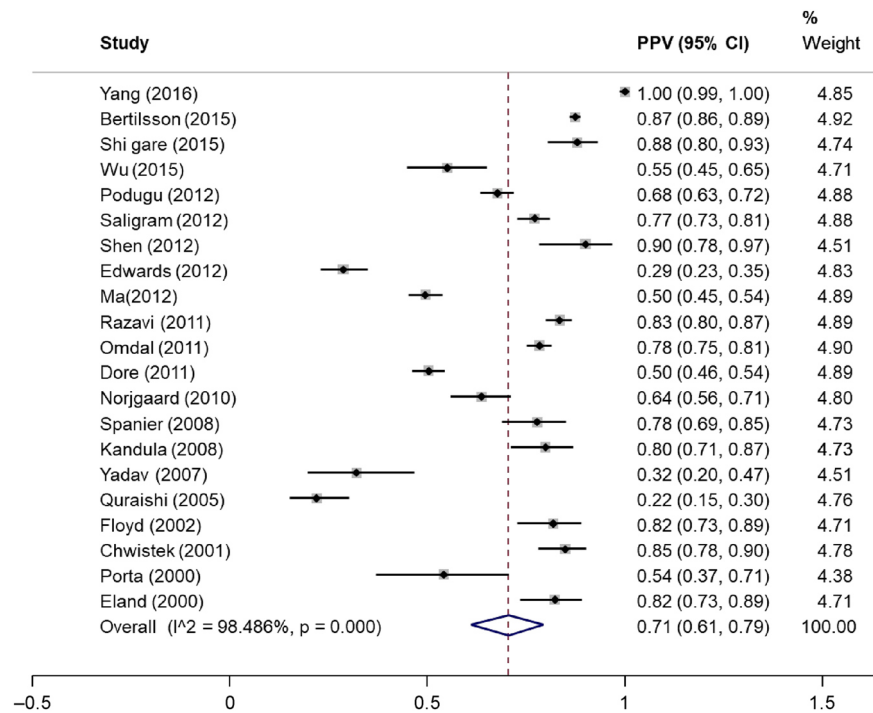


Fig. 2 Pooled positive predictive value of ICD codes in identifying patients with acute pancreatitis

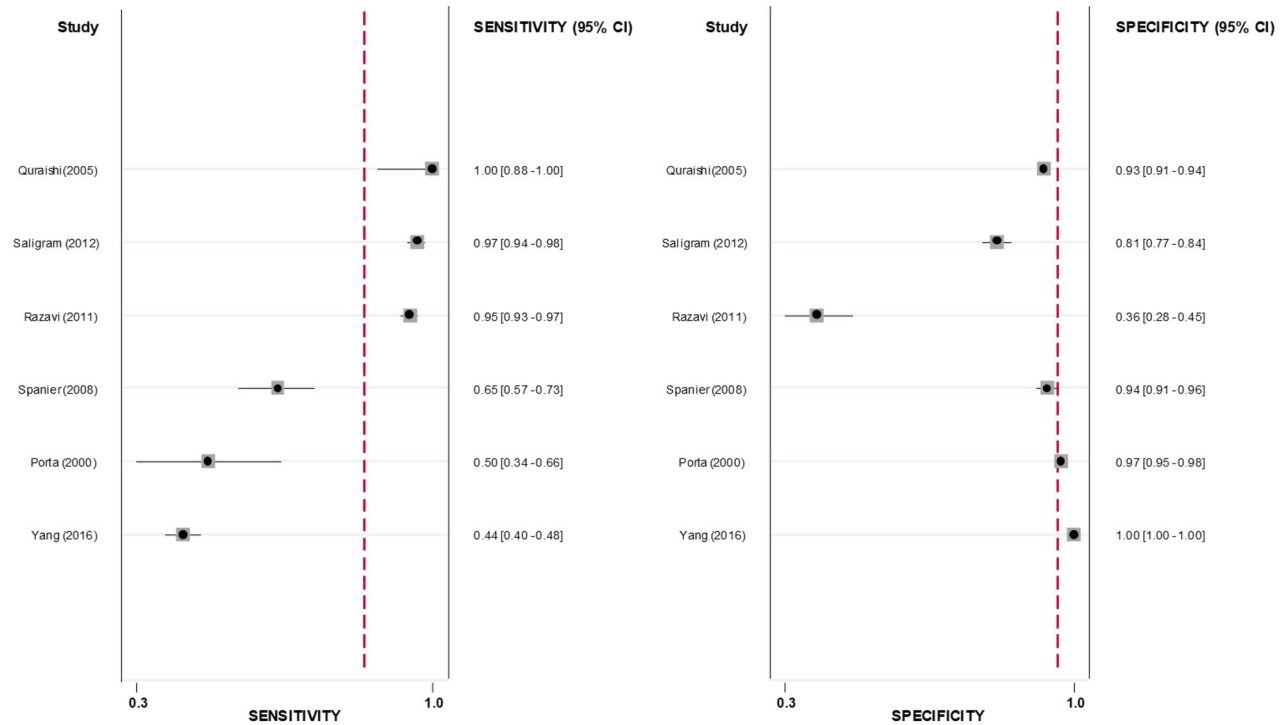


Fig. 3 Pooled sensitivity and specificity of ICD codes in identifying patients with acute pancreatitis

incident episode of AP as opposed to recurrent AP and when cases are validated with the use of Atlanta definition. Specifically, the subgroup analysis according to versions of ICD showed that ICD-10 codes yield a

10% higher PPV than that of ICD-9 codes, and this is likely a reflection of improvements in diagnostic methods⁴¹. ICD-10 also requires the input of aetiology of AP²⁴, which would require more confidence in the diagnosis. It

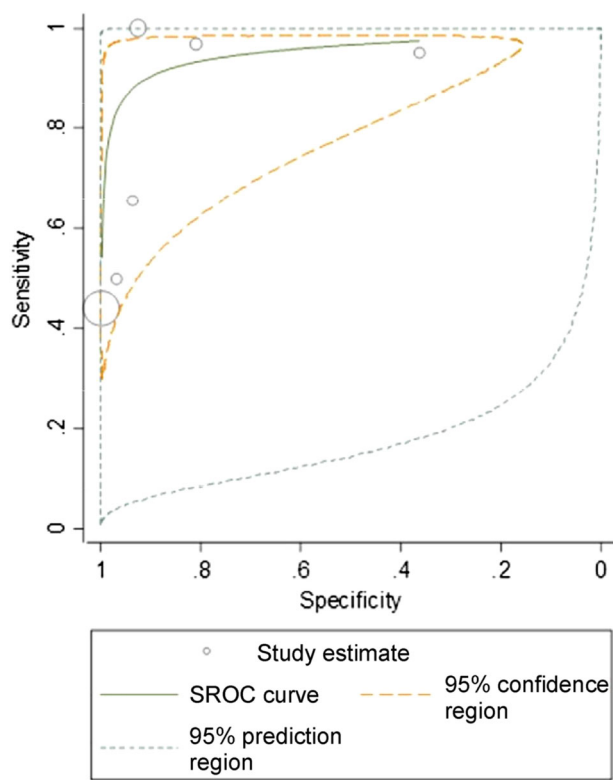


Fig. 4 Summary receiver operating characteristic (SROC) curve of sensitivity and specificity of ICD codes in identifying patients with acute pancreatitis

is assuring that ICD-10 is now the most commonly used version of ICD⁴², and improvement of PPV of the ICD codes for AP is expected in the future. The sensitivity analysis limited to cases of only incident episode of AP showed a 7% higher PPV in comparison with the overall AP cohort. This suggests that misdiagnosis may occur when re-admitted patients with previous pancreatitis are assumed to have another episode of pancreatitis⁴³. Analysis of cases validated with the use of Atlanta definition yielded the PPV of 0.79. Although this is an improvement in comparison with the overall estimate, it is worrying that 21% cases are diagnosed with AP when, in fact, they do not have it. The other noteworthy finding is that the PPV of diagnostic codes is lower in children, with a PPV of just 0.68. Of note, our study did not find PPV of AP to be improved in the subgroup analysis of primary coding position (0.75) in comparison to primary or secondary coding position (0.81). The value of PPV in primary or secondary coding positions may be higher than that of primary coding position alone because the diagnosis of AP was more confidently made when in conjunction with another related diagnosis, such as cholelithiasis.

Given the generally moderate PPV values of ICD codes for AP and CP, the main clinical implication of the present study is that overdiagnosing of pancreatitis is frequent. Patients with a previous history of AP are likely to be re-admitted with the coding of an episode of AP again⁴⁴. This episode may be a continuation of a previous

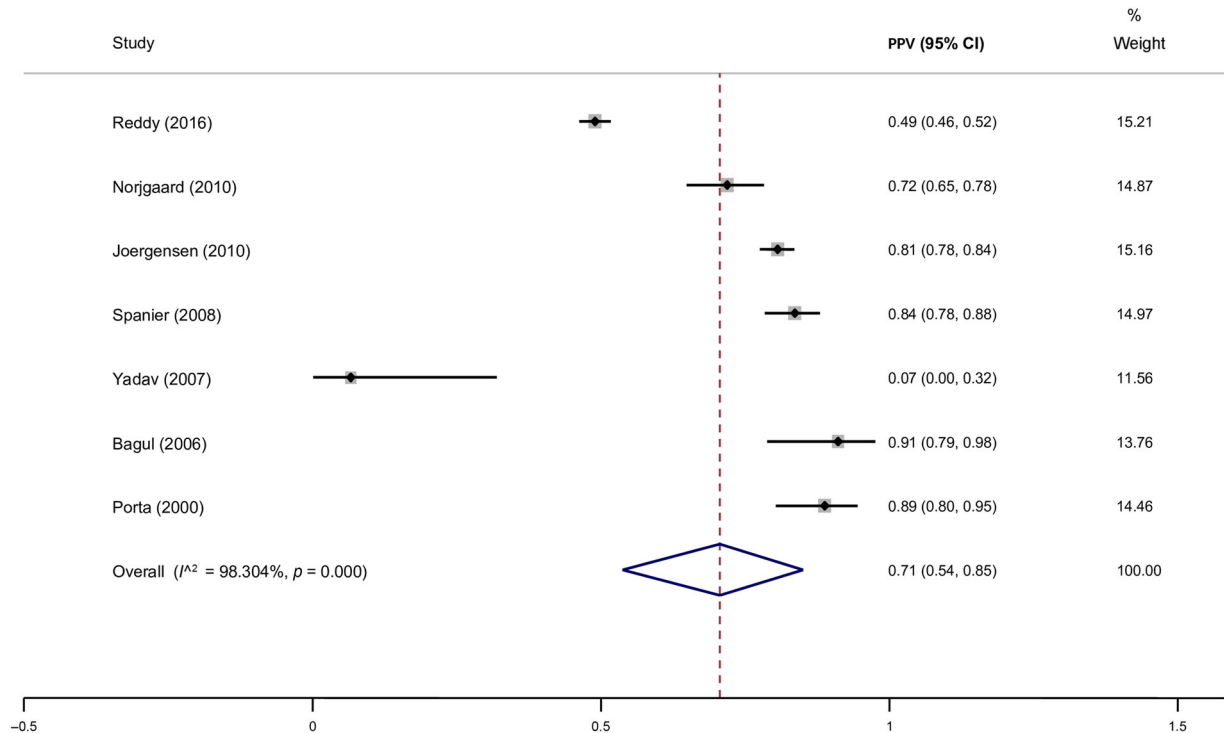


Fig. 5 Pooled positive predictive value of ICD codes in identifying patients with chronic pancreatitis

inadequately treated episode or it could be a different pathology at all⁴³. One previous code of AP predisposes a patient to more likely receive future pancreatitis diagnostic codes²⁸. Advances in serum testing have allowed detection of more mild cases of AP, but has also led to more overdiagnosing⁴⁵. The diagnosis of early CP remains a significant challenge. One component of the diagnostic criteria for CP is histology, which is often unavailable at the time of coding⁴⁶. The diagnosis, thus, becomes predominantly based on imaging modalities⁴⁶.

The main immediate implication for research is that a correction factor may need to be employed to estimate accurately the real burden of pancreatitis in the studies that used ICD codes. Leong and colleagues suggested a formula that uses specificity and sensitivity to give a corrected prevalence⁴⁷. This formula may be useful for correcting the prevalence of CP rather than AP. Ley and colleagues, as well as Esposito and colleagues, proposed the use of PPV itself as a correction factor for incidence and this would be more appropriate for AP^{48,49}. While development of more accurate diagnostic codes is anticipated in the future, the pooled PPV value of 0.71 in the present study can be used to derive corrected incidence of AP in the existing literature. There are also other ways to improve on accuracy of epidemiological estimates in the field of Pancreatology. Participants can be recruited in future studies by searching for the unique patient rather than for the episode²⁸. Exclusion of patients with a previous pancreatitis diagnosis can increase the PPV as these cases tend to have a higher chance of a misdiagnosed readmission²⁸. The requirement of elevated pancreatic enzyme levels above a three-time threshold, as suggested by current guidelines, may further increase the accuracy of ICD codes²⁸.

The limitations of the present study need to be acknowledged. First, the included studies came from different countries and from hospitals of various size, which may have contributed to heterogeneity. Second, the validation criteria used in the primary studies were not standardised. Third, PPV as a measure of diagnostic accuracy is affected by disease prevalence⁵⁰. Given that CP is a much less common disease than AP, PPV for CP may have been low due to its relatively low prevalence⁵. Last, inclusion of primary studies was restricted to English, and this may have led to a language bias.

In conclusion, the overall diagnostic accuracy of ICD codes for pancreatitis is suboptimal. It is higher when the codes are applied to incident episode of AP and to adults, as well as when ICD-10 is used. The correction factor of 0.71 can be used to estimate accurately the burden of AP in studies using administrative databases. In the future, new diagnostic criteria may need to be developed for patients with recurrent AP and CP.

Acknowledgements

This study was part of the Clinical and epidemiological investigations in Metabolism, nutrition, and pancreatic diseases (COSMOS) program.

Author details

¹Department of Surgery, University of Auckland, Auckland, New Zealand. ²CIBER Epidemiology and Public Health (CIBERESP), Universidad Francisco de Vitoria, Madrid, Spain. ³Division of Gastroenterology, Hepatology & Nutrition, University of Pittsburgh Medical Center, Pittsburgh, USA. ⁴Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK

Competing interests

Guarantor of the article: Maxim S. Petrov.

Specific author contributions: Data collection: Amy Y. Xiao and Marianne L. Tan; analysis of the data: Amy Y. Xiao, Maria N. Plana and Javier Zamora; drafting the paper: Amy Y. Xiao; critical reviewing of the paper for intellectual content: Maria N. Plana, Dhiraj Yadav, Javier Zamora and Maxim S. Petrov; study supervision: Maxim S. Petrov.

Financial support: COSMOS is supported in part by the Royal Society of New Zealand (Rutherford Discovery Fellowship to Associate Professor Petrov), which played no role in the study design, collection, analysis or interpretation of data or writing of the paper.

Potential competing interests: None.

Received: 19 August 2018 Accepted: 6 September 2018

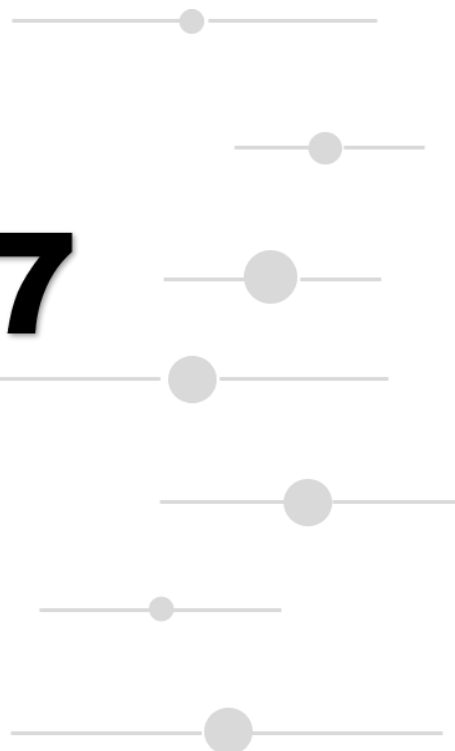
Published online: 04 October 2018

References

1. Peek, N., Holmes, J. H. & Sun, J. Technical challenges for big data in biomedicine and health: data sources, infrastructure, and analytics. *Yearb. Med. Inform.* **9**, 42–47 (2014).
2. Belle, A. et al. Big data analytics in healthcare. *BioMed. Res. Int.* **2015**, 370194 (2015).
3. Greene, C. S., Tan, J., Ung, M., Moore, J. H. & Cheng, C. Big data bioinformatics. *J. Cell. Physiol.* **229**, 1896–1900 (2014).
4. Genta, R. M. & Sonnenberg, A. Big data in gastroenterology research. *Nat. Rev. Gastroenterol. & Hepatol.* **11**, 386–390 (2014).
5. Xiao, A. Y. et al. Global incidence and mortality of pancreatic diseases: a systematic review, meta-analysis, and meta-regression of population-based cohort studies. *Lancet Gastroenterol. & Hepatol.* **1**, 45–55 (2016).
6. Sankaran, S. J. et al. Frequency of progression from acute to chronic pancreatitis and risk factors: a meta-analysis. *Gastroenterology* **149**, 1490–1500 (2015).
7. Petrov, M. S., Windsor, J. A. & Lévy, P. New international classification of acute pancreatitis: more than just 4 categories of severity. *Pancreas* **42**, 389–391 (2013).
8. Quraishi, E. R. et al. Acute pancreatitis in patients on chronic peritoneal dialysis: an increased risk? *Am. J. Gastroenterol.* **100**, 2288–2293 (2005).
9. Petrov, M. S. Therapeutic implications of oxidative stress in acute and chronic pancreatitis. *Curr. Opin. Clin. Nutr. Metab. Care* **13**, 562–568 (2010).
10. Bae, J.-M. Overdiagnosis: epidemiologic concepts and estimation. *Epidemiol. Health* **37**, e2015004 (2015).
11. Whiting, P., Rutjes, A. W., Reitsma, J. B., Bossuyt, P. M. & Kleijnen, J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *Bmc. Med. Res. Methodol.* **3**, 25 (2003).
12. Reitsma, J. B. et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J. Clin. Epidemiol.* **58**, 982–990 (2005).
13. StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP. <https://www.stata.com/support/faqs/resources/citing-software-documentation-faqs/>
14. Eland, I. A., Sturkenboom, M. J., Wilson, J. H. & Stricker, B. H. Incidence and mortality of acute pancreatitis between 1985 and 1995. *Scand. J. Gastroenterol.* **35**, 1110–1116 (2000).

15. Porta, M. et al. Validity of the hospital discharge diagnosis in epidemiologic studies of biliopancreatic pathology. PANKRAS II Study Group. *Eur. J. Epidemiol.* **16**, 533–541 (2000).
16. Chwistek, M., Roberts, I. & Amoateng-Adjepong, Y. Gallstone pancreatitis: a community teaching hospital experience. *J. Clin. Gastroenterol.* **33**, 41–44 (2001).
17. Floyd, A., Pedersen, L., Nielsen, G. L., Thorlacius-Ussing, O. & Sorensen, H. T. Risk of acute pancreatitis in users of azathioprine: a population-based case-control study. *Am. J. Gastroenterol.* **98**, 1305–1308 (2003).
18. Yadav, D., Eigenbrodt, M. L., Briggs, M. J., Williams, D. K. & Wiseman, E. J. Pancreatitis: prevalence and risk factors among male veterans in a detoxification program. *Pancreas* **34**, 390–398 (2007).
19. Kandula, L. & Lowe, M. E. Etiology and outcome of acute pancreatitis in infants and toddlers. *J. Pediatr.* **152**, 106–110, 10.e1 (2008).
20. Spanier, B. W., Schreuder, D., Dijkgraaf, M. G. & Bruno, M. J. Source validation of pancreatitis-related hospital discharge diagnoses notified to a national registry in the Netherlands. *Pancreatol.* **8**, 498–503 (2008).
21. Nojgaard, C., Bendtsen, F., Matzen, P. & Becker, U. The aetiology of acute and chronic pancreatitis over time in a hospital in Copenhagen. *Dan. Med. Bull.* **57**, A4103 (2010).
22. Dore, D. D., Chaudhry, S., Hoffman, C. & Seeger, J. D. Stratum-specific positive predictive values of claims for acute pancreatitis among commercial health insurance plan enrollees with diabetes mellitus. *Pharmacoeconom. Drug. Saf.* **20**, 209–213 (2011).
23. Omdal, T. et al. Time trends in incidence, etiology, and case fatality rate of the first attack of acute pancreatitis. *Scand. J. Gastroenterol.* **46**, 1389–1398 (2011).
24. Razavi, D., Ljung, R., Lu, Y., Andren-Sandberg, A. & Lindblad, M. Reliability of acute pancreatitis diagnosis coding in a national patient register: a validation study in Sweden. *Pancreatol.* **11**, 525–532 (2011).
25. Ma, M. H. et al. Risk factors associated with biliary pancreatitis in children. *J. Pediatr. Gastroenterol. Nutr.* **54**, 651–656 (2012).
26. Edwards, M. A., Graham, G. & Smith, J. *Can We Reliably Predict Normal Acid Base Status in Acute Pancreatitis Without An Arterial Blood Gas?* (Official Journal of the Society for Academic Emergency Medicine, Dublin, Ireland, 2012; 747 14th International Conference on Emergency Medicine; 27 June 2012).
27. Shen, H. N., Lu, C. L. & Li, C. Y. The effect of gastrointestinal bleeding on outcomes of patients with acute pancreatitis: a national population-based study. *Pancreatol.* **12**, 331–336 (2012).
28. Saligram, S., Lo, D., Saul, M. & Yadav, D. Analyses of hospital administrative data that use diagnosis codes overestimate the cases of acute pancreatitis. *Clin. Gastroenterol. Hepatol.: Off. Clin. Pract. J. Am. Gastroenterol. Assoc.* **10**, 805–11.e1 (2012).
29. Podugu, A. et al. (AP). *Positive Predictive Value (PPV) of ICD-9 Discharge Diagnosis of Acute Pancreatitis* (pp. 5104. The American Journal of Gastroenterology, Las Vegas, 2012). 77th Annual Scientific Meeting of the American College of Gastroenterology; 19 Oct 2012.
30. Wu, B. U., Pandol, S. J. & Liu, I. L. Simvastatin is associated with reduced risk of acute pancreatitis: findings from a regional integrated healthcare system. *Gut* **64**, 133–138 (2015).
31. Shinagare, A. B. et al. Use of CT and MRI in emergency department patients with acute pancreatitis. *Abdom. Imaging* **40**, 272–277 (2015).
32. Bertilsson, S. & Kalaitzakis, E. Acute pancreatitis and use of pancreatitis-associated drugs: a 10-year population-based cohort study. *Pancreas* **44**, 1096–1104 (2015).
33. Yang, A. L., Ondrejko, A. & Vege, S. S. *Accuracy and limitations of ICD-9 codes in the identification of patients with acute pancreatitis*. (W.B Saunders, San Diego, 2016; S499 Gastroenterology Conference: Digestive Disease Week 2016).
34. Bagul, A. & Siriwardena, A. K. Evaluation of the Manchester classification system for chronic pancreatitis. *JOP: J. Pancreas* **7**, 390–396 (2006).
35. Joergensen, M., Brusgaard, K., Cruger, D. G., Gerdes, A. M. & de Muckadell, O. B. Incidence, prevalence, etiology, and prognosis of first-time chronic pancreatitis in young patients: a nationwide cohort study. *Dig. Dis. Sci.* **55**, 2988–2998 (2010).
36. Reddy, N. G., Nangia, S. & DiMaggio, M. J. The chronic pancreatitis international classification of diseases, ninth revision, clinical modification code 577.1 is inaccurate compared with criterion-standard clinical diagnostic scoring systems. *Pancreas* **45**, 1276–1281 (2016).
37. McCormick, N., Bhole, V., Lacaille, D. & Avina-Zubieta, J. A. Validity of diagnostic codes for acute stroke in administrative databases: a systematic review. *PLoS. One.* **10**, e0135834 (2015).
38. McCormick, N., Lacaille, D., Bhole, V. & Avina-Zubieta, J. A. Validity of myocardial infarction diagnoses in administrative databases: a systematic review. *PLoS. One.* **9**, e92286 (2014).
39. McCormick, N., Lacaille, D., Bhole, V. & Avina-Zubieta, J. A. Validity of heart failure diagnoses in administrative databases: a systematic review and meta-analysis. *PLoS. One.* **9**, e104519 (2014).
40. Fiest, K. M., Jette, N., Quan, H., St. Germaine-Smith, C. & Metcalfe, A. et al. Systematic review and assessment of validated case definitions for depression in administrative data. *BMC Psychiatry* **14**, 289 (2014).
41. Pendharkar, S. A., Mathew, J. & Petrov, M. S. Age- and sex-specific prevalence of diabetes associated with diseases of the exocrine pancreas: a population-based study. *Dig. Liver. Dis.* **49**, 540–544 (2017).
42. Fung, K. W. et al. Preparing for the ICD-10-CM transition: automated methods for translating ICD Codes in clinical phenotype definitions. *EGEMS* **4**, 1211 (2016).
43. Stigliano, S., Sternby, H., de Madaria, E., Capurso, G. & Petrov, M. S. Early management of acute pancreatitis: a review of the best evidence. *Dig. Liver. Dis.* **49**, 585–594 (2017).
44. Lindkvist, B., Appelros, S., Manjer, J. & Borgstrom, A. Trends in incidence of acute pancreatitis in a Swedish population: is there really an increase? *Clin. Gastroenterol. Hepatol.* **2**, 831–837 (2004).
45. DeSouza, S. V. et al. Pancreas volume in health and disease: a systematic review and meta-analysis. *Expert Rev. Gastroenterol. Hepatol.* **12**, 757–766 (2018).
46. Lévy, P., Domínguez-Muñoz, E., Imrie, C., Löhr, M. & Maisonneuve, P. Epidemiology of chronic pancreatitis: burden of the disease and consequences. *United European Gastroenterology J.* **2**, 345–354 (2014).
47. Leong, A. et al. Systematic review and meta-analysis of validation studies on a diabetes case definition from health administrative records. *PLoS. One.* **8**, e75256 (2013).
48. Ley, B. et al. Code-based diagnostic algorithms for idiopathic pulmonary fibrosis. Case validation and improvement. *Ann. Am. Thorac. Soc.* **14**, 880–887 (2017).
49. Esposito, D. B. et al. Idiopathic pulmonary fibrosis in United States automated claims. Incidence, prevalence, and algorithm validation. *Am. J. Respir. Crit. Care. Med.* **192**, 1200–1207 (2015).
50. Parikh, R. et al. Understanding and using sensitivity, specificity and predictive values. *Indian J. Ophthalmol.* **56**, 45–50 (2008).

Capítulo 7



Capítulo 7

En este capítulo se presenta el estudio de revisión sistemática realizado con el objetivo de evaluar la exactitud diagnóstica de la determinación de DNACf (del inglés cell-free, ADN fetal libre) circulante en sangre materna para la detección prenatal de trisomías en los cromosomas 21 (Síndrome de Down), 18 (Síndrome de Edwards) y 13 (Síndrome de Patau) (T21, T18 y T13) en embarazo gemelar.

El diagnóstico prenatal permite en ocasiones detectar si el feto sufre de algún defecto congénito. Uno de los diagnósticos disponibles para la mujer embarazada es el cribado combinado del primer trimestre para las trisomías autosómicas más frecuentes que combina edad materna con determinaciones ecográficas fetales (translucencia nucal y/o otras) y niveles de hormonas placentarias maternas (PAPP-A o proteína plasmática

asociada al embarazo y la Beta-hCG). En función de los niveles de riesgo del cribado actualmente se ofrece una prueba invasiva diagnóstica (amniocentesis o biopsia corial) o la determinación de ADN fetal en embarazo único. El análisis de fragmentos de ADN fetal presentes en el plasma materno es una prueba no invasiva que sólo requiere de una extracción de sangre de la madre. Esta prueba se puede realizar a partir de la semana 10 de embarazo para detectar trisomías en el feto.

En embarazo único, la determinación del ADN fetal libre en sangre materna ha demostrado ser efectivo en el cribado de trisomías autosómicas (T21, T18 y T13). En un metanálisis previo publicado en 2017 se obtuvo una tasa de detección o sensibilidad promedio del 99.7% (IC 95% 99.1; 99.9) y una tasa de falsos positivos (TFP) de 0.04% (IC 95% 0.02; 0.07) en la detección de T21, sobre un total de 1,963 casos de T21 y 223,932 embarazos sin T21. La sensibilidad y TFP del ADN fetal para el diagnóstico de T18 fue de un 97.9% (IC 95% 94.9; 99.1) y del 0.04% (IC 95% 0.03; 0.07) respectivamente, sobre un total de 563 casos de T18 y 222,013 embarazos no afectados de T18. La sensibilidad y TFP del ADN fetal para el diagnóstico de T13 fue de un 99.0% (IC 95% 65.8; 100) y del 0.04% (IC 95% 0.02; 0.07) respectivamente, sobre un total de 119 casos de T13 y 212,883 embarazos de fetos no afectados de T13. Sin embargo, la evidencia en embarazo general es más limitada (Gil et al. 2017).

7.1 Objetivos

Evaluar la exactitud diagnóstica de la determinación de ADN fetal libre en sangre materna para la detección prenatal de T21, T18 y T13 en embarazo gemelar.

7.2 Metodología

El estudio incluye un estudio primario prospectivo de exactitud diagnóstica y una revisión sistemática en la que se incluyeron los resultados de dicho estudio. Los datos para el estudio prospectivo se obtuvieron a partir del cribado de T21, T18 y T13 en embarazo gemelar entre las semanas 10+0 a 14+1 de gestación. Se incluyeron dos poblaciones: la primera, de mujeres que acudieron sin derivación del especialista al Centro de Medicina Fetal de Londres o al Hospital Universitario Brugmann de Bruselas y, la segunda, las mujeres seleccionadas para la prueba de ADN fetal después de la prueba combinada rutinaria del primer trimestre en uno de los dos hospitales del Servicio Nacional de Salud de Reino Unido.

Se realizó una búsqueda en las bases de datos de MEDLINE, EMBASE, CENTRAL y en ClinicalTrials.gov y la Plataforma Internacional de Registro de Ensayos Clínicos (ICTRP) de la Organización Mundial de la Salud para identificar todas las publicaciones revisadas por pares sobre la validación clínica o la implementación de las pruebas de ADN fetal para el cribado de T21, T18 y T13 en el embarazo gemelar. Se evaluó el riesgo de sesgo en los estudios incluidos mediante la herramienta QUADAS-2 (Whiting et al. 2011). Se utilizó el modelo bivariante para la síntesis de resultados y las varianzas de los logit de sensibilidad y especificidad como medidas de heterogeneidad.

7.3 Resumen de resultados

El estudio primario incluyó datos de 997 embarazos gemelares. La prueba de ADN fetal clasificó correctamente 16 (94.1%) de los 17 casos de T21, 9 (90.0%) de los 10

casos de T18, uno (50.0%) de los dos casos de T13 y 962 (99.4%) de los 968 casos sin ninguna de las tres trisomías.

La búsqueda bibliográfica identificó 7 estudios adicionales para su inclusión. En los 8 estudios que evaluaban el cribado de T21, se registraron 56 embarazos gemelares con diagnóstico de T21 y 3,718 embarazos sin T21; la sensibilidad y la TFP promedio del ADN fetal fue de 98.2% (IC 95% 83.2; 99.8) y de 0.05% (IC 95%, 0.01; 0.26), respectivamente. En los 5 estudios que evaluaban el screening de T18, se registraron 18 casos con T18 y 3,143 embarazos sin T18, la sensibilidad y la TFP promedio del ADN fetal fue de 88.9% (IC 95% 64.8; 97.2) y 0.03% (IC 95% 0.00; 0.33), respectivamente. En el caso de la T13 (n=3 estudios), sólo hubo tres casos afectados y dos (66.7) de ellos fueron detectados por la prueba de ADN fetal con una TFP del 0.19% (5/2,569).

7.4 Conclusiones

El rendimiento de la prueba de ADN fetal libre para la trisomía 21 en el embarazo gemelar es similar al observado en el embarazo único y es superior al de la prueba combinada del primer trimestre o a la prueba bioquímica del segundo trimestre mostrando mayor sensibilidad y menor tasa de falsos positivos. El número de casos de trisomías 18 y 13 es demasiado pequeño para evaluar con precisión el rendimiento diagnóstico de la prueba de ADN fetal.

7.5 Publicación

Se incluye como cuerpo del capítulo 7 el texto completo del estudio de revisión sistemática publicado en la revista *Ultrasound in Obstetrics & Gynecology* (Factor de

impacto año 2019: 5.571; rango: Q1-primer decil, 5/82 de la categoría Obstetrics and Gynecology):

Gil MM, Galeva S, Jani J, Konstantinidou L, Akolekar R, Plana MN, Nicolaides KH. Screening for trisomies by cfDNA testing of maternal blood in twin pregnancy: update of The Fetal Medicine Foundation results and meta-analysis. *Ultrasound Obstet Gynecol.* 2019 Jun;53(6):734-742.

El número de citas de esta publicación (07/07/2021) es de 41 citas en la plataforma Scopus y 61 en Google Scholar.



Screening for trisomies by cfDNA testing of maternal blood in twin pregnancy: update of The Fetal Medicine Foundation results and meta-analysis

M. M. GIL^{1,2,3} , S. GALEVA^{1,4}, J. JANI⁵ , L. KONSTANTINIDOU¹, R. AKOLEKAR^{4,6} ,
M. N. PLANA^{7,8} and K. H. NICOLAIDES¹

¹Fetal Medicine Research Institute, King's College Hospital, London, UK; ²Obstetrics and Gynecology Department, Hospital Universitario de Torrejón, Torrejón de Ardoz, Madrid, Spain; ³School of Health Sciences, Universidad Francisco de Vitoria, Pozuelo de Alarcón, Madrid, Spain; ⁴Department of Fetal Medicine, Medway Maritime Hospital, Gillingham, UK; ⁵Department of Obstetrics and Gynecology, University Hospital Brugmann, Université Libre de Bruxelles, Brussels, Belgium; ⁶Institute of Medical Sciences, Canterbury Christ Church University, Chatham, UK; ⁷Department of Preventive Medicine and Public Health, Hospital Príncipe de Asturias, Alcalá de Henares, Madrid, Spain; ⁸CIBER Epidemiology and Public Health (CIBERESP), Madrid, Spain

KEYWORDS: cell-free DNA; first-trimester screening; meta-analysis; non-invasive prenatal testing; trisomy 21; trisomy 18; trisomy 13; twin pregnancy

ABSTRACT

Objectives To report on the routine clinical implementation of cell-free DNA (cfDNA) analysis of maternal blood for trisomies 21, 18 and 13 in twin pregnancy and to define the performance of the test by combining our results with those identified in a systematic review of the literature.

Methods The data for the prospective study were derived from screening for trisomies 21, 18 and 13 in twin pregnancies at 10+0 to 14+1 weeks' gestation. Two populations were included; first, self-referred women to the Fetal Medicine Centre in London or Brugmann University Hospital in Brussels and, second, women selected for the cfDNA test after routine first-trimester combined testing at one of two National Health Service hospitals in England. This dataset was used to determine the performance of screening for the three trisomies. Search of MEDLINE, EMBASE, CENTRAL (The Cochrane Library), ClinicalTrials.gov and the World Health Organization International Clinical Trials Registry Platform (ICTRP) was carried out to identify all peer-reviewed publications on clinical validation or implementation of maternal cfDNA testing for trisomies 21, 18 and 13 in twin pregnancy. A meta-analysis was then performed using our data and those in the studies identified by the literature search.

Results In our dataset of 997 twin pregnancies with a cfDNA result and known outcome, the test classified correctly 16 (94.1%) of the 17 cases of trisomy 21, nine

(90.0%) of the 10 cases of trisomy 18, one (50.0%) of the two cases of trisomy 13 and 962 (99.4%) of the 968 cases without any of the three trisomies. The literature search identified seven relevant studies, excluding our previous papers because their data are included in the current study. In the combined populations of our study and the seven studies identified by the literature search, there were 56 trisomy-21 and 3718 non-trisomy-21 twin pregnancies; the pooled weighted detection rate (DR) and false-positive rate (FPR) were 98.2% (95% CI, 83.2–99.8%) and 0.05% (95% CI, 0.01–0.26%), respectively. In the combined total of 18 cases of trisomy 18 and 3143 non-trisomy-18 pregnancies, the pooled weighted DR and FPR were 88.9% (95% CI, 64.8–97.2%) and 0.03% (95% CI, 0.00–0.33%), respectively. For trisomy 13, there were only three affected cases and two (66.7%) of these were detected by the cfDNA test at a FPR of 0.19% (5/2569).

Conclusions The performance of cfDNA testing for trisomy 21 in twin pregnancy is similar to that reported in singleton pregnancy and is superior to that of the first-trimester combined test or second-trimester biochemical testing. The number of cases of trisomies 18 and 13 is too small for accurate assessment of the predictive performance of the cfDNA test. Copyright © 2019 ISUOG. Published by John Wiley & Sons Ltd.

INTRODUCTION

In singleton pregnancies, cell-free DNA (cfDNA) analysis of maternal blood provides effective screening

Correspondence to: Prof. K. H. Nicolaides, Fetal Medicine Research Institute, King's College Hospital, 16–20 Windsor Walk, Denmark Hill, London SE5 8BB, UK (e-mail: kypros@fetalmedicine.com)

Accepted: 14 March 2019

for trisomies 21, 18 and 13. A recent meta-analysis of clinical validation studies reported that, in the combined total of 1963 cases of trisomy 21 and 223 932 non-trisomy-21 singleton pregnancies, the pooled weighted detection rate (DR) was 99.7% (95% CI, 99.1–99.9%) and the false-positive rate (FPR) was 0.04% (95% CI, 0.02–0.07%); in a total of 563 cases of trisomy 18 and 222 013 unaffected pregnancies, the pooled weighted DR and FPR were 97.9% (95% CI, 94.9–99.1%) and 0.04% (95% CI, 0.03–0.07%), respectively, and, in a total of 119 cases of trisomy 13 and 212 883 unaffected singleton pregnancies, the pooled weighted DR and FPR were 99.0% (95% CI, 65.8–100%) and 0.04% (95% CI, 0.02–0.07%), respectively¹. In contrast to singleton pregnancies, data on cfDNA testing in twins are very limited; the meta-analysis reported that only five studies had examined twin pregnancies prospectively and, in a total of 24 cases of trisomy 21 and 1111 non-trisomy-21 cases, the DR was 100% and FPR was 0%¹.

In previous studies, we reported our data on cfDNA testing for trisomies in twins. In the first study, which included stored and prospectively collected samples, the cfDNA test classified correctly 11 of the 12 cases of trisomy 21, the one case of trisomy 18, the one case of trisomy 13 and all 241 non-trisomic pregnancies². In the second study, we reported the results of prospective screening in twin pregnancies; the cfDNA test classified correctly 11 of the 12 cases of trisomy 21, all five cases of trisomy 18 and all 334 non-trisomic pregnancies³. In the third study, we reported performance of cfDNA testing in 417 prospectively examined pregnancies; the cfDNA test identified correctly all eight cases of trisomy 21, three of the four cases of trisomy 18, not the single case of trisomy 13 and 403 (99.8%) of the 404 non-trisomic pregnancies⁴. In the fourth study, we performed cfDNA testing in pregnancies identified by the combined test as being at intermediate or high risk for trisomy; the cfDNA test identified correctly all six cases of trisomy 21, two of the three cases of trisomy 18 and all 206 non-trisomic pregnancies⁵.

The objectives of this study are, first, to report our updated experience on prospective first-trimester screening for trisomies 21, 18 and 13 in twins by cfDNA testing and, second, to carry out a meta-analysis of all studies on cfDNA testing in twin pregnancies published up to 9 March 2019.

METHODS

Update of The Fetal Medicine Foundation results

Study design and participants

The data for this study were derived from prospective screening for trisomies 21, 18 and 13 in twin pregnancies at 10 + 0 to 14 + 1 weeks' gestation. Two populations were included; first, self-referred women to the Fetal Medicine Centre in London, which is a private clinic⁶, or the Brugmann University Hospital in Brussels, which

is a public hospital, and, second, women selected for the cfDNA test after routine first-trimester combined testing in one of two National Health Service hospitals in England^{5,7}. The patients were examined between October 2012 and January 2018. The study was approved by the appropriate ethics committees (NREC reference 13/LO/0885, NREC reference 19/HRA/0576, CE 2014/5).

We recorded maternal characteristics and medical history, including maternal age and racial origin (white, black, South Asian, East Asian or mixed), method of conception (natural or assisted conception requiring the use of ovulation drugs or by *in-vitro* fertilization), cigarette smoking during pregnancy (yes or no) and parity (parous or nulliparous if no previous pregnancy ≥ 24 weeks' gestation). An ultrasound scan was carried out to determine gestational age from the measurement of crown–rump length⁸ of the larger fetus and chorionicity by examining the junction of the intertwin membrane with the placenta⁹.

Women provided written informed consent and maternal blood (20 mL) was collected into either Cell-Free DNA BCT® tubes (Streck, Omaha, NE, USA) or Roche Cell-Free DNA Collection Tubes (Roche, Pleasanton, CA, USA). These were shipped via courier to Ariosa Diagnostics, Inc. (San Jose, CA, USA) where they were processed within 7 days after collection. Targeted cfDNA testing for fetal trisomy was performed using the Harmony™ prenatal test, as described previously^{10–13}. Harmony uses Digital ANalysis of Selected Regions (DANSR) assays targeting sequences on chromosomes 13, 18 and 21 for chromosome quantitation and single-nucleotide polymorphisms on chromosomes 1–12 for fetal-fraction measurement. Products of the DANSR assays can be quantified using either next-generation sequencing or a custom microarray; both were used during the course of this study. The data were analyzed using the fetal fraction-optimized risk of trisomy evaluation (FORTE) algorithm, which calculates probability scores for fetal trisomy, with $> 1\%$ considered to be high probability. In cases in which the cfDNA test did not provide a result, the parents were offered repeat testing or to rely on the results of the combined test in deciding whether to have an invasive test. In cases with a high-risk result on the cfDNA test, the parents were advised to consider having invasive fetal karyotyping before deciding on the further management of their pregnancy.

Patient characteristics, results of the investigations and pregnancy outcome were recorded in a database. The outcomes were divided into, first, trisomy 21, 18 or 13 if the karyotype on analysis of chorionic villi, amniotic fluid or neonatal blood demonstrated the relevant trisomy in one or both fetuses, second, no trisomy 21, 18 or 13 if the karyotype was normal or both neonates were phenotypically normal, third, no known karyotype in both fetuses because the pregnancy resulted in termination, embryo reduction, miscarriage or stillbirth and no karyotyping of fetal tissue was carried out, and, fourth, outcome unknown because the pregnancy was lost to follow-up.

Statistical analysis

Descriptive data are presented as median (interquartile range (IQR)) for continuous variables and as n (%) for categorical variables.

Systematic review and meta-analysis

Literature search and study selection

Searches of MEDLINE via PubMed, EMBASE and CENTRAL (The Cochrane Library) were performed to identify clinical validation or implementation studies on maternal cfDNA testing in screening for aneuploidy in twin pregnancy; additionally, ClinicalTrials.gov and the World Health Organization International Clinical Trials Registry Platform (ICTRP) were searched for ongoing or recently completed trials. The study period was from January 2011, when the first such study was published, to 9 March 2019; the initial search was performed on 8 December 2018 and this was updated with autoalerts in MEDLINE. A list of relevant citations was generated from these databases using the search strategies detailed in Appendix S1. This review was registered in PROSPERO international database for systematic reviews (reference: CRD42019121506).

The abstracts of citations were examined by two reviewers (M.M.G., S.G.) to identify all potentially relevant articles, which were then examined in full-text form. Reference lists of relevant original and review articles were searched manually for additional reports. Agreement on potential relevance was reached by consensus and by consultation with a third reviewer (K.H.N.).

The inclusion criteria were peer-reviewed study reporting on clinical validation or implementation of maternal cfDNA testing in screening for aneuploidy in twin pregnancy, in which data on pregnancy outcome were provided for at least 85% of the study population. Proof-of-principle articles and studies in which the laboratory scientists carrying out the tests were aware of fetal karyotype or pregnancy outcome were excluded. We also excluded case-control studies because they tend to introduce an optimistic bias to the estimates of diagnostic performance.

Data extraction and meta-analysis

Data regarding sample size, gestational age at analysis, method used for cfDNA testing and DR or sensitivity and FPR or specificity for non-mosaic trisomies 21, 18 and 13 were obtained from each study included in the systematic review and documented in contingency tables. In the construction of these tables, we used the results from the cfDNA test and excluded cases of known mosaicism and those in which the test failed to give a result. In the calculation of FPR, we included all euploid cases and cases with aneuploidy other than the one under investigation. Authors were contacted when clarification was required in the interpretation of their data.

We extracted data from the primary studies to obtain the four cell values of a diagnostic 2×2 table to calculate test accuracy measures of DR and FPR. The analyses were stratified according to type of aneuploidy (trisomy 21 or trisomy 18). We calculated DR and FPR with corresponding 95% CI for individual studies and displayed them in forest plots to investigate heterogeneity. We pooled the DR and FPR estimates using bivariate random-effects regression models. The bivariate model assumes that logit transformations of DR and FPR are correlated negatively and follow a bivariate normal distribution¹⁴. We computed the positive and negative likelihood ratios from the pooled estimates of DR and FPR. Heterogeneity among studies was quantified with the variance of the logit of accuracy indices as estimated by the bivariate model.

Publication bias was not analyzed given the limited power of available tests and the uncertainty about interpreting statistical evidence of funnel plot asymmetry as necessarily implying publication bias¹⁵. For trisomy 13, there was an insufficient number of cases for meaningful meta-analysis and we therefore computed average DR and FPR values.

We conducted statistical analyses using the metandi and midas commands in Stata software¹⁶.

RESULTS

Update of The Fetal Medicine Foundation results

Study population

A total of 1122 twin pregnancies had cfDNA testing, but 125 (11.1%) of these were excluded from further analysis either because the cfDNA test did not provide a result ($n = 52$), the pregnancy ended in termination, miscarriage or stillbirth with no known karyotype ($n = 45$) or there was loss to follow-up ($n = 28$).

Of the 997 cases included in the study, 854 (85.7%) were dichorionic and 143 (14.3%) were monochorionic; the median maternal age was 38.0 (IQR, 34.5–41.0) years, the median maternal weight was 69.0 (IQR, 60.4–82.6) kg and the median gestational age at sampling was 12.1 (IQR, 10.7–12.9) weeks. Maternal racial origin was white in 772 (77.4%) pregnancies, South Asian in 65 (6.5%), East Asian in 32 (3.2%), black in 104 (10.4%) and mixed in 24 (2.4%). Conception was natural in 766 (76.8%) pregnancies and after use of assisted reproductive techniques in 231 (23.2%).

The study population of 997 pregnancies included 17 with trisomy 21, 10 with trisomy 18, two with trisomy 13 and 968 without trisomy 21, 18 or 13; one case of trisomy 18 was a monochorionic twin pregnancy in which both fetuses were affected, and all the other trisomic cases were a dichorionic pregnancy in which only one fetus was trisomic and the cotwin was non-trisomic.

Performance of screening

The cfDNA test classified correctly 16 (94.1%) of the 17 cases of trisomy 21, nine (90.0%) of the 10

cases of trisomy 18, one (50.0%) of the two cases of trisomy 13 and 962 (99.4%) of the 968 cases without any of the three trisomies. One case each of trisomy 21, trisomy 18 and trisomy 13 was classified as normal. In the non-trisomic group, four cases were classified as trisomy 13, one as trisomy 18 and one as trisomy 21 and, therefore, the combined FPR was 0.62% (6/968).

Systematic review and meta-analysis

Data sources

The search identified 329 potentially relevant citations, but 320 were excluded because they were not relevant, a conference abstract rather than a peer-reviewed paper, a review article or opinion, a study not on twins, a case-control study, a study on clinical implementation of cfDNA testing in screening for aneuploidy in which pregnancy outcome data were provided for < 85% of the study population or a proof-of-principle study reporting

laboratory techniques rather than clinical validation of a predefined method of maternal blood cfDNA analysis (Figure 1). In total, nine relevant studies were identified^{4,5,17–23} but two of these^{4,5} were excluded from the meta-analysis because their data are included in the updated Fetal Medicine Foundation results presented above. The characteristics of the current study and the seven identified by the literature search are summarized in Table 1.

Methodological quality of selected studies

The methodological quality of the selected studies, assessed by the quality assessment tool for diagnostic accuracy studies (QUADAS-2)²⁴, is illustrated in Figure 2. This tool comprises four domains; each one is assessed in terms of risk of bias and the first three are also assessed in terms of concerns regarding applicability.

Risk of bias. The first domain relates to patient selection. A study was considered to be at low risk of bias if the cfDNA test was carried out in a consecutive or random sample of patients; all the studies were classified as being at high risk of bias because the samples were not explicitly stated to have been either consecutive or selected at random. The second domain relates to the index test; all included studies were considered to be at low risk of bias because the cfDNA test was carried out and the results given by the laboratory without prior knowledge of the fetal karyotype or pregnancy outcome. The third domain relates to the reference standard; all included studies were considered to be at low risk of bias because the method of diagnosing the chromosomal abnormality under investigation, including karyotyping or neonatal examination, was accepted to be true. The fourth domain relates to flow and timing. A study was considered to be at low risk of bias if, first, in the calculation of performance of screening, all patients in the study population had both a result from the cfDNA test and pregnancy outcome and, second, if the method of classifying the outcome result (invasive testing or clinical examination) was the same in all cases in the study population. All but one study¹⁸ were classified as being at high risk of bias because cfDNA testing was not carried out or did not provide results in

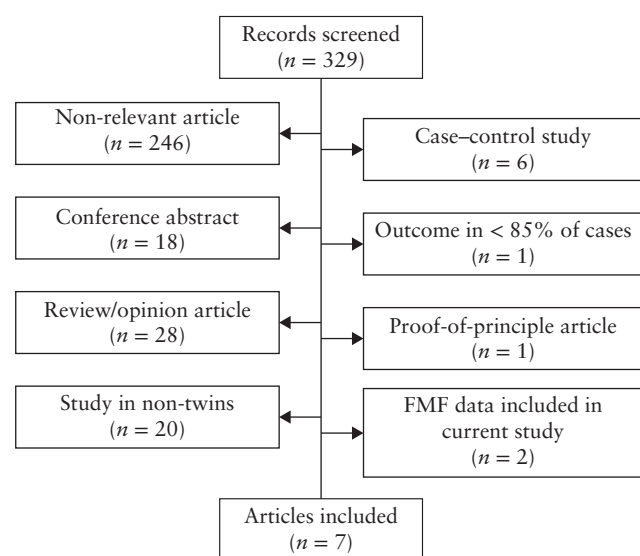


Figure 1 Flowchart summarizing selection from literature of studies for inclusion in systematic review. FMF, Fetal Medicine Foundation.

Table 1 Summary of characteristics of studies reporting on cell-free DNA (cfDNA) analysis of maternal blood in screening for trisomies (T) 21, 18 and 13 in twin pregnancy

Study	Aneuploidy studied	n	Monochorionic (n (%))	T21 (n)	T18 (n)	T13 (n)	Outcome known (%)	cfDNA method	GA (weeks)	Population
Lau (2013) ¹⁷	T21	12	2 (16.7)	1	—	—	100	MPSS	13 (11–20)	High risk
Huang (2014) ¹⁸	T21,T18	189	33 (17.5)	9	2	—	100	MPSS	19 (11–36)	High risk
Tan (2016) ¹⁹	T21	510	— (3.2)*	4	—	—	90	MPSS	12 (11–28)	Mixture
Du (2017) ²⁰	T21	92	39 (42.4)	2	—	—	100	MPSS	18 (14–23)	High risk
Le Conte (2018) ²¹	T21,T18,T13	418	86 (20.6)	3	1	0	85	MPSS	16 (10–35)	Mixture
Yang (2018) ²²	T21,T18	432	95 (22.0)	4	1	—	91	MPSS	> 9	Mixture
Yu (2019) ²³	T21,T18,T13	1157	308 (26.6)	16	4	1	99	MPSS	18 (8–30)	Mixture
Current study	T21,T18,T13	997	143 (14.3)	17	10	2	94	Targeted	11 (10–14)	Mixture

Only first author is given for each study. Numbers reported are those after exclusion of cases without cfDNA test result or pregnancy outcome. Gestational age (GA) is given as mean (range) or actual value. *Value in original sample before exclusion for failed results and no follow-up. MPSS, massively parallel shotgun sequencing.

all cases and/or there was no complete follow-up and/or the method of determining outcome was not the same in all cases.

Concerns regarding applicability. In the context of screening for fetal aneuploidy by cfDNA analysis of maternal blood in the general population, the first domain relates to patient selection and all the studies were classified as being at high risk of concerns regarding applicability because, in these studies, the test was not carried out in the general population but in a mixture of low- and high-risk pregnancies, in which some of the patients had had another screening test before opting for cfDNA testing. In terms of the second and third domains (index test and reference standard, respectively), all studies

were considered as being at low risk of concerns regarding applicability.

Method of analyzing samples

The studies included in the meta-analysis used one of two methods for analysis of cfDNA in maternal blood: massively parallel shotgun sequencing or targeted analysis (either by next-generation sequencing or by a custom microarray) (Table 1).

Nature of studies

All studies included in this meta-analysis were prospective; three studies were in high-risk pregnancies and five examined a mixture of high-risk and routine populations (Table 1). The proportion of mono chorionic twin pregnancies ranged from 3.2% to 42.4%.

Meta-analysis and performance of screening for aneuploidy

The DR and FPR for each study, pooled weighted data and heterogeneity between studies (variance of the logit sensitivity and specificity) are provided in Tables 2 and 3; sensitivity and specificity are illustrated in Figures 3 and 4. Heterogeneity between studies was very low.

Trisomy 21. A total of eight studies reported on the performance of screening by cfDNA analysis for trisomy 21 in a combined total of 56 cases of trisomy 21 and 3718 non-trisomy-21 twin pregnancies (Table 2 and Figure 3). Among individual studies, the DR varied between 94.1% and 100% and the FPR varied between 0% and 0.24%. The pooled weighted DR and FPR were 98.2% (95% CI, 83.2–99.8%) and 0.05% (95% CI, 0.01–0.26%), respectively.

Trisomy 18. A total of five studies reported on the performance of screening by cfDNA analysis for trisomy 18 in a combined total of 18 cases of trisomy 18 and 3143 non-trisomy-18 twin pregnancies (Table 3 and Figure 4). Among individual studies, the DR varied between 50.0%

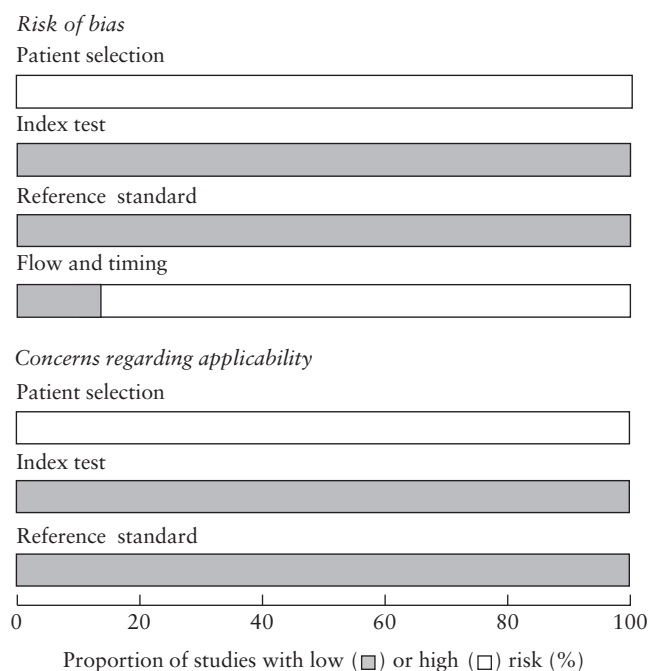


Figure 2 Summary of quality of included studies reporting on cell-free DNA testing for trisomies 21, 18 and 13, assessed using Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) checklist.

Table 2 Studies reporting on application of cell-free DNA analysis of maternal blood in screening for trisomy 21 in twin pregnancy

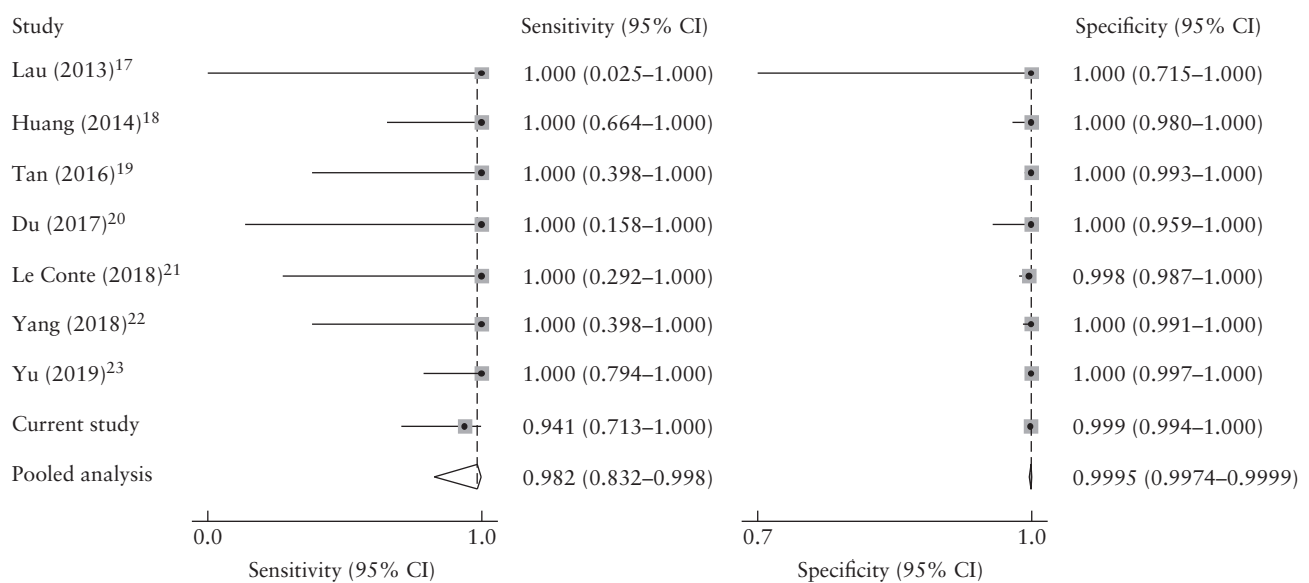
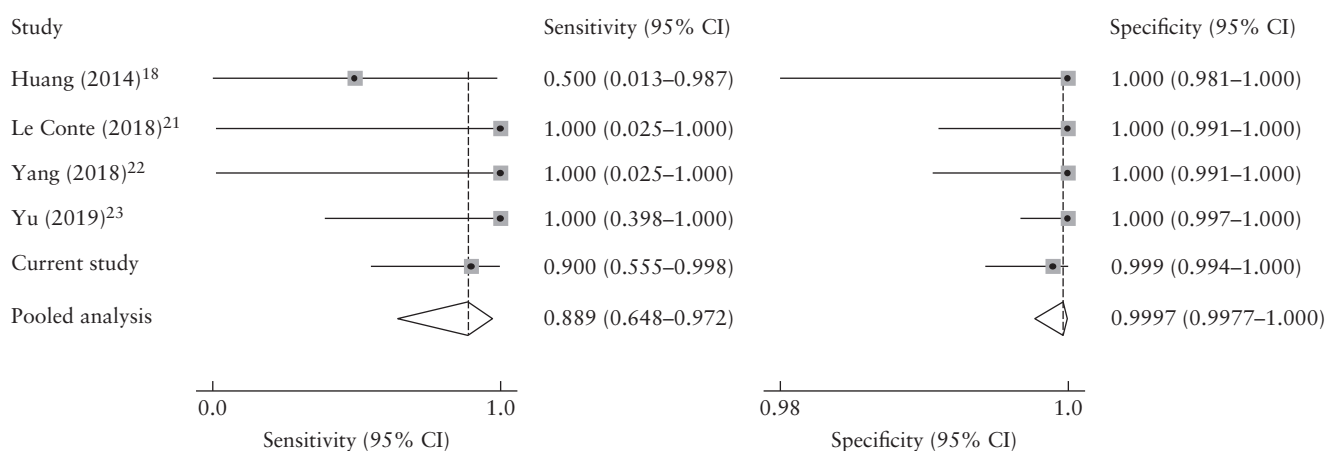
Study	Trisomy 21		Non-trisomy 21	
	Total (n)	Detection rate (n (%), 95% CI)	Total (n)	False-positive rate (n (%), 95% CI)
Lau (2013) ¹⁷	1	1 (100, 2.50–100)	11	0 (0, 0–28.49)
Huang (2014) ¹⁸	9	9 (100, 66.4–100)	180	0 (0, 0–2.03)
Tan (2016) ¹⁹	4	4 (100, 39.8–100)	506	0 (0, 0–0.73)
Du (2017) ²⁰	2	2 (100, 15.8–100)	89	0 (0, 0–4.06)
Le Conte (2018) ²¹	3	3 (100, 29.2–100)	415	1 (0.24, 0.01–1.34)
Yang (2018) ²²	4	4 (100, 39.8–100)	396	0 (0, 0–0.93)
Yu (2019) ²³	16	16 (100, 79.4–100)	1141	0 (0, 0–0.32)
Current study	17	16 (94.1, 71.3–100)	980	1 (0.10, 0–0.57)
Pooled analysis (% (95% CI))*		98.2 (83.2–99.8)		0.05 (0.01–0.26)
Heterogeneity assessment		0.020		0.011
Positive likelihood ratio (95% CI)				1837 (369–9149)
Negative likelihood ratio (95% CI)				0.018 (0.002–0.190)

Only first author is given for each study. Cases with mosaicism were excluded from calculations. *Bivariate random-effects model.

Table 3 Studies reporting on application of cell-free DNA analysis of maternal blood in screening for trisomy 18 in twin pregnancy

Study	Trisomy 18		Non-trisomy 18	
	Total (n)	Detection rate (n (%), 95% CI)	Total (n)	False-positive rate (n (%), 95% CI)
Huang (2014) ¹⁸	2	1 (50.0, 1.3–98.7)	187	0 (0, 0–1.95)
Le Conte (2018) ²¹	1	1 (100, 2.5–100)	417	0 (0, 0–0.88)
Yang (2018) ²²	1	1 (100, 2.5–100)	399	0 (0, 0–0.92)
Yu (2019) ²³	4	4 (100, 39.8–100)	1153	0 (0, 0–0.32)
Current study	10	9 (90.0, 55.5–99.8)	987	1 (0.10, 0–0.56)
Pooled analysis (% (95% CI))*		88.9 (64.8–97.2)		0.03 (0.00–0.33)
Heterogeneity assessment		0		0
Positive likelihood ratio (95% CI)		2774 (388–19 823)		
Negative likelihood ratio (95% CI)		0.111 (0.030–0.411)		

Only first author is given for each study. Cases with mosaicism were excluded from calculations. *Bivariate random-effects model.

**Figure 3** Forest plots of sensitivity and specificity with 95% CI and pooled weighted summary statistics using bivariate random-effects model in assessing performance of cell-free DNA analysis in screening for trisomy 21 in twin pregnancy. Only first author is given for each study.**Figure 4** Forest plots of sensitivity and specificity with 95% CI and pooled weighted summary statistics using bivariate random-effects model in assessing performance of cell-free DNA analysis in screening for trisomy 18 in twin pregnancy. Only first author is given for each study.

and 100% and the FPR varied between 0% and 0.10%. The pooled weighted DR and FPR were 88.9% (95% CI, 64.8–97.2%) and 0.03% (95% CI, 0.00–0.33%), respectively.

Trisomy 13. A total of three studies reported on the performance of screening by cfDNA analysis for trisomy 13 in a combined total of three cases of trisomy 13 and 2569 non-trisomy-13 twin pregnancies. In our study, one of the two affected cases was detected by cfDNA testing at a FPR of 0.4% (4/995). In the second study²³, the one affected case was detected at a FPR of 0% (0/1156). In the third study²¹, there were no cases of trisomy 13 but one false-positive result. In the combined results, the DR was 66.7% (2/3) and FPR was 0.19 (5/2569).

DISCUSSION

Principal findings

The results of our study and the meta-analysis of cfDNA testing of maternal blood in twin pregnancies suggest that the performance of the test for trisomy 21 may be similar to that in singleton pregnancies. In the combined total of 56 trisomy-21 and 3718 non-trisomy-21 twin pregnancies, the pooled weighted DR and FPR were 98.2% (95% CI, 83.2–99.8%) and 0.05% (95% CI 0.01–0.26%), respectively; in our meta-analysis of studies in singleton pregnancies, the pooled weighted DR in 1963 cases of trisomy 21 was 99.7% (95% CI, 99.1–99.9%) and the FPR in 223 932 non-trisomy-21 pregnancies was 0.04% (95% CI, 0.02–0.07%)¹. In the combined total of 18 cases of trisomy 18 and 3143 non-trisomy-18 pregnancies, the pooled weighted DR and FPR were 88.9% (95% CI, 64.8–97.2%) and 0.03% (95% CI, 0.00–0.33%), respectively; in our meta-analysis of studies in singleton pregnancies, the pooled weighted DR in 563 cases of trisomy 18 was 97.9% (95% CI, 94.9–99.1%) and the FPR in 222 013 non-trisomy-18 pregnancies was 0.04% (95% CI, 0.03–0.07%)¹. The number of twin pregnancies with trisomy 13 ($n = 3$) was too small for accurate assessment of DR. The average FPR for trisomy 13 of 0.19% (5/2569) seems slightly higher than the values reported in singleton pregnancy (0.04%; 95% CI, 0.02–0.07%)¹.

In our study, the method of cfDNA testing was targeted but, in all other studies, massively parallel shotgun sequencing was used. Similarly, our study was confined to pregnancies in the first trimester, whereas the other studies included pregnancies in the second, and some in the third, trimesters. Although there was no obvious difference in performance of screening between our study and the other studies, the small number of cases prevented meaningful subgroup analyses, including chorionicity, cfDNA method for analysis, background risk or gestational age at testing.

This study has not addressed the issue of cfDNA test failure because we have reported recently our experience in a larger cohort including both singleton and twin pregnancies²⁵. In that study, we found that important contributors to cfDNA test failure are increased maternal

weight, conception by *in-vitro* fertilization, black or South Asian racial origin, dichorionicity, nulliparity, low gestational age and low serum pregnancy-associated plasma protein-A and free beta human chorionic gonadotropin. Test failure after first sampling in dichorionic twins was 3.3-times higher than in singletons, but to a great extent this excess failure rate could be attributed to the fact that a considerably higher proportion of twins were conceived by *in-vitro* fertilization and more women were nulliparous.

Comparison with previous meta-analyses in twin pregnancies

Our previous meta-analysis on the performance of cfDNA testing for fetal aneuploidy in clinical validation or implementation studies included five studies in twin pregnancies; in a total of 24 cases of trisomy 21 and 1111 unaffected cases, the DR was 100% and FPR was 0%¹. Another meta-analysis examined four studies that included both singletons and twins and three studies that included only twins; it was not possible to extract the number of twin pregnancies that were evaluated but the authors reported that the DR of trisomy 21 in twins was 89.4% (95% CI, 75–96%) and the FPR was 0.4%²⁶.

Liao *et al.* conducted a meta-analysis of studies reporting on cfDNA testing in twin pregnancy²⁷. They included 10 studies, of which one was retrospective, three were a mixture of retrospective and prospective and six were prospective; five were cohort studies, two were case–control studies and three were a mixture of cohort and case–control studies. The authors did not set any criteria on the degree of follow-up. In a combined total of 69 cases of trisomy 21, the DR was 99% (95% CI, 92–100%), in 13 cases of trisomy 18 the DR was 85% (95% CI, 55–98%), in three cases of trisomy 13 the DR was 100% and in 2008 euploid pregnancies the FPR was 0.05%. Our meta-analysis included only prospective cohort studies with follow-up in at least 85% of cases to avoid reporting bias. Case–control studies were excluded because they tend to introduce an optimistic bias to the estimates of diagnostic performance.

Implications for clinical practice

This meta-analysis provides good evidence that the performance of cfDNA testing for trisomy 21 in twin pregnancies may be similar to that in singletons. In this respect, the performance of the cfDNA test is superior, both in terms of higher DR and substantially lower FPR, to that of the first-trimester combined test or second-trimester biochemical testing²⁸. This is particularly important in the case of dichorionic twins in which both the incidence of aneuploidy and the invasive procedure-related risk of pregnancy loss are increased compared to in singletons. If the pregnancies are discordant for aneuploidy and the parents choose selective feticide, the subsequent risk of miscarriage or early preterm birth increases with gestational age at feticide²⁹; in this respect it would be

preferable to offer screening leading to prenatal diagnosis in the first than in the second trimester.

A positive or high-risk cfDNA result should be confirmed by invasive testing. In the case of high risk for trisomy 21 on first-trimester combined screening and positive cfDNA result for trisomy 21, the diagnostic test can be chorionic villus sampling. In the case of trisomy 18 or 13, a positive result should be followed by a detailed ultrasound examination and, if the characteristic defects associated with the trisomy are detected, chorionic villus sampling can be carried out; if no defects are detected on the scan, the preferred diagnostic test is amniocentesis to avoid an erroneous result due to placenta-confined mosaicism.

On the other hand, a negative or low-risk cfDNA result reassures that the fetus is unlikely to be affected by the trisomy under investigation. The posterior risk for a given patient can be obtained by multiplying the prior risk by the negative likelihood ratios calculated in this meta-analysis; the risk for trisomies 21 and 18 is reduced by a factor of 56 and 9, respectively. For example, if prior screening by the combined test had shown that the risk for trisomy 21 was 1 in 10 and cfDNA testing gives a low-risk result, the chance that the fetus is affected is 1 in 560; in contrast, if the risk for trisomy 18 from the combined test was one in two and cfDNA testing gives a low-risk result, the chance that the fetus is affected is one in 18.

Limitations

Contrary to our previous meta-analysis in singleton pregnancies, the number of published studies analyzing the performance of cfDNA testing in twin pregnancies is limited and, consequently, the number of affected cases included in this meta-analysis is considerably smaller. However, the results reported in the literature for trisomies 21 and 18 present low heterogeneity and are therefore likely to represent the true performance. There was an insufficient number of trisomy-13 cases to assess accurately performance.

On assessment of the quality of the included studies, all were considered to be at high risk of selection bias and at high risk of concerns regarding applicability in relation to patient selection because they were not performed as part of routine primary screening but were carried out in preselected populations. However, the ability to detect aneuploidy with cfDNA analysis is dependent upon assay precision and fetal DNA percentage in the sample rather than the prevalence of the disease in the study population. Most studies were also classified as being at high risk of bias in relation to flow and timing. This is essentially because cfDNA testing did not provide results in all cases, follow-up was incomplete or the method of determining outcome was not the same in all cases. However, such criticisms could be applied to any clinical study; all methods of traditional screening occasionally fail to give a result and no screening study in pregnancy can have complete follow-up, especially because some women miscarry and karyotyping is not carried out.

Conclusions

Performance of cfDNA testing for trisomy 21 in twin pregnancies is similar to that reported in singleton pregnancies and is superior to that of the first-trimester combined test or second-trimester biochemical testing. The number of cases of trisomies 18 and 13 was too small for accurate assessment of predictive performance of the cfDNA test.

ACKNOWLEDGMENTS

This study was supported by a grant from The Fetal Medicine Foundation (UK Charity No: 1037116). The cost of collection and analysis of some of the samples for the cfDNA test was covered by Ariosa Diagnostics, Inc., San Jose, CA, USA. This organization had no role in study design, data collection, data analysis, data interpretation or writing of the report.

REFERENCES

- Gil MM, Accurti V, Santacruz B, Plana MN, Nicolaides KH. Analysis of cell-free DNA in maternal blood in screening for aneuploidies: updated meta-analysis. *Ultrasound Obstet Gynecol* 2017; 50: 302–314.
- del Mar Gil M, Quezada MS, Bregant B, Syngelaki A, Nicolaides KH. Cell-free DNA analysis for trisomy risk assessment in first-trimester twin pregnancies. *Fetal Diagn Ther* 2014; 35: 204–211.
- Bevilacqua E, Gil MM, Nicolaides KH, Ordoñez E, Cirigliano V, Dierckx H, Willems PJ, Jani JC. Performance of screening for aneuploidies by cell-free DNA analysis of maternal blood in twin pregnancies. *Ultrasound Obstet Gynecol* 2015; 45: 61–66.
- Sarno L, Revello R, Hanson E, Akolekar R, Nicolaides KH. Prospective first-trimester screening for trisomies by cell-free DNA testing of maternal blood in twin pregnancy. *Ultrasound Obstet Gynecol* 2016; 47: 705–711.
- Galeva S, Konstantinidou L, Gil MM, Akolekar R, Nicolaides KH. Routine first-trimester screening for fetal trisomies in twin pregnancies: cell-free DNA test contingent on results from the combined test. *Ultrasound Obstet Gynecol* 2019; 53: 208–213.
- Quezada MS, Gil MM, Francisco C, Orós G, Nicolaides KH. Screening for trisomies 21, 18 and 13 cell-free DNA analysis of maternal blood at 10–11 weeks' gestation and the combined test at 11–13 weeks. *Ultrasound Obstet Gynecol* 2015; 45: 36–41.
- Gil MM, Revello R, Poon LC, Akolekar R, Nicolaides KH. Clinical implementation of routine screening for fetal trisomies in the UK NHS: cell-free DNA test contingent on results from first-trimester combined test. *Ultrasound Obstet Gynecol* 2016; 47: 45–52.
- Robinson HP, Fleming JE. A critical evaluation of sonar crown rump length measurements. *Br J Obstet Gynaecol* 1975; 82: 702–710.
- Sepulveda W, Sebire NJ, Hughes K, Odibo A, Nicolaides KH. The lambda sign at 10–14 weeks of gestation as a predictor of chorionicity in twin pregnancies. *Ultrasound Obstet Gynecol* 1996; 7: 421–423.
- Sparks AB, Struble CA, Wang ET, Song K, Oliphant A. Noninvasive prenatal detection and selective analysis of cell-free DNA obtained from maternal blood: evaluation for trisomy 21 and trisomy 18. *Am J Obstet Gynecol* 2012; 206: 319.e1–9.
- Ashoor G, Syngelaki A, Wang E, Struble C, Oliphant A, Song K, Nicolaides KH. Trisomy 13 detection in the first trimester of pregnancy using a chromosome-selective cell-free DNA analysis method. *Ultrasound Obstet Gynecol* 2013; 41: 21–25.
- Struble CA, Syngelaki A, Oliphant A, Song K, Nicolaides KH. Fetal fraction estimate in twin pregnancies using directed cell-free DNA analysis. *Fetal Diagn Ther* 2014; 35: 199–203.
- Juneau K, Bogard PE, Huang S, Mohseni M, Wang ET, Ryvkin P, Kingsley C, Struble CA, Oliphant A, Zahn JM. Microarray-based cell-free DNA analysis improves noninvasive prenatal testing. *Fetal Diagn Ther* 2014; 36: 282–286.
- 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. *J Clin Epidemiol* 2005; 58: 982–990.
- Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. *Ann Intern Med* 2008; 149: 889–897.
- StataCorp. 2015. *Stata Statistical Software: Release 14*. College Station, TX, USA: StataCorp LP.
- Lau TK, Jiang F, Chan MK, Zhang H, Lo PS, Wang W. Non-invasive prenatal screening of fetal Down syndrome by maternal plasma DNA sequencing in twin pregnancies. *J Matern Fetal Neonatal Med* 2013; 26: 434–437.

18. Huang X, Zheng J, Chen M, Zhao Y, Zhang C, Liu L, Xie W, Shi S, Wei Y, Lei D, Xu C, Wu Q, Guo X, Shi X, Zhou Y, Liu Q, Gao Y, Jiang F, Zhang H, Su F, Ge H, Li X, Pan X, Chen S, Chen F, Fang Q, Jiang H, Lau TK, Wang W. Noninvasive prenatal testing of trisomies 21 and 18 by massively parallel sequencing of maternal plasma DNA in twin pregnancies. *Prenat Diagn* 2014; **34**: 335–340.
19. Tan Y, Gao Y, Lin G, Fu M, Li X, Yin X, Du J, Li J, Li W, Peng H, Yuan Y, Chen F, Jiang F, Zhang H, Lu G, Gong F, Wang W. Noninvasive prenatal testing (NIPT) in twin pregnancies with treatment of assisted reproductive techniques (ART) in a single center. *Prenat Diagn* 2016; **36**: 672–679.
20. Du E, Feng C, Cao Y, Yao Y, Lu J, Zhang Y. Massively Parallel Sequencing (MPS) of Cell-Free Fetal DNA (cffDNA) for Trisomies 21, 18, and 13 in twin pregnancies. *Twin Res Hum Genet* 2017; **20**: 242–249.
21. Le Conte G, Letourneau A, Jani J, Kleinfinger P, Lohmann L, Costa JM, Benachi A. Cell-free fetal DNA analysis in maternal plasma as screening test for trisomies 21, 18 and 13 in twin pregnancy. *Ultrasound Obstet Gynecol* 2018; **52**: 318–324.
22. Yang J, Qi Y, Hou Y, Guo F, Peng H, Wang D, Haoxin OY, Wang Y, Huang H, Yin A. Performance of non-invasive prenatal testing for trisomies 21 and 18 in twin pregnancies. *Mol Cytogenet* 2018; **11**: 4–9.
23. Yu W, Lv Y, Yin S, Liu H, Li X, Liang B, Kong L, Liu C. Screening of fetal chromosomal aneuploidy diseases using noninvasive prenatal testing in twin pregnancies. *Expert Rev Mol Diagn* 2019; **19**: 189–196.
24. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011; **155**: 529–536.
25. Galeva S, Gil MM, Konstantinidou L, Akolekar R, Nicolaides KH. First-trimester screening for trisomies by cfDNA testing of maternal blood in singleton and twin pregnancies: factors affecting test failure. *Ultrasound Obstet Gynecol* 2019. DOI: 10.1002/uog.20290.
26. Taylor-Phillips S, Freeman K, Geppert J, Agbebiyi A, Uthman OA, Madan J, Clarke A, Quenby S, Clarke A. Accuracy of non-invasive prenatal testing using cell-free DNA for detection of Down, Edwards and Patau syndromes: a systematic review and meta-analysis. *BMJ Open* 2016; **6**: e010002.
27. Liao H, Liu S, Wang H. Performance of non-invasive prenatal screening for fetal aneuploidy in twin pregnancies: a meta-analysis. *Prenat Diagn* 2017; **37**: 874–882.
28. Audibert F, Gagnon A. No. 262-Prenatal screening for and diagnosis of aneuploidy in twin pregnancies. *J Obstet Gynaecol Canada* 2017; **39**: e347–e361.
29. Evans MI, Goldberg JD, Dommergues M, Wapner RJ, Lynch L, Dock BS, Horenstein J, Golbus MS, Rodeck CH, Dumez Y, Holzgreve W. Efficacy of second-trimester selective termination for fetal abnormalities: International collaborative experience among the world's largest centers. *Am J Obstet Gynecol* 1994; **171**: 90–94.

SUPPORTING INFORMATION ON THE INTERNET

The following supporting information may be found in the online version of this article:



Appendix S1 Search strategies



A video abstract of this article is available online

7.6 Reanálisis de la heterogeneidad en las revisiones sistemáticas publicadas

En el momento del desarrollo de las revisiones sistemáticas incluidas en los Capítulos 5, 6 y 7 todavía no se habían implementado las medidas adicionales para complementar la evaluación de la heterogeneidad en las revisiones de exactitud diagnóstica que se han propuesto en esta tesis. En la Tabla 5 del Capítulo 4 se han mostrado los resultados del reanálisis realizado para obtener las medidas de heterogeneidad propuestas (área de la elipse de predicción y Odds ratio mediano para sensibilidad y especificidad)

En la revisión sobre la exactitud diagnóstica de la pulsioximetría (Plana et al. 2018) (Capítulo 5) se calcularon y publicaron las varianzas de los logit de sensibilidad como medida de cuantificación de la heterogeneidad. Esta revisión mostraba gráficamente una amplia variabilidad en sensibilidad y prácticamente nula en especificidad. En cambio, las varianzas de los logit para sensibilidad y especificidad fueron de 0.1 y 2 respectivamente. Esta discrepancia entre lo observado gráficamente y las varianzas es fruto de la escala logit de estas varianzas. Los Odds ratio medianos, al calcularse a partir de estas mismas varianzas, dan unos valores de 1.4 y 3.9 para la sensibilidad y especificidad respectivamente. Si calculamos el área de la elipse de predicción al 95% corresponde a 0.008, o lo que es lo mismo, representa un 0.8% del área total del espacio ROC. Podríamos concluir que estamos ante un escenario de baja heterogeneidad (Figura 14).

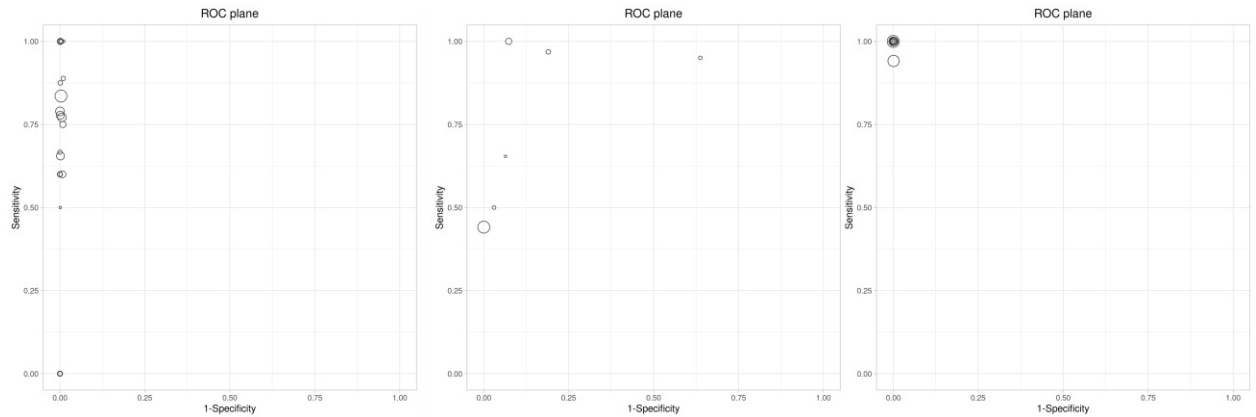
En la revisión sistemática sobre la exactitud diagnóstica de los códigos CIE (Xiao et al. 2018) (Capítulo 6) se calcularon y publicaron las varianzas de los logit de sensibilidad y especificidad como medidas de heterogeneidad. Esta revisión mostraba

gráficamente mayor variabilidad en la especificidad que en la sensibilidad las varianzas de los logit fueron de 2.6 y 9.9 para sensibilidad y especificidad.

Si se transforman estas varianzas a escala de OR, obtenemos un Odds Ratio mediano de 4.7 y 19.9 para sensibilidad y especificidad respectivamente. La interpretación es que hay un incremento de 4.7 veces de ser un “verdadero positivo”, en mediana, si un individuo enfermo cambiara de un estudio con menor sensibilidad a otro estudio con mayor sensibilidad. Por otro lado, hay un incremento de 19.9 veces de ser un “verdadero negativo”, en mediana, si un individuo no enfermo cambiara de un estudio con menor especificidad a un estudio con mayor especificidad. Por último, el área de la elipse de predicción al 95% corresponde a 0.815, o lo que es lo mismo, representa un 81.5% del área total del espacio ROC. Podríamos concluir que estamos ante un escenario de alta heterogeneidad (Figura 14).

En la revisión sobre la exactitud diagnóstica del DNA fetal (Gil et al. 2019) (Capítulo 7) se incluyeron en la publicaron las varianzas de los logit de sensibilidad y especificidad como medidas de cuantificación de la heterogeneidad. Esta revisión mostraba gráficamente que no había variabilidad ni en sensibilidad ni en especificidad. Las varianzas para ambos índices fueron cercanas a 0 para ambos índices y los MOR de 1 para ambos índices como es esperable. Podríamos concluir que estamos ante un escenario de ausencia de heterogeneidad (Figura 14).

Figura 14. Representación gráfica en el plano ROC de la sensibilidad y especificidad de las revisiones incluidas en los capítulos 5, 6 y 7.

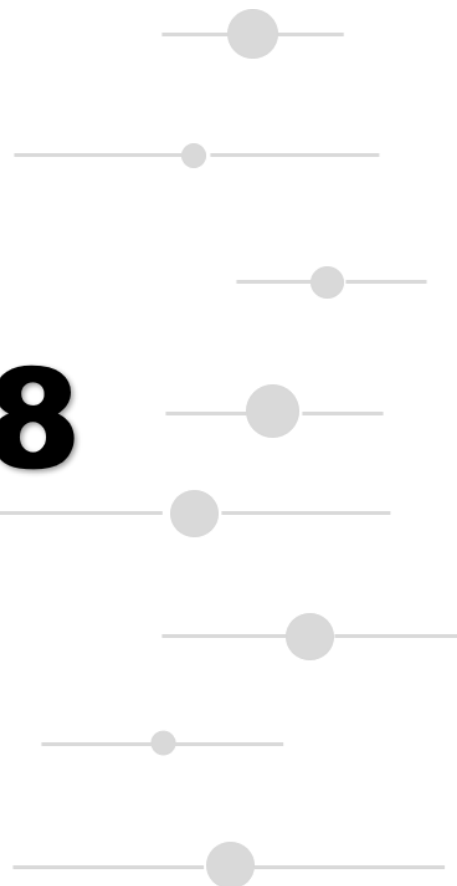


Plana et al. 2018

Xiao et al. 2018

Gil et al. 2019

Capítulo 8



Capítulo 8

En este capítulo se presenta una guía elaborada para realizar la lectura crítica de revisiones sistemáticas de estudios de exactitud diagnóstica. El objetivo es dotar a los profesionales de habilidades en valoración crítica de este tipo de estudios utilizando metodologías de la medicina basada en la evidencia y la epidemiología clínica.

En la guía se abordan inicialmente los elementos clave que permiten evaluar cuán válidos son los resultados de la revisión sistemática. En segundo lugar, se muestran aspectos a considerar para valorar críticamente la metodología estadística empleada en el metanálisis y la interpretación de los resultados. Por último, se aborda cómo evaluar críticamente la aplicabilidad o validez externa de los resultados de la revisión.

La guía está aceptada para su publicación como capítulo de libro en: Cabello Juan B, editor. Lectura crítica de la evidencia clínica, 2.^a ed. Barcelona: Elsevier; 2022. A continuación, se incluye la guía como cuerpo del capítulo 8.

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, Hira 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 desgarrando 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 fáciles 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.

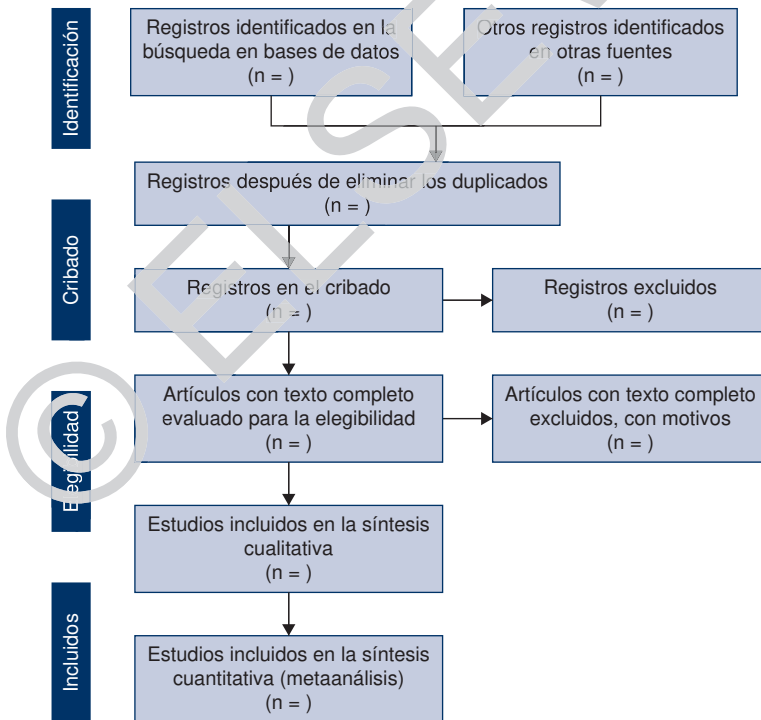


Figura 14.1 Diagrama de flujo propuesto por la declaración PRISMA para ilustrar el proceso de búsqueda y selección de estudios para la revisión. (Tomado de Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred Reporting Items for Systematic Reviews and MetaAnalyses: The PRISMA Statement. 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 señalado 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

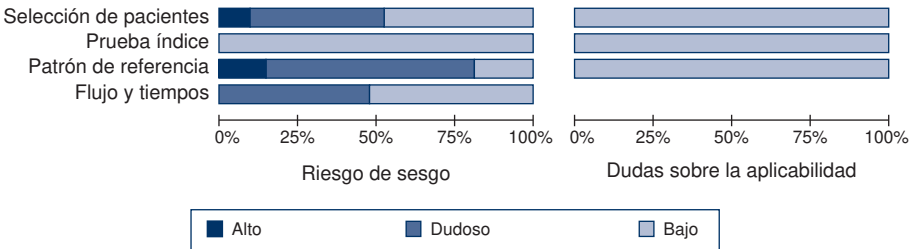


Figura 14.2 Representación gráfica del resultado de la evaluación del riesgo de sesgo de los estudios incluidos en una revisión sistemática de pruebas diagnósticas.

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 contruidos 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

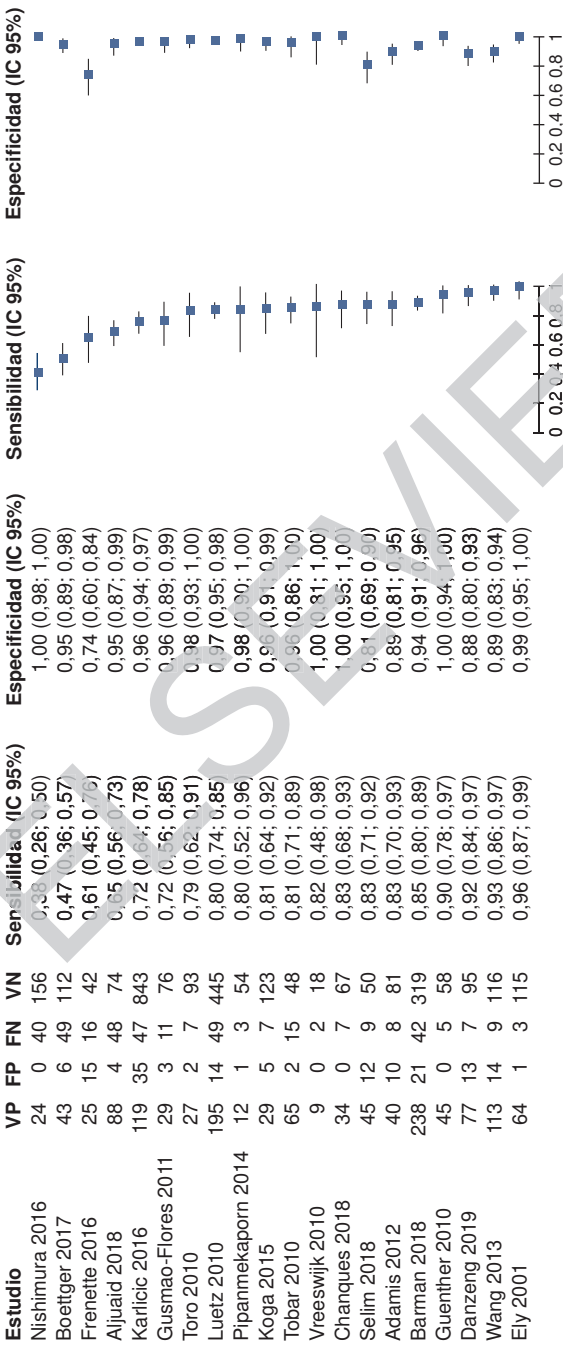


Figura 14.3 Representación gráfica en forma de forest plot de los resultados de sensibilidad y especificidad emparejados.

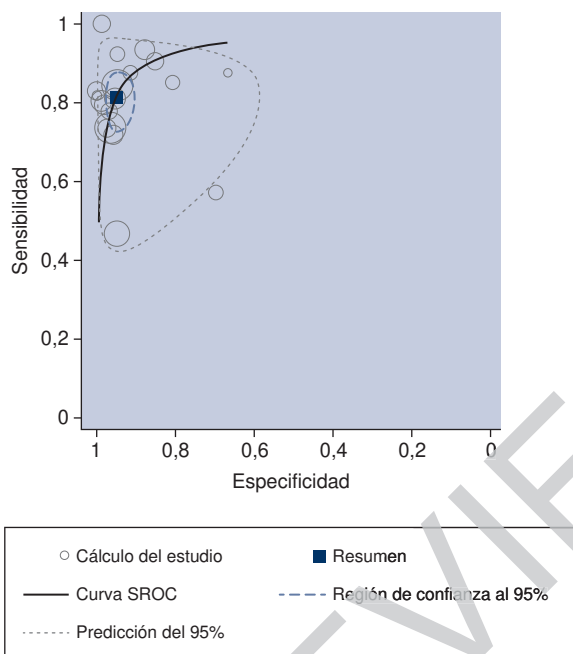


Figura 14.4 Representación en un plano ROC de los resultados de exactitud diagnóstica de los estudios individuales en una revisión sistemática. Se representa el punto promedio de sensibilidad y especificidad junto con las elipses de confianza y predicción y la curva SROC resumen.

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

¿Cuáles 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 clínicos 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 bivalente 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 bivalente 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 bivariable 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?	Sí ✓	No sé	No
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.	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 (<i>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 and the Association Internationale pour la Recherche et L'Enseignement en Neurosciences</i> criterios, 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?	Sí ✓	No sé	No
PISTA: el mejor tipo de estudio es el que: • Se dirige a la pregunta objeto de revisión. • Tiene un diseño apropiado para la pregunta.	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?	Sí ✓	No sé	No
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.	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? <i>PISTA: revisa:</i> • ¿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
5. Si los resultados de los diferentes estudios han sido mezclados para obtener un resultado combinado, ¿era razonable hacer eso? <i>PISTA: considera si:</i> • 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 ✓

(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 (tabla 3) considerando las variables región (América, Asia, Europa) y ámbito de reclutamiento de pacientes (comunitario, clínica, hospital, atención primaria, otros).
- Curva HSROC (efigura 1) del test MMSE en la detección de demencia, incluyendo resultado combinado de sensibilidad y especificidad y elipse de confianza.
- Plano ROC (efigura 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?

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.

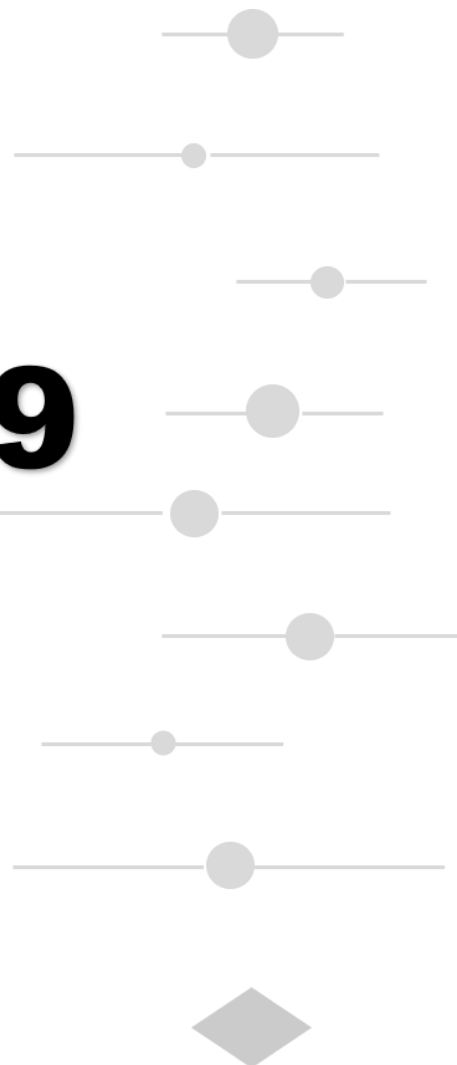
Bibliografía

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, Bossel 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. Leeftang 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, Bossel 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:

Plans 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.

Capítulo 9



Discusión

En esta tesis doctoral se han revisado los métodos estadísticos para el metanálisis de estudios de exactitud diagnóstica. La investigación sobre diagnóstico tiene una serie de peculiaridades que determina que deban emplearse modelos de análisis estadístico complejos, como son los modelos mixtos, multinivel o jerárquicos. Entre estas peculiaridades se tiene el carácter bidimensional de los resultados de los estudios de exactitud diagnóstica, la frecuente correlación entre los índices de sensibilidad y especificidad, la frecuente presencia de elevada heterogeneidad en los resultados y la distinta precisión de medida, con la que se obtienen los índices de sensibilidad y especificidad fruto de los a veces asimétricos tamaños muestrales de enfermos y no enfermos en los estudios primarios.

La heterogeneidad en este tipo de revisiones es una constante. El grupo Cochrane de metodología en diagnóstico (*Cochrane Screening and Diagnostic Tests Methods Group*) propone evaluar la magnitud de la heterogeneidad observada de forma gráfica representando los puntos de sensibilidad y especificidad en un plano ROC y a partir de la elipse de predicción. Cochrane también propone utilizar la estimación numérica de las varianzas de los logit para cuantificar esta heterogeneidad.

Poniendo el foco en las revisiones sistemáticas del propio grupo Cochrane bajo la premisa de que están sujetas a un mayor rigor metodológico, se han evaluado los distintos métodos de abordaje de la heterogeneidad utilizados en sus revisiones. Los resultados muestran que la descripción y cuantificación de la heterogeneidad en estas revisiones es muy variable. Es habitual que las revisiones incluyan una descripción gráfica de la heterogeneidad y planteen como objetivo explorar fuentes de heterogeneidad, sin embargo, sólo un número limitado de las revisiones sistemáticas analizadas dan una cuantificación numérica de la heterogeneidad en términos de varianza de los logit, probablemente debido a los problemas de interpretabilidad referidos. Muchas revisiones hacen una valoración subjetiva de la heterogeneidad observada pero no es frecuente que se reporte si los resultados del análisis de la heterogeneidad han influido en la decisión de dónde poner el foco del metanálisis (sin en la curva sROC o en el punto resumen) ni tampoco hay una clara contextualización ni discusión de los resultados de exactitud diagnóstica en base a la heterogeneidad observada.

Con objeto de facilitar una mejor caracterización de la heterogeneidad, en esta tesis doctoral se proponen dos nuevas medidas de cuantificación de la heterogeneidad

complementarias a las varianzas de los logit. Por un lado, se propone calcular el área de la elipse de predicción que permite interpretar la heterogeneidad como una proporción de la superficie total del espacio ROC. Un 24% de las revisiones analizadas representan gráficamente las elipses de predicción. Si además de representar gráficamente esta elipse de predicción se calculara el área que representa se estaría dando una aproximación de la heterogeneidad mucho más robusta y objetiva.

Además del área de la elipse de predicción y, con la finalidad de mejorar la interpretación de las varianzas de los logit, se propone transformar dichas varianzas a escala de OR y calcular un OR mediano para sensibilidad y para especificidad. El OR mediano para sensibilidad se interpreta como el incremento en el odds de ser un “verdadero positivo”, en mediana, si un individuo enfermo cambiara de un estudio con menor sensibilidad a otro estudio con mayor sensibilidad. De forma similar, el OR mediano de especificidad se interpreta como el incremento en el odds de ser un “verdadero negativo”, en mediana, si un individuo no enfermo cambiara de un estudio con menor especificidad a un estudio con mayor especificidad.

No se persigue simplificar el análisis de heterogeneidad recomendando una medida única de cuantificación de la heterogeneidad, sino enfatizar que la evaluación de la heterogeneidad es relevante, compleja y multifacética, y debe apoyarse tanto en la descripción gráfica como en medidas objetivas de cuantificación como las varianzas de los logit, y complementarse con las medidas que se proponen. Las revisiones sistemáticas incluidas en esta tesis se han reanalizado utilizando las medidas adicionales de heterogeneidad propuestas. Este reanálisis aporta un valor añadido al análisis de la

heterogeneidad y se propone que se haga extensivo a futuras revisiones sistemáticas que evalúen pruebas diagnósticas.

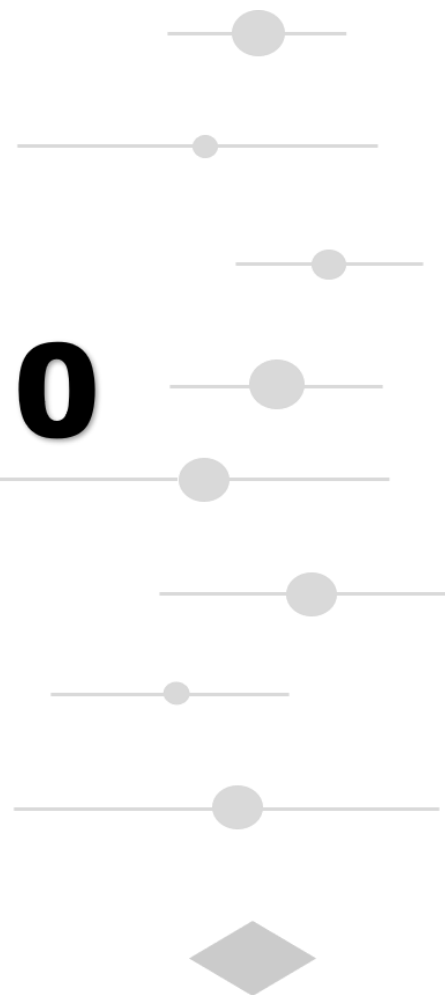
Una dificultad a la que se enfrentan los profesionales que realizan este tipo de estudios de revisión, generalmente clínicos, es que la ejecución de los modelos de metanálisis recomendados requiere de un cierto grado de experiencia en el manejo de paquetes estadísticos avanzados. Para facilitar que estos modelos de metanálisis puedan ejecutarse de forma sencilla incorporando también las medidas de cuantificación desarrolladas en esta investigación, se ha desarrollado y validado una aplicación web interactiva como producto de transferencia que actualiza los métodos de un software de análisis previo ampliamente utilizado para el metanálisis de pruebas diagnósticas (MetaDisC®) (Zamora et al. 2006). Este producto de transferencia da pleno contenido al planteamiento de la tesis que opta a Mención Industrial, y que se ha realizado bajo un convenio de colaboración entre la UCM y el IRYCIS.

Existe otra aplicación web que implementa el mismo modelo bivalente lineal generalizado mixto (MetaDTA) (Freeman et al. 2019) para realizar metanálisis de estudios de exactitud diagnóstica. A diferencia de la aplicación MetaDiSc 2.0, esta aplicación MetaDTA no permite realizar análisis de subgrupos ni metarregresión y sólo incluye las varianzas de los logit de sensibilidad y especificidad como medidas de cuantificación de la heterogeneidad. La aplicación MetaDiSc 2.0 presenta un mejor control de errores, incorpora la posibilidad de hacer los análisis de forma univariante e implementa además de las varianzas de los logit, el área de la elipse de predicción, los OR medianos de sensibilidad y especificidad y el I² bivalente propuesto por Zhou (2014). Los paquetes estadísticos MIDAS y METANDI de STATA tampoco permiten realizar

análisis de metarregresión. En futuras actualizaciones de la aplicación está previsto incorporar el modelo jerárquico HSROC y la posibilidad de graficar los resultados de la evaluación de la calidad de los estudios realizada con la herramienta QUADAS-2.

La discusión clínica sobre los resultados de las revisiones sistemáticas incluidas en esta tesis se circunscribe a lo reflejado en el apartado de discusión de cada una de las publicaciones incluidas en la tesis. Dos de las revisiones sistemáticas publicadas han contribuido a modificar la práctica clínica habitual. Tanto la pulsioximetría en el cribado postnatal de cardiopatías críticas congénitas como la determinación en sangre materna de DNA fetal en el cribado prenatal de trisomías en embarazo gemelar se están incorporando progresivamente a la cartera de servicios de los hospitales como pruebas de rutina para el cribado neonatal y prenatal.

Capítulo 10



Conclusiones

10.1 Conclusiones metodológicas

1. La heterogeneidad en las revisiones sistemáticas Cochrane de evaluación de exactitud diagnóstica se describe de forma subjetiva y no cuantitativa. Esta valoración subjetiva, en ocasiones, es inconsistente con medidas objetivas de cuantificación de la heterogeneidad.
2. El área de la elipse de predicción permite interpretar la heterogeneidad como una proporción del área total en el plano ROC. Los Odds Ratio medianos para la sensibilidad y especificidad pueden facilitar la interpretación de las varianzas de los logit de ambos índices. La incorporación de estas medidas adicionales en el apartado de resultados de las revisiones sistemáticas de estudios de exactitud diagnóstica, permitirá una mejor caracterización de la heterogeneidad.
3. La aplicación web MetaDiSc 2.0 es una herramienta válida que permite a usuarios sin conocimientos estadísticos avanzados, realizar metanálisis de estudios de exactitud diagnóstica según los estándares metodológicos recomendados.

4. La aplicación web MetaDiSc 2.0 es una herramienta útil para calcular las medidas alternativas de cuantificación de la heterogeneidad propuestas permitiendo realizar análisis de comparación estadística de subgrupos mediante técnicas de metarregresión que sólo se incluyen en paquetes estadísticos avanzados.
5. La guía de lectura crítica de revisiones sistemáticas de diagnóstico permite acercar a los profesionales sanitarios la metodología de la medicina basada en la evidencia y la epidemiología clínica facilitándoles la interpretación crítica de este tipo de estudios.
6. Se confirma la importancia de la metodología de metanálisis como herramienta para facilitar la incorporación de los resultados de investigación clínica en la práctica médica.

10.2 Conclusiones clínicas

1. La pulsioximetría es una prueba diagnóstica muy específica y moderadamente sensible para la detección de cardiopatías críticas congénitas en recién nacidos asintomáticos con una baja tasa de falsos positivos. La evidencia disponible apoya la introducción del cribado con pulsioximetría para la detección de cardiopatías críticas congénitas de rutina en los recién nacidos asintomáticos de forma previa al alta.
2. La exactitud diagnóstica de los códigos CIE para la identificación de pacientes con pancreatitis no es óptima.
3. El rendimiento de la prueba de ADN fetal libre para la trisomía 21 en el embarazo gemelar es similar al observado en el embarazo único y es superior al de la prueba combinada del primer trimestre o a la prueba bioquímica del segundo trimestre mostrando mayor sensibilidad y menor tasa de falsos positivos. El número de casos de trisomías 18 y 13 es demasiado pequeño para evaluar con precisión el rendimiento diagnóstico de la prueba de ADN fetal.

RESUMEN

La metodología de revisión sistemática y metanálisis es la más robusta para sintetizar la evidencia disponible sobre la exactitud de una prueba diagnóstica y obtener una estimación global de su sensibilidad y especificidad. La presencia de variabilidad entre estudios o heterogeneidad en estas revisiones es una constante y su análisis es el objeto de este trabajo de tesis. Los objetivos son: *i.* evaluar cómo se explora y cuantifica la heterogeneidad en este tipo de estudios; *ii.* proponer medidas adicionales de cuantificación de la heterogeneidad; *iii.* desarrollar un software para el desarrollo de metanálisis de estudios de exactitud diagnóstica; *iv.* aplicar esta metodología a diferentes escenarios clínicos y *v.* desarrollar una guía de lectura crítica de revisiones sistemáticas de pruebas diagnósticas.

Los resultados muestran que la heterogeneidad se describe de forma subjetiva y no cuantitativa en las revisiones sistemáticas Cochrane de evaluación de pruebas diagnósticas. De las 91 revisiones sistemáticas con metanálisis analizadas en el Capítulo 3, solo 26 aportaban medidas de heterogeneidad, en 21 (23%) metanálisis se ilustraba con la elipse de predicción y únicamente 5 comunicaban las varianzas de los logit de sensibilidad y especificidad. Se proponen dos medidas adicionales de cuantificación de la heterogeneidad, el área de la elipse de predicción que permite interpretar la heterogeneidad como una proporción del área total del plano ROC y el Odds ratio mediano de la sensibilidad y especificidad que facilita la interpretación de las varianzas de los logit de sensibilidad y especificidad.

Los resultados clínicos de las pruebas evaluadas son los siguientes:

La pulsioximetría, analizada en el Capítulo 5, es una prueba muy específica, 99.9% (IC 95% 99.7; 99.9) y moderadamente sensible, 76.3% (IC 95% 69.5; 82.0) para la detección de cardiopatías críticas congénitas en recién nacidos asintomáticos con una baja tasa de falsos positivos (TFP), 0.14% (IC 95% 0.07; 0.22) (n=19 estudios). La TFP fue significativamente inferior cuando la prueba se realizó pasadas 24 horas tras el nacimiento, 0.06% (IC 95% 0.03; 0.13) (p=0.027).

El promedio ponderado de sensibilidad y especificidad de los códigos de la Clasificación Internacional de Enfermedades (CIE) para la identificación de pacientes con pancreatitis aguda, analizado en el Capítulo 6, fue de 0.85 (IC 95% 0.59; 0.96) y 0.96 (IC 95% 0.65; 1.00) respectivamente (n=6 estudios). La sensibilidad y especificidad para la identificación de pacientes con pancreatitis crónica fue de 0.75 (IC 95% 0.71; 0.80) y 0.94 (IC 95% 0.93; 0.96) respectivamente (n=2 estudios).

La exactitud diagnóstica del ADN fetal para el cribado de Trisomía 21 en embarazo gemelar evaluada en el Capítulo 7 (n=8 estudios) es similar al observado en embarazo único con una sensibilidad de 98.2% (IC 95% 83.2; 99.8) y una TFP de 0.05% (IC 95% 0.01; 0.26). Para Trisomía 18 (n=5 estudios), la sensibilidad y TFP fue de 88.9% (IC 95% 64.8; 97.2) y 0.03% (IC 95% 0.00; 0.33), respectivamente. Para Trisomía 13 (n=3 estudios), la sensibilidad y TFP fue de 66.7% y 0.19% respectivamente.

La aplicación web MetaDiSc 2.0 desarrollada y descrita en el Capítulo 4 es una herramienta gratuita y válida que permite a los usuarios sin conocimientos estadísticos avanzados realizar metanálisis de estudios de exactitud diagnóstica según los estándares metodológicos recomendados, incorporando las medidas de heterogeneidad

propuestas. La guía de lectura crítica desarrollada y descrita en el Capítulo 8 facilita la interpretación de los estudios de revisiones sistemáticas de exactitud diagnóstica.

SUMMARY

The methodology of systematic review (SR) and metanalysis is the most robust for pooling the available evidence on the accuracy of a diagnostic test and obtaining an overall estimate of its sensitivity and specificity. The analysis of the heterogeneity in Diagnostic Test Accuracy Reviews (DTA) is the goal of this thesis.

The objectives are: i. to evaluate the methods used to quantify heterogeneity in Cochrane DTA reviews; ii. To propose alternative measures to quantify heterogeneity; iii. To develop a web application to summarise diagnostic test accuracy results; iv. To apply this methodology to different clinical scenarios and v. To develop a guide for critical appraisal of DTA reviews.

Current reporting of heterogeneity is patchy in DTA Cochrane reviews. In 21 of 91 (23%) metanalysis analysed in Chapter 3, the 95% prediction ellipse was included but without calculating its area. In 5/91 (5%) SRs, heterogeneity was quantified reporting the variances of the logit transformed of both sensitivity and specificity. We propose complementary statistics to improve reporting of heterogeneity in DTA systematic reviews. The area of the 95% prediction ellipse in the ROC plane allows interpreting heterogeneity as a proportion of the total area of the ROC space. The median OR is a measure that can be obtained for both accuracy indices separately (i.e., sensitivity and specificity) and are derived from their (logit) variances.

The clinical results of the tests evaluated are as follows:

The overall sensitivity of pulse oximetry for detection of critical congenital heart defects, analysed in Chapter 5, was 76.3% (95% CI 69.5; 82.0). Specificity was 99.9% (95% CI 99.7; 99.9) with a false positive rate (FPR) of 0.14% (95% CI 0.07; 0.22) (n=19 studies). The FPR was significantly lower when newborn pulse oximetry was performed longer than 24 hours after birth (0.06%, 95% CI 0.03; 0.13) (p=0.027).

The pooled estimates of sensitivity and specificity of International Classification of Disease (ICD) codes, analysed in Chapter 6, for acute pancreatitis were 0.85 (95% CI 0.59; 0.96) and 0.96 (95% CI 0.65; 1.00) (n=6 studies). The pooled estimates of sensitivity and specificity of ICD codes for chronic pancreatitis were 0.75 (95% CI 0.71; 0.80) and 0.94 (95% CI 0.93; 0.96) respectively (n=2 studies).

The pooled estimates of sensitivity and FPR of cfDNA, analysed in Chapter 7, for trisomy 21 were 98.2% (95% CI 83.2; 99.8) and 0.05% (95% CI 0.01; 0.26), respectively (n=8 studies). For Trisomy 18, the pooled estimates of sensitivity and FPR were 88.9% (95% CI 64.8; 97.2) and 0.03% (95% CI 0.00; 0.33), respectively (n=5 studies). For Trisomy 13, there were only three affected cases and two (66.7%) of these were detected by the cfDNA test at a FPR of 0.19% (n=3 studies).

We have developed a free and valid web application to perform diagnostic test accuracy meta-analyses (Chapter 4). The application implements the recommended statistical methods and incorporates the new heterogeneity measures proposed in this thesis. The guide for the critical appraisal of DTA reviews facilitates the interpretation of these studies (Chapter 8).

ANEXOS

Se presenta como Anexo I la rutina en R para calcular el área de la elipse de predicción. En el Anexo II figura el artículo en preparación sobre la aplicación MetaDiSc 2.0 para su envío a la revista *BMC Medical Research Methodology*. Por último, en el Anexo III figura el certificado de registro de propiedad intelectual de la aplicación MetaDiSc 2.0.

Anexo I

```
# Title:          Area of the prediction ellipse
# Author:         Maria N Plana
# Date:          08/08/2020
# Objective:      Calculate the area of the prediction ellipse to quantify
between-study heterogeneity in meta-analysis of diagnostic test accuracy
studies.
#Acknowledgement: First part of the code has been modelled on previous
work by Christopher Partlett and Yemisi Takwoingi. We have adapted it to
perform the code of area of the prediction ellipse (second part).

#####

### Libraries

#####

install.packages("lme4")

install.packages("pracma")

library(lme4)

library(pracma)

#####

#### Read and prepare the data

#####

## Set your working directory to the appropriate drive where you saved
the .csv file ##

## Before performing analysis, data should be prepared as csv file as
follows

## tp: true positives
```

```

## fp: false positives

## tn: true negatives

## fn: false negatives

## study is the identifier for each study

## Read in the data from the .csv file

(X = read.csv("DATA.csv")) ## Your file name now is X

X$n1 <- X$tp+X$fn

X$n0 <- X$fp+X$tn

X$true1 <- X$tp

X$true0 <- X$tn

X$study <- 1: n    ## n has to be fixed previously as the total number of
studies included

## Reshape the data from wide to long format

Y = reshape (X, direction = "long", varying = list(c("n1" , "n0") , c(
"true1","true0" ) ) , timevar = "sens" , times = c(1,0) , v.names =
c("n","true") )

## Sort data by study to cluster the 2 records per study together

Y = Y[order(Y$id),]

Y$spec<- 1-Y$sens

#####

#### FIRST PART: Perform generalized linear mixed effects model glmer

#####

```

```

## Perform meta-analysis

(Me_Y= glmer(formula = cbind( true , n - true ) ~ 0 + sens + spec +
(0+sens + spec|study), data = Y , family = binomial ) )

(me_Y = summary(Me_Y))

## Extract the coefficients of the model

u = me_Y$coeff[1,1]      ## E(logit(sen))

v = me_Y$coeff[2,1]      ## E(logit(spe)

SELogitSen = me_Y$coeff[1,2]  ## standard error E(logit(sen))

SELogitSpe = me_Y$coeff[2,2]  ## standard error E(logit(spe))

covAB = me_Y$vcov[1,2]      ##Covariance between logits

VarLogitSen= me_Y$varcor$study[1,1]  ##Var(logit(sen))

VarLogitSpe = me_Y$varcor$study[2,2]  ##Var(logit(spe))

corAB = me_Y$varcor$study[1,2]

N=me_Y$ngrps

#####

#### SECOND PART: calculate the area of the prediction ellipse ####

#####

phi <- seq(0, 2*pi, len=5000)

SEPA <- sqrt(VarLogitSen+SELogitSen^2)

SEPB <- sqrt(VarLogitSpe+SELogitSpe^2)

sAB<-corAB*sqrt(VarLogitSen)*sqrt(VarLogitSpe)

rho<-(sAB+covAB)/(SEPA*SEPB)

```

```

## boundary constant of the ellipse

croot <- sqrt(2*qf(0.95,2,N-2))

a<-SEPA*croot

b<-SEPB*croot

sen <- exp(u)/(1+exp(u))

esp <- exp(v)/(1+exp(v))

y <- u + a * cos(phi + acos(rho))      ## Sensitivity (logit scale)

x <- v + b * cos(phi)                  ## Specificity (logit scale)

specificity <- 1-(exp(x)/(1+exp(x)))    ## ROC scale

sensitivity <- exp(y)/(1+exp(y))        ## ROC scale

Area=round(polyarea(specificity, sensitivity), 3)

## Graph the ROC plane and add the area

par(pty="s")

plot (specificity, sensitivity, type = "l", lwd = 5, lty = "dotted", col
= floor(runif(1,1,16))),

xlim = c(0, 1), ylim = c(0, 1), xlab = "1-Specificity", ylab =
"Sensitivity",

main = "SROC Plane", sub = "Prediction Ellipse")

points((1-esp), sen)

legend('bottomleft', legend=parse(text=sprintf('paste(Area, \' =
%s\') ',Area)), bty='n')

```

Anexo II

Meta-DiSc 2.0: An updated shiny web version for meta-analysis of test accuracy data

M Nieves Plana, Ingrid Arevalo-Rodriguez, Silvia Fernández, Javier Soto, Martin Fabregate, Marta Roqué, Teresa Pérez, Javier Zamora

ABSTRACT

This paper presents MetaDiSC 2.0, an updated version of the widely used MetaDiSc software. The novel Shiny-based application implements the more robust and current recommended statistical models to summarise results of test accuracy studies in diagnostic reviews, including the bivariate model, subgroup and metaregression analyses. In addition, the novel measures of variability implemented in the application represents an added value for improving the description and reporting of the heterogeneity. Here, we introduce a user interface description and show the metanalysis results using a data set of a published review. The web application is freely available at <https://ciberisciii.shinyapps.io/MetaDiSc2/>.

BACKGROUND

The evaluation of the role and properties of diagnostic tools have raised as a priority for health policy and global decision making, driven mainly by the development of new technologies for well-known diseases, as well as the emerging of new deleterious conditions affecting important population groups (1, 2). The diagnostic evidence about the accuracy of a test for detecting a target condition of interest can be appraised using

systematic approaches following standardised methodologies (3, 4). Briefly, the diagnostic studies focus on estimating the ability of the index tool to identify the subjects with or without the condition of interest (3); then, the evidence synthesis requires the knowledge of two quantities, the test sensitivity and specificity and the trade-off between them (5). The statistical approach to be used depends on the choice between estimating accuracy for a common threshold (i.e. an average operating point), or an expected curve across many thresholds (i.e. a summary ROC curve) (6-8), using commercial software packages with the analytical characteristics needed for the fitting of complex hierarchical models.

We recently found that the statistical synthesis of accuracy data was one of the methods more frequently omitted during the development of rapid reviews of diagnostic tests. Therefore, it would be a potential bottleneck for the extended evaluation of diagnostic tools (9). For several years, the free-to-use Meta-DiSc software was one of the statistical programs widely used for meta-analysis of diagnostic data, with more than 1300 citations in peer-reviewed scientific articles (10). This software implemented the statistical methods recommended at the time of its development, including the linear model proposed by Littenberg and Moses and the univariate I-square index to quantify the heterogeneity. Currently, hierarchical models are the recommended choice to overcome the limitations of previous statistical approaches (3). Due to these recent developments, it is imperative the update of MetaDiSc with more statistical rigorous methodologies, which account for the relationship between sensitivity and specificity, as well as the heterogeneity in test accuracy reviews.

Objective

To develop a web application (app) to summarise diagnostic test accuracy results. The tool updates the statistical methods included in MetaDiSc (10).

Software

We have developed a web app using the statistical software R (<https://www.r-project.org>). Shiny is an R package that allows web applications and interactive graphics from the integrated development environment for R, Rstudio (<http://shiny.rstudio.com/>). Shiny applications contain a user-interface definition and a server script.

The application uses the *glmer* function of the *lme4* package (11) for fitting generalised linear mixed effect models to produce summary points of sensitivity and specificity and a HSROC curve. Positive and negative likelihood ratios (LR) and Diagnostic Odds Ratio (DOR) estimates were derived from sensitivity and specificity parameters. Delta method in *msm* package (12) was used to compute the standard error of the derived parameters. Forest plots and ROC plots have been implemented through the functionalities of the *ggplot2* and *plotly* packages (13, 14). The application allows performing subgroup and metaregression analysis with the *lmeta* package (15). The area of the 95% prediction ellipse was calculated applying the *polyarea* function of the *pracma* package (16). A random univariate effects model run if there were few studies (less than four) to estimate between-study variances or when the hierarchical model did not converge. MetaDiSc 2.0 is freely available on the shinyapps server (<https://ciberisciii.shinyapps.io/MetaDiSc2/>).

Interface

The user interface gets an intuitive and easy to use design. Left lateral panel organises the app in four main menus: *file upload*, *graphical description*, *metanalysis* and *summary of findings*. In addition, a video demonstrating the complete application is provided in a final user guide menu.

File upload

Through the first menu, the app lets to import data from a .csv or .xlsx file. The file must contain the specified names (TP, FP, TN, FN) in columns for the number of true positives, false positives, true negatives and false negatives and a unique study identity (ID). File may also contain additional columns that will be identified for the program as covariates (Figure 1). Figures included in this demonstration using a data set of a systematic review published by our group (17).

Graphical description

The second menu of the app evaluates the heterogeneity scenario of our review through a graphical description before performing the metanalysis displaying forest-plots to represent results of individual studies (Figure 2) and a ROC plot representing the individual sensitivity and specificity and the error bars. Studies in the forest plot will be shown as ordered in the same way as in the excel file uploaded to the application.

Forest-plots and ROC plane plots are downloadable in .png and .svg formats. Users can also apply this graphical description for the categories of the different covariates included in the dataset.

Metanalysis menu

The app performs statistical pooling of sensitivities, specificities, likelihood ratios and diagnostic odds ratios as currently recommended method using a bivariate random-effects model (Figure 3). The option of the bivariate model tab shows the outputs in the following tabs: statistics, sROC curve, subgroup analysis and sensitivity analysis. In the statistics tab, users will find *i.* summary statistics as pooled results of accuracy indices (sensitivity and specificity, positive and negative likelihood ratios, diagnostic odds ratio and false positive rate); *ii.* Model parameters (logit sensitivity, logit specificity, standard errors, logits variances, covariance and correlation) for RevMan users to transfer MetaDisc Results into the Cochrane Review manager system; *iii.* Heterogeneity results (logit variances and median OR for sensitivity and specificity (18), I² bivariate (19) and the area of the 95% prediction ellipse (18)). The option of performing an HSROC model will be implemented in future updates. A univariate random-effects model tab is available. Although separate pooling is not the recommended method because fails to account for the trade-off between sensitivity and specificity, we have included this option because univariate models in some instances have a role to play in DTA reviews. This is the case for example when few studies are available and/or when it is difficult to estimate all parameters of the bivariate model, or the model has convergence problems or when our focus is on only one accuracy indices (*i.e.* sensitivity or specificity).

A figure of the ROC plane can be visualised and downloaded in the corresponding tab. Users can select different display options to include in the figure (summary point, confidence ellipse, prediction ellipse, sROC curve, points). Subgroup and Sensitivity

analyses tabs only will be active if the data set contains one or more covariates. In a dropdown menu, users can select the covariate for the analysis.

Summary of findings

It is also possible to calculate the absolute number of false positive and false negative test results observed in a population, to describe the absolute impact of a test in a population with a given prevalence. Users can download a figure that shows the outcomes (TP, FP, FN, TN) when a chosen number of people with a pre-test probability are tested using natural frequencies, which are generally easier to understand.

Results and discussion

As an example, we use a dataset taken from a systematic review evaluating the diagnostic accuracy of pulse oximetry as a screening method for detecting critical congenital heart defects (CCHD) in asymptomatic newborn infants (17). The metanalysis included nineteen studies. The overall sensitivity of pulse oximetry for detection of CCHD was 76.3% (95% confidence interval 69.5 to 82.0). Specificity was 99.9% (95% CI 99.7 to 99.9), with a false-positive rate of 0.14% (95% CI 0.07 to 0.22). We measured total between-study variability in sensitivity and in specificity through variances of the random effects for $\text{logit}(\text{sensitivity})$, $\text{logit}(\text{specificity})$, and their covariance of the bivariate model. We also provided confidence and prediction ellipses. MetaDiSc 2.0 allows now to extend the heterogeneity description, including the area of the 95% ellipse prediction (18), medians odds ratio for sensitivity and specificity (18) and I² bivariate (19) (Figure 3). For assessing potential sources of heterogeneity, we performed subgroup analyses and meta-regression using a METADAS SAS macro that estimates parameters for the model with SAS Proc NLMIXED (20). We tested our application to replicate the subgroup

analysis for the covariate test of timing, summary estimates of sensitivity and specificity of studies that performed screening after 24 hours were 73.6% (95% CI 62.8 to 82.1) and 99.9% (95% CI 99.9 to 100). For studies that performed screening within 24 hours, summary estimates of sensitivity and specificity were 79.5% (95% CI 70.0 to 86.6) and 99.6% (95%CI 99.1 to 99.8). The relative specificity for detection of CCHD was higher when newborn pulse oximetry was done after 24 hours from birth (Figure 4).

We have further evaluated the app, replicating the analysis of four systematic reviews published in the literature (21-24). The comparison of the numerical results obtained with MetaDiSc and the results obtained with other software (Metadas in SAS (20) and metandi in Stata (25) are shown in tables 1 and 2.

The philosophy of our project has been to provide users without advanced statistical expertise and faithful to the original version of MetaDiSc a friendly and updated application. It is available another shiny web application, MetaDTA (26), developed by the UK National Institute for Health Research (NIHR) Complex Review Support Unit in 2019, to conduct metanalyses for DTA reviews. MetaDiSc 2.0 has a strength over the MetaDTA software (26), the capacity to perform metaregression analyses and calculate additional measures to quantify heterogeneity such as the area of the 95% prediction ellipse showed in a SROC curve.

The development of this web application has been led by the Clinical Biostatistics Unit of the Ramon y Cajal Research Institute (IRYCIS). The unit has a broad experience in diagnostic test synthesis research focused to support informed decision-making about health. This application constitutes a collaborative innovative project for knowledge transfer between the Ramón y Cajal Institute, IRYCIS and the Complutense University in

Madrid. It is also supported for an intramural project funded by the Ramón y Cajal Hospital on "*Rapid diagnostic reviews for decision making in healthcare: analysis of critical points and software development*". The Biomedical Research Networking Center in Epidemiology and Public Health (CIBEREsp) funds the subscription to the shinyapps.io platform where the developed application is hosted.

Conclusion

We have developed an updated version of MetaDiSc for performing diagnostic test accuracy meta-analyses. All computational algorithms have been validated through comparison with different statistical tools and published meta-analyses. Finally, we evaluated the tool usability by external collaborators with large experience in evidence synthesis.

Availability and requirements

Project name: MetaDiSc 2.0

Project home page: <https://ciberisciii.shinyapps.io/MetaDiSc2/>

Archived version:

Operating system(s): Platform independent

Programming language: R

Other requirements: Internet browser

License: Freeware for academic use

Any restrictions to use by non-academics: None

REFERENCES

1. Knottnerus J, Frank B. The evidence base of clinical diagnosis. 2nd ed. London: BMJ Books; 2009.
2. Bossuyt PM. Testing COVID-19 tests faces methodological challenges. *Journal of Clinical Epidemiology*.
3. Deeks J, Bossuyt P, Gatsonis CE. *Cochrane handbook for systematic reviews of diagnostic test accuracy*. London: The Cochrane Collaboration; 2010.
4. Moher D, Stewart L, Shekelle P. All in the Family: systematic reviews, rapid reviews, scoping reviews, realist reviews, and more. *Syst Rev*. 2015;4:183.
5. 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.
6. 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. *J Clin Epidemiol*. 2005;58(10):982-90.
7. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *J Clin Epidemiol*. 2006;59(12):1331-2; author reply 2-3.
8. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med*. 2001;20(19):2865-84.
9. Arevalo-Rodriguez I, Steingart KR, Tricco AC, Nussbaumer-Streit B, Kaunelis D, Alonso-Coello P, et al. Current methods for development of rapid reviews about diagnostic tests: an international survey. *BMC Medical Research Methodology*. 2020;20(1):115.
10. Zamora J, Abraira V, Muriel A, Khan K, Coomarasamy A. Meta-DiSc: a software for meta-analysis of test accuracy data. *BMC Med Res Methodol*. 2006;6:31.
11. Bates D, Mächler M, Bolker B, S W. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*. 2015;67(1):1-48.
12. Jackson CH. Multi-State Models for Panel Data: The msm Package for R. *Journal of Statistical Software*. 2011;38(8):1-29.
13. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York 2016. Available from: <https://ggplot2.tidyverse.org>.
14. Sievert C. *Interactive Web-Based Data Visualization with R, plotly, and shiny*. CRC: Chapman and Hall; 2020. Available from: <https://plotly-r.com>.
15. Zeileis A, Hothorn T. Diagnostic Checking in Regression Relationships. *R News*. 2002;2(3):7-10.
16. Borchers HW. *pracma: Practical Numerical Math Functions 2021* [2.3.3:[Available from: <https://CRAN.R-project.org/package=pracma>.
17. Plana MN, Zamora J, Suresh G, Fernandez-Pineda L, Thangaratinam S, Ewer AK. Pulse oximetry screening for critical congenital heart defects. *Cochrane Database Syst Rev*. 2018;3(3):Cd011912.
18. Plana MN, Pérez T, Zamora J. New measures improved the reporting of heterogeneity in diagnostic test accuracy reviews: a metaepidemiological study. *J Clin Epidemiol*. 2021;131:101-12.
19. Zhou Y, Dendukuri N. Statistics for quantifying heterogeneity in univariate and bivariate meta-analyses of binary data: the case of meta-analyses of diagnostic accuracy. *Stat Med*. 2014;33(16):2701-17.
20. METADAS: A SAS macro for meta-analysis of diagnostic accuracy studies. User guide version 1.0 beta. December 2008. Available at: <http://srdta.cochrane.org/en/clib.html>. Accessed July 3, 2009.
21. Fahey MT, Irwig L, Macaskill P. Meta-analysis of Pap test accuracy. *Am J Epidemiol*. 1995;141(7):680-9.
22. Honest H, Bachmann LM, Gupta JK, Kleijnen J, Khan KS. Accuracy of cervicovaginal fetal fibronectin test in predicting risk of spontaneous preterm birth: systematic review. *Bmj*. 2002;325(7359):301.

23. Scheidler J, Hricak H, Yu KK, Subak L, Segal MR. Radiological evaluation of lymph node metastases in patients with cervical cancer. A meta-analysis. *Jama*. 1997;278(13):1096-101.
24. Verde PE. Meta-analysis of diagnostic test data: a bivariate Bayesian modeling approach. *Stat Med*. 2010;29(30):3088-102.
25. Harbord RM, Whiting P. metandi: Meta-analysis of diagnostic accuracy using hierarchical logistic regression. *Stata Journal*. 2009;9(2):211-29.
26. Freeman SC, Kerby CR, Patel A, Cooper NJ, Quinn T, Sutton AJ. Development of an interactive web-based tool to conduct and interrogate meta-analysis of diagnostic test accuracy studies: MetaDTA. *BMC Med Res Methodol*. 2019;19(1):81.

Figure 1 corresponde a Figura 6, Capítulo 4

Figure 2 corresponde a Figura 8, Capítulo 4

Figure 3 corresponde a Figura 9, Capítulo 4

Figure 4 corresponde a Figura 11, Capítulo 4

Table 1 corresponde a Tabla 6, Capítulo 4

Table 2 corresponde a Tabla 7, Capítulo 4

Anexo III

Figura como Anexo III el certificado de registro de propiedad intelectual de la aplicación MetaDiSc 2.0.

CERTIFICADO DE INSCRIPCIÓN
DE LA OBRA

METADISC 2.0

Prueba tecnológica

Realizada el **26 de octubre de 2021** a las **9:59 UTC**. A instancia de: **Fundación para la Investigación Biomédica del Hospital Universitario Ramón y Cajal** (User ID: 1609022147992) ("el poseedor"). Que demuestra la posesión del archivo digital aportado en la fecha indicada, a los efectos protección y defensa de derechos de propiedad intelectual sobre el principio "Prior tempore, potior iure".

Identificador de la prueba 2110269633104.

Constituida por los siguientes archivos digitales:



Fichero aportado por "el poseedor".

Nombre del fichero: MetaDiSc2.0_251021.zip

Tamaño del fichero: 2443464 bytes

Huella digital SHA1: a9250077a8910f63b72aac8397ad6b18ec245c04

Huella digital SHA256:

08f50a3e851d0bc185847721572d627218595d516acbb24506fcc80d708c5416

Huella digital SHA512:

030c2580818008ccdf8a6ebfb58ab752376061cf8bee818b4295d531d907c72132cdb454420c22741
184ada1905c1533112d60c7f9dec8d7c70338d9fc905cd5



Fichero de identificación¹ contiene las huellas digitales del fichero aportado.

Nombre del fichero: 2110269633104.txt



Fichero de garantía¹ sellados de tiempo del fichero de identificación adjuntos al documento.

Sellado de tiempo interno de Safe Creative: timestamp-safecreative.asn

Autoridad de sellado de tiempo:

C=ES,O=Safe Stamper,OU=Safe Stamper,CN=safestamper.com,E=dev@safecreative.org

Sellado de tiempo externo: timestamp-external.asn

Autoridad de sellado de tiempo:

C=ES,O=Firmaprofesional S.A.,organizationIdentifier=VATES-A62634068,
CN=FIRMAPROFESIONAL CLOUD QUALIFIED TSU - 2020

¹ Ficheros embebidos en este documento pdf.



METADISC 2.0

IDENTIFICADOR: 2110269633104

Safe Creative certifica que, a 26-oct-2021 a las 10:07 UTC, consta la siguiente información en el registro electrónico de propiedad intelectual

Prueba de la obra:

Realizada el 26-oct-2021 9:59 UTC a instancia de: Fundación para la Investigación Biomédica del Hospital Universitario Ramón y Cajal ("el poseedor")

Que demuestra la posesión del archivo digital aportado por el poseedor en la fecha indicada, a los efectos protección y defensa de derechos de propiedad intelectual sobre el principio "Prior tempore, potior iure".

Inscripciones declarativas de derechos:

Titular: Fundación para la Investigación Biomédica del Hospital Universitario Ramón y Cajal % 51.00	Derechos de: Titular	Fecha-hora: 26-oct-2021 9:59
Titular: Consortio Centro de Investigación Médica en Red - CIBERESP % 39.00	Derechos de: Titular	Fecha-hora: 26-oct-2021 9:59
Titular: Universidad Politécnica de Madrid - UPM % 10.00	Derechos de: Titular	Fecha-hora: 26-oct-2021 9:59
Titular: Javier Zamora romero - FIBIOHRC y CIBERESP % 50.00	Derechos de: Autor	Fecha-hora: 26-oct-2021 9:59
Titular: María Nieves Plana Farrás - FIBIOHRC y CIBERESP % 25.00	Derechos de: Autor	Fecha-hora: 26-oct-2021 9:59
Titular: Silvia Fernández García - FIBIOHRC % 10.00	Derechos de: Autor	Fecha-hora: 26-oct-2021 9:59
Titular: Javier Soto Pérez-Olivares - UPM % 10.00	Derechos de: Autor	Fecha-hora: 26-oct-2021 9:59
Titular: Ingrid Arévalo Rodríguez - FIBIOHRC % 5.00	Derechos de: Autor	Fecha-hora: 26-oct-2021 9:59



REFERENCIAS

BIBLIOGRÁFICAS

Arends LR, Hamza TH, Van Houwelingen JC, Heijenbrok-Kal MH, Hunink MGM, Stijnen T. Bivariate random effects meta-analysis of ROC curves. *Medical Decision Making*. 2008;28(5):621–38.

Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Statistics in medicine*. 1995;14(4):395-411.

Borchers HW. *pracma: Practical Numerical Math Functions* 2021. Available from: <https://CRAN.R-project.org/package=pracma>.

Borm GF, Lemmers O, Fransen J, Donders R. The evidence provided by a single trial is less reliable than its statistical analysis suggests. *Journal of clinical epidemiology*. 2009;62(7):711-5.e1.

Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ (Clinical research ed)*. 2006;332(7549):1089-92.

- Campbell JM, Klugar M, Ding S, Carmody DP, Hakonsen SJ, Jadotte YT, et al.
Diagnostic test accuracy: methods for systematic review and meta-analysis.
International journal of evidence-based healthcare. 2015;13(3):154-62.
- Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. Journal of clinical epidemiology.
2006;59(12):1331-2; author reply 2-3.
- Cochrane AL. 1972. *Effectiveness and efficiency: Random reflections on health services*. Nuffield Trust.
- Dahabreh IJ, Trikalinos TA, Lau J, Schmid C. An Empirical Assessment of Bivariate Methods for Meta-Analysis of Test Accuracy. Rockville (MD): Agency for Healthcare Research and Quality (US); 2012 2012 Nov.
- Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. BMJ (Clinical research ed). 2001;323(7305):157-62.
- DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled clinical trials.
1986;7(3):177-88.
- Bates D, Maechler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software. 2015; 67(1), 1-48.
- Fahey MT, Irwig L, Macaskill P. Meta-analysis of Pap test accuracy. Am J Epidemiol.
1995 Apr 1;141(7):680-9.
- Fisher RA. 1932. *Statistical Methods for Research Workers*. London.

- Flicker L, Logiudice D, Carlin JB, Ames D. The predictive value of dementia screening instruments in clinical populations. *International journal of geriatric psychiatry*. 1997;12(2):203-9.
- Foxlee N, Stone JC, Doi SA. A comparison of univariate and bivariate models in meta-analysis of diagnostic accuracy studies. *Int J Evid Based Healthc*. 2015 Mar;13(1):28-34.
- Freedman LS. Evaluating and comparing imaging techniques: a review and classification of study designs. *The British journal of radiology*. 1987;60(719):1071-81.
- Freeman SC, Kerby CR, Patel A, Cooper NJ, Quinn T, Sutton AJ. Development of an interactive web-based tool to conduct and interrogate meta-analysis of diagnostic test accuracy studies: MetaDTA. *BMC Med Res Methodol*. 2019 Apr 18;19(1):81.
- Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Medical decision making : an international journal of the Society for Medical Decision Making*. 1991;11(2):88-94.
- Furukawa TA, Streiner DL, Hori S. Discrepancies among megatrials. *Journal of clinical epidemiology*. 2000;53(12):1193-9.
- Gatsonis C. Design of evaluations of imaging technologies: development of a paradigm. *Academic radiology*. 2000;7(9):681-3.
- Gil MM, Accurti V, Santacruz B, Plana MN, Nicolaides KH. Analysis of cell-free DNA in maternal blood in screening for aneuploidies: updated meta-analysis. *Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology*. 2017;50(3):302-14.

- Gil MM, Galeva S, Jani J, Konstantinidou L, Akolekar R, Plana MN, et al. Screening for trisomies by cfDNA testing of maternal blood in twin pregnancy: update of The Fetal Medicine Foundation results and meta-analysis. *Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology*. 2019;53(6):734-42.
- Glas AS, Lijmer JG, Prins MH, Bossel GJ, Bossuyt PMM. The diagnostic odds ratio: a single indicator of test performance. *Journal of clinical epidemiology*. 2003;56(11):1129-35.
- Glass GV. 1976. «Primary, secondary and meta-analysis of research». *Educ Res* 5: 3-9.
- Goehring C, Perrier A, Morabia A. Spectrum bias: a quantitative and graphical analysis of the variability of medical diagnostic test performance. *Statistics in medicine*. 2004;23(1):125-35.
- Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*. 1986;134(6):587-94.
- Hamza TH, van Houwelingen HC, Stijnen T. The binomial distribution of meta-analysis was preferred to model within-study variability. *Journal of clinical epidemiology*. 2008;61(1):41-51.
- Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.

- Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JAC. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics*. 2007;8(2):239-51.
- Harbord RM, Whiting P, Sterne JAC, Egger M, Deeks JJ, Shang A, et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *Journal of Clinical Epidemiology*. 2008;61(11):1095-103
- Harbord RM, Whiting P. metandi: Meta-analysis of diagnostic accuracy using hierarchical logistic regression. *Stata Journal*. 2009;9(2):211-29.
- Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Statistics in medicine*. 1998;17(8):841-56.
- Haynes B. Of studies, syntheses, synopses, summaries, and systems: the "5S" evolution of information services for evidence-based healthcare decisions. *Evidence-based nursing*. 2007;10(1):6-7.
- Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*. 2002;21(11):1539-58.
- Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ (Clinical research ed)*. 2003;327(7414):557-60.
- Honest H, Bachmann LM, Gupta JK, Kleijnen J, Khan KS. Accuracy of cervicovaginal fetal fibronectin test in predicting risk of spontaneous preterm birth: systematic review. *British Medical Journal* 2002; 325(7359):301–304.
- Houn F, Bright RA, Bushar HF, Croft BY, Finder CA, Gohagan JK, et al. Study design in the evaluation of breast cancer imaging technologies. *Academic radiology*. 2000;7(9):684-92.

- Ioannidis JP, Cappelleri JC, Lau J. Issues in comparisons between meta-analyses and large trials. *JAMA*. 1998;279(14):1089-93.
- Jackson CH (2011). "Multi-State Models for Panel Data: The msm Package for R." *Journal of Statistical Software*, 38(8), 1–29.
- Jackson D. The power of the standard test for the presence of heterogeneity in meta-analysis. *Statistics in medicine*. 2006;25(15):2688-99.
- Jackson D, White IR, Riley RD. Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. *Statistics in Medicine*. 2012;31(29):3805-20.
- Kassirer JP. Diagnostic reasoning. *Annals of internal medicine*. 1989;110(11):893-900.
- Kassirer JP. Our stubborn quest for diagnostic certainty. A cause of excessive testing. *The New England journal of medicine*. 1989;320(22):1489-91.
- Knottnerus JA, van Weel C, Muris JWM. Evaluation of diagnostic procedures. *BMJ (Clinical research ed)*. 2002;324(7335):477-80.
- Kobberling J TH, Windeler J, eds. Memorandum for the evaluation of diagnostic measures. *Journal of clinical chemistry and clinical biochemistry Zeitschrift für klinische Chemie und klinische Biochemie*. 1990;28(12):873-9.
- Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Annals of internal medicine*. 1992;117(2):135-40.
- Ledley RS LL. Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason. *Science*. 1959;130:9-21.

- Leeflang MMG, Bossuyt PMM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *Journal of clinical epidemiology*. 2009;62(1):5-12.
- Leeflang MMG, Deeks JJ, Rutjes AWS, Reitsma JB, Bossuyt PMM. Bivariate meta-analysis of predictive values of diagnostic tests can be an alternative to bivariate meta-analysis of sensitivity and specificity. *Journal of Clinical Epidemiology*. 2012;65(10):1088-97.
- LeLorier J, Grégoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *The New England journal of medicine*. 1997;337(8):536-42.
- Lijmer JG, Mol BW, Heisterkamp S, Bossel 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.
- Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Statistics in medicine*. 2002;21(11):1525–37.
- Lijmer JG, Leeflang M, Bossuyt PM. *Proposals for a Phased Evaluation of Medical Tests. Medical Tests-White Paper Series*. Rockville (MD): Agency for Healthcare Research and Quality (US); 2009.
- Loop JW, Lusted LE. American College of Radiology Diagnostic Efficacy Studies. *AJR American journal of roentgenology*. 1978;131(1):173-9.
- Lusted LB. Decision-making studies in patient management. *The New England journal of medicine*. 1971;284(8):416-24.

- Ma X, Nie L, Cole SR, Chu H. Statistical methods for multivariate meta-analysis of diagnostic tests: An overview and tutorial. *Statistical methods in medical research*. 2016;25(4):1596-619.
- Macaskill P, GGC, 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. Available from: <http://srdta.cochrane.org/>. 2010.
- Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Statistics in medicine*. 1993;12(14):1293-316.
- Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Annals of internal medicine*. 2002;137(7):598-602.
- Nordic Cochrane Centre, The Cochrane Collaboration Review Manager (RevMan). Version 5.4. Copenhagen: Nordic Cochrane Centre, The Cochrane Collaboration, 2020.
- Nyaga VN, Arbyn M, Aerts M. Metaprop: a Stata command to perform meta-analysis of binomial data. *Arch Public Health*. 2014 Nov 10;72(1):39. doi: 10.1186/2049-3258-72-39.
- O'Connor PW, Tansay CM, Detsky AS, Mushlin AI, Kucharczyk W. The effect of spectrum bias on the utility of magnetic resonance imaging and evoked potentials in the diagnosis of suspected multiple sclerosis. *Neurology*. 1996;47(1):140-4.

Pearson K. Report on Certain Enteric Fever Inoculation Statistics. Br Med J. 1904 Nov 5;2(2288):1243-6.

Pepe MS. Evaluating technologies for classification and prediction in medicine. Statistics in medicine. 2005;24(24):3687-96.

Plana MN, Zamora J, Suresh G, Fernandez-Pineda L, Thangaratinam S, Ewer AK. Pulse oximetry screening for critical congenital heart defects. The Cochrane database of systematic reviews. 2018;3:CD011912.

Plana MN, Pérez T, Zamora J. New measures improved the reporting of heterogeneity in diagnostic test accuracy reviews: a metaepidemiological study. J Clin Epidemiol. 2021 Mar;131:101-112.

R Core Team (2020). R: A language and environment for statistical computing. <https://www.R-project.org/>. R foundation for statistical Computing, Vienna, Austria.

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. Journal of Clinical Epidemiology. 2005;58(10):982-90.

Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I² in assessing heterogeneity may mislead. BMC Medical Research Methodology. 2008;8(1).

Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Statistics in medicine. 2001;20(19):2865–84.

- Rutjes AWS, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PMM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ: Canadian Medical Association journal = journal de l'Association medicale canadienne*. 2006;174(4):469-76.
- Sackett DL, Haynes RB. The architecture of diagnostic research. *BMJ (Clinical research ed)*. 2002;324(7336):539-41.
- Scheidler J, Hricak H, Yu KK, Subak L, Segal MR. Radiological evaluation of lymph node metastases in patients with cervical cancer. A meta-analysis. *JAMA*. 1997 Oct 1;278(13):1096-101.
- van der Schouw YT, Verbeek AL, Ruijs SH. Guidelines for the assessment of new diagnostic tests. *Invest Radiol*. 1995 Jun;30(6):334-40.
- Schuetz GM, Zacharopoulou NM, Schlattmann P, Dewey M. Meta-analysis: noninvasive coronary angiography using computed tomography versus magnetic resonance imaging. *Ann Intern Med*. 2010 Feb 2;152(3):167-77.
- Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, Williams JW Jr, Kunz R, Craig J, Montori VM, Bossuyt P, Guyatt GH; GRADE Working Group. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ*. 2008 May 17;336(7653):1106-10.
- StataCorp. 2021. Stata Statistical Software: Release 17. College Station, TX: StataCorp LLC.
- Sterne JA, Jüni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Statistics in medicine*. 2002;21(11):1513–24.

- Stirzaker D. 2003. *Elementary Probability Cambridge, UK: Cambridge University Press; 22nd edn; 54-55.*
- Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in medicine.* 2004;23(9):1351-75.
- Takwoingi Y, Deeks JJ. MetaDAS: A SAS macro for meta-analysis of diagnostic accuracy studies. User Guide Version 1.3. 2010 July. Available from: <http://srdta.cochrane.org/>
- Taube SE, Jacobson JW, Lively TG. Cancer diagnostics: decision criteria for marker utilization in the clinic. *American journal of pharmacogenomics : genomics-related research in drug development and clinical practice.* 2005;5(6):357-64.
- Van den Bruel A, Cleemput I, Aertgeerts B, Ramaekers D, Buntinx F. The evaluation of diagnostic tests: evidence on technical and diagnostic accuracy, impact on patient outcome and cost-effectiveness is needed. *Journal of clinical epidemiology.* 2007;60(11):1116-22.
- Van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. *Statistics in medicine.* 1993;12(24):2273-84.
- Verde PE. Meta-analysis of diagnostic test data: a bivariate Bayesian modeling approach. *Statistics in Medicine* 2010;29(30):3088–3102.
- White IR. 2011. «Multivariate random-effect meta-regression: updates to mvmeta». *The Stata Journal* 11: 255-70.

- Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol*. 2003 Nov 10;3:25.
- Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine*. 2011;155(8):529-36.
- Willis BH. Spectrum bias--why clinicians need to be cautious when applying diagnostic test studies. *Family practice*. 2008;25(5):390-6.
- Xiao AY, Tan ML, Plana MN, Yadav D, Zamora J, Petrov MS. The Use of International Classification of Diseases Codes to Identify Patients with Pancreatitis: A Systematic Review and Meta-analysis of Diagnostic Accuracy Studies. *Clinical and translational gastroenterology*. 2018;9(10):191.
- Yates F. CW. The analysis of groups of experiments. 1938;28:556-80. *Journal of Agricultural Science*. 1938;28:556-80.
- Zamora J, Abaira V, Muriel A, Khan K, Coomarasamy A. Meta-DiSc: a software for meta-analysis of test accuracy data. *BMC Medical Research Methodology*. 2006;6(1).
- Zeileis A, Hothorn T (2002). "Diagnostic Checking in Regression Relationships." *R News*, 2(3), 7–10. <https://CRAN.R-project.org/doc/Rnews/>.
- Zhou Y, Dendukuri N. Statistics for quantifying heterogeneity in univariate and bivariate meta-analyses of binary data: the case of meta-analyses of diagnostic accuracy. *Stat Med*. 2014 Jul 20;33(16):2701-17.

Zweig MH, Robertson EA. Why we need better test evaluations. Clin Chem. 1982 Jun;28(6):1272-6.

