Revisión Sistemática y Metaanálisis de Estudios de Diagnóstico y Pronóstico: una Guía

Marcos R. de Sousa[1,2] e Antonio Luiz P. Ribeiro[1,2]

Serviço de Cardiologia do Hospital das Clínicas da Universidade Federal de Minas Gerais (UFMG)[1], Programa de Pós-Graduação (Doutorado) em Clínica Médica da Faculdade de Medicina da Universidade Federal de Minas Gerais (UFMG)[2], Belo Horizonte, MG - Brasil

Resumen

Las revisiones sistemáticas con metaanálisis de estudios de exámenes diagnósticos o de factores pronósticos son herramientas de investigación aún en fase de desarrollo. El presente trabajo tiene como objetivo describir la metodología de revisión sistemática y de metaanálisis de esta clase de estudios, paso a paso. Se hizo la revisión de la literatura sobre el tema, compilándose las recomendaciones con objeto de organizarse el texto en:

a) Introducción;

b) Detalle de los ocho pasos a seguirse;

c) Formato de publicación de la revisión sistemática con metaanálisis; y

d) Conclusión.

Se describieron los métodos de revisión sistemática con detalle, y se analizó de manera crítica los métodos de compilación estadística de los resultados, con énfasis en la utilización de la curva *Summary Receiver Operator Characteristic*. Se suministró referencia para los detalles de cada técnica estadística utilizada en el metaanálisis. Concluimos que las revisiones sistemáticas con metaanálisis de exámenes diagnósticos o de factores pronósticos son valiosas en la compilación de datos de diversos estudios sobre el mismo tema, reduciéndose sesgos y aumentándose el poder estadístico de la investigación primaria.

Introducción

Se denomina revisión sistemática de la literatura la revisión planeada de la literatura científica, que usa métodos sistemáticos para identificarse, seleccionarse y evaluarse críticamente estudios relevantes sobre una cuestión claramente planteada. El objetivo de la sistematización es reducirse los posibles sesgos que pudieran ocurrir en una revisión no-sistemática[1], tanto los sesgos observados en el formato de revisión de la literatura y en la selección de los artículos como aquellos detectados por la evaluación crítica de cada estudio. El metaanálisis es el método estadístico utilizado en la revisión sistemática para integrar los resultados de los estudios incluidos y aumentar el poder estadístico de la investigación primária[2]. Aunque existan metaanálisis publicadas en 1904 y 1955[2], el término metaanálisis ha sido utilizado por primera vez por Glass, en 1976, para indicar el análisis estadístico de los resultados de los análisis de muchos estudios individuales, con el propósito de integrarse los hallazgos[3]. A veces, el término metaanálisis se utiliza como sinónimo de revisión sistemática, cuando la revisión incluye metaanálisis[4]. Aunque en portugués se utiliza la palabra *metanálise* (*segmentación no etimológica* en español) como sinónimo de *meta-análise* (*metaanálisis* o *meta-análisis* en español), ambas significan cosas distintas. En el contexto de la Lingüística Histórica de los estudios de Lengua Portuguesa, la palabra *metanálise* está referida a la segmentación no etimológica de un vocablo, una locución o un enunciado, que han sido interpretados por los hablantes de la lengua de modo distinto ante su origen.

Los estudios de pruebas diagnósticas y pronósticas son antiguos en la literatura médica, pero su aplicación de metodología estadística a las pruebas diagnósticas y de evaluación pronóstica se desarrolló tras su aplicación en estudios terapéuticos[5]. De igual manera, la estandarización del formato de publicación de los estudios diagnósticos[6] tuvo lugar casi una década luego del mismo proceso haber ocurrido en los estudios terapéuticos[7]. Los principales conceptos estadísticos esenciales al estudio de los métodos de diagnóstico y evaluación pronóstica están detallados en la Figura 1 y se utilizaron en el texto que sigue a continuación.

Existen diferencias importantes entre los metaanálisis de estudios de intervención terapéutica –para los cuales hay manuales ya publicados–, y el metaanálisis de factores pronósticos o de examen diagnósticos, más recientes y menos estandarizados en comparación con los otros[8]. Los metaanálisis de estudios que comparan intervenciones o tratamientos generalmente incluyen estudios aleatorizados, con dos grupos semejantes, además de evaluar la misma intervención, en general comparada con placebo o con tratamiento convencional. Ya los metaanálisis de estudios de factores pronósticos o de exámenes diagnósticos enfrentan retos distintos, como puntos de corte diferentes para el resultado positivo o negativo de un examen o evaluación de exámenes que han sido realizados en estudios prospectivos para estudio de intervenciones terapéuticas. En la década de 1990, surgieron nuevas técnicas estadísticas de combinación de estudios de exámenes diagnósticos[9-11]. Desde 1994, cuando se

Palabras-clave

Metodología, diagnóstico, pronóstico, metaanálisis, literatura de revisión.

Correspondencia: Marcos R. de Sousa •
Rua Aristides Duarte, 39/601 - Barroca - 30410-040 - Belo horizonte, MG - Brasil
E-mail: mrsousa@cardiol.br
Artículo enviado el 01/01/08; revisado recibido el 23/01/08; aceptado el 14/02/08.

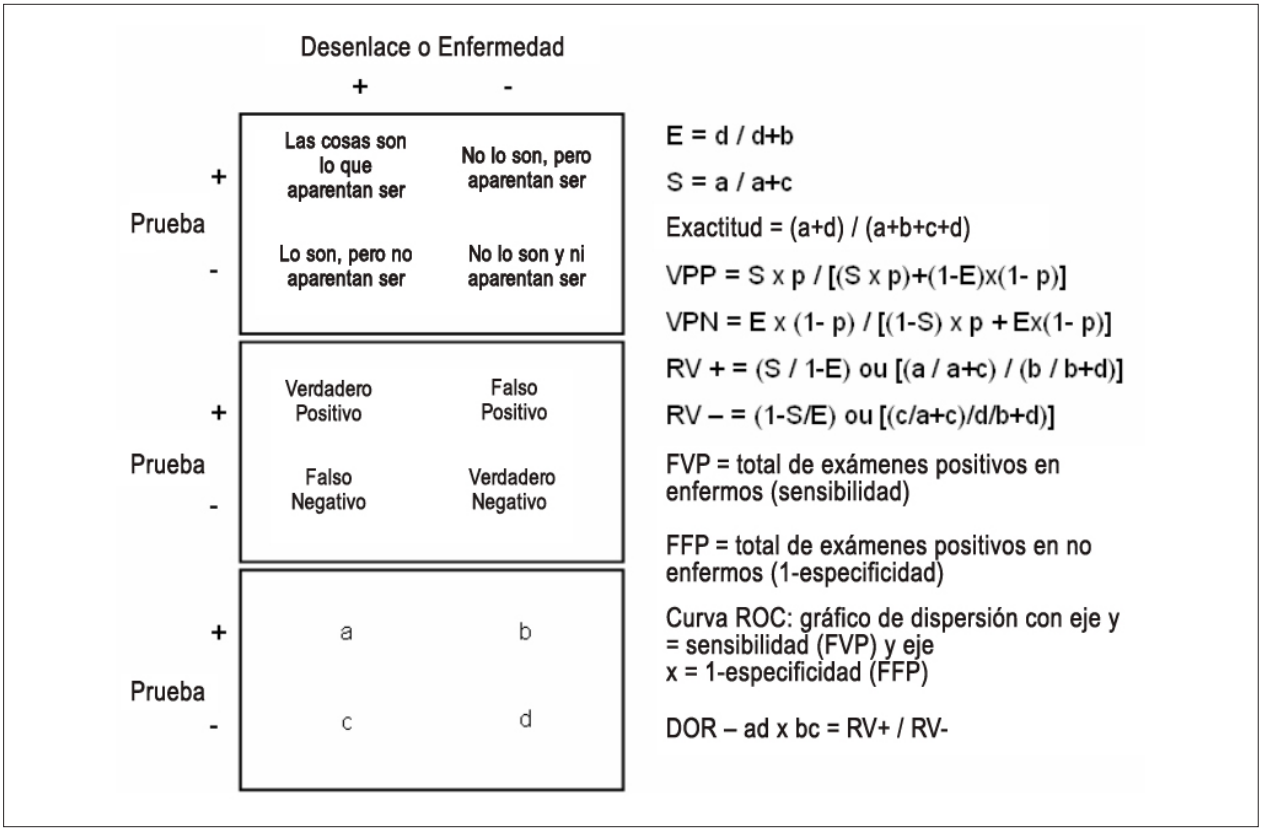

Figura 1 - *Conceptos y medidas de desempeño de una prueba diagnóstica o pronóstica. Toda decisión clínica está basada, conscientemente o no, en probabilidad. Pruebas diagnósticas se pueden utilizar para evaluar tanto la presencia o ausencia de enfermedad, como la gravedad del cuadro clínico, y también para monitorear la respuesta a una intervención y para estimarse el pronóstico. a - Número de resultados verdadero-positivos (VP); b - Número de resultados falso-positivos (FP); c - Número de resultados falso-negativos (FN); d - Número de resultados verdadero-negativos (VN). Especificidad (E) - probabilidad de examen negativo en los no enfermos; Sensibilidad (S) - probabilidad de examen positivo en los enfermos; Exactitud del examen - proporción de resultados correctos; Valor predictivo negativo (VPN) - probabilidad de no existir la enfermedad en personas con prueba negativa; Valor predictivo positivo (VPP) - probabilidad de enfermedad en personas con prueba positiva; P en la fórmula de los valores predictivos significa prevalencia de la enfermedad en la población; Razón de verosimilitud de una prueba positiva (RV+) - mide la mayor probabilidad de la prueba ser positiva en los enfermos en comparación a los no enfermos; Razón de verosimilitud de una prueba negativa (RV-) - mide la mayor probabilidad de la prueba ser negativa en los enfermos en comparación a los no enfermos; Fracción de verdadero-positivos (FVP) - total de exámenes positivos en enfermos; Fracción de falso-positivos (FFP) - total de exámenes positivos en no enfermos; Curva ROC – curva* Receiver Operator Characteristic *- se la utiliza para compararse un examen con el resultado continuo en relación con un "patrón oro" o un desenlace. Se trata de un gráfico de dispersión con eje y = sensibilidad (FVP) y eje x = 1-especificidad (FFP). El punto del gráfico en la esquina superior más alta es el punto ideal de desempeño del examen, con sensibilidad = 100% y especificidad = 100%; DOR - razón de ventajas de diagnóstico, difícil de interpretarse clínicamente, pero muy útil desde el punto de vista estadístico para evaluarse el desempeño global de la prueba y también muy útil en el metaanálisis, porque ayuda en la construcción de la curva sROC (summary ROC, resultados agrupados de diversos estudios en la forma de curva ROC).*

Tabla 1 – Pasos para la revisión sistemática y el metaanálisis[12,19]

1. Definir claramente la cuestión a plantearse.
2. Buscar en diversas fuentes todos los estudios confiables que tratan del tema.
3. Desde criterios claros de inclusión y de exclusión, seleccionar los estudios y evaluar su calidad.
4.Recolectar los datos de cada estudio y presentarlos de forma clara.
5. Evaluar la heterogeneidad entre los estudios.
6. Calcular los resultados de cada estudio (y combinarlos, si fuera apropiado), estimándose el desempeño diagnóstico.
7. Evaluar el efecto de la variación de la validez de cada estudio en las estimativas de desempeño diagnóstico.
8. Interpretar los resultados, evaluándose el límite de generalización de la revisión y/o el metaanálisis, de acuerdo con las características de los pacientes.

publicó la directriz para el metaanálisis de estudios de exámenes diagnósticos[12], surgieron varias publicaciones diferentes con críticas y proposiciones en aspectos específicos de cada etapa del proceso. El uso de metaanálisis para exámenes diagnósticos y pronósticos está todavía en fase de desarrollo, pero viene ganando cada vez más importancia[1,3,8].

El objetivo de esta revisión es detallar la literatura disponible y definir una guía para la realización, paso a paso, de revisión sistemática y, si fuera apropiado, de metaanálisis de estudios diagnósticos y pronósticos. A continuación, revisaremos los pasos necesarios, detallados en la Tabla 1.

Definir claramente la cuestión a plantearse.

Especificar claramente la prueba diagnóstica o pronóstica en cuestión, la enfermedad en estudio, como se realizó el diagnóstico y en que contexto se planteó la cuestión. De

modo general, el examen en cuestión se lo compara a un patrón-oro para el diagnóstico de la enfermedad. Sin embargo, los métodos estadísticos utilizados para el metaanálisis de exámenes diagnósticos pueden tener una aplicación bien más amplia[13]. En los casos de exámenes pronósticos, se los puede avaluar mediante el desenlace muerte, la respuesta al tratamiento o, teóricamente, cualquier variable dicotómica de interés que se refiera al pronóstico de largo plazo[13]. Además de ello, se debe también aclarar si se realizará una comparación de pruebas[12].

Buscar en diversas fuentes todos los estudios confiables que tratan del tema

Se recomienda ampliarse al máximo las fuentes de búsqueda. Buscar en publicaciones gubernamentales, comisiones de ética, resúmenes en anales de congresos, tesis, y también averiguar en bases electrónicas (MEDLINE, EMBASE, LILACS etc.)[3]. Además de las fuentes de búsqueda de estudios, es importante consultar la biblioteca de revisiones Cochrane (www.bvs.br) para verificar si dicha revisión ya está hecha. Aunque la intención no sea de utilizar datos no publicados, el contacto con investigadores de estudios en marcha o no publicados puede ser importante[3]. Para la averiguación en la base de datos MEDLINE, vale especificar claramente el procedimiento de búsqueda en la literatura con términos de búsqueda citados, tomando en consideración criterios de inclusión y exclusión[12]. La forma de investigar con términos de búsqueda puede interferir en la sensibilidad de la revisión sistemática[14]. Es importante averiguar términos descriptores *MeSH* (*Medical Subject Headings*, vocabulario en Lengua Inglesa usado para indexar artículos, disponible en: http://

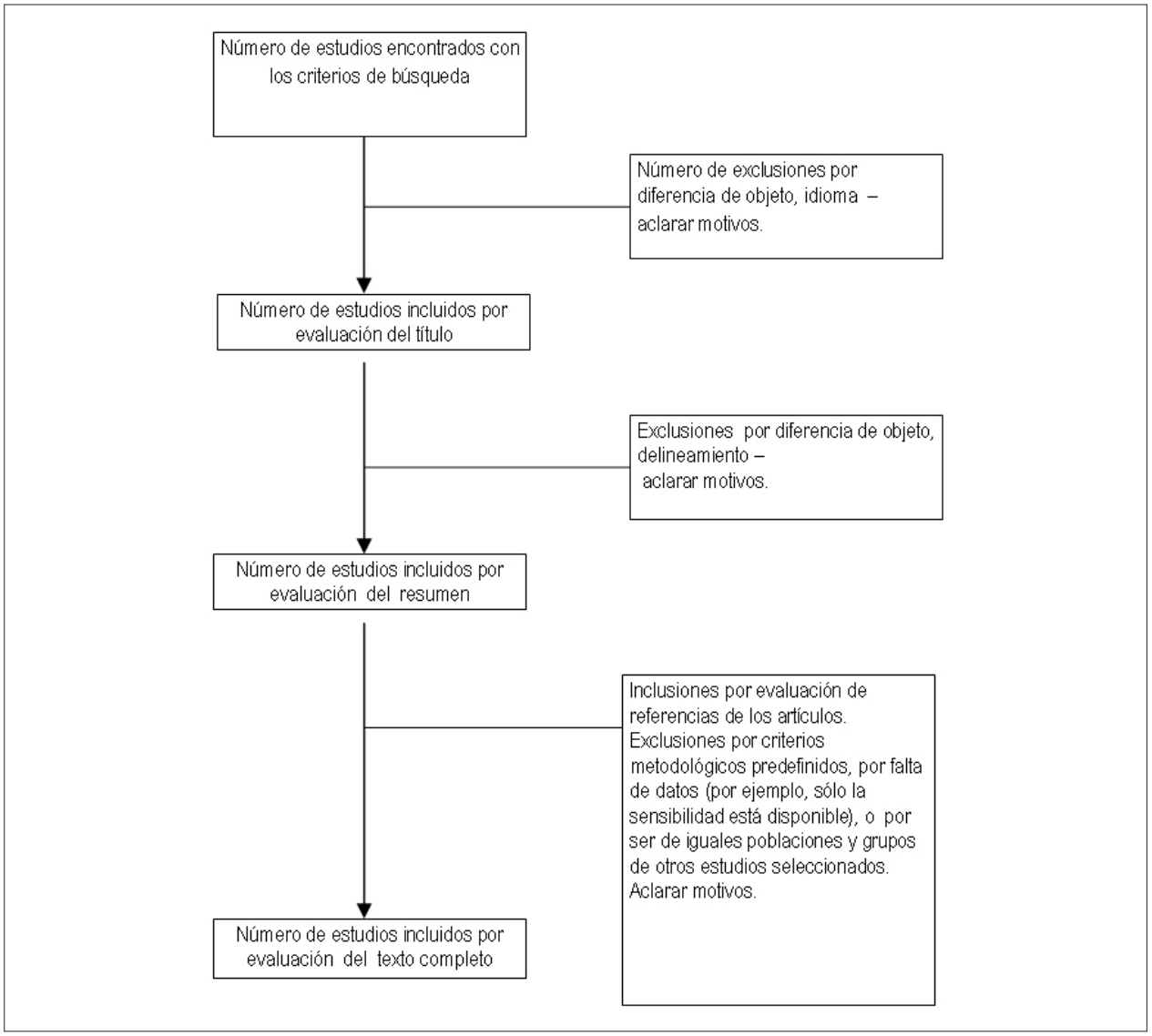

Figura 2 - Proceso de búsqueda y selección de artículos[6].

www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=mesh) para auxiliar en la investigación. Se obtiene la mejor estrategia, en general, mediante la combinación de los términos *MeSH* utilizados con palabras textuales[14]. Para estudios de marcadores pronósticos, se sugiere aumentar la sensibilidad por medio de la asociación del tema de investigación con los descriptores: (incidence[MeSH] OR mortality[MeSH] OR follow-up studies[MeSH] OR prognos*[Text Word] OR predict*[Text Word] OR course*[Text Word])[14]. Es importante también dejar claro como fue el proceso de revisión de la literatura (fig. 2).

El sesgo de publicación es la tendencia de se publicar más frecuentemente los estudios con resultados positivos en comparación a los de resultados negativos. Ello sucede sobre todo en revistas de mayor impacto y de lengua inglesa[3]. Dicho hecho se debe tanto al autor como al editor que presentan resistencias en publicar estudios con resultados negativos. Estudios con muestras muy reducidas presentan mayor probabilidad de sesgo de publicación, motivo por el que algunos autores preconizan la exclusión de ellos[3,15]. Para disminuirse la posibilidad de sesgo de publicación, se deben ampliar al máximo las fuentes de búsqueda. Un método de búsqueda de estudios de intervención terapéutica –difícil de aplicarse en estudios de exámenes diagnósticos o pronósticos– consiste en averiguar la existencia de estudios registrados, pero no ya publicados, en comisiones de ética o en registros gubernamentales (por ejemplo, www.clinicaltrials.gov); buscando así por sus resultados[16]. Otra fuente que se puede utilizar son los resúmenes en anales de congresos, en los que se puede reconocerse los trabajos presentados y los todavía no publicados[17].

Una forma estadística de evaluarse el sesgo de publicación es mediante la utilización del gráfico de dispersión en embudo, embudo invertido o "árbol de navidad" (*funnel plot*)[16]. Ese gráfico tiene como premisa que el tamaño de la muestra es el más fuerte correlato del sesgo de publicación[3,15] (Gráf. 1) Se puede evaluar la simetría de modo objetivo a través de métodos estadísticos[16]. La apariencia asimétrica sugiere que hubo sesgo de publicación, con tendencia de la distribución de las razones de ventajas para una parte, en general la parte "más positiva", ya que los "negativos" no habrían sido publicados.

Seleccionar estudios por medio de criterios claros de inclusión y de exclusión, evaluándose así la calidad de los estudios

Idealmente, dos investigadores deben buscar y evaluar los estudios de forma independiente. Se puede utilizar la prueba estadística Kappa para evaluarse la concordancia entre los dos investigadores. Explicar como las discordancias entre ellos han sido solucionadas, lo que en general se hace mediante un acuerdo y con base en la opinión de un tercero investigador experimentado. Listar claramente las

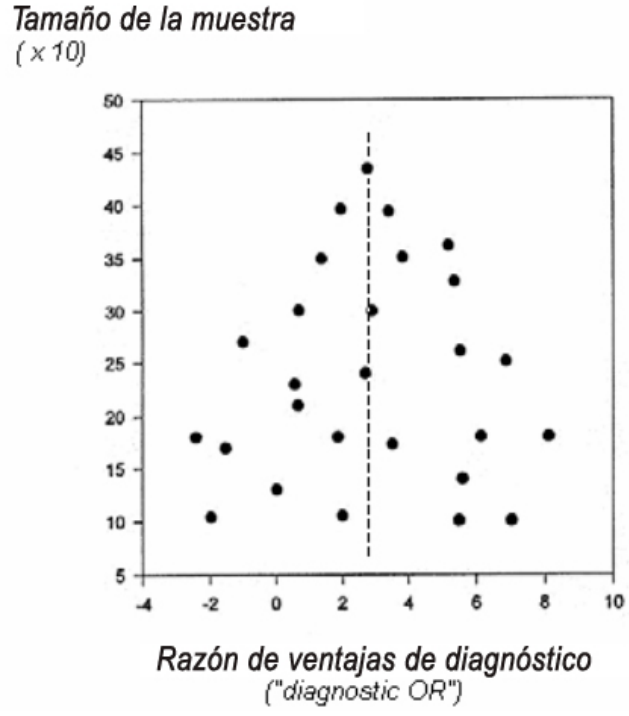

Gráfico 1 -Ejemplo de gráfico de dispersión en embodo invertido o "árbol de navidad". Cada punto del gráfico representa un estudio con su razón de ventajas de diagnóstico y el tamaño de la muestra. La forma de embodo simétrico sugiere la ausencia de sesgo de publicación (en ese caso, un embodo invertido, denominado asimismo de "árbol de navidad"). Se puede presentarlo invirtiéndose el orden de los valores del eje vertical, así quedaría en forma de embodo. Una asimetría con concentración de estudios en el lado derecho (el lado de valores mayores de razón de ventajas de diagnóstico) sugiere un sesgo de publicación, con menos estudios negativos publicados.

características de cada estudio primario así como los resultados de cada uno de ellos[12]. El trabajo puede ser facilitado si el estudio ha sido publicado de acuerdo con la estandarización STARD (*Standards for Reporting of Diagnostic Accuracy* - www.consort-statement.org/stardstatement.htm); formulado para garantizar más clareza, rigor metodológico y posibilidad de comparación de los estudios de métodos diagnósticos[18]. Los requisitos de calidad se deben conferir (tab. 2).

Recolectar los datos de cada estudio y presentarlos de forma clara

Tablas de comparación de los estudios son muy útiles para averiguarse las diferencias clínicas y metodológicas entre los estudios (tab. 2). Comparar estudios evaluando la distribución por edad, sexo, forma de diagnóstico o selección de pacientes, covariables relevantes, tiempo de seguimiento y tamaño de la muestra[6]. Para obtención de los datos a combinarse, recolectar los valores originales de falso y verdadero-positivos, falso y verdadero-negativos. Eventualmente, esos datos pueden estimarse a partir de valores de sensibilidad, especificidad y de los valores de ocurrencia del desenlace o examen de referencia[12].

Evaluar la heterogeneidad entre los estudios

Antes de realizar la combinación estadística (metaanálisis) de los estudios, es fundamental evaluar la heterogeneidad entre ellos. Es importante determinar[19]:

- ¿Por qué los resultados cambiaron entre los estudios?
- ¿El cambio fue al azar?
- ¿El cambio ha sido utilizado en función de diferencias metodológicas?

Para contestar a esas preguntas, son necesarios criterios metodológicos y estadísticos de evaluación de heterogeneidad.

Los criterios metodológicos están referidos a la forma de selección, al delineamento y a la comparación de características clínicas de los pacientes incluidos en cada estudio. Tablas que detallan esos requisitos son necesarias para permitir una comparación entre los estudios y deben estar explicitadas en la revisión sistemática. Desde el punto de vista metodológico, son muchas las fuentes de heterogeneidad entre los estudios: el azar, las diferencias de delineamento, la forma de selección de pacientes, las diferencias en las intervenciones terapéuticas aplicadas y en la forma en que se evaluaron los exámenes[19]. Otra causa de heterogeneidad importante y exclusiva de los estudios de exámenes diagnósticos y pronósticos es la variación en los puntos de cortes para los valores de referencia del examen en cuestión. Aun en estudios aleatorizados para intervención terapéutica, puede existir heterogeneidad porque la aleatorización no se volcó hacia el examen en cuestión sino hacia la intervención terapéutica. Estudios retrospectivos son debilitados a causa de su riesgo de sesgo de selección. El sesgo de verificación (*verification bias*; *ascertainment bias*; *work-up bias*) sucede cuando la indicación del examen patrón-oro es influenciada por el resultado del examen investigado: por ejemplo, si la probabilidad de someterse a la cineangiocoronariografía ("patrón-oro") fuera mayor en aquellos con prueba ergométrica

Tabla 2 - Lista de aspectos a conferirse en la evaluación de los estudios de diagnóstico y pronóstico durante la revisión sistemática y el metaanálisis

Distribución por sexo y edad de la población estudiada[29].
Fecha de inclusión y período de seguimiento del estudio[29].
Prueba de referencia estandarizada, adecuación del patrón oro elegido, evaluándose si éste no lleva la clasificación equivocada del *status* de enfermedad[13].
Aspectos técnicos de la realización del examen.
Evaluar el grado de pérdida de datos *(missing data)*.
Resultados originales de falso y verdadero-positivos, falso y verdadero-negativos. Eventualmente, esos datos pueden estimarse a partir de valores de sensibilidad, especificidad y de los valores positivos y negativos del desenlace o examen de referencia.
Valores de referencia para el examen patrón-oro y para el examen en investigación, de forma clara y representativa de la patología en cuestión[12, 29].
El intervalo de confianza y el error estándar para las medidas de desempeño del examen[29].
El número de evaluadores y su entrenamiento para el examen en cuestión y el patrón- oro[29].
Presencia de sesgo de revisión: verificar si el resultado del examen en el estudio fue evaluado de modo "enmascarado" para desenlaces y otros exámenes (interpretación independiente).
Presencia de sesgo de verificación: el examen de referencia puede haber sido realizado de preferencia en pacientes con pruebas positivas, lo que es más frecuente cuando los exámenes considerados patrón-oro son invasivos. En ese caso, la elección de pacientes para realizar la prueba patrón oro no es aleatória[12].
Si la prueba de referencia ha sido aplicada a todos los pacientes. Caso el examen en investigación y el patrón oro no hayan sido aplicados a todos los pacientes –lo que es el ideal–, se debe evaluar si la elección de pacientes para las pruebas ocurrió de modo aleatorio, disminuyéndose la probabilidad de sesgo[3].
Presencia de sesgo de espectro clínico: ausencia da representación del espectro clínico de la enfermedad estudiada en la población del estudio. Evaluar datos demográficos y clínicos de los pacientes, tales como edad, sexo, raza, características clínicas, presencia de síntomas, estadio de la enfermedad, duración y comorbidades. La prevalencia de la condición en la población estudiada ofrece visión más amplia del espectro, circunstancias y potencial de generalización.
En los exámenes de triage, puede haber sesgo de exceso de diagnóstico (cuando se detecta una enfermedad que podría evolucionar de forma asintomática), sesgo de exceso de representación (para enfermedades que evolucionan con progresión lenta, haciéndolas "sobresalirse" más a causa del método triage) y sesgo de detección precoz (superestima los efectos de beneficio clínico)[13].

positiva que en aquellos con prueba negativa. El análisis del examen investigado debe ser idealmente enmascarado para otras pruebas y para el desenlace. El sesgo causado por el espectro de fases de la enfermedad (*spectrum bias*) provoca variaciones en la sensibilidad y en la especificidad del examen investigado, por comparar poblaciones con fases diferentes de una misma enfermedad: algunos estudios con la mayoría de los pacientes en una fase leve e inicial y otros estudios con pacientes en fase avanzada de la enfermedad[19]. La Tabla 2 resume los aspectos metodológicos a evaluarse. Los métodos para averiguar la heterogeneidad estadística de los estudios se abordarán en el próximo tópico, con la explicación de la forma de combinación (metaanálisis) de resultados de estudios.

Calcular los resultados por medio de metaanálisis, estimando el desempeño diagnóstico

Al utilizarse el metaanálisis, es posible suministrar un sumario agrupado del desempeño diagnóstico (tab. 3). En la dirección electrónica http://www.hrc.es/investigacion/metadisc_en.htm se pude encontrar un *software* gratuito[20] para realización de metaanálisis de exámenes diagnósticos o de exámenes pronósticos. Otros *softwares* y programas especializados que utilizan un abordaje por modelos de regresión binomial basados en razón de verosimilitud y en el teorema de Bayes están disponibles en la dirección electrónica: www.mrc-bsu.cam.ac.uk/bugs/[13]. Estos últimos permiten una evaluación de covariables que influencian el desempeño del examen.

Los métodos de combinación calculan los promedios ponderados de los resultados de los estudios. Dichos métodos se dividen usualmente en dos categorías: métodos con efectos fijos y métodos con efectos aleatorios. En la combinación que utiliza métodos con efectos fijos, se atribuye un valor a cada estudio que es el inverso de la varianza (1/v) del estudio. Métodos de combinación con efectos aleatorios atribuyen un valor a cada estudio que es el inverso de la varianza sumada a la heterogeneidad (1/v + h). De modo simplificado, es como si los métodos con efectos fijos consideraran que la variabilidad entre los estudios ocurrió sólo al azar e ignoraran la heterogeneidad entre ellos[15]. Ya los métodos con efectos aleatorios incorporan un poco de la heterogeneidad entre los estudios en los resultados. De esa manera, se generan resultados combinados con mayor intervalo de confianza. A pesar de tener esa ventaja y ser más recomendados, los métodos con efectos aleatorios son criticados por atribuir un mayor valor a estudios menores[15].

Tabla 3 - Formas de detallarse el desempeño de la prueba por medio de metaanálisis

1. Combinación de sensibilidades y especificidades
2. Combinación de razones de verosimilitud positiva y negativa
3. Combinación de razones de ventajas diagnóstica (o de diagnóstico)
4. Escores de efectividad diagnóstica (o medida del tamaño del efecto)
5. Curvas sROC (*summary ROC* o curva ROC común)

Como es muy común en estudios de exámenes diagnósticos que la variabilidad de resultados no sea solamente al azar –ya que la variabilidad se puede causar de modo explícito o implícito por la variación del punto de corte–, las estimativas de variabilidad suministradas por los modelos de efectos aleatorios son particularmente importantes[10]. Al utilizarse más frecuentemente métodos con efectos aleatorios, las formas de metaanálisis de estudios de exámenes diagnósticos o de factores pronósticos están presentadas en la Tabla 3. Para cada uno de los métodos, se discutirá también el modo de averiguarse la heterogeneidad entre los estudios.

Combinación de sensibilidades y especificidades

Los métodos usados para combinación estadística de sensibilidades y especificidades de los estudios son los mismos usados para la comparación de proporciones. Se combinan la sensibilidad y la especificidad de los estudios en un valor integrado de todos los estudios (*pooling*) por el promedio simple o ponderado (por el tamaño de la muestra o inverso de la varianza de cada estudio). En metaanálisis de estudios diagnósticos y pronósticos, es muy común el autor integrar conjuntamente las sensibilidades y especificidades obtenidas en cada estudio. Sin embargo, frecuentemente eso no es adecuado a causa de la diferencia de umbral o punto de corte del examen en cuestión, explícita o implicitamente[6]. Existe una relación de dependencia entre el punto de corte y la sensibilidad y la especificidad. Un ejemplo de variación explícita en el punto de corte sería la definición de puntos de corte distintos y explícitos en el estudio, por dos estudios distintos, para determinar si el examen era positivo o negativo. Ya la variación implícita ocurriría, por ejemplo, cuando se realiza el examen en estudios con diferencias poblacionales que determinan sensibilidades y especificidades distintas[19]. Dichas diferencias implícitas o explícitas entre los estudios se dicen "efecto de umbral". Ese efecto se puede evaluar mediante la correlación de Spearman entre la sensibilidad y la especificidad encontradas en los diversos estudios incluidos. Cuando hay "efecto de umbral", generalmente hay correlación fuerte e inversa[21]. Aumentando la sensibilidad, generalmente hay disminución de la especificidad. Al integrar matemáticamente (*pooling*) la sensibilidad y especificidad, es necesario utilizarse un método que tome en consideración esa interdependencia entre sensibilidad y especificidad[7]. Además de los umbrales diagnósticos afectaren el desempeño de la prueba, es importante observar si son solamente los umbrales o si hay también problemas metodológicos del estudio que determinan la variación del desempeño[7]. La evaluación de heterogeneidad estadística de los valores de sensibilidad y especificidad obtenidos en los diversos estudios se puede realizar a través de pruebas Mann-Whitney U, Prueba Z, meta-regresión o por modelos de regresión logística[19]; y aún la prueba χ^2 con k-1 grados de libertad (donde k es el número de estudios incluidos). A causa de todos los problemas citados, combinaciones de sensibilidades y especificidades raramente son maneras apropiadas de combinarse resultados.

Combinación de razones de verosimilitud positiva y negativa

La razón de verosimilitud de una prueba positiva (RV+) mide

la mayor probabilidad de la prueba ser positiva en los enfermos en comparación a los no enfermos. La razón de verosimilitud de una prueba negativa (RV-) mide la mayor probabilidad de la prueba ser negativa en los enfermos en comparación a los no enfermos. 1) Los métodos de combinación de razones de verosimilitud pueden ser métodos con efectos fijos –como Mantel-Haenszel o varianza invertida–, y más frecuentemente métodos con efectos aleatorios, como el método de DerSimonian y Laird. Los análisis utilizan combinaciones de razones de verosimilitud tras aplicación de transformación logarítmica[19]. La razón de verosimilitud combinada tiene una ventaja de poder analizar exámenes cuyo resultado es una variable continua o con muchas categorías, evitándose pérdidas de información al dicotomizarse la variable. Otra ventaja es que la *odds* o ventaja postprueba de la enfermedad, una vez que el examen fue positivo, puede calcularse por medio de la fórmula: *odds postprueba = odds preprueba x razón de verosimilitud*[12]. Se debe convertir la ventaja (odds) para probabilidad (v=p/1-p y p=v/1+v – donde v es ventaja p es probabilidad). Así que la probabilidad postprueba = ventaja (*odds*) postprueba/(ventaja (*odds*) postprueba + 1) (http://www.cebm.net/index.aspx?o=1043).

La heterogeneidad de los resultados de razón de verosimilitud de los diversos estudios puede ser evaluada por medio de pruebas univariadas, Pruebas Z y Prueba χ^2. Un método interesante de evaluación de heterogeneidad es el de la estadística Q de Cochrane ($Q = \Sigma w_i(\theta_i - \theta)^2$, donde w_i es el valor que se atribuye al estudio en el metaanálisis (por tamaño de la muestra, por inversión o tamaño de la varianza) y θ es el logaritmo de la razón de verosimilitud promedio y θ_i es el valor del logaritmo de la razón de verosimilitud de cada estudio)[19]. El valor de Q sigue la distribución del χ^2 bajo la hipótesis de que la razón de verosimilitud es la misma para todos los estudios. Otra medida de heterogeneidad que se puede obtener desde ese valor Q es la estadística I^2, que se denomina como la medida de inconsistencia, obtenida mediante la fórmula:

$$I^2 = \frac{(Q - gl)}{Q} \times 100\%$$

Donde gl es el número de grados de libertad (número de estudios menos uno). Esa estadística describe el porcentaje de variabilidad del efecto que, a su vez, es consecuencia de la heterogeneidad y no del azar[19,22]. Cuando I^2 presenta un valor superior a un 50%, se considera que hay heterogeneidad substancial[19]. Hay que prestar la atención para no confundirse la estadística Q de Cochrane –para evaluación de heterogeneidad de valores de razón de similitud– con la medida Q sumarizada que se detalla a continuación, para evaluar globalmente la eficacia de un examen en un metaanálisis.

Razón de ventajas de diagnóstico o *diagnostic odds ratio*

La razón de ventajas de diagnóstico es una combinación estadística de la sensibilidad, la especificidad y los valores de razón de verosimilitud positiva y negativa. Ella es difícil de aplicarse clínicamente, pero muy útil por motivos diversos: a) Es una medida estadística de desempeño global de la prueba; b) Se la puede obtener fácilmente mediante el producto cruzado de la Tabla 2 x 2 (fig. 1); c) Es frecuentemente constante a despecho del punto de corte utilizado para el examen en los diversos estudios; d) Es útil en la construcción del intervalo de confianza de la curva sROC, descrita a continuación[23]. Indica también la razón de verosimilitud positiva dividida par la negativa. Los valores de razón de ventajas de diagnóstico de cada estudio se pueden combinar por medio de métodos de efectos fijos –tales como Mantel-Haenszel–, y métodos de efectos aleatorios (DerSimonian y Laird)[23]. En estudios epidemiológicos para factores de riesgo de enfermedades raras o poco frecuentes, la razón de ventajas tiene valor próximo al riesgo relativo. En el caso de estudios diagnósticos, las razones de ventajas generalmente son diferentes numéricamente del riesgo relativo, porque resultados positivos no son eventos raros[13].

Scores de efectividad o *diagnostic effectiveness scores*

El score de efectividad cuantifica el grado de superposición de resultados entre enfermos y no enfermos. Se puede interpretarlo como el número de desviaciones estándar, al separarse el promedio entre las dos curvas de distribución (enfermos y no enfermos, por ejemplo) de resultados que se comportan como variable continua. Se lo puede obtener por medio de fórmula propia de cálculo o a partir de la razón de ventajas de diagnóstico[23,24]. Es la medida de la distancia estandarizada entre los promedios de dos poblaciones –también denominada medida del tamaño del efecto o medida de efectividad, que también se puede evaluar por medio de modelos de efectos fijos o aleatorios[17]. Es una medida cuantitativa que se puede usar para comparar métodos diagnósticos o para sumarizar resultados de estudios en metaanálisis. Para más detalles de su obtención, sugerimos el trabajo de Hasselblad y Hedges[24], que revisa el método. Así como la curva sROC, descrita a continuación, el score de efectividad suministra una descripción de la separación de dos distribuciones de resultados de exámenes (entre enfermos y no enfermos), independientemente del modo de distribución de los resultados.

Curvas sROC o curvas ROC comunes o sumarizadas – *sROC curves*

Se pueden utilizar los gráficos de dispersión para evaluar la heterogeneidad entre los estudios. El gráfico de dispersión en el espacio ROC presenta los estudios en los ejes FVP vs. FFP (fig. 1). Se puede notar que la curva ROC fue creada para resultados de exámenes que se comportan como variable continua. Sin embargo, en ese caso, cada punto es el resultado combinado de FVP y FFP de cada estudio. Si los estudios utilizaron puntos de corte diferentes, se espera que esa elección determine mayor o menor sensibilidad. O si la sensibilidad y la especificidad de los estudios variaron por causas implícitas, influenciadas por otras covariables[11], se supone que los estudios se complementarían para ilustrar el desempeño diagnóstico del examen en diferentes espectros de formas clínicas o poblaciones. Si en ese gráfico, al unirse

los puntos que representan los estudios, se origina una curvatura semejante a una curva ROC, más probablemente la diferencia entre los estudios es causada por el punto de corte del valor de referencia del examen. Esta es otra forma de evaluación del efecto de umbral[19]. Leves divergencias pueden ocurrir al azar, pero se presupone que otras clases de sesgos (selección, delineamiento etc.) aumentarían la variabilidad observada y causarían una configuración más dispersiva de la representación de los estudios[19]. Los diagramas de bosque (*forest plots*) y el gráfico de Galbraith también facilitan en la visualización de la heterogeneidad entre los estudios[19]. Así que, al presentar los estudios en los gráficos o observar la distribución de los resultados en el espacio ROC, se suministra una idea de heterogeneidad. Además de ello, tras esa evaluación de heterogeneidad, se puede utilizar el espacio ROC para construir una curva ajustada que combina (metaanálisis) los resultados de los estudios, la curva sROC, descrita a continuación.

La curva sROC (curva ROC común o sumarizada –*summary ROC*: *sROC*) es la estimativa de una curva ROC común ajustada para los resultados de los estudios en el espacio ROC[12]. La curva sROC es recomendada para evaluarse el desempeño de una prueba diagnóstica, a partir de un metaanálisis[25]. Destacamos la curva sROC como la mejor opción de metaanálisis cuando hay variación en el punto de corte del valor de referencia del examen o cuando existen variaciones implícitas o explícitas en los estudios que generen diferencias de sensibilidad y especificidad[13,19,23] (Gráf. 2). A causa de esas variaciones frecuentemente encontradas en este tipo de estudios, los promedios de sensibilidad y especificidad de los diversos estudios no expresan un buen desempeño del examen[13].

Se puede obtener la curva a partir de la razón de ventajas de diagnóstico (descrita en el ítem "Razón de ventajas de diagnóstico") tomando en consideración la magnitud de la heterogeneidad entre los estudios. La razón de ventajas de diagnóstico global es muy robusta para heterogeneidad y es homogénea cuando no sufre variaciones relacionadas al punto de corte del examen en estudio[25]. El margen de error estándar de la curva es adecuada cuando los estudios son homogéneos y se mostró una aproximación razonable para estudios heterogéneos[25].

El área bajo la curva (ABC o *area under the curve* – AUC) y el índice Q son sumarios útiles de la curva[13,25]. El área bajo la curva se puede utilizar si tomarse en cuenta la premisa de que los datos presentan distribución bilogística con varianza igual y si hubiera homogeneidad entre los estudios en la estimativa de razón de ventajas de diagnóstico[19]. En este caso, se utiliza el modelo de Moses[11,25] que restringe el análisis sólo a los puntos (estudios) ubicados en la región de interés del espacio ROC. Ello teóricamente podría superestimar el desempeño de la prueba y, por eso, no es aceptado por todos los autores[13]. Rutter y Gatsonis[26] propusieron métodos para cálculo de una curva sROC, tomando en consideración las variaciones entre los estudios no sólo por el umbral de corte, sino también por medio de modelos yerárquicos[13,23]. La utilización del área bajo la curva sROC presenta el riesgo de extrapolación hacia más allá de los datos de sensibilidad

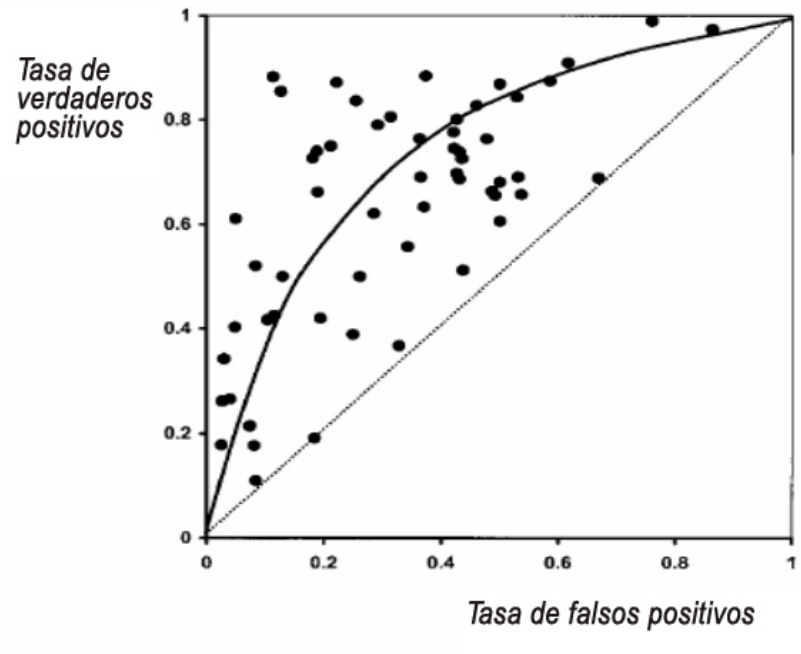

Gráfico 2 - *Curva sROC hipotética; La curva sROC revela la distribución de los estudios (cada estudio es un punto del gráfico de dispersión) y mantiene la relación de interdependencia entre sensibilidad y especificidad, que varía conforme el umbral de corte, implícita o explícitamente.*

y especificidad suministrados por los estudios, a menos que cada estudio haya suministrado una curva ROC, y que ellas sean realmente semejantes[19,23,26]. Ello porque curvas de formas diferentes presentan áreas diferentes. Para construcción del modelo de regresión lineal que precede la curva ROC, existe un debate acerca de la utilización o no utilización de modelos ponderados por la varianza y por el tamaño de la muestra (*n*) de los estudios. La mejor opción es construir las dos curvas (una con modelos ponderados y otra sin incluir el valor de la varianza y del *n*) y compararselas[13]. A pesar de dichas limitaciones y dudas teóricas, el área bajo la curva sROC es uno de los métodos más robustos y útiles para sumarizar los datos de estudios diagnósticos.

Como alternativa para evaluar globalmente la prueba sumarizando la curva sROC, se sugiere la medida Q sumarizada, que evalúa el punto de la curva sROC donde sensibilidad y especificidad son iguales. El valor de Q no cambia de acuerdo con la heterogeneidad y es bastante robusto[25]. Equivale al punto de simetría de la curva ROC[13]. La medida Q, con valores entre 0,5 y 1,0 (cuanto mayor, mejor), es una medida global de eficacia de la prueba[11]. Esa medida sumarizada revela lo cuanto más próximo está el "hombro" de la curva de la esquina superior izquierda[11]. Si menor o igual a 0,5, la prueba no contribuye para la evaluación, y cuanto más próxima de 1,0, mejor el desempeño de la prueba[13]. Así como el área bajo la curva, también se evalúa globalmente la eficacia de la prueba. Si se evaluar diez estudios como mínimo, la distribución de Q es gaussiana (normal)[23]. Se puede utilizar el valor de Q para comparar métodos o verificar sesgos, separándose los estudios con problemas metodológicos en subgrupos y comparando su valor de Q con el valor de Q de los otros subgrupos de estudios[23]. El error estándar de la ABC y el error de Q son próximos numéricamente[25]. Cuando el intervalo de confianza del valor de Q o del ABC pasa por 0,5, el examen no presenta desempeño significativo y no contribuye para la evaluación de la enfermedad.

Evaluar el efecto de la variación de la validez de cada estudio en las estimativas de desempeño diagnóstico

Al evaluar la validez interna y externa de cada estudio y de los resultados combinados, es necesario decidir sobre como lidiar con la heterogeneidad encontrada. Hay cuatro opciones para lidiar con la heterogeneidad entre los estudios e interpretar variaciones de resultados: 1. Ignorar la heterogeneidad y utilizar métodos con efectos fijos; 2. Utilizar pruebas estadísticas de heterogeneidad (son poco sensibles) y no combinar resultados de haber heterogeneidad; 3. Incorporar la heterogeneidad por la utilización de métodos con efectos aleatorios; ó 4. Explicar las diferencias por medio de análisis de subgrupos de estudios o de meta-regresión, incluyendo covariables en el análisis. Al utilizar el metaanálisis, es posible determinar si las estimativas de desempeño dependen de las características de delineamiento del estudio. Se separan subgrupos de estudios por característica de delineamiento, y se los analiza separadamente y en conjunto. De esa manera, se evalúa lo cuanto cambia la diferencia de delineamiento en el desempeño del examen. Asimismo es posible determinar si el desempeño diagnóstico difiere en subgrupos definidos por características del paciente o del examen, utilizando la misma técnica descrita[27]. De ese modo, es posible identificar áreas para investigación adicional[12].

Por ejemplo, un subgrupo de estudios presenta sesgo de verificación (cuando se someten al método patrón-oro solamente los positivos y los negativos con sospecha clínica), lo que frecuentemente subestima la prueba. Otro subgrupo de estudios presenta sesgo de revisión (no evaluar el examen de forma disfrazada para otras pruebas y desenlaces), lo que tiende a superestimar la prueba. Agrupar esos estudios en scores de calidad ni siempre es apropiado. Se los pueden analizar separadamente, por clase de falla metodológica, analizando lo que la falla provoca en el desempeño del examen[27]. Así, los resultados de medidas globales de desempeño pueden ser comparados en cada subgrupo. De esa manera, se puede evaluar también el efecto de la variación de las características de los pacientes y de la prueba en las estimativas de desempeño[12].

En metaanálisis comparativo de pruebas, es fundamental que las pruebas hayan sido realizadas en los mismos pacientes; o por lo menos que los pacientes hayan sido aleatorizados para someterse a cada prueba[12]. Sin embargo, en la mayoría de las veces, es impracticable o antiético realizar todos los exámenes o exámenes invasivos en todos los pacientes, lo que constituye este asunto controvertido[28]. En la comparación de exámenes, hay técnicas de construcción de curvas sROC de los exámenes aisladamente y también en combinación, evaluándose si la combinación de los exámenes aumenta el desempeño diagnóstico o pronóstico[13,29].

Interpretar los resultados, evaluándose el límite de generalización del metaanálisis, de acuerdo con las características de los pacientes

Evaluar el límite de generalización de los resultados, de acuerdo con las características clínicas de los pacientes estudiados, en comparación a la población-blanco de la aplicación del metaanálisis, o la relación entre el desempeño del examen y el año de la publicación[12]. Concluir sobre posibles aplicaciones en poblaciones específicas. Además de ello, generar nuevas hipótesis a investigarse es un importante aporte.

Comentarios sobre el formato de publicación del metaanálisis

Según la conferencia *Quality of Reporting of Meta-analysis* (QUOROM)[6] para publicación de metaanálisis de estudios de intervención terapéutica, se debe, en la publicación de resultados de metaanálisis de estudios diagnósticos y pronósticos, describir detalladamente la metodología, explicitando cada etapa del proceso[18]. El título debe identificar el trabajo como metaanálisis o como revisión sistemática. Se debe estructurar el resumen con la descripción de los aspectos que siguen: la cuestión clínica, las fuentes y bases de datos, los métodos de revisión y selección de la literatura y de síntesis cuantitativa de los datos de forma reproductible, los resultados con estimativas y intervalos de confianza, y la conclusión

con los resultados principales. La introducción debe contextualizar y fundamentar el objetivo. La metodología debe detallar las fuentes y el modo de búsqueda, el período e idioma, los criterios de selección de los estudios, la manera de evaluación de sesgo de publicación, la evaluación de la calidad y validad metodológicas de los estudios, el modo de extracción de los datos idealmente por dos investigadores, las características de los estudios, la manera de evaluación de la heterogeneidad y el modo de sintetizar matemáticamente los datos. Los resultados deben presentar el flujo de la revisión de acuerdo con la figura 2, las características de los estudios[29,30] evaluándose la distribución por edad, sexo, modo de diagnóstico o selección de pacientes, covariables relevantes, tiempo de seguimiento, tamaño de la muestra[6] (tab. 2), y las estimativas de desempeño diagnóstico o pronóstico, con los debidos intervalos de confianza. En la discusión, sumarizar los puntos-clave, discutir las inferencias clínicas con base en la validez interna y externa, interpretar los resultados enfocando la totalidad de las evidencias, describir las limitaciones y los potenciales sesgos, específicamente el sesgo de publicación, y sugerir estudios futuros[6].

Conclusión

Revisiones sistemáticas de la literatura de una cuestión claramente formulada, con técnica de búsqueda y selección de artículos bien planeada, son herramientas extremamente útiles en investigación sobre métodos diagnósticos o pronósticos. En algunos casos, es posible compilarse los datos por medio de técnicas estadísticas, aumentando el poder de las estimaciones de desempeño diagnóstico del examen en la investigación primaria. Mediante el análisis crítico de los sesgos, esas técnicas suministran informaciones que pueden ser útiles para la práctica clínica y para la formulación de cuestiones a probarse en nuevos estudios.

Agradecimientos

Agradecemos la revisión del texto y las sugerencias de la Prof. Carisi A. Polanczyk, de la Universidad Federal de Rio Grande do Sul. El trabajo fue parcialmente financiado por el Consejo Nacional de Desarrollo Científico y Tecnológico (CNPq) y por la Coordinación de Perfeccionamiento de Personal de Nivel Superior (Capes).

Potencial Conflicto de Intereses

Declaro no haber conflicto de intereses pertinentes.

Fuentes de Financiación

El presente estudio fue parcialmente financiado por CNPq y CAPES.

Vinculación Académica

Este artículo forma parte de tesis de Doctorado de Marcos Roberto de Sousa, por la Universidad Federal de Minas Gerais.

Referencias

1. Halligan S. Systematic reviews and meta-analysis of diagnostic tests. Clin Radiol. 2005; 60 (9): 977-9.
2. Sacks HS, Reitman D, Pagano D, Kupelnick B. Meta-analysis: an update. Mt Sinai J Med. 1996; 63 (3-4): 216-24.
3. Zhou A, Obuchowski N, McClish D. Issues in meta-analysis for diagnostic tests. In: Zhou A, Obuchowski N, McClish D, eds. Statistical methods in diagnostic medicine. New York: Wiley & Sons, Inc; 2002. p. 222-40.
4. Alderson P GS, Higgins JPT (eds.). Cochrane Reviewers' Handbook 4.2.2 updated March 2004. Chichester, UK: John Wiley & Sons, Inc; 2004.
5. Knottnerus JA. The evidence base of clinical diagnosis. London: BMJ Publishing Group; 2002.
6. Moher DCD, Eastwood S, Olkin I, Rennie D, Stroup DF, for the QUOROM Group*. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Lancet. 1999; 354: 1896-900.
7. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. JAMA. 1996; 276: 637-9.
8. Higgins JPT, Green S (eds). Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 updated September 2006. Chichester: John Wiley & Sons, Inc; 2006.
9. Velanovich V. Meta-analysis for combining Bayesian probabilities. Med Hypotheses. 1991; 35 (3): 192-5.
10. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making. 1993; 13 (4): 313-21.
11. Moses LE SD, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med. 1993; 12 (14): 1293-316.
12. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med. 1994; 120 (8): 667-76.
13. Pepe MS. The statistical evaluation of medical tests for classification and prediction. New York: Oxford University Press Inc; 2003.
14. Wilczynski NL, Haynes RB, Hedges Team. Developing optimal search strategies for detecting clinically sound prognostic studies in MEDLINE: an analytic survey. BMC Med. 2004; 2: 23.
15. Moayyedi P. Meta-analysis: can we mix apples and oranges? Am J Gastroenterol. 2004; 99 (12): 2297-301.
16. Egger M, Smith GD. Bias in location and selection of studies. BMJ. 1998; 316: 61-6.
17. Vaitkus PT, Brar C. N-acetylcysteine in the prevention of contrast-induced nephropathy: publication bias perpetuated by meta-analyses. Am Heart J. 2007; 153 (2): 275-80.
18. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM,

et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. The Standards for Reporting of Diagnostic Accuracy Group. Croat Med J. 2003; 44 (5): 639-50.

19. Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. Health Technol Assess. 2005; 9 (12): 1-113, iii.

20. 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.

21. Devillé WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt DA, et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Med Res Methodol. 2002; 2: 9.

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

23. Zhou A, Obuchowski N, McClish D. Statistical methods for meta-analysis. In: Zhou A, Obuchowski N, McClish D (eds). Statistical methods in diagnostic medicine. New York: John Wiley & Sons, Inc; 2002. p. 396-417.

24. Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. Psychol Bull. 1995; 117 (1): 167-78.

25. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. Stat Med. 2002; 21 (9): 1237-56.

26. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001; 20: 2865-84.

27. Irwig LMP, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. J Clin Epidemiol. 1995; 48 (1): 119-30.

28. Kertai MD, Boersma E, Bax JJ, Heijenbrok-Kal MH, Hunink MG, L'Talien GJ, et al. A meta-analysis comparing the prognostic accuracy of six diagnostic tests for predicting perioperative cardiac risk in patients undergoing major vascular surgery. Heart. 2003; 89 (11): 1327-34.

29. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ. 2006; 174 (4): 469-76.

30. Rassi A Jr, Rassi A, Rassi SG. Predictors of mortality in chronic chagas disease: a systematic review of observational studies. Circulation. 2007; 115: 1101-8.