Revisão Sistemática e Meta-análise de Estudos de Diagnóstico e Prognóstico: um Tutorial

Systematic Review and Meta-analysis of Diagnostic and Prognostic Studies: a Tutorial

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

Resumo

As revisões sistemáticas com meta-análise de estudos de exames diagnósticos ou de fatores prognósticos são ferramentas de pesquisa ainda em fase de desenvolvimento. O objetivo do presente texto é descrever a metodologia de revisão sistemática e de meta-análise deste tipo de estudos, passo a passo. Foi feita a revisão da literatura sobre o tema, compilando as recomendações e organizando o texto em:

a) Introdução,

b) Setalhamento dos oito passos a serem seguidos,

c) Forma de publicação da revisão sistemática com meta-análise e

d) Conclusão.

Foram descritos os métodos de revisão sistemática de forma detalhada, com análise crítica dos métodos de compilação estatística dos resultados, com ênfase na utilização da curva *Summary Receiver Operator Characteristic*. Forneceu-se referência para os detalhes de cada técnica estatística utilizada na meta-análise. Concluímos que as revisões sistemáticas com meta-análise de exames diagnósticos ou de fatores prognósticos são valiosas na compilação de dados de vários estudos sobre o mesmo tema, reduzindo vieses e aumentando o poder estatístico da pesquisa primária.

Introdução

Denomina-se revisão sistemática da literatura a revisão planejada da literatura científica, que usa métodos sistemáticos para identificar, selecionar e avaliar criticamente estudos relevantes sobre uma questão claramente formulada. O objetivo da sistematização é reduzir possíveis vieses que ocorreriam em uma revisão não-sistemática[1], tanto os vieses observados na forma de revisão da literatura e na seleção dos artigos quanto aqueles detectados pela avaliação crítica de cada estudo. Meta-análise é o método estatístico utilizado na revisão sistemática para integrar os resultados dos estudos incluídos e aumentar o poder estatístico da pesquisa primária[2]. Embora existam meta-análises publicadas em 1904 e 1955[2], o termo meta-análise foi utilizado pela primeira vez por Glass, em 1976, para indicar a análise estatística dos resultados das análises de muitos estudos individuais, com o propósito de integrar os achados[3]. Às vezes, o termo meta-análise é utilizado como sinônimo de revisão sistemática, quando a revisão inclui meta-análise[4]. Embora ocasionalmente usadas como sinônimos, metanálise e meta-análise têm definições diferentes. Metanálise é um recurso da lingüística, que significa a segmentação não-etimológica de um vocábulo, locução ou enunciado, que foram interpretados pelos falantes de forma diversa daquela determinada por sua origem.

Os estudos de testes diagnósticos e prognósticos são antigos na literatura médica, mas a aplicação de metodologia estatística aos testes diagnósticos e de avaliação prognóstica desenvolveu-se depois de sua aplicação em estudos terapêuticos[5]. Do mesmo modo, a padronização da forma de publicação dos estudos diagnósticos[6] ocorreu quase uma década após o mesmo processo ter ocorrido nos estudos terapêuticos[7]. Os principais conceitos estatísticos essenciais ao estudo dos métodos de diagnóstico e avaliação prognóstica estão listados na figura 1 e serão utilizados no texto a seguir.

Existem diferenças importantes entre meta-análises de estudos de intervenção terapêutica, para as quais existem manuais já publicados, e meta-análises de fatores prognósticos ou de exames diagnósticos, mais recentes e menos padronizadas que sobre as primeiras[8]. Meta-análises de estudos comparando intervenções ou tratamentos geralmente incluem estudos aleatorizados, com dois grupos semelhantes, avaliando a mesma intervenção, em geral comparada com placebo ou com tratamento convencional. Já meta-análises de estudos de fatores prognósticos ou de exames diagnósticos enfrentam diferentes desafios, como pontos de corte diferentes para o resultado positivo ou negativo de um exame ou avaliação de exames que foram realizados em estudos prospectivos para estudo de intervenções terapêuticas. Na década de 1990, surgiram novas técnicas estatísticas de combinação de estudos de exames diagnósticos[9-11]. Desde 1994, quando foi publicada diretriz para meta-análise de estudos de exames diagnósticos[12], surgiram várias publicações diferentes com críticas e proposições em aspectos específicos de cada etapa

Palavras-chave

Metodologia, diagnóstico, prognóstico, meta-análise, literatura de revisão.

Correspondência: Marcos R. de Sousa •
Rua Aristides Duarte, 39/601 - Barroca - 30410-040 - Belo horizonte, MG - Brasil
E-mail: mrsousa@cardiol.br
Artigo enviado em 01/01/08; revisado recebido em 23/01/08; aceito em 14/02/08.

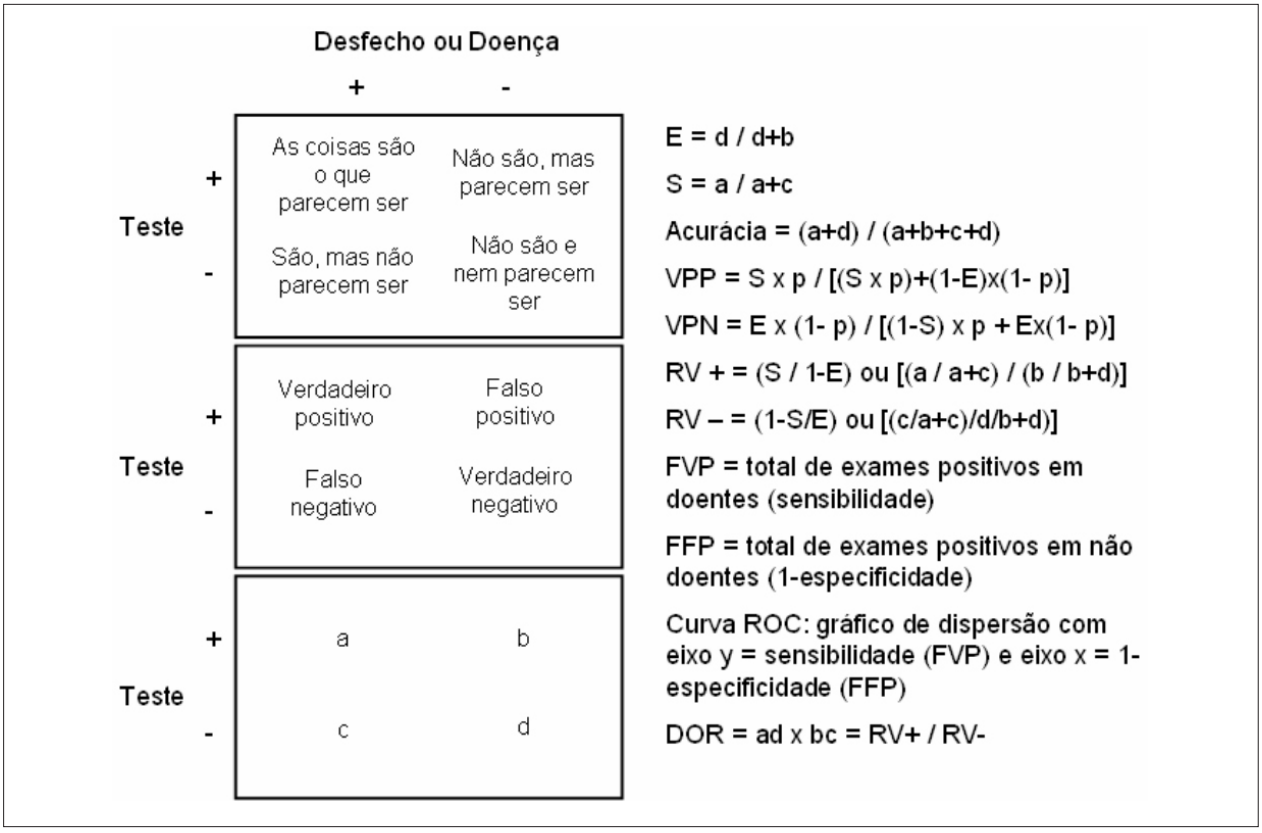

Fig. 1 - *Conceitos e medidas de desempenho de um teste diagnóstico ou prognóstico; Toda decisão clínica é baseada, conscientemente ou não, em probabilidade; Testes diagnósticos podem ser utilizados para avaliar presença ou ausência de doença, para avaliar a gravidade do quadro clínico, para monitorar a resposta a uma intervenção e para estimar o prognóstico; a - número de resultados verdadeiro-positivos (VP); b - número de resultados falso-positivos (FP); c - número de resultados falso-negativos (FN); d - número de resultados verdadeiro-negativos (VN); Especificidade (E) - probabilidade de exame negativo nos não-doentes; Sensibilidade (S) - probabilidade de exame positivo nos doentes; Acurácia do exame - proporção de resultados corretos; Valor preditivo negativo (VPN) - probabilidade de não haver a doença em pessoas com teste negativo; Valor preditivo positivo (VPP) - probabilidade de doença em pessoas com teste positivo; P - na fórmula dos valores preditivos, significa prevalência da doença na população; Razão de verossimilhança de um teste positivo (RV+) - mede o quão mais provável ser o teste positivo nos doentes que nos não-doentes; Razão de verossimilhança de um teste negativo (RV-) - mede o quão mais provável ser o teste negativo nos doentes que nos não-doentes; Fração de verdadeiro-positivos (FVP) - total de exames positivos em doentes; Fração de falso-positivos (FFP) - total de exames positivos em não-doentes; Curva ROC - curva Receiver Operator Characteristic. É usada para comparar um exame com resultado contínuo em relação a um "padrão-ouro" ou a um desfecho. Trata-se de um gráfico de dispersão com eixo y = sensibilidade (FVP) e eixo x = 1-especificidade (FFP). O ponto do gráfico no canto mais alto superior esquerdo é o ponto ideal de desempenho do exame, com sensibilidade = 100% e especificidade = 100%; DOR: razão de chances de diagnóstico, difícil de ser interpretada clinicamente, mas muito útil do ponto de vista estatístico para avaliar o desempenho global do teste e também muito útil na meta-análise, porque ajuda na construção da curva sROC (summary ROC, resultados agrupados de vários estudos na forma de curva ROC).*

Tabela 1 – Passos para a revisão sistemática e meta-análise[12,19]

1. Definir claramente a questão a ser formulada.
2. Buscar em diversas fontes todos os estudos confiáveis, abordando a questão.
3. A partir de critérios claros de inclusão e de exclusão, selecionar os estudos e avaliar sua qualidade.
4. Coletar os dados de cada estudo e apresentá-los de forma clara.
5. Avaliar a heterogeneidade entre os estudos.
6. Calcular os resultados de cada estudo (e combiná-los, se apropriado), estimando o desempenho diagnóstico.
7. Avaliar o efeito da variação da validade de cada estudo nas estimativas de desempenho diagnóstico.
8. Interpretar os resultados, avaliando o quanto se pode generalizar da revisão e/ou meta-análise, conforme as características dos pacientes.

do processo. O uso de meta-análise para exames diagnósticos e prognósticos ainda está em fase de desenvolvimento, mas vem ganhando cada vez mais importância[1,3,8].

O objetivo desta revisão é sumarizar a literatura disponível, definindo tutorial para a realização, passo a passo, de revisão sistemática e, se apropriada, meta-análise de estudos diagnósticos e prognósticos. A seguir, revisaremos os passos necessários, listados na tabela 1.

Definir claramente a questão a ser formulada

Especificar claramente o teste diagnóstico ou prognóstico em questão, a doença em estudo, como foi realizado o diagnóstico e em qual o contexto foi formulada a questão. Geralmente, o exame em questão é comparado com um

padrão-ouro para o diagnóstico da doença, mas os métodos estatísticos utilizados para meta-análise de exames diagnósticos podem ter aplicação bem mais ampla[13]. Nos casos de exames prognósticos, o exame pode ser avaliado pelo desfecho morte, resposta ao tratamento ou, teoricamente, qualquer variável dicotômica de interesse referindo-se a prognóstico de longo prazo[13]. Esclarecer também se será realizada comparação de testes[12].

Buscar em diversas fontes todos os estudos confiáveis abordando a questão

Recomenda-se ampliar ao máximo as fontes de busca. Buscar em publicações governamentais, comissões de ética, resumos em anais de congressos, teses, além da busca em bases eletrônicas (MEDLINE, EMBASE, LILACS etc.)[3]. Além das fontes de busca de estudos, é importante consultar a biblioteca de revisões Cochrane (www.bvs.br) para verificar se tal revisão já foi realizada. Mesmo se não for utilizar dados não publicados, o contato com pesquisadores de estudos em andamento ou não publicados pode ser importante[3]. Para a busca na base de dados MEDLINE, especificar claramente o procedimento de busca na literatura com termos de busca citados, com critérios de inclusão e exclusão explicitados[12]. A forma de pesquisar com termos de busca pode interferir com a sensibilidade da revisão sistemática[14]. É importante buscar termos descritores *MeSH* (*Medical Subject Headings*, vocabulário em língua inglesa usado para indexar artigos (disponível em: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=mesh) para auxiliar na pesquisa. A melhor estratégia geralmente é obtida pela combinação dos termos MeSH utilizados com palavras textuais[14]. Para estudos de marcadores prognósticos, sugere-se aumentar a sensibilidade por meio da associação do tema de pesquisa

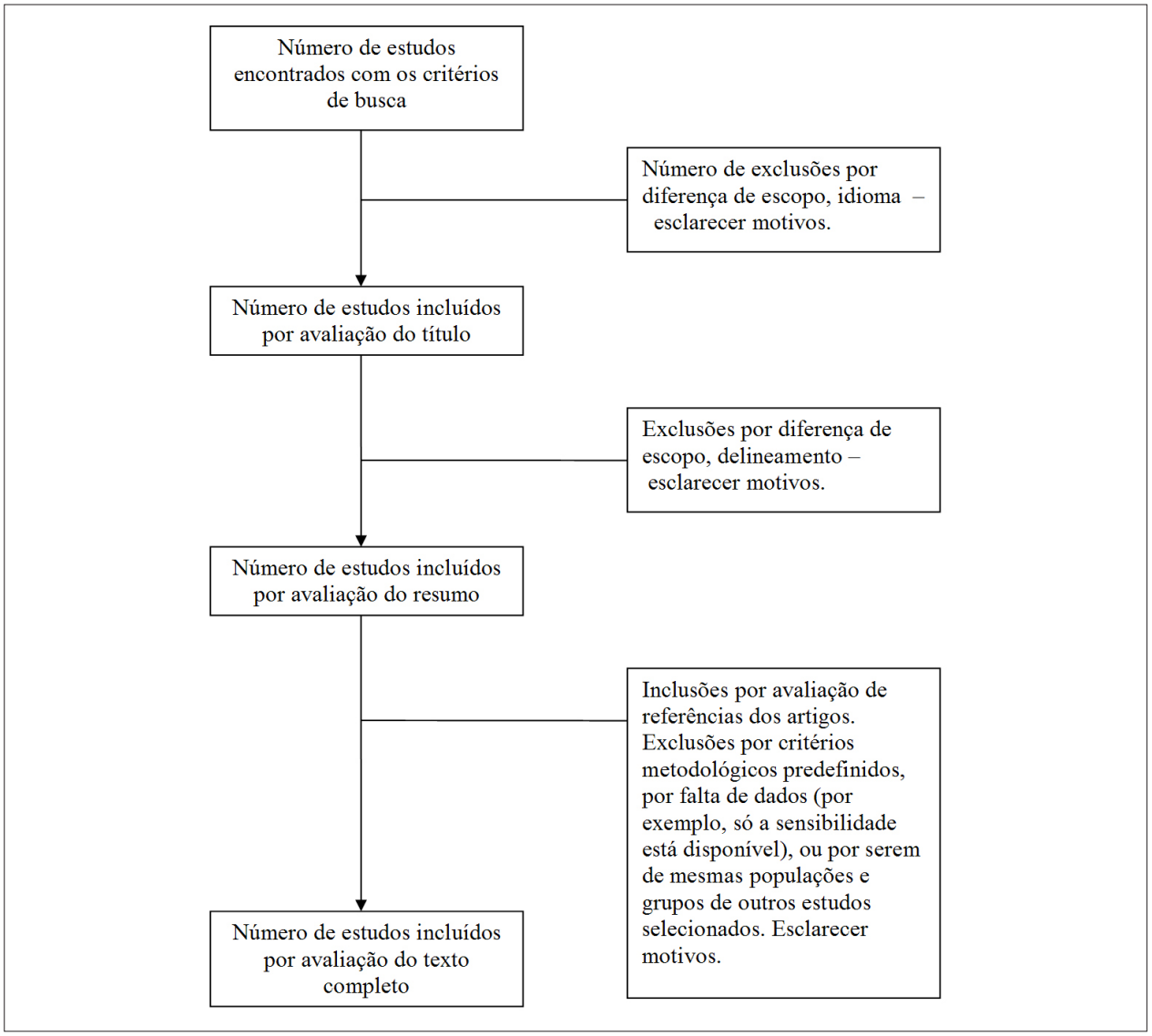

Fig. 2 - Processo de busca e seleção de artigos[6].

com os descritores: (incidence[MeSH] OR mortality[MeSH] OR follow-up studies[MeSH] OR prognos*[Text Word] OR predict*[Text Word] OR course*[Text Word])[14]. Deixar claro como foi o processo de revisão da literatura (fig. 2).

O viés de publicação é a tendência de os estudos com resultados positivos serem mais freqüentemente publicados que estudos com resultados negativos, especialmente em revistas de maior impacto e em língua inglesa[3]. Ocorre habitualmente porque tanto o autor como o editor apresentam resistências em publicar estudos com resultados negativos. Estudos com amostras muito pequenas apresentam maior chance de viés de publicação, motivo pelo qual alguns autores preconizam que sejam excluídos[3,15]. Para reduzir a possibilidade de viés de publicação, as fontes de busca devem ser ampliadas ao máximo. Um método de busca de estudos de intervenção terapêutica, difícil de ser aplicado em estudos de exames diagnósticos ou prognósticos, é averiguar a existência de estudos registrados, mas não publicados, em comissões de ética ou em registros governamentais (por exemplo, www.clinicaltrials.gov), procurando por seus resultados[16]. Outra fonte que pode ser utilizada são os resumos em anais de congressos, onde podem ser reconhecidos trabalhos apresentados e não publicados[17].

Uma forma estatística de avaliar o viés de publicação é pelo uso do gráfico de dispersão em funil, funil invertido ou "árvore de natal" (*funnel plot*)[16]. Esse gráfico tem como premissa que o tamanho da amostra é o mais forte correlato do viés de publicação[3,15] (Gráf. 1). A simetria pode ser avaliada objetivamente por meio de métodos estatísticos[16]. A aparência assimétrica sugere que houve viés de publicação, com tendência da distribuição das razões de chances para um lado, geralmente o lado "mais positivo", já que os "negativos" não teriam sido publicados.

Cada ponto do gráfico representa um estudo com sua razão de chances de diagnóstico e tamanho amostral. A forma de funil simétrico sugere ausência de viés de publicação (nesse caso, um funil invertido, também chamado de "árvore de natal"). Pode-se apresentá-lo invertendo a ordem dos valores do eixo vertical, assim ficaria em forma de funil. Uma assimetria com concentração de estudos no lado direito (o lado de valores maiores de razão de chances de diagnóstico) sugere viés de publicação, com menos estudos negativos publicados.

Selecionar estudos por meio de critérios claros de inclusão e de exclusão, avaliando a qualidade dos estudos

Idealmente, dois pesquisadores devem buscar e avaliar os estudos de forma independente. O teste estatístico Kappa pode ser utilizado para avaliar a concordância entre os dois pesquisadores. Explicar como as discordâncias entre eles foram resolvidas, o que geralmente é feito por um acordo e com base na opinião de um terceiro pesquisador experiente. Listar claramente as características de cada estudo primário

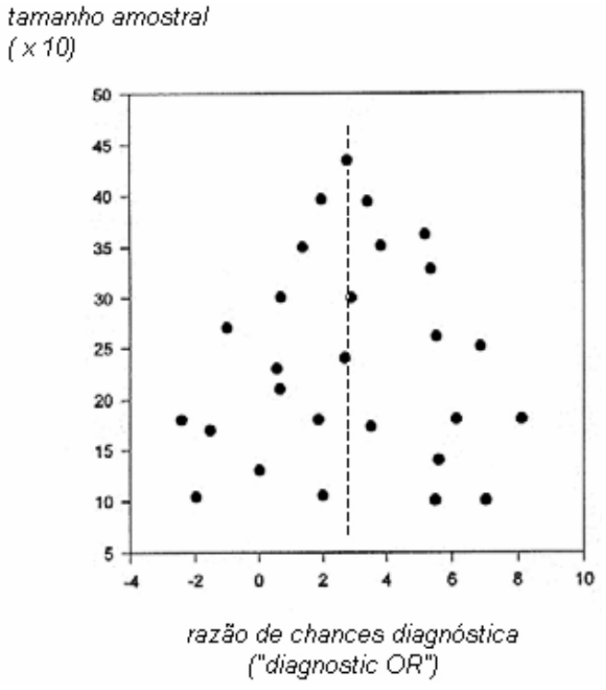

Gráfico 1 -Exemplo de gráfico de dispersão em funil invertido ou "árvore de natal"; Cada ponto do gráfico representa um estudo com sua razão de chances de diagnóstico e tamanho amostral. A forma de funil simétrico sugere ausência de viés de publicação (nesse caso, um funil invertido, também chamado de "árvore de natal"). Pode-se apresentá-lo invertendo a ordem dos valores do eixo vertical, assim ficaria em forma de funil. Uma assimetria com concentração de estudos no lado direito (o lado de valores maiores de razão de chances de diagnóstico) sugere viés de publicação, com menos estudos negativos publicados.

e os resultados de cada um[12]. O trabalho pode ser facilitado se o estudo foi publicado de acordo com a padronização STARD (*Standards for Reporting of Diagnostic Accuracy* - www.consort-statement.org/stardstatement.htm), formulado para garantir mais clareza, rigor metodológico e possibilidade de comparação dos estudos de métodos diagnósticos[18]. Os quesitos de qualidade devem ser conferidos (tab. 2).

Coletar os dados de cada estudo e apresentá-los de forma clara

Tabelas de comparação dos estudos são muito úteis para averiguar as diferenças clínicas e metodológicas entre os estudos (tab. 2). Comparar estudos avaliando a distribuição por idade, sexo, forma de diagnóstico ou seleção de pacientes, co-variáveis relevantes, tempo de seguimento e tamanho da amostra[6]. Para obtenção dos dados a serem combinados, coletar os valores originais de falso e verdadeiro-positivos, falso e verdadeiro-negativos. Eventualmente, esses dados podem ser estimados a partir de valores de sensibilidade, especificidade e os valores de ocorrência do desfecho ou do exame de referência[12].

Avaliar a heterogeneidade entre os estudos

Antes de realizar a combinação estatística (meta-análise) dos estudos, é fundamental avaliar a heterogeneidade entre eles. É importante determinar[19]:

- Por que os resultados variaram entre os estudos?
- A variação foi ao acaso?
- A variação foi causada por diferenças metodológicas?

Para responder a essas perguntas, são necessários critérios metodológicos e estatísticos de avaliação de heterogeneidade.

Os critérios metodológicos se referem à forma de seleção, ao delineamento e à comparação de características clínicas dos pacientes incluídos em cada estudo. Tabelas demonstrando esses quesitos são necessárias para permitir comparação entre os estudos e devem estar explicitadas na revisão sistemática. Do ponto de vista metodológico, as fontes de heterogeneidade entre os estudos são muitas: o acaso, as diferenças de delineamento, a forma de seleção de pacientes, as diferenças nas intervenções terapêuticas aplicadas e na forma em que os exames foram avaliados[19]. Outra causa de heterogeneidade importante e exclusiva dos estudos de exames diagnósticos e prognósticos é a variação nos pontos de cortes para os valores de referência do exame em questão. Mesmo em estudos aleatorizados para intervenção terapêutica, pode existir heterogeneidade porque a aleatorização não foi voltada para o exame em questão e sim para a intervenção terapêutica. Estudos retrospectivos são enfraquecidos por causa de seu risco de viés de seleção. O viés de verificação (*verification bias*; *ascertainment bias*; *work-up bias*) ocorre quando a indicação do exame padrão-ouro é influenciada pelo resultado do exame investigado: por exemplo, se a probabilidade de ser submetido à cineangiocoronariografia ("padrão-ouro") for maior naqueles com teste ergométrico positivo do que naqueles com teste negativo. A análise do exame investigado deve ser idealmente mascarada para outros testes e para o desfecho. O viés causado pelo espectro

Tabela 2 - Lista de aspectos a serem conferidos na avaliação dos estudos de diagnóstico e prognóstico durante a revisão sistemática e meta-análise

Distribuição por sexo e idade da população estudada[29].
Data de inclusão e período de seguimento do estudo[29].
Teste de referência padronizado, adequação do padrão-ouro escolhido, avaliando se este não leva à classificação equivocada do *status* de doença[13]
Aspectos técnicos da realização do exame.
Avaliar o grau de perda de dados (*missing data*).
Resultados originais de falso e verdadeiro-positivos, falso e verdadeiro-negativos. Eventualmente, esses dados podem ser estimados a partir de valores de sensibilidade, especificidade e os valores positivos e negativos do desfecho ou exame de referência
Valores de referência para o exame padrão-ouro e para o exame em investigação, de forma clara e representativa da patologia em questão[12, 29]
O intervalo de confiança e o erro padrão para as medidas de desempenho do exame[29].
O número de avaliadores e seu treinamento para o exame em questão e o padrão-ouro[29].
Presença de viés de revisão: verificar se o resultado do exame no estudo foi avaliado de forma mascarada para desfechos e outros exames (interpretação independente).
Presença de viés de verificação: o exame de referência pode ter sido realizado preferencialmente em pacientes com testes positivos, o que é mais freqüente quando os exames considerados padrão-ouro são invasivos. Nesse caso, a escolha de pacientes para realizar o teste padrão-ouro não é aleatória[12].
Se o teste de referência foi aplicado a todos os pacientes. Caso o exame em investigação e o padrão-ouro não tenham sido aplicados a todos os pacientes, o que é ideal, avaliar se a escolha de pacientes para os testes ocorreu aleatoriamente, diminuindo a chance de viés[3].
Presença de viés de espectro clínico: ausência da representação do espectro clínico da doença estudada na população do estudo. Avaliar dados demográficos e clínicos dos pacientes, tais como idade, sexo, raça, características clínicas, presença de sintomas, estágio da doença, duração e comorbidades. A prevalência da condição na população estudada oferece visão mais ampla do espectro, circunstâncias e potencial de generalização.
Nos exames de triagem, pode haver viés de excesso de diagnóstico (quando uma doença que poderia evoluir de forma assintomática é detectada), viés de excesso de representação (para doenças que evoluem com progressão lenta, fazendo-as "aparecer" mais por causa da triagem) e viés de detecção precoce (superestima os efeitos de benefício clínico)[13].

de fases da doença (*spectrum bias*) provoca variações na sensibilidade e na especificidade do exame investigado, por comparar populações com fases diferentes de uma mesma doença: alguns estudos com a maioria dos pacientes numa fase leve e inicial e outros estudos com pacientes em fase avançada da doença[19]. A tabela 2 resume os aspectos metodológicos a serem avaliados. Os métodos para averiguar a heterogeneidade estatística dos estudos serão abordados no próximo tópico, com a explicação da forma de combinação (meta- análise) de resultados de estudos.

Calcular os resultados por meio de meta-análise, estimando o desempenho diagnóstico

Utilizando meta-análise, é possível fornecer um sumário agrupado do desempenho diagnóstico (tab. 3). No endereço eletrônico http://www.hrc.es/investigacion/metadisc_en.htm pode ser encontrado um *software* gratuito[20] para realização de meta-análise de exames diagnósticos ou de exames prognósticos. Outros *softwares* e programas especializados utilizando abordagem por modelos de regressão binomial baseados em razão de verossimilhança e no teorema de Bayes estão disponíveis no endereço eletrônico: www.mrc-bsu.cam.ac.uk/bugs/[13]. Estes últimos permitem avaliação de co-variáveis que influenciam o desempenho do exame.

Os métodos de combinação calculam médias ponderadas dos resultados dos estudos. Tais métodos são usualmente divididos em duas categorias: métodos com efeitos fixos e métodos com efeitos aleatórios. Na combinação utilizando métodos com efeitos fixos, atribui-se um peso a cada estudo que é o inverso da variância (1/v) do estudo. Métodos de combinação com efeitos aleatórios atribuem um peso a cada estudo que é o inverso da variância somada à heterogeneidade (1/v + h). De forma simplificada, é como se os métodos com efeitos fixos considerassem que a variabilidade entre os estudos ocorreu apenas pelo acaso e ignorassem a heterogeneidade entre eles[15]. Já os métodos com efeitos aleatórios incorporam um pouco da heterogeneidade entre os estudos nos resultados. Assim, geram resultados combinados com maior intervalo de confiança. Apesar de terem essa vantagem e serem mais recomendados, os métodos com efeitos aleatórios são criticados por atribuírem maior peso a estudos menores[15].

Como é muito comum em estudos de exames diagnósticos que a variabilidade de resultados não seja apenas pelo acaso, já que a variabilidade pode ser causada explícita ou implicitamente pela variação do ponto de corte, as estimativas de variabilidade fornecidas pelos modelos de efeitos aleatórios são particularmente importantes[10]. Utilizando mais freqüentemente métodos com efeitos aleatórios, as formas de meta-análise de estudos de exames diagnósticos ou de fatores prognósticos estão apresentadas na tabela 3. Para cada um dos métodos, será discutido também como se averiguar a heterogeneidade entre os estudos.

Tabela 3 - Formas de sumarizar o desempenho do teste por meio de meta-análise.

1. Combinação de sensibilidades e especificidades
2. Combinação de razões de verossimilhança positiva e negativa
3. Combinação de razões de chances diagnóstica (ou de diagnóstico)
4. Escores de efetividade diagnóstica (ou medida do tamanho do efeito)
5. Curvas sROC (*summary ROC* ou curva ROC comum)

Combinação de sensibilidades e especificidades

Os métodos usados para combinação estatística de sensibilidades e especificidades dos estudos são os mesmos usados para a comparação de proporções. Combinam-se a sensibilidade e a especificidade dos estudos em um valor integrado de todos os estudos (*pooling*) pela média simples ou ponderada (pelo tamanho da amostra ou inverso da variância de cada estudo). Em meta-análises de estudos diagnósticos e prognósticos, é muito comum o autor integrar conjuntamente as sensibilidades e especificidades obtidas em cada estudo. Porém, freqüentemente isso não é adequado por causa da diferença de limiar ou ponto de corte do exame em questão, explícita ou implicitamente[6]. Existe uma relação de dependência entre o ponto de corte e a sensibilidade e a especificidade. Um exemplo de variação explícita no ponto de corte seria quando dois estudos diferentes definiram por pontos de corte diferentes e explícitos no estudo para determinar se o exame era positivo ou negativo. Já a variação implícita ocorreria, por exemplo, quando o exame é realizado em estudos com diferenças populacionais que determinam sensibilidades e especificidades diferentes[19]. Tais diferenças implícitas ou explícitas entre os estudos são chamadas de "efeito de limiar". Esse efeito pode ser avaliado pela correlação de Spearman entre a sensibilidade e a especificidade encontradas nos diversos estudos incluídos. Quando há "efeito de limiar", geralmente há correlação forte e inversa[21]. Aumentando a sensibilidade, geralmente há diminuição da especificidade. Ao integrar matematicamente (*pooling*) a sensibilidade e especificidade, é necessário utilizar um método que leve em consideração essa interdependência entre sensibilidade e especificidade[7]. Além de os limiares diagnósticos afetarem o desempenho do teste, é importante observar se são apenas os limiares ou se há também problemas metodológicos do estudo que determinam a variação do desempenho[7]. A avaliação de heterogeneidade estatística dos valores de sensibilidade e especificidade obtidos nos diversos estudos pode ser realizada por meio dos testes Mann-Whitney U, teste Z, meta-regressão ou por modelos de regressão logística[19] e ainda o teste do χ^2 com k-1 graus de liberdade (onde k é o número de estudos incluídos). Por causa de todos os problemas citados, combinações de sensibilidades e especificidades raramente são maneiras apropriadas de combinar resultados.

Combinação de razões de verossimilhança positiva e negativa

A razão de verossimilhança de um teste positivo (RV+) mede o quão mais provável de ser o teste positivo nos doentes que nos não-doentes. A razão de verossimilhança de um teste negativo (RV-) mede o quão mais provável de ser o teste negativo nos doentes que nos não-doentes (fig. 1). Os métodos

de combinação de razões de verossimilhança podem ser métodos com efeitos fixos, como Mantel-Haenszel ou variância invertida, e mais freqüentemente por meio de métodos com efeitos aleatórios, como o método de DerSimonian e Laird. As análises utilizam combinações de razões de verossimilhança após aplicação de transformação logarítmica[19]. A razão de verossimilhança combinada tem a vantagem de poder analisar exames cujo resultado é uma variável contínua ou com muitas categorias, evitando perdas de informação ao dicotomizar a variável. Outra vantagem é que a *odds* ou chance pós-teste da doença, uma vez que o exame deu positivo, pode ser calculada pela fórmula: *odds* pós-teste = *odds* pré-teste x razão de verossimilhança[12]. Chance (*odds*) deve ser convertida para probabilidade (c=p/1-p e p=c/1+c – onde c é chance e p é probabilidade). Então a probabilidade pós-teste = chance (*odds*) pós-teste/(chance (*odds*) pós-teste + 1) (http://www.cebm.net/index.aspx?o=1043).

A heterogeneidade dos resultados de razão de verossimilhança dos diversos estudos pode ser avaliada por meio de testes univariados, testes z e teste do χ^2. Um método interessante de avaliação de heterogeneidade é o da estatística Q de Cochrane ($Q = \Sigma w_i(\theta_i - \theta)^2$, onde w_i é o peso atribuído ao estudo na meta-análise (por tamanho de amostra, por inversão ou tamanho da variância) e θ é o logaritmo da razão de verossimilhança média e θ_i é o valor do logaritmo da razão de verossimilhança de cada estudo)[19]. O valor de Q segue a distribuição do χ^2 sob a hipótese de que a razão de verossimilhança é a mesma para todos os estudos. Outra medida de heterogeneidade que pode ser obtida a partir desse valor Q é a estatística I^2, que é chamada de medida de inconsistência, obtida pela fórmula:

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

onde gl é o número de graus de liberdade (número de estudos menos um). Essa estatística descreve a porcentagem de variabilidade do efeito que é devida à heterogeneidade e não por acaso[19,22]. Quando I^2 apresenta valor acima de 50%, considera-se que há heterogeneidade substancial[19]. Cuidado para não confundir a estatística Q de Cochrane para avaliação de heterogeneidade de valores de razão de verossimilhança com a medida Q sumarizada descrita a seguir, para avaliar globalmente a eficácia de um exame em uma meta-análise.

Razão de chances de diagnóstico ou diagnostic odds ratio

A razão de chances de diagnóstico é uma combinação estatística da sensibilidade, especificidade e dos valores de razão de verossimilhança positiva e negativa. Ela é difícil de ser aplicada clinicamente, mas útil por vários motivos:

a) É uma medida estatística de desempenho global do teste;

b) Pode ser facilmente obtida pelo produto cruzado da tabela 2 x 2 (fig. 1);

c) É freqüentemente constante a despeito do ponto de corte utilizado para o exame nos diversos estudos;

d) É útil na construção do intervalo de confiança da curva sROC, descrita a seguir[23].

Indica também a razão de verossimilhança positiva dividida pela negativa. Os valores de razão de chances de diagnóstico de cada estudo podem ser combinados por meio de métodos de efeitos fixos, tais como Mantel-Haenszel e métodos de efeitos aleatórios (DerSimonian e Laird)[23]. Em estudos epidemiológicos para fatores de risco de doenças raras ou pouco freqüentes, a razão de chances tem valor próximo ao risco relativo. No caso de estudos diagnósticos, as razões de chances geralmente são diferentes numericamente do risco relativo, porque resultados positivos não são eventos raros[13].

Escores de efetividade ou *diagnostic effectiveness scores*

O escore de efetividade quantifica o grau de sobreposição de resultados entre doentes e não-doentes, e pode ser interpretado como o número de desvios padrão separando a média entre as duas curvas de distribuição (doentes e não-doentes, por exemplo) de resultados que se comportam como variável contínua. Ele pode ser obtido por meio de fórmula própria de cálculo ou a partir da razão de chances de diagnóstico[23,24]. É a medida da distância padronizada entre as médias de duas populações - também chamada de medida do tamanho do efeito ou medida de efetividade, que também pode ser avaliada por meio de modelos de efeitos fixos ou aleatórios[17]. É uma medida quantitativa que pode ser usada para comparar métodos diagnósticos ou para sumarizar resultados de estudos em meta-análises. Para mais detalhes de sua obtenção, sugerimos o trabalho de Hasselblad e Hedges[24], que faz uma revisão do método. Assim como a curva sROC, descrita a seguir, o escore de efetividade fornece uma descrição da separação de duas distribuições de resultados de exames (entre doentes e não-doentes), independentemente da forma de distribuição dos resultados.

Curvas sROC ou curvas ROC comuns ou sumarizadas - sROC curves

Os gráficos de dispersão podem ser usados para avaliar a heterogeneidade entre os estudos. O gráfico de dispersão no espaço ROC apresenta os estudos nos eixos FVP vs. FFP (fig. 1). Note-se que a curva ROC foi criada para resultados de exames que se comportam como variável contínua. Mas, nesse caso, cada ponto é o resultado combinado de FVP e FFP de cada estudo. Se os estudos utilizaram pontos de corte diferentes, espera-se que essa escolha determine maior ou menor sensibilidade. Ou se a sensibilidade e a especificidade dos estudos variaram por causas implícitas, influenciadas por outras co-variáveis[11], supõe-se que os estudos se complementariam para ilustrar o desempenho diagnóstico do exame em diferentes espectros de formas clínicas ou populações. Se nesse gráfico, unindo-se os pontos que representam os estudos, surge uma curvatura semelhante a uma curva ROC, mais provavelmente a diferença entre os estudos é causada pelo ponto de corte do valor de referência do exame. Esta é outra forma de avaliação do efeito de limiar[19]. Leves divergências podem ocorrer ao acaso, mas pressupõe-se que outros tipos de vieses (seleção, delineamento etc.) aumentariam a variabilidade observada e

causariam uma configuração mais dispersiva da representação dos estudos[19]. Os gráficos de dispersão em floresta (*forest plots*) e o gráfico de Galbraith também facilitam na visualização da heterogeneidade entre os estudos[19]. Assim, ao apresentar os estudos nos gráficos ou observar a distribuição dos resultados no espaço ROC, fornece-se uma idéia de heterogeneidade. Alem disso, após essa avaliação de heterogeneidade, pode-se utilizar o espaço ROC para construir uma curva ajustada que combina (meta-análise) os resultados dos estudos, a curva sROC, descrita a seguir.

A curva sROC (curva ROC comum ou sumarizada - *summary ROC*: *sROC*) é a estimativa de uma curva ROC comum ajustada para os resultados dos estudos no espaço ROC[12]. A curva sROC é recomendada para avaliar o desempenho de um teste diagnóstico, a partir de uma meta-análise[25]. Destacamos a curva sROC como a melhor opção de meta-análise quando há variação no ponto de corte do valor de referência do exame ou quando existem variações implícitas ou explícitas nos estudos que gerem diferenças de sensibilidade e especificidade[13,19,23] (Gráf. 2). Por causa dessas variações freqüentemente encontradas neste tipo de estudos, as médias de sensibilidade e especificidade dos diversos estudos não refletem bem o desempenho do exame[13].

A curva pode ser obtida a partir da razão de chances de diagnóstico (descrita no item "Razão de chances de diagnóstico") considerando-se a magnitude da heterogeneidade entre os estudos. A razão de chances de diagnóstico global é muito robusta para heterogeneidade e é homogênea quando não sofre variações relacionadas ao ponto de corte do exame em estudo[25]. A margem de erro padrão da curva é adequada quando os estudos são homogêneos e mostrou-se ser uma aproximação razoável para estudos heterogêneos[25].

A área sob a curva (*area under the curve* - AUC) e o índice Q são sumários úteis da curva[13,25]. A área sob a curva pode ser utilizada se considerar a premissa de que os dados apresentam distribuição bilogística com variância igual e se houver homogeneidade entre os estudos na estimativa de razão de chances de diagnóstico[19]. Neste caso, utiliza-se o modelo de Moses[11,25] que restringe a análise apenas aos pontos (estudos) localizados na região de interesse do espaço ROC, o que teoricamente poderia superestimar o desempenho do teste e, assim, não é aceito por todos autores[13]. Rutter e Gatsonis[26] propuseram métodos para cálculo de uma curva sROC, levando em consideração as variações entre os estudos não apenas pelo limiar de corte, mas por meio de modelos hierárquicos[13,23]. O uso da área sob a curva sROC apresenta o risco de extrapolação além dos dados de sensibilidade e especificidade fornecidos pelos estudos, a menos que cada estudo tenha fornecido uma curva ROC, e que elas sejam realmente semelhantes[19,23,26]. Isso porque curvas de formas diferentes apresentam áreas diferentes. Para construção do modelo de regressão linear que precede a curva ROC, existe debate sobre utilizar ou não modelos ponderados pela variância e pelo tamanho da amostra (*n*) dos estudos. A

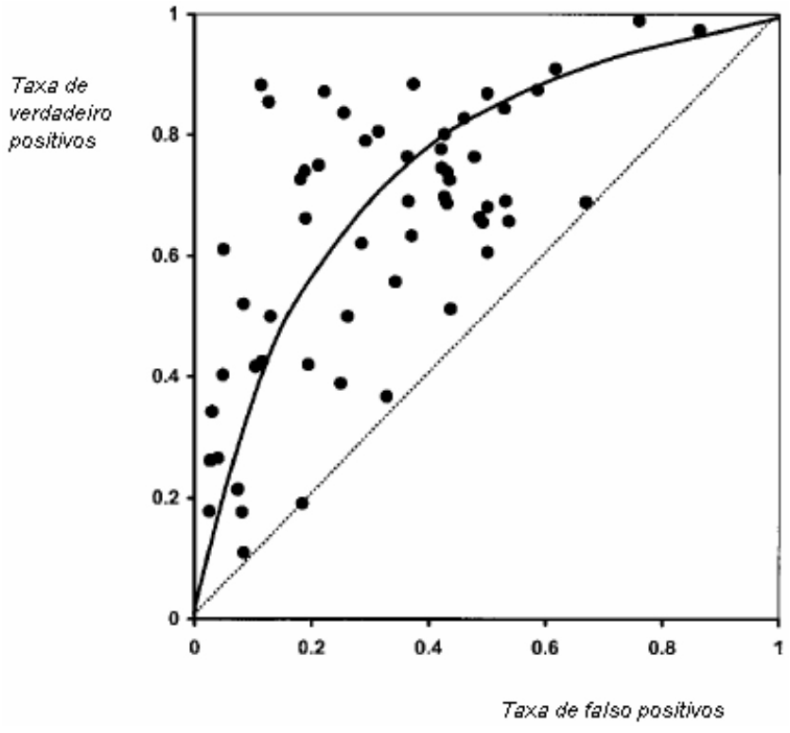

Gráfico 2 -Curva sROC hipotética; A curva sROC mostra a distribuição dos estudos (cada estudo é um ponto do gráfico de dispersão) mantendo a relação de interdependência entre sensibilidade e especificidade, que varia conforme o limiar de corte, implícita ou explicitamente.

melhor opção é construir as duas curvas (uma com modelos ponderados e outra sem incluir o peso da variância e do *n*) e compará-las[13]. Apesar de tais limitações e dúvidas teóricas, a área sob a curva sROC é um dos métodos mais robustos e úteis para sumarizar os dados de estudos diagnósticos.

Como alternativa para avaliar globalmente o teste sumarizando a curva sROC, sugere-se a medida Q sumarizada, que avalia o ponto da curva sROC onde sensibilidade e especificidade são iguais. O valor de Q não varia conforme a heterogeneidade e é bastante robusto[25]. Equivale ao ponto de simetria da curva ROC[13]. A medida Q, com valores entre 0,5 e 1,0 (quanto maior, melhor), é uma medida global de eficácia do teste[11]. Essa medida sumarizada mostra o quanto mais próximo está o "ombro da curva" do canto superior esquerdo[11]. Se menor ou igual a 0,5, o teste não contribui para a avaliação, e quanto mais próxima de 1,0, melhor o desempenho do teste[13]. Assim como a área sob a curva, também avalia globalmente a eficácia do teste. Se forem avaliados pelo menos dez estudos, a distribuição de Q é gaussiana (normal)[23]. O valor de Q pode ser usado para comparar métodos ou verificar vieses, separando os estudos com problemas metodológicos em subgrupos e comparando seu valor de Q com o valor de Q dos outros subgrupos de estudos[23]. O erro padrão da AUC e o erro padrão de Q são próximos numericamente[25]. Quando o intervalo de confiança do valor de Q ou da AUC passam pelo 0,5, o exame não apresenta desempenho significativo e não contribui para a avaliação da doença.

Avaliar o efeito da variação da validade de cada estudo nas estimativas de desempenho diagnóstico

Ao avaliar a validade interna e externa de cada estudo e dos resultados combinados, é necessário decidir sobre como lidar com a heterogeneidade encontrada. Existem quatro opções para lidar com a heterogeneidade entre os estudos e interpretar variações de resultados:

1) Ignorar a heterogeneidade e utilizar métodos com efeitos fixos;

2) Utilizar testes estatísticos de heterogeneidade (são pouco sensíveis) e não combinar resultados se houver heterogeneidade;

3) Incorporar a heterogeneidade pelo uso de métodos com efeitos aleatórios; ou

4) Explicar as diferenças por meio de análises de subgrupos de estudos ou de meta-regressão, incluindo co-variáveis na análise.

Utilizando meta-análise, é possível determinar se as estimativas de desempenho dependem das características de delineamento do estudo. Separam-se subgrupos de estudos por característica de delineamento, analisando-os separadamente e em conjunto, avaliando-se em que magnitude a diferença de delineamento altera no desempenho do exame. Também é possível determinar se o desempenho diagnóstico difere em subgrupos definidos por características do paciente ou do exame utilizando a mesma técnica descrita[27]. Dessa maneira, é possível identificar áreas para pesquisa adicional[12].

Por exemplo, um subgrupo de estudos apresenta viés de verificação (quando submete ao método padrão-ouro apenas os positivos mais os negativos com suspeita clínica), o que freqüentemente subestima o teste. Outro subgrupo de estudos apresenta viés de revisão (não avaliar o exame de forma mascarada para outros testes e para desfechos), o que tende a superestimar o teste. Agrupar esses estudos em escores de qualidade nem sempre é apropriado. Eles podem ser analisados separadamente por tipo de falha metodológica, analisando o que a falha provoca no desempenho do exame[27]. Assim, os resultados de medidas globais de desempenho podem ser comparados em cada subgrupo. Dessa forma, pode-se também avaliar o efeito da variação das características dos pacientes e do teste nas estimativas de desempenho[12].

Em meta-análise comparativa de testes, é fundamental que os testes tenham sido realizados nos mesmos pacientes, ou pelo menos que os pacientes tenham sido aleatorizados para serem submetidos a cada teste[12]. Porém, na maioria das vezes, é impraticável ou antiético realizar todos os exames ou exames invasivos em todos os pacientes, sendo este tema controverso[28]. Na comparação de exames, existem técnicas de construção de curvas sROC dos exames isoladamente e em combinação, avaliando se a combinação dos exames aumenta o desempenho diagnóstico ou prognostico[13,29].

Interpretar os resultados avaliando o quanto se pode generalizar da meta-análise, conforme as características dos pacientes

Avaliar o quanto os resultados podem ser generalizados, conforme as características clínicas dos pacientes estudados em comparação com a população-alvo da aplicação da meta-análise ou a relação entre o desempenho do exame e o ano da publicação[12]. Concluir sobre possíveis aplicações em populações específicas. Além disso, gerar novas hipóteses a serem pesquisadas é uma importante contribuição.

Comentários sobre a forma de publicação da meta-análise

Em analogia com a conferência *Quality of Reporting of Meta-analysis* (QUOROM)[6] para publicação de meta-análises de estudos de intervenção terapêutica, deve-se, na publicação de resultados de meta-análise de estudos diagnósticos e prognósticos, descrever detalhadamente a metodologia, deixando explícita cada etapa do processo[18]. O título deve identificar o trabalho como meta-análise ou como revisão sistemática. O resumo deve ser estruturado com descrição dos seguintes aspectos: a questão clínica, as fontes e bases de dados, os métodos de revisão e seleção da literatura e de síntese quantitativa dos dados de forma reprodutível, os resultados com estimativas e intervalos de confiança, e a conclusão com os resultados principais. A introdução deve contextualizar e fundamentar o objetivo. A metodologia deve detalhar as fontes e a forma de busca, o período e idioma, os critérios de seleção dos estudos, a forma de avaliação de viés de publicação, a avaliação de qualidade e validade metodológica dos estudos, a forma de extração dos dados

idealmente por dois pesquisadores, as características dos estudos, a forma de avaliação da heterogeneidade e a forma de sintetizar matematicamente os dados. Os resultados devem apresentar o fluxo da revisão conforme a figura 2, as características dos estudos[29,30] avaliando a distribuição por idade, sexo, forma de diagnóstico ou seleção de pacientes, co-variáveis relevantes, tempo de seguimento, tamanho da amostra[6] (tab. 2), e as estimativas de desempenho diagnóstico ou prognóstico, com os devidos intervalos de confiança. Na discussão, sumarizar os pontos-chave, discutir as inferências clínicas com base na validade interna e externa, interpretar os resultados à luz da totalidade das evidências, descrever as limitações e os potenciais vieses, especialmente o viés de publicação, e sugerir estudos futuros[6].

Conclusão

Revisões sistemáticas da literatura de uma questão claramente formulada, com técnica de busca e seleção de artigos bem planejada, são ferramentas extremamente úteis em pesquisa sobre métodos diagnósticos ou prognósticos. Em alguns casos, é possível compilar os dados por meio de técnicas estatísticas, aumentando o poder das estimativas de desempenho diagnóstico do exame na pesquisa primária. Por meio da análise crítica dos vieses, essas técnicas fornecem informações que podem ser úteis para a prática clínica e para a formulação de questões a serem testadas em novos estudos.

Agradecimentos

Agradecemos a revisão do texto e as sugestões da Prof. Carisi A. Polanczyk, da Universidade Federal do Rio Grande do Sul. O trabalho foi parcialmente financiado pelo Conselho Nacional de desenvolvimento Científico e Tecnológico (CNPq) e pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes).

Potencial Conflito de Interesses

Declaro não haver conflito de interesses pertinentes.

Fontes de Financiamento

O presente estudo foi parcialmente financiado por CNPq e CAPES.

Vinculação Acadêmica

Este artigo é parte de tese de Doutorado de Marcos Roberto de Sousa pela Universidade Federal de Minas Gerais.

Referências

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.