

MINISTÉRIO DA SAÚDE

**DIRETRIZES METODOLÓGICAS: Elaboração de revisão
sistemática e metanálise de estudos de acurácia diagnóstica**

Brasília – DF
2014



MINISTÉRIO DA SAÚDE
Secretaria de Ciência, Tecnologia e Insumos Estratégicos
Departamento de Ciência e Tecnologia

DIRETRIZES METODOLÓGICAS

Elaboração de revisão sistemática e metanálise de estudos de acurácia diagnóstica

Brasília – DF
2014



MINISTÉRIO DA SAÚDE
Secretaria de Ciência, Tecnologia e Insumos Estratégicos
Departamento de Ciência e Tecnologia

DIRETRIZES METODOLÓGICAS

Elaboração de revisão sistemática e metanálise de estudos de acurácia diagnóstica

Brasília – DF
2014



2014 Ministério da Saúde.

Todos os direitos reservados. A coleção institucional do Ministério da Saúde pode ser acessada, na íntegra, na Biblioteca Virtual em Saúde do Ministério da Saúde: <www.saude.gov.br/bvs>. O conteúdo desta e de outras obras da Editora do Ministério da Saúde pode ser acessado na página: <<http://editora.saude.gov.br>>.



Esta obra é disponibilizada nos termos da Licença Creative Commons – Atribuição – Não Comercial – Sem Derivações 4.0 Internacional. É permitida a reprodução parcial ou total desta obra, desde que citada a fonte.

Este trabalho foi desenvolvido no âmbito do termo de cooperação nº 47 entre o Departamento de Ciência e Tecnologia e a Organização Panamericana da Saúde

Tiragem: 1ª edição – 2014 – 4.000 exemplares

Elaboração, distribuição e informações:

MINISTÉRIO DA SAÚDE
Secretaria de Ciência, Tecnologia e Insumos Estratégicos
Departamento Ciência e Tecnologia
Coordenação-Geral de Gestão do Conhecimento
SQN Quadra 2 Projeção C, térreo sala O3
CEP: 70712-902 – Brasília/DF
Tel: (61) 3410-4118
Site: www.saude.gov.br
E-mail: ats.decit@saude.gov.br

Supervisão Geral:

Carlos Augusto Gabrois Gadelha (SCTIE/MS)
Antônio Carlos Campos de Carvalho (Decit/SCTIE/MS)
Jorge Otávio Maia Barreto (Decit/SCTIE/MS)

Organização:

Nashira Campos Vieira (Anvisa)
Roberta Moreira Wichmann (Decit/SCTIE/MS)

Elaboração de texto:

Anna Maria Buehler (Hcor)
Mabel Figueiró (Hcor)
Frederico Rafael Moreira (Hcor)
Alexandre Biasi Cavalcanti (Hcor)
André Sasse (Hcor)
Otávio Berwanger (Hcor)

Revisão Técnica:

Cláudia Cristina de Aguiar Pereira (ENSP/Fiocruz)
Gabriela Vilela de Brito (Decit/SCTIE/MS)
Marina Gonçalves de Freitas (Decit/SCTIE/MS)
Nashira Campos Vieira (Anvisa)
Roberta Moreira Wichmann (Decit/SCTIE/MS)

Editoração:

Eliana Carlan (Decit/SCTIE/MS)
Jessica Rippel (Decit/SCTIE/MS)

Design Gráfico:

Gustavo Veiga e Lins (Decit/SCTIE/MS)

Normalização:

Francisca Martins Pereira (CGDI/ Editora MS)

Impresso no Brasil/Printed in Brazil

Ficha Catalográfica

Brasil. Ministério da Saúde. Secretaria de Ciência, Tecnologia e Insumos Estratégicos. Departamento de Ciência e Tecnologia.

Diretrizes metodológicas : elaboração de revisão sistemática e metanálise de estudos de acurácia diagnóstica / Ministério da Saúde, Secretaria de Ciência, Tecnologia e Insumos Estratégicos, Departamento de Ciência e Tecnologia. – Brasília : Editora do Ministério da Saúde, 2014.
116 p. : il.

ISBN 978-85-334-2129-5

1. Tecnologia em saúde. 2. Acurácia diagnóstica. 3. Ministério da Saúde. I. Título.

CDU 614

Catálogo na fonte – Coordenação-Geral de Documentação e Informação – Editora MS – OS 2014/0211

Títulos para indexação

Em inglês: Methodological guideline: development of systematic review and meta-analysis of diagnostic accuracy studies.

Em espanhol: Directriz metodológica: desarrollo de revision sistemática y metanálisis de precisión diagnóstica

/// LISTA DE QUADROS

Quadro 1 – Exemplo de questão de pesquisa de estudos de testes diagnósticos	24
Quadro 2 – Variáveis consideradas para elaboração de ficha clínica de extração dos dados	38

/// LISTA DE TABELAS

Tabela 1 – Tabela de contingência 2 x 2	47
Tabela 2 – Estruturação da discussão da Revisão Sistemática	61
Tabela 3 – Fatores que diminuem a qualidade da evidência de estudos de acurácia diagnóstica e como eles diferem dos demais critérios de classificação para outras intervenções	64

LISTA DE FIGURAS

Figura 1 – Apresentação tabular dos resultados do QUADAS-2 para os estudos incluídos	43
Figura 2 – Apresentação gráfica dos resultados do QUADAS-2 para os estudos incluídos	43
Figura 3 – Resultado da busca na LILACS	51
Figura 4 – Exemplo hipotético de uma curva SROC	52
Figura 5 – Exemplo hipotético de uma SROC utilizando o modelo bivariado	54
Figura 6 – Exemplo hipotético de uma HSROC	55

LISTA DE SIGLAS E ABREVIATURAS

AHRQ – Agency for Healthcare Research and Quality
ARIF – Aggressive Research Intelligence Facility
Bireme – Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde
BVS – Biblioteca Virtual em Saúde
CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
C-EBLM – Evidence-Based Laboratory Medicine
CENTRAL – The Cochrane Central Register of Controlled Trials
CESU – Continental Europe Regional Support Unit
CINAHL – Cumulative Index to Nursing and Allied Health Literature
CRD – Centre for Reviews and Dissemination
CRDTAS – Cochrane Register of Diagnostic Test Accuracy Studies
DARE – Database of Abstracts of Reviews of Effects
DeCS – Descritores em Ciências da Saúde
Decit – Departamento de Ciência e Tecnologia
EPC – Evidence-based Practice Center
GRADE – Grades of Recommendation, Assessment, Development and Evaluation
HIRU – Health Information Research Unit
HTA – Health Technology Assessment Database
IATS – Instituto de Avaliações Econômicas
ISSG – Information Specialists’ Subgroup
IBICT – Instituto Brasileiro de Ciência e Tecnologia
LILACS – Literatura científica e técnica da América Latina e Caribe
MEDLINE – Medical Literature Analysis and Retrieval System Online
MEDION – Meta-analyses van Diagnostisch Onderzoek
MeSH – Medical Subject Headings
NTIS – National Technical Information Service
PROSPERO – International Prospective Register of Ongoing Systematic Reviews
QUADAS – Quality Assessment of Diagnostic Accuracy Studies
Rebrats – Rede Brasileira de Avaliação de Tecnologia em Saúde
Rebec – Registro Brasileiro de Ensaio Clínicos
SAS – Statistical Analysis System
SPSS – Statistical Package for Social Science
SUS – Sistema Único de Saúde
RCD – Razão de Chances Diagnóstica
ROC – Receiver operating characteristic
SciELO – Scientific Electronic Library Online
SRDTA – Systematic Review Diagnostic Test Accuracy
STARD – Standards for Reporting of Diagnostic Accuracy
STATA – Data Analysis and Statistical Software
TRIP – Turning Research into Practice
UKSU – Regional Support Unit
USA – United States of America

SUMÁRIO

APRESENTAÇÃO	17
1 INTRODUÇÃO ÀS REVISÕES SISTEMÁTICAS	19
1.1 Definição de Revisão Sistemática	19
1.2 Definição de Metanálise	19
1.3 Importância das Revisões Sistemáticas de acurácia diagnóstica	20
1.4 Limitações das Revisões Sistemáticas de estudos de acurácia diagnóstica	20
1.5 Recursos necessários	21
ETAPAS FUNDAMENTAIS DE UMA REVISÃO SISTEMÁTICA	23
2 PLANEJAMENTO DA REVISÃO SISTEMÁTICA	23
2.1 Definição da questão de pesquisa estruturada no formato do acrônimo PIRO	23
2.2 Definição dos critérios de elegibilidade	25
2.3 Revisão de literatura: justificativa para a revisão sistemática	25
2.4 Documentação da metodologia: redação de protocolo	25
3 CONDUÇÃO DA REVISÃO SISTEMÁTICA	27
3.1 Busca de potenciais estudos elegíveis	27
3.2 Fontes de evidências	28
3.3 Pesquisa em fontes de dados não publicados, em andamento e literatura cinzenta	29
3.4 Elaboração da estratégia de busca nas várias bases de dados	30
3.4.1 Seleção dos termos para a busca	31
3.5 Filtros de busca para estudos de acurácia de testes diagnósticos	32
3.6 Relatando o processo de busca	33
3.7 Avaliação da elegibilidade dos estudos	34
3.7.1 Avaliação da elegibilidade: Triagem dos artigos pela leitura de título e resumo	34
3.7.2 Avaliação da elegibilidade pela leitura do manuscrito em texto completo e ficha clínica padronizada	35
3.8 Extração de dados	36
3.8.1 Processo de extração de dados	36
3.8.2 Elaboração da ficha clínica padronizada	37
3.8.2.1 Características dos estudos incluídos e dados descritivos	37
3.8.2.2 Avaliação do risco de viés dos estudos incluídos	39
3.8.2.3 Extração dos dados quantitativos	44
3.9 Resultados	44
3.9.1 Apresentação dos dados descritivos	44
3.9.2 Sumário dos efeitos do tratamento nos estudos, cálculo e apresentação da metanálise	45
3.9.2.1 Tipos de variáveis e medidas de desempenho em estudos diagnósticos	46
3.9.2.2 Modelos de Análise	48
3.9.2.3 Métodos Estatísticos	49

3.9.2.3.1 Metanálises individuais de sensibilidade e especificidade (summary operating point)	49
3.9.2.3.2 Curva ROC sumária (SROC)	50
3.9.2.3.3 Modelos Hierárquicos de metanálise de estudos de acurácia diagnóstica	53
a) Modelo bivariado	53
b) Modelo HSROC.....	54
3.9.2.4 Programas para cálculo de metanálise	55
3.9.3 Investigando a heterogeneidade	57
3.10 Avaliação do risco de viés na Revisão Sistemática	58
3.10.1 Avaliando o viés de publicação	59
4 RELATO E APLICABILIDADE DOS RESULTADOS	61
4.1 Estruturando a discussão e conclusão	61
4.1.1 Estruturando a discussão	61
4.1.2 Estruturando a conclusão	61
4.1.2.1 Implicações para prática clínica	61
4.1.2.2 Implicações para pesquisas futuras	62
5 ETAPAS OPCIONAIS:	63
5.1 Avaliação da qualidade da evidência	63
6 CONCLUSÕES DA DIRETRIZ	67
REFERÊNCIAS	69
APÊNDICES	75
APÊNDICE 1: Construção da estratégia de busca: símbolos e operadores booleanos	77
APÊNDICE 2: Gerenciadores de referências: o que são e como utilizá-los	81
APÊNDICE 3: Tipos de estudos de acurácia diagnóstica	85
APÊNDICE 4: Medidas de desempenho dos testes diagnósticos	89
ANEXOS	97
ANEXO A: Tabela das principais bases de dados e respectivos acessos eletrônicos	99
ANEXO B: Exemplos de construções de estratégia de busca.....	103
ANEXO C: Fluxo de seleção dos artigos da revisão sistemática	107
ANEXO D: Exemplo de ficha clínica para revisões sistemáticas	109
ANEXO E: Principais fontes de viés e variabilidade nos estudos de acurácia diagnóstica	111
ANEXO F: QUADAS-2	113

1 APRESENTAÇÃO

A revisão sistemática é uma revisão de literatura científica que utiliza uma metodologia padrão para encontrar, avaliar e interpretar todas as pesquisas relevantes disponíveis para uma questão particular de pesquisa, área do conhecimento ou fenômeno de interesse. As revisões sistemáticas pretendem apresentar uma estimativa mais correta sobre uma questão de pesquisa, por meio de uma metodologia confiável, rigorosa e auditável.

A partir da Oficina de Prioridades de Pesquisa em Saúde – “Editais Temáticos” – realizada em março em 2006, foi estabelecido que as diretrizes metodológicas eram fundamentais para orientar a elaboração de revisões sistemáticas, pois teriam como objetivo padronizar e equalizar a qualidade na elaboração e condução de revisões sistemáticas. O público alvo seriam pesquisadores interessados neste tipo de estudo, inclusive para atender as demandas do Ministério da Saúde.

Partindo desta iniciativa, o Departamento de Ciência e Tecnologia – Decit criou o Grupo de Trabalho de Desenvolvimento e Padronização Metodológica da Rebrats, que, dentre outras atribuições, ficaram responsáveis pela Elaboração de Diretrizes Metodológicas. O grupo é formado por profissionais de especialidades diversas e de Instituições, como hospitais de excelência, universidades federais, fundações e técnicos de secretarias e departamentos do Ministério da Saúde.

Em reunião, estabeleceu-se que seriam publicadas as diretrizes de Avaliações Econômicas em Saúde, de Revisões Sistemáticas e Metanálises de Ensaios Clínicos Randomizados, de Estudos de Acurácia Diagnóstica, de Estudos Observacionais e de Estudos Qualitativos. Coube ao HCor – Hospital do Coração de São Paulo – elaborar as diretrizes sobre Revisões Sistemáticas e Metanálises dos diferentes tipos de estudo.

A diretriz apresentada neste documento foi baseada no manual especializado da Cochrane¹. Adicionalmente, foram utilizados artigos de revisão metodológica publicados em bases de informações científicas e referenciados ao longo do texto.

O objetivo deste documento é apresentar os principais conceitos necessários à condução de uma revisão sistemática de estudos de acurácia diagnóstica, utilizando uma linguagem mais simples e acessível para profissionais da área da saúde, que estejam ou não familiarizados com essa metodologia.

A diretriz aborda as três fases da revisão sistemática: planejamento, condução e relato da revisão sistemática. Ela não leva em consideração o impacto do tipo de questão no processo da revisão e não esgota todos os mecanismos necessários para entender e realizar uma metanálise. O detalhamento de alguns conceitos fundamentais é apresentado em forma de apêndices.

A estrutura desta Diretriz é constituída de acordo com as seguintes seções:

1. Seção 1 promove a introdução às revisões sistemáticas como um método de pesquisa.
2. Seção 2 discute o planejamento de uma revisão sistemática.
3. Seção 3 discute os estágios envolvidos na condução da revisão sistemática e seus resultados.
4. Seção 4 discute como relatar a revisão sistemática.
5. Seção 5 apresenta etapas opcionais que complementam os dados da revisão sistemática

A importância dessa diretriz é orientar e padronizar a elaboração e condução de uma revisão sistemática de qualidade, principalmente devido à lacuna que encontramos na literatura nacional desse tipo de estudo, além da baixa qualidade metodológica. Por outro lado, cada vez mais estudos com esse delineamento têm sido conduzidos e publicados em literatura científica, de modo que se faz necessária a compreensão dos principais fundamentos dessa metodologia. As diretrizes são guias práticos das principais etapas que estruturam a elaboração de revisões sistemáticas.

1 INTRODUÇÃO ÀS REVISÕES SISTEMÁTICAS

1.1 Definição de Revisão Sistemática

Revisão sistemática é um tipo de estudo secundário que sumariza as evidências provenientes de estudos primários conduzidos para responder uma questão específica de pesquisa. Utiliza um processo de revisão de literatura abrangente, imparcial e reprodutível, que localiza, avalia e sintetiza o conjunto de evidências dos estudos científicos para obter uma visão geral e mais precisa da estimativa do efeito da intervenção².

1.2 Definição de Metanálise

Metanálise é uma análise estatística que combina os resultados de dois ou mais estudos independentes, gerando uma única estimativa de efeito³. Dada à natureza dos estudos de testes diagnósticos, na qual não existe uma única estatística que represente adequadamente a concordância entre o teste diagnóstico e o padrão de referência, a escolha do método de metanálise vai depender de como as medidas de desempenho são apresentadas entre os estudos⁴. Neste sentido, alguns modelos de metanálises podem ser bastante complexos e necessitar de programas estatísticos e profissionais capacitados. A escolha acerca de qual método é o mais apropriado para realizar a metanálise está diretamente relacionada ao grau de heterogeneidade entre os estudos de testes diagnósticos, incluídos na revisão sistemática.

A metanálise da acurácia de testes diagnósticos fornece estimativas da média da acurácia diagnóstica de um ou mais testes, a incerteza desta média e a variabilidade dos estudos em torno dessa média. Ainda, permite descrever como a acurácia varia em função de diferentes valores de corte e outras características do estudo. Ajuda a interpretar resultados conflitantes entre estudos, uma vez que permite identificar quais diferenças são reais, quais são explicadas pelo acaso e quais podem ser explicadas pelas características conhecidas dos estudos. A precisão da estimativa normalmente aumenta com a quantidade de dados, conferindo maior poder à metanálise para detectar diferenças reais na acurácia entre testes de estudos individuais e podendo gerar estimativas mais precisas de sensibilidade e especificidade esperadas¹.

Assim como as revisões sistemáticas de ensaios clínicos randomizados, a metanálise deve ser considerada apenas quando os estudos são similares, tanto em relação ao perfil de pacientes incluídos, quanto em relação ao protocolo de estudo. Mesmo que esses critérios sejam preenchidos, ainda pode haver heterogeneidade importante entre os estudos que torne inapropriado sumarizar o desempenho dos testes em uma única estimativa de efeito⁵. Nesse caso, apenas se apresenta o desempenho dos testes nos vários estudos, de forma sistemática.

1.3 Importância das Revisões Sistemáticas de acurácia diagnóstica

Os testes diagnósticos são utilizados para que o profissional da saúde possa discriminar se um indivíduo tem ou não uma doença ou condição particular em populações consideradas suspeitas para a doença. Geralmente, esses estudos são realizados em amostras pequenas de casos, especialmente quando a doença é rara. Nesses casos, tendem a fornecer estimativas com imprecisões consideráveis.

Portanto, o aumento da precisão da estimativa de desempenho de um teste é desejado e conseguido com o cálculo da metanálise.

Adicionalmente, a metanálise de estudos de testes diagnósticos permite investigar a consistência do desempenho do teste entre diferentes delineamentos de estudos diagnósticos, em diferentes perfis de população, conceitos esses que serão aprofundados ao longo desse manual.

1.4 Limitações das Revisões Sistemáticas de estudos de acurácia diagnóstica

Várias são as limitações das revisões sistemáticas de estudos de acurácia diagnóstica. Por serem permitidas variações no delineamento de estudo, muitas vezes a combinação dos achados desses diferentes estudos não é indicada e pode gerar estimativas enviesadas acerca da acurácia do teste.

A depender do espectro da doença e da maneira como os pacientes foram incluídos nos diferentes estudos, as estimativas de desempenho podem não ser aplicáveis em todos os perfis de população e, portanto, tais diferenças devem ser exploradas.

A maioria dos sumários estatísticos para o cálculo da metanálise de estudos diagnósticos necessita dos números que compõem a tabela de contingência 2×2 . Quando estes dados não estão explícitos, deve-se calculá-los a partir do sumário estatístico das medidas de desempenho do teste relatadas. Na ausência de algum dado necessário, o contato com o autor do estudo deve ser realizado, o que nem sempre é efetivo, além de demandar tempo.

Adicionalmente, por haver mais de uma opção de método para o cálculo de metanálise, algumas utilizando modelos matemáticos mais complexos, pode haver a necessidade de consultoria de uma pessoa especializada, como um estatístico, além da aquisição de programas estatísticos comerciais, que nem sempre estão disponíveis.

A limitação mais relevante das revisões sistemáticas de estudos de acurácia diagnóstica está pautada na questão inerente ao próprio objetivo do estudo, que é determinar as medidas de desempenho do teste. Tais desfechos não podem ser considerados de importância para o paciente. Nesse sentido, os desfechos avaliados

nos estudos de acurácia diagnóstica se comportam como os desfechos substitutos (*surrogate outcomes*), muitas vezes utilizados nos demais delineamentos. É importante destacar que estudos diagnósticos podem ser investigados em desfechos relevantes, como eventos clínicos ou para definições de condutas clínicas que são baseadas nos resultados de exames. Nesses casos, o delineamento de estudo preferencial é o ensaio clínico randomizado, utilizando as diferentes abordagens diagnósticas como intervenção e controle. O ensaio clínico randomizado difere do delineamento clássico dos estudos de acurácia diagnóstica, que têm natureza transversal, conforme será discutido ao longo desse manual.

1.5 Recursos necessários

O conjunto de recursos, conhecimentos e habilidades são decisivos para a realização de uma revisão sistemática de forma adequada. O revisor principal deve fazer uma avaliação para identificar a necessidade do auxílio que irá precisar. O tempo é fator predominante e vai depender muito da quantidade de literatura disponível sobre a questão de pesquisa. É importante planejar, em cronograma, todas as etapas, como treinamentos, reuniões, desenvolvimento do protocolo, busca dos artigos, seleção dos artigos, fichas clínicas, extração e análise dos dados, etc. Adicionalmente, deve-se prever no cronograma um tempo necessário à troca de informação com os autores dos manuscritos, a fim de obter dados essenciais que podem não estar relatados no manuscrito.

Devem ser considerados os seguintes recursos:

Recursos Financeiros: além de gastos previsíveis de materiais de consumo e escritório, papel e impressora, alguns outros gastos devem ser programados. Por exemplo, dependendo da escolha do sumário estatístico da metanálise, pode ser necessária a aquisição de programas estatísticos específicos.

Ainda, pode haver a necessidade de contratar um professor de idioma, para tradução de artigos em língua não dominada pela equipe, ou haver a necessidade de comprar artigos em texto completo, não disponíveis para os revisores. Assim, sugere-se a elaboração de um orçamento, contemplando todos os itens e respectivos valores.

Esses recursos financeiros podem ser próprios ou financiados por agências de fomento à pesquisa, instituições responsáveis pela avaliação tecnológica ou envolvidas na elaboração de diretrizes para prática clínica.

Recursos Humanos: devem ser discriminadas quantas pessoas irão compor a equipe da revisão, em quais etapas, quem será o coordenador principal da revisão, a necessidade de consultoria de especialistas no assunto ou profissional da informação (bibliotecário e/ou estatístico).

Recursos Materiais: computadores, internet, materiais de escritório, telefone, fax, papel, impressão, fotocópias, local para reuniões, programas de computador (como gerenciadores de referência) e programas estatísticos (SPSS, SAS, STATA, por exemplo).

As habilidades e conhecimentos dos revisores são fundamentais para a garantia da qualidade metodológica da revisão. Assim, são necessárias noções básicas de metodologia de revisões sistemáticas, estatística, epidemiologia, conhecimento clínico sobre a questão de pesquisa, noções de informática e domínio de, pelo menos, a língua inglesa. Além disso, a equipe deve saber utilizar as ferramentas necessárias à condução da revisão, como os gerenciadores de referência, os programas que geram a metanálise, as peculiaridades de busca nas várias bases de dados, além de habilidades de redação de protocolo e do manuscrito.

Eventualmente algumas dessas habilidades podem não ser dominadas pelo grupo que conduzirá a revisão sistemática. Portanto, pode ser necessário recorrer a uma consultoria externa, com especialistas, e esse investimento deve estar previsto em orçamento.

ETAPAS FUNDAMENTAIS DE UMA REVISÃO SISTEMÁTICA

2 PLANEJAMENTO DA REVISÃO SISTEMÁTICA

2.1 Definição da questão de pesquisa estruturada no formato do acrônimo PIRO

Em Pesquisa Clínica, independente do delineamento de estudo, a(s) questão(ões) de pesquisa a ser(em) investigada(s) deve(m) ser clara(s) e objetiva(s). Para questões de pesquisa que envolvem intervenções, o usual é estruturá-la de acordo com os componentes do acrônimo PICO, em que cada letra representa um componente da questão: P = população; I = intervenção; C = controle e O = desfecho, tradução do termo em inglês *outcome*. Para questões de pesquisa de estudos de acurácia diagnóstica, a adaptação dessa estrutura sugere o acrônimo PIRO, que corresponde as seguintes definições dos domínios:

P – População: Os testes diagnósticos têm desempenhos diferentes em diferentes populações⁶. Adicionalmente, variam com o espectro da doença, com o cenário clínico ou de acordo com a interpretação dos resultados. Portanto, é importante definir claramente a patologia ou condição clínica de interesse, o grupo populacional, o grau de severidade da doença, se dependente de idade, sexo, raça, nível sócio-educacional, bem como o cenário clínico de interesse, se cirúrgico, hospitalar, na comunidade ou outros.

I – Teste índice (*index test*): O teste índice é o teste que terá o desempenho avaliado. Pode incluir diferentes tecnologias ou alguma alternativa menos invasiva ou mais barata que o teste de referência. Um novo teste diagnóstico pode ser proposto para atuar na triagem de uma doença; para substituir o padrão de referência atual ou mesmo para ser utilizado como teste adicional ao padrão de referência utilizado, melhorando a acurácia do padrão atual.

Frequentemente o conceito tradicional da acurácia de um teste implica na dicotomização dos dados em resultados que são classificados como positivo ou negativo. Assim, qualquer revisão sistemática da acurácia de um teste terá que considerar limiares diagnósticos para cada teste índice incluídos.

R – Padrão de referência (*reference standard*): o equivalente do controle para questões de pesquisa da acurácia de testes diagnósticos é chamado de padrão de referência. Ele é o melhor teste disponível na atualidade e com quem o teste índice será comparado. Entretanto, o padrão de referência não está restrito a um único teste diagnóstico. Ele pode envolver mais de um teste ou alguma sintomatologia ou sinais no paciente, ou mesmo um período de seguimento para confirmação do diagnóstico. A seleção do padrão de referência é o ponto crítico para validar um estudo de acurácia e a definição do limiar diagnóstico de positividade faz parte da definição do padrão de referência.

O – Desfecho (*outcome*): Os desfechos de um estudo diagnóstico primário podem ser muitos, a depender da intenção do teste. Em termos de importância, idealmente os estudos envolvendo testes diagnósticos deveriam investigar o impacto de uma estratégia diagnóstica em desfechos que são importantes para o paciente. Nesse sentido, o delineamento ideal desse estudo seria um ensaio clínico randomizado clássico, onde os pacientes seriam randomizados para serem submetidos ao teste índice ou ao padrão de referência e as estratégias seriam comparadas em termos de risco para desfechos clinicamente relevantes associados à morbidade, mortalidade ou qualidade de vida. Outros desfechos possíveis seriam os que envolvem respostas emocionais, sociais, cognitivas e comportamentais decorrentes dos resultados dos testes, efeitos éticos ou legais e custos associados aos testes.

Em revisões sistemáticas de estudos de acurácia diagnóstica, o principal objetivo é apresentar as medidas de desempenho do teste entre os diferentes estudos que o avaliaram e, portanto, os desfechos são limitados aos parâmetros que avaliam esse desempenho. Se realizado o cálculo de metanálise, aumenta-se a precisão destas estimativas. Para tanto, é necessário obter os dados que irão compor a tabela de contingência 2 x 2, da qual derivam todas as medidas de desempenho do teste. A tabela descreve a relação entre os resultados do teste índice em um determinado limiar diagnóstico e o estado da doença (se presente ou ausente), definido pela aplicação do padrão de referência. A tabela inclui o número de verdadeiros positivos (aqueles que têm a doença e o teste foi positivo), falso positivos (aqueles que não têm a doença, mas o teste índice foi positivo), falso negativos (aqueles que têm a doença, mas o teste índice foi negativo) e os verdadeiros negativos (aqueles que não têm a doença e o teste foi negativo). A partir desses dados, qualquer estatística utilizada para determinar a acurácia de testes diagnósticos pode ser calculada (sensibilidade, especificidade, valores preditivos positivo e negativo, razões de verossimilhança positiva e negativa, razão de chances diagnóstica), bem como o cálculo da metanálise dos estudos.

O quadro 1 exemplifica uma questão de pesquisa estruturada para estudos de acurácia de testes diagnósticos.

Quadro 1 – Exemplo de questão de pesquisa de estudos de testes diagnósticos

Ressonância Magnética *versus* Tomografia Computadorizada na detecção de lesões vasculares agudas em pacientes com sintomas de Acidente Vascular Cerebral:

P = Pacientes com sintomas de Acidente Vascular Cerebral

I = Ressonância Magnética

R = Tomografia Computadorizada

O = Detecção de lesões vasculares agudas

2.2 Definição dos critérios de elegibilidade

Assim como para inclusão e exclusão de pacientes nos estudos clínicos primários, os critérios de elegibilidade de estudos em revisões sistemáticas devem ser definidos *a priori* e registrados em protocolo (seção 2.4).

Os critérios de elegibilidade complementam a questão de pesquisa estruturada. Pode-se estabelecer, por exemplo, a inclusão de pacientes que tenham certo grau de gravidade da doença, ou limitados a uma região geográfica com uma prevalência estabelecida de doença.

2.3 Revisão de literatura: justificativa para a revisão sistemática

Ao se definir uma questão de pesquisa para a revisão sistemática podemos nos deparar com a situação em que essa pergunta já tenha sido respondida de forma definitiva por algum estudo ou revisão sistemática prévia.

Assim, deve-se fazer uma breve busca na literatura. Para isto, pode-se recorrer ao *site* oficial da Colaboração Cochrane⁷, onde há uma base de revisões sistemáticas de acurácia de testes diagnósticos, na Cochrane Library⁸ e na Biblioteca Cochrane, disponível na BVS⁹ – Biblioteca Virtual em Saúde – Bireme – Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde, além de fontes adicionais como o CRD¹⁰ – *Centre for Reviews and Dissemination*, que são especializados em revisões sistemáticas e já incluem revisões específicas de testes diagnósticos. Essas bases contêm referências de revisões que estão em andamento (acesso ao protocolo registrado via PROSPERO).

Outra fonte essencial é a busca no PubMed/MEDLINE¹¹ por meio do *Clinical Queries*, que é uma ferramenta da base, apresentada na parte média-inferior da página inicial. No *Clinical Queries* os resultados são categorizados por tipo dos estudos: revisões sistemáticas, estudos clínicos ou estudos em genética.

Recomendamos também a busca no Sisrebrats¹², que é o banco de estudos da Rede Brasileira de Avaliação de Tecnologias em Saúde (Rebrats), composta de instituições que atuam na Avaliação de Tecnologias em Saúde (ATS) de acordo com temas que são de prioridade do SUS.

Adicionalmente, outras fontes como o *Sumsearch* e o *TripDatabase*, são úteis na busca de evidências prévias.

2.4 Documentação da metodologia: redação de protocolo

Assim como em qualquer estudo clínico primário, o protocolo de uma revisão sistemática deve ser elaborado *a priori* e tem como objetivo registrar de forma clara e

transparente todo o processo que envolve a realização da revisão sistemática, bem como definir as análises que serão realizadas. Isso é necessário, pois minimiza o risco de erros sistemáticos ou vieses, introduzidos por decisões que são influenciadas pelos achados.

O protocolo descreve as etapas realizadas na revisão e pode ser estruturado em introdução e justificativa da revisão, objetivos e metodologia. Em metodologia, devem-se registrar os critérios de elegibilidade definidos, as bases de dados a serem pesquisadas, a definição da estratégia de busca, o processo de triagem e seleção dos artigos, o processo de extração de dados, o plano de análises estatísticas, os desfechos e as análises de sensibilidade de interesse. O relato dos resultados, discussão e conclusões fazem parte da redação do manuscrito.

No caso específico de metanálise de estudos de acurácia diagnóstica, nem sempre é possível estabelecer o plano de análise estatística de forma conclusiva, já que as análises vão depender, em alguma extensão, do tipo e da quantidade de dados reportados nos estudos. Entretanto, todos os parâmetros possíveis devem ser determinados.

Uma sugestão de modelo de protocolo é o disponível no programa *Review Manager*, que estrutura a elaboração de um protocolo de revisão sistemática nos moldes exigidos para publicação da revisão na *Cochrane Library*.

O protocolo da revisão deve idealmente ser elaborado por um grupo de revisores com experiência tanto na área do conhecimento clínico, quanto técnico em metodologia de revisões sistemáticas.

Desde setembro de 2004, editores dos mais prestigiados periódicos médicos anunciaram que recusariam publicar pesquisas patrocinadas por companhias farmacêuticas a menos que os estudos fossem registrados em base de dados públicas antes de iniciados¹³. Esta medida visava garantir que o estudo fosse conduzido de acordo com o plano inicial do protocolo, independente dos resultados obtidos serem positivos ou não. Em revisões sistemáticas, já existem bases de dados que aceitam o registro do protocolo da revisão. Entretanto, esse registro ainda não é exigido pelos principais periódicos. O PROSPERO (*International Prospective Register of Ongoing Systematic Reviews*) é um novo banco internacional de registros de revisões sistemáticas, ativo desde 22 de fevereiro de 2011. Além do registro do protocolo, essa base permite pesquisar quais estudos estão em andamento.

DIRETRIZ: PLANEJAMENTO DA REVISÃO SISTEMÁTICA:

- 1) Formular questão de pesquisa estruturada de acordo com o acrônimo PIRO.
- 2) Buscar revisões prévias na literatura.
- 3) Redigir o protocolo.

3 CONDUÇÃO DA REVISÃO SISTEMÁTICA

3.1 Busca de potenciais estudos elegíveis

A busca por estudos de acurácia de testes diagnósticos deve ser abrangente o suficiente para assegurar a recuperação das evidências disponíveis para a questão de pesquisa. Acima de tudo, deve ser reproduzível para garantir sua validade.

A ajuda de um bibliotecário especialista em elaborar estratégias de busca para revisões sistemáticas, se possível, é de grande valia para auxiliar na construção desta etapa.

Estratégias de busca para identificar estudos de acurácia de testes diagnósticos não irão se traduzir em apenas um tipo de desenho de estudo e os termos utilizados devem ser escolhidos em função da questão de pesquisa estruturada.

A busca deve ser feita em fontes de dados publicados como MEDLINE, Embase, e também em fontes de dados não publicados, literatura cinzenta e estudos em andamento. Na Biblioteca Cochrane já existe uma base de dados de revisões da acurácia de testes diagnósticos publicadas, a SRDTA¹ (*Systematic Review Diagnostic Test Accuracy*) e também pela iniciativa da Colaboração Cochrane o CRDTAS¹⁴ (*Cochrane Register of Diagnostic Test Accuracy Studies*), ainda em desenvolvimento.

Segundo Golder¹⁵ e Whiting¹⁶, a busca somente pelo MEDLINE não é considerada adequada para revisões sistemáticas, pois pode levar a um potencial viés pela perda de estudos não contemplados pela base.

Recomendamos também que os autores pesquisem em outras fontes de registro de estudos e revisões da acurácia de testes diagnósticos, como o MEDION¹⁷ (*Meta-analyses van Diagnostisch Onderzoek*), C-EBLM¹⁸ (*Evidence-Based Laboratory Medicine*), ARIF¹⁹ (*Aggressive Research Intelligence Facility*) Database, HTA²⁰ (*Health Technology Assessment Database*), DARE²¹ (*Database of Abstracts of Reviews of Effects*), entre outras. O Anexo A apresenta uma tabela com o endereço eletrônico das principais bases de dados eletrônicas.

O processo da busca, que pode ser extenuante, mas necessário, irá garantir que os autores reduzam o risco para vieses como o de publicação. Existem menos evidências sobre o impacto do viés de publicação em revisões sistemáticas de estudos de acurácia diagnóstica do que em revisões de ensaios clínicos randomizados²².

Pesquisas recentes relatam que para minimizar possíveis vieses de publicação é recomendado realizar a busca em várias bases de dados eletrônicas¹⁶, como citado, e também usar outros métodos como lista de referências, especialistas na área, busca manual e fontes de literatura cinzenta.²³

3.2 Fontes de evidências

A busca por potenciais estudos sempre se inicia pelas grandes bases de dados eletrônicas, como MEDLINE e Embase. Pesquisas revelaram^{16,24,25} que MEDLINE e Embase são boas fontes de estudos de teste de acurácia diagnóstica. Indicamos também a pesquisa na LILACS (Literatura científica e técnica da América Latina e Caribe) e também outras bases como: SCOPUS, Web of Science, CINAHL, Australasian Medical Index, Chinese Biomedical Literature Database, outras bases de dados especializadas nacionais e regionais, anais de congressos, bancos de teses, etc.

Assim como para os ensaios clínicos randomizados temos o CENTRAL (*The Cochrane Central Register of Controlled Trials*), a Colaboração Cochrane está desenvolvendo um registro para estudos de teste diagnóstico, o CRDTAS (*Cochrane Register of Diagnostic Test Accuracy Studies*).

Nem todas as bases de dados eletrônicas são de livre acesso. Porém, as que são, nem sempre garantem o acesso ao manuscrito em forma de texto completo. Normalmente elas fornecem apenas o acesso à citação completa e ao resumo.

O MEDLINE é uma das mais importantes bases de dados internacionais. Contém mais de 23 milhões de citações de resumos e referências de artigos no campo da medicina, biomedicina, ciências da vida e áreas correlatas, disponíveis pelo PubMed, um serviço da Biblioteca Nacional de Medicina dos Estados Unidos. É uma base de dados que inclui citações datadas a partir de 1950, pode ser acessado livremente através do PubMed, ou por outras plataformas como o OVID (acesso via assinatura) e BVS (Bireme, livre acesso).

O Embase tem ênfase na literatura europeia, atualmente com mais de 24 milhões de citações, ultrapassando, com isto, o MEDLINE em seu conteúdo²⁶.

Realizar a busca na base de dados Embase é fundamental. Jadad e col.²⁷ demonstraram que a sobreposição dos artigos em ambas as bases, Embase e MEDLINE, é de apenas 34%. É um produto comercializado pela editora Elsevier e seu acesso é mediante assinatura.

A LILACS, coordenada pela Bireme, é o mais importante e abrangente índice da literatura científica e técnica da América Latina e Caribe, composta por 27 países. Uma vantagem é que uma grande parte dos artigos está disponível em forma de texto completo através do Scielo. É uma fonte essencial de informação quando a questão de pesquisa envolve testes diagnósticos de doenças tropicais, por exemplo²⁸.

Bases de dados de teses e dissertações também são importantes fontes de dados publicados. Normalmente não estão indexadas nas grandes bases de dados bibliográficas, como MEDLINE ou Embase. No Brasil dispomos do Banco de Teses da CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), da Biblioteca

Digital Brasileira de Teses e Dissertações do IBICT (Instituto Brasileiro de Ciência e Tecnologia) e de alguns bancos de teses coordenados por grandes universidades. Ainda, é possível pesquisar em bases internacionais de teses e dissertações, como por exemplo, a *ProQuest Dissertations e Theses Database*²⁹.

A busca na internet muitas vezes se faz necessária, mas os revisores precisam estar atentos às fontes consultadas. O Google Acadêmico é uma boa ferramenta de busca, pois permite integrar várias fontes na pesquisa em um só lugar. Recupera artigos revisados por especialistas, editores, teses, livros, resumos e artigos de editoras acadêmicas, organizações profissionais, bibliotecas de pré-publicações, universidades e outras entidades acadêmicas³⁰.

O *TRIP database (Turning Research into Practice)*³¹ também é uma ferramenta de busca e traz os resultados por categorias como: diretrizes, revisões, artigos do MEDLINE, informações para pacientes, capítulos de livros, etc.

O *Medion database* é uma iniciativa de um pequeno grupo de pesquisadores das Universidades de Maastricht (Holanda) e Leuven (Bélgica) que tem especial interesse em estudos e revisões de teste diagnóstico. A base inclui resumos e referências destes estudos. A base é estruturada em três partes:

1. Metodologia para a condução de revisões sistemáticas de estudos de testes diagnósticos;
2. Revisões sistemáticas publicadas de estudos de testes diagnósticos;
3. Revisões sistemáticas de estudos sobre testes genéticos¹⁷.

O *Evidence-Based Laboratory Medicine (C-EBLM)* é uma base de dados que integra o trabalho do comitê *International Federation of Clinical Chemistry and Laboratory*, especializados em conhecimento e metodologia para estudos e revisões sistemáticas de teste diagnóstico¹⁸.

A *ARIF databases* é uma base que inclui revisões sistemáticas sobre vários escopos sendo um deles de revisões de teste diagnóstico. Oferece uma base "*ARIF Methodology database*" que pode ajudar os autores na aplicação da metodologia para revisões sistemáticas¹⁹.

3.3 Pesquisa em fontes de dados não publicados, em andamento e literatura cinzenta

Muitos estudos são finalizados, mas não são publicados. As principais causas dessa associação seriam resultados não significativos ou negativos. A identificação destes estudos não é fácil, principalmente no contexto dos estudos diagnósticos, que são menos explorados. Uma possível maneira é por meio da busca em bases de dados específicas que registram protocolos de estudos que serão conduzidos. Como exemplo, temos o *Clinical Trials* e, no Brasil, o Registro Brasileiro de Ensaio Clínicos – Rebec

(disponível a partir de dezembro de 2010). O limitante é que essas bases de registro de protocolos de estudo priorizam os estudos de intervenção, principalmente os ensaios clínicos randomizados. A partir da disponibilidade do CRDTAS (*Cochrane Register of Diagnostic Test Accuracy Studies*), espera-se aprimorar o registro de estudos de teste diagnóstico e poder investigar melhor o viés de publicação para esse tipo de estudo.

A literatura cinzenta é aquela que não é publicada formalmente em fontes como livros e periódicos. Segundo o *Handbook* da *Cochrane*³² (cap. 6, seção 6.2.1.8), a literatura cinzenta também deve ser considerada no processo de busca das evidências. Anais de congressos (*Conference Proceedings*) são fontes importantes de evidências e podem conter dados de estudos que não foram incluídos e que podem alterar os resultados da revisão. Resumos de congressos e outros tipos de literatura cinzenta correspondem a aproximadamente 10% de estudos referenciados pelas revisões da *Cochrane*³².

O acesso à literatura cinzenta pode ser feito pela busca manual em revistas que não estão indexadas nas grandes bases de dados e, também, pela presença dos revisores em congressos e eventos na área. Nesse contexto, procura-se ativamente identificar esses estudos nas apresentações do evento ou via publicação de resumos. Algumas revistas publicam esse tipo de literatura e podem estar indexadas nas bases de dados eletrônicas disponíveis. Como exemplos dessas bases, podemos citar o *ISI Web of Knowledge*, o *British Library Inside*, o *BMC Meeting Abstracts*.

Temos ainda bases de dados específicas que cobrem a literatura cinzenta como o *OpenGrey* (anteriormente chamado de *OpenSIGLE*) sistema de informação da literatura cinzenta Européia e o *National Technical Information Service* (NTIS) que oferece acesso aos resultados de pesquisas governamentais e não governamentais dos EUA e estão disponíveis via internet³³.

O contato com colegas pesquisadores, especialistas no assunto da revisão, autores renomados e que se destacam em estudos que foram selecionados, indústrias farmacêuticas ou empresas de interesse na área também são de grande valia na busca de estudos não publicados.

Recomendamos também verificar a lista de referências dos estudos incluídos na revisão, a fim de detectar alguma evidência relevante que pode não ter sido recuperada pela estratégia de busca.

A lista de referências utilizadas para elaboração de diretrizes na área do conhecimento em estudo também pode ser utilizada. Nesse sentido, O *Guideline International Network* é uma boa opção para busca.

3.4 Elaboração da estratégia de busca nas várias bases de dados

A elaboração da estratégia de busca é a componente chave de qualquer revisão

sistemática. São as definições dos termos apropriados que irão tornar a busca mais sensível que específica, garantindo que resgataremos toda a evidência disponível. O conhecimento dos mecanismos de busca nas várias bases de dados se faz necessário, já que diferem em alguns recursos entre si.

Desde 2003 a Colaboração Cochrane vem trabalhando em um guia para ajudar os autores na elaboração do processo de busca para revisões da acurácia de testes diagnósticos. No Handbook da Cochrane para revisões da acurácia de testes diagnósticos, os autores encontrarão informações detalhadas que poderão ajudá-los no processo de elaboração da estratégia de busca. Uma dessas orientações é que se evite o uso de filtros para o desenho de estudo. Isso porque é provável que se perca estudos relevantes em razão de alguns fatores, como a inconsistência na indexação, variações no desenho do estudo ou na natureza prospectiva ou retrospectiva³³.

Além da Cochrane, os autores podem contar com a ajuda de dois grupos europeus que trabalham com a finalidade de ajudar autores de revisões de teste diagnóstico. São eles: o *Continental Europe Regional Support Unit* (CESU), em funcionamento desde fevereiro de 2007 e o *UK Regional Support Unit* (UKSU), desde setembro de 2007³³.

A elaboração da estratégia de busca deve ser pensada a partir dos componentes da questão de pesquisa estruturada no formato PIRO. Entretanto, é importante ressaltar que a busca das evidências deve ser mais sensível que específica. Nesse sentido, sugere-se não estabelecer termos para os componentes do desfecho, sob o risco de excluir evidências relevantes. Definir termos para o padrão de referência também é uma questão delicada, já que diferentes estudos que avaliam o mesmo teste índice podem ter estabelecido padrões de referência diferentes. Adicionalmente, o padrão de referência pode não ser composto de um único teste, ou mesmo ser composto com um conjunto de sintomas. Neste caso, fica ainda mais complexo estabelecer tais parâmetros. Por outro lado, a definição de termos apenas para o tipo de paciente e do teste índice pode deixar a busca muito sensível, a ponto de recuperar um grande número de referências que possa inviabilizar a triagem. Portanto, é preciso ter em mente que não existe uma regra para a construção da estratégia de busca. Deve-se realizar as tentativas necessárias, a fim de definir a mais adequada para a questão da pesquisa.

3.4.1 Seleção dos termos para a busca

Sempre que possível, deve-se utilizar o vocabulário controlado, que é o descritor de assunto. O descritor de assunto é um termo específico em cada base, que representa o assunto da pesquisa na qual o artigo foi classificado (indexado). Para o MEDLINE esse vocabulário chama-se MeSH (*Medical Subject Headings*); para o Embase, o Emtree e, para a LILACS, o DeCS (descritores em ciências da saúde).

Sugerimos que a seleção dos termos sempre se inicie pelos descritores do MeSH no MEDLINE. Para a busca no Embase, deve-se identificar o Emtree correspondente

ao MeSH, já que nem sempre um termo MeSH corresponde ao mesmo termo no Emtree. Entretanto, dependendo do termo, o Embase referencia o Emtree equivalente ao MeSH.

Abaixo, um exemplo de como o mesmo termo descritor de assunto pode variar nas bases:

Ex.: Termo no MeSH: *magnetic resonance imaging*

Ex.: Termo no Emtree: *nuclear magnetic resonance imaging*

A estratégia de busca não pode ficar restrita somente aos descritores de assunto. Ela deve ser a mais sensível possível e deve englobar também o vocabulário não controlado, que seria a utilização de palavras de texto, sinônimos, siglas, termos relacionados, palavras chave e variações de grafia. Isto vai garantir a recuperação de artigos mais antigos, pois a indexação de alguns assuntos só foi introduzida posteriormente. Para garantir uma boa recuperação das informações na LILACS recomendamos que a estratégia de busca seja elaborada com a soma dos termos utilizando o operador booleano OR nas três línguas que predominam na base: português, espanhol e inglês.

Uma estratégia que permite aumentar a sensibilidade da busca na base MEDLINE é incluir os chamados “*entry terms*”, que se encontram listados dentro da definição do termo MeSH. Estes termos representam sinônimos, indexações prévias ou derivações do assunto, que contribuem para sensibilização da estratégia. A quantidade destes termos varia de termo MeSH para termo MeSH.

No Anexo B encontra-se um exemplo desta estratégia de sensibilização da busca.

Com a definição dos termos a serem utilizados, os resultados da busca devem ser combinados utilizando os operadores booleanos, especialmente o “OR” e o “AND”. Uma explicação mais detalhada da construção da estratégia de busca e utilização dos operadores booleanos encontra-se no Apêndice 1.

3.5 Filtros de busca para estudos de acurácia de testes diagnósticos

No decorrer deste capítulo já fizemos algumas considerações sobre utilizar ou não filtros de busca para estudos da acurácia de testes diagnósticos. Sabemos que a Colaboração Cochrane trabalha arduamente nesse assunto tentando identificar e padronizar termos e estratégias que traduzam esse tipo de estudo, principalmente nas grandes bases de dados como MEDLINE e Embase. A base de Revisões Sistemáticas de Estudos de Teste Diagnóstico da Cochrane e o *Handbook for DTA Reviews* disponibilizam revisões sistemáticas Cochrane e os autores podem recorrer e utilizar essas revisões verificando as etapas de elaboração e metodologia aplicada, além do próprio manual, que é um excelente guia.

Existem muitos estudos realizados com o objetivo de avaliar a utilização de filtros específicos para recuperar estudos diagnósticos. O maior limitante dessa utilização está pautado na possibilidade de haver diversos delineamentos de estudo com o objetivo diagnóstico. Até a utilização do termo “randomizado” em estudos de acurácia diagnóstica não está inadequada, uma vez que é possível randomizar a ordem dos testes a que o paciente será submetido.

Existem grupos ou entidades que estão trabalhando no desenvolvimento de filtros, como o *UK InterTASC Information Specialists' Subgroup* (ISSG) e o *HIRU – Health Information Research Unit*, da *McMaster University*.

Temos um movimento mundial em esforços para melhorar vários aspectos dos estudos da acurácia de testes diagnósticos, como relato dos dados, identificação do estudo nas bases de dados de literatura médica, filtros de busca, etc. O *Standards for Reporting of Diagnostic Accuracy* (STARD), iniciativa de um grupo de pesquisadores e editores liderados por Bossuyt³⁴, procura ajudar autores a melhorar a qualidade de relato deste tipo de estudo, mas o impacto dessa diretriz na qualidade dos estudos publicados ainda é incerto.

Portanto, a utilização de filtros para recuperar os estudos de acurácia diagnóstica ainda não é um consenso entre os estudiosos da área. A recomendação para os autores é que utilizem como primeira opção para construção da estratégia de busca a utilização de termos para a condição da doença e para o teste índice. A utilização de termos para o padrão de referência só é encorajada se o padrão de referência for consensual e estiver consolidado na literatura.

3.6 Relatando o processo de busca

Deve-se descrever em protocolo e relatar na seção de metodologia do manuscrito a lista de todas as bases de dados que serão pesquisadas, com suas respectivas datas de abrangência (ex.: MEDLINE de 1950 a 24 de Abril de 2014). Demais fontes de busca, como literatura cinzenta, contatos, busca manual, etc, também devem estar relatadas. Esses dados irão ajudar a compor o fluxo de seleção dos artigos, que faz parte da apresentação dos resultados.

As estratégias de busca utilizadas para cada base de dados devem constar como anexo da revisão. É importante este registro para futuras atualizações da revisão sistemática e para garantir a reprodutibilidade da busca.

DIRETRIZ: ESTRATÉGIA DE BUSCA

- 1) Buscar evidência disponível nas três bases fundamentais MEDLINE, Embase, e LILACS e em outras bases eletrônicas específicas do assunto.
- 2) Buscar evidência proveniente de literatura cinzenta.
- 3) Definir quais termos serão utilizados para os itens P e I da questão estruturada no formato PIRO. A utilização de termos para o padrão de referência só deve ser utilizada se o mesmo tiver definição consensual consolidada na literatura. As utilizações de termos para o item “O” do desfecho, bem como de filtros para estudos diagnósticos devem ficar restritas às situações onde o assunto é amplamente estudado e precisa-se restringir o número de potenciais artigos elegíveis.
- 4) Combinar os termos utilizando operadores booleanos.
- 5) Registrar toda a estratégia de busca em cada base, informando a data do acesso.

3.7 Avaliação da elegibilidade dos estudos

O processo de avaliação da elegibilidade passa por uma etapa de triagem dos artigos, com leitura de título e resumo (quando disponível), e uma etapa de confirmação, pela leitura do manuscrito em forma de texto completo.

As etapas estão descritas abaixo:

3.7.1 Avaliação da elegibilidade: triagem dos artigos pela leitura de título e resumo

A busca de estudos em todas as possíveis fontes de dados gera um número muito maior de artigos do que os que realmente serão elegíveis pelos critérios estabelecidos. Isto ocorre porque a estratégia de busca é elaborada preconizando a sensibilidade em detrimento à especificidade.

Uma leitura rápida do título e resumo permite realizar uma triagem destas referências e descartar um grande número de referências que não se enquadram nos critérios de elegibilidade estabelecidos pela revisão.

Não existe a obrigatoriedade de utilizar um gerenciador de referências para a triagem dos artigos, mas as facilidades dessa ferramenta em relação à organização das referências, praticidade e otimização de tempo são inquestionáveis e devem preferencialmente ser utilizadas. Existem inúmeros programas gratuitos e comerciais, nas mais diversas plataformas de acesso. O Apêndice 2 fornece mais informações sobre essas ferramentas.

O somatório dos artigos recuperados em todas as bases de dados deve ser registrado para constar na elaboração do fluxo de seleção dos artigos, que faz parte da apresentação dos resultados. Ainda, um mesmo artigo pode estar indexado

em mais de uma base de dados, de modo que teremos referências duplicadas no arquivo. O número de artigos em duplicata também deve ser apresentado como resultados no fluxo de seleção dos artigos (Anexo C). Após o registro desses valores, excluem-se os artigos em duplicata. Esse arquivo é então dividido pela(s) dupla(s) de revisores, que devem trabalhar individualmente, excluindo os artigos que claramente não preenchem os critérios de elegibilidade e mantendo os possivelmente incluídos.

A triagem pela leitura de título e resumo deve ser realizada por dupla de revisores, de maneira independente. Muitas vezes o resumo não está disponível e, nesse caso, se o título for sugestivo de inclusão, o artigo permanece na base e passa para a etapa seguinte, de avaliação da elegibilidade pela leitura do texto completo.

As discordâncias entre os revisores devem ser resolvidas por consenso. Em alguns casos, pequenas divergências entre os revisores podem ser desconsideradas, visto que estas referências terão sua elegibilidade confirmada na etapa seguinte. Por outro lado, se o resultado entre a dupla for muito divergente, vale a pena que os revisores resolvam as discordâncias, para não gastar tempo de forma desnecessária, recuperando os textos completos de artigos que não são elegíveis na triagem. A decisão acerca do modo como as discordâncias devem ser resolvidas deve estar pautada na disponibilidade da equipe e cronograma da revisão: às vezes é mais fácil resolver entre a dupla, que já leu o estudo previamente, do que envolver uma terceira pessoa, que teria que iniciar o trabalho, sujeito à sua disponibilidade de agenda. A concordância entre os revisores pode ser mensurada usando a estatística Kappa de Cohen³⁵. Quanto mais próximo de um for esta estatística, maior é a concordância entre os revisores.

3.7.2 Avaliação da elegibilidade pela leitura do manuscrito em texto completo e ficha clínica padronizada

Todos os artigos que foram triados na fase anterior têm sua elegibilidade confirmada pela leitura mais detalhada do estudo, através do texto completo do artigo. Assim como na etapa de triagem, a confirmação da elegibilidade é realizada por uma dupla de revisores, de modo independente.

Nessa etapa, a razão primária da exclusão deve ser registrada para compor o fluxo de seleção dos artigos (Anexo C). Para guiar essa etapa, utiliza-se uma ficha clínica padronizada, que contém basicamente os critérios de elegibilidade estabelecidos. Esta ficha deve conter uma folha de rosto com a identificação do título da revisão (um mesmo revisor pode realizar mais de uma revisão sistemática simultaneamente), nome do revisor e identificação do estudo, ou acrônimo do estudo, autor e jornal. O Anexo D apresenta um modelo de folha de rosto de ficha clínica.

Havendo discordância entre os revisores, estas devem ser resolvidas ou por consenso ou por um terceiro revisor. Ao final do processo, teremos, finalmente, o total de estudos que são de fato elegíveis e que vão compor a revisão sistemática. A concordância entre os revisores pode ser mensurada pela estatística Kappa de Cohen³⁶.

DIRETRIZ: AVALIAÇÃO DA ELEGIBILIDADE DOS ARTIGOS

- 1) Somar os resultados de busca de todas as bases.
- 2) Preferencialmente, utilizar um gerenciador de referências para avaliação da elegibilidade.
- 3) Remover as duplicatas dos artigos.
- 4) Triar os artigos pela leitura de título e resumo (quando disponível).
- 5) A triagem dos artigos deve ser realizada por dupla de revisores, de forma independente. As discordâncias podem ser resolvidas por consenso.
- 6) Confirmar a elegibilidade dos artigos triados pela leitura do texto completo do artigo.
- 7) Na etapa de confirmação da elegibilidade, utilizar a ficha clínica contendo os critérios de elegibilidade, a fim de registrar os motivos de exclusão.
- 8) A confirmação da elegibilidade deve ser realizada por dupla de revisores, de forma independente.
- 9) Resolver as discordâncias por consenso ou por meio de um terceiro revisor.
- 10) Se desejável, aplicar o teste estatístico Kappa para quantificar a concordância entre os revisores.

3.8 Extração de dados

3.8.1 Processo de extração de dados

Consideram-se dados de uma revisão sistemática quaisquer informações sobre o estudo, incluindo detalhes de métodos, participantes, cenário clínico, testes utilizados, dados para desfechos e resultados. Sua extração é sempre guiada por uma ficha clínica padronizada, elaborada previamente. Assim como na etapa de seleção dos artigos, a extração de dados também é realizada por dupla de revisores de maneira independente. A obtenção dos dados que compõem a tabela 2x2 é particularmente importante, uma vez que, geralmente, estes dados ou não são reportados^{36, 37}, ou estão incompletos. Os dados da tabela 2x2 podem precisar ser calculados a partir das estimativas de desempenho de um teste, utilizando suas definições matemáticas. Entretanto, algumas destas medidas de desempenho necessárias para derivar os demais cálculos podem não estar relatadas e, nestes casos, será necessário contatar os autores do estudo primário, o que nem sempre é uma tarefa concluída com sucesso, além de demandar tempo.

Uma vez extraídos os dados, as informações são confrontadas entre os revisores. Havendo discordâncias nos dados coletados, estas podem ser resolvidas ou por consenso entre a dupla ou por consulta de um terceiro revisor.

Para extração correta e padronizada dos dados, sugere-se o agendamento de um treinamento prévio entre todos os revisores que participarão da coleta de dados, para instruções.

3.8.2 Elaboração da ficha clínica padronizada

A elaboração da ficha clínica é fundamental para a produção dos resultados. Ela deve permitir coletar dados de todas as variáveis que forem consideradas importantes para interpretação e aplicabilidade dos resultados. Por isso, deve ficar a cargo de pessoas familiarizadas com a patologia de interesse, com o teste diagnóstico e com o método padrão de referência atualmente utilizado. Se necessário, pode-se consultar um especialista na área.

O quadro 2 resume as principais variáveis que devem compor a ficha clínica de extração de dados. Assim como na ficha clínica de avaliação de elegibilidade, a ficha clínica de extração de dados deve possuir uma folha de rosto no mesmo formato. Com a coleta de dados das variáveis de interesse geram-se resultados descritivos e quantitativos.

3.8.2.1 Características dos estudos incluídos e dados descritivos

Apresentar na seção de resultados uma tabela de “Características dos estudos incluídos” é mandatório em manuscritos de revisões sistemáticas. Esta tabela permite a comparabilidade das variáveis entre os estudos que podem afetar a magnitude do desempenho do teste diagnóstico. Além disso, auxiliam na interpretação crítica e validação externa dos resultados.

A variação das características de base dos pacientes ou características metodológicas entre os estudos é chamada de diversidade clínico-metodológica e pode ser fonte de inconsistência entre os achados. Em metanálises de ensaios clínicos randomizados, se a diversidade clínico-metodológica é significativa, mas as análises estatísticas que permitem quantificar a heterogeneidade não a demonstram, podemos, então, atribuir maior validade externa à intervenção. Em caso de metanálises de estudos de testes diagnósticos, essa assertiva não se aplica, já que as estimativas de desempenho do teste serão afetadas pela diversidade. Em estudos de acurácia de testes diagnósticos, não se indica realizar metanálise de estudos com diversidades importantes, se as fontes não forem exploradas, sob o risco de estimar erroneamente um desempenho do teste diagnóstico. Como exemplo, podemos realizar a metanálise de estudos que apresentem boa qualidade metodológica *versus* qualidade inadequada, ou de acordo com o delineamento do estudo. Se forem realizadas análises de sensibilidade, as mesmas devem estar previstas *a priori*, em protocolo.

Estudos que avaliem um mesmo teste diagnóstico podem ter diferentes padrões de referência estabelecidos. É importante descrever em ficha clínica essa característica do estudo, bem como as definições utilizadas para esses parâmetros. Estas

informações são importantes, pois podem impactar na estimativa de desempenho do teste e, portanto, devem ser avaliadas.

Quadro 2 – Variáveis consideradas para elaboração de ficha clínica de extração dos dados

Participantes

Número de pacientes no estudo

Média de idade

Proporção de homens e mulheres

Histórico de doenças

Comorbidades

Parâmetros clínicos de interesse para a situação clínica (pressão arterial, dados de exames de base, índice de massa corporal, etc.)

Métodos

Delineamento dos estudos incluídos

Espectro da doença da população incluída

Método de aplicação do teste índice e padrão de referência

Intervenção (teste índice)

Descrição do(s) teste(s) a ser(em) avaliado(s)

Metodologia do teste

Características dos equipamentos para realização do teste (se aplicável)

Resultados

Para cada desfecho: coleta de variáveis categóricas e/ou numéricas.

Medidas sumárias do desempenho do teste índice (sensibilidade, especificidade, valores preditivos, razões de verossimilhança, etc.)

Limiares de positividade

Dados necessários para preenchimento da tabela de contingência 2X2.

Número de pacientes excluídos por conta de resultados inconclusivos/ indeterminados.

Padrão de referência

Descrição do(s) teste(s) considerado(s) padrão de referência (tipo do teste, necessidade de seguimento ou sinais e sintomas, se o padrão de referência for composto, etc.), bem como limiares de positividade ou definições de categorização da positividade

Fonte: elaboração própria.

3.8.2.2 Avaliação do risco de viés dos estudos incluídos

O acesso à qualidade metodológica dos estudos incluídos em uma revisão sistemática é importante independente do desenho do estudo primário. Particularmente em estudos de acurácia diagnóstica, existem questões específicas em termos de desenho que são diferentes da abordagem de estudos de intervenção. Se essas não forem identificadas, comprometem o resultado do estudo.

O desenho de um estudo de acurácia diagnóstica é transversal (ou *cross-sectional*) por definição, no qual uma série de pacientes consecutivos com a suspeita da doença ou condição-alvo é submetida ao teste índice. Então, todos esses pacientes devem, também, ser submetidos ao mesmo padrão de referência. A ordem da aplicação dos testes índice e padrão de referência podem ou não ser randomizada. O teste índice e o padrão de referência devem ser interpretados por pessoas que estejam cegas em relação aos resultados do outro teste e as várias medidas de concordância são calculadas, como por exemplo, sensibilidade, especificidade, razão de verossimilhança, razão de chances diagnóstica, curvas ROC, entre outras. Adicionalmente, o tempo decorrente entre a aplicação dos testes não pode se estender a ponto de alterar o grau de gravidade da doença. Esse desenho clássico permite variações, incluindo diferenças na maneira com que os pacientes são selecionados para o estudo, de como são interpretados os testes índices e padrão de referência, a natureza prospectiva ou retrospectiva da coleta dos resultados, entre outras. Algumas dessas diferenças enviesam os resultados do estudo, enquanto outras limitam a aplicabilidade dos resultados em diferentes perfis de pacientes. O Apêndice 3 descreve melhor os desenhos de estudos de acurácia diagnóstica mais utilizados.

Estudos que contêm vieses produzirão estimativas de desempenho do teste que diferem do seu valor real. Por outro lado, variabilidades surgem de diferenças entre os estudos, por exemplo, em termos de população incluída, cenário, protocolo de teste ou definição da condição alvo do teste³⁴, e se referem à generalização do uso do teste diagnóstico. Assim, apesar da variabilidade não acarretar em estimativas de desempenho enviesadas, ela limita a aplicabilidade do teste na prática clínica e é uma importante consideração quando se avaliam estudos de acurácia diagnóstica, principalmente no contexto de uma revisão sistemática.

Em revisões sistemáticas de estudos de acurácia diagnóstica, além de se distinguir entre viés e variabilidade, é necessário que essas informações sejam sistematicamente apresentadas. Se as diferenças nas características dos estudos limita a aplicabilidade do teste na prática clínica, os vieses do estudo têm impacto direto na estimativa de desempenho do teste, levando à obtenção de resultados errôneos. Portanto, a avaliação estruturada da qualidade metodológica é chave na validade interna dos estudos de acurácia incluídos na revisão sistemática.

Nesse sentido, é importante a distinção entre fonte de viés e variabilidade entre os estudos. Entretanto, isso nem sempre é fácil, e a utilização de diferentes definições e referências entre os estudos complica ainda mais essa questão.

De maneira geral, os vieses mais comumente apresentados em estudos diagnósticos podem ser resumidos conforme as descrições abaixo.

Viés de Verificação (ou *work-up bias*): Viés de verificação ocorre sempre que a amostra dos pacientes no estudo é submetida ao teste índice, mas nem todos os pacientes são submetidos ao teste padrão de referência. Assim, a acurácia do teste é relatada apenas para os pacientes que tiveram o estado de doença validado pelo padrão de referência⁶. Essa situação é comum de ocorrer quando o teste padrão de referência envolve algum procedimento invasivo, como biópsia, por exemplo. Assim, a grande parte dos pacientes que será submetida à verificação pelo padrão de referência terá resultados positivos do teste índice. A sensibilidade (teste positivo quando a doença está presente), portanto, estará elevada nestes casos. Como apenas uma pequena parcela dos pacientes que tiveram o resultado do teste índice negativo serão submetidos ao padrão de referência, poucos pacientes que não têm a doença terão tido o resultado do teste índice negativo. Assim, a especificidade (ausência da patologia em pacientes com resultado de teste negativo) estará subestimada.

Viés de Incorporação: Ocorre quando o teste índice faz parte do padrão de referência, ou seja, o teste índice e o padrão de referência não são independentes, levando a superestimação da sensibilidade e especificidade³⁸. É importante notar que o padrão de referência é assumido como tendo 100% de acurácia, o que não representa a realidade. Por exemplo, o diagnóstico de metástase de câncer de fígado nunca poderá ser definitivamente determinado até a realização de uma autópsia. Em alguns contextos o padrão de referência é composto de informação clínica ou por uma bateria de outros testes³⁹. Mesmo o mais definitivo padrão de referência pode ser considerado inaccurado, o que leva a super ou subestimação da verdadeira acurácia do novo teste⁴⁰.

Viés de Inspeção: Ocorre quando o conhecimento prévio do resultado de um dos exames influencia a interpretação do outro. Para evitar este viés, é necessário que o estudo seja cego, em que ambos os testes são interpretados sem o conhecimento das características clínicas dos pacientes ou do resultado do outro teste, para assegurar que apenas a contribuição diagnóstica do teste é que está sendo avaliada.

Desenho do estudo: Estudos diagnósticos de acurácia são, por definição, estudos transversais. O delineamento clássico do estudo diagnóstico de acurácia permite incluir uma população sob o risco de apresentar a situação clínica e, portanto, representativa da utilização do teste na prática. Em epidemiologia, esse desenho de estudo pode ter a terminologia de estudos de coorte, porque os indivíduos são incluídos antes do desfecho final (presença ou ausência da condição alvo) ser

conhecido. Variações desse desenho clássico são permitidas de estudos de acurácia diagnóstica. Nos casos em que o *status* da doença é previamente conhecido antes do paciente ser submetido ao teste índice, esses estudos são conhecidos como estudos de caso-controle diagnóstico, conceito esse que é diferente do delineamento caso-controle em epidemiologia. Um detalhamento dos diferentes desenhos de estudos de acurácia de testes diagnósticos e seu impacto nas estimativas de desempenho encontra-se no Apêndice 3.

Outros exemplos de fontes de viés e variabilidade entre os estudos são apresentados no Anexo E.

Para avaliarmos o risco de viés em estudos de acurácia diagnóstica, e de maneira análoga às revisões sistemáticas de ensaios clínicos randomizados, a utilização de escores de qualidade não é recomendada. É preferível que essas características metodológicas sejam avaliadas de forma individual, por domínio, que representa as etapas de como o estudo foi elaborado e conduzido⁴¹. Escores de avaliação da qualidade metodológica de estudos diagnósticos não atribuem pesos distintos para itens que são fontes de viés e variabilidade entre os estudos envolvidos⁴². Assim, acabam por não ser adequados para avaliar o risco de viés nos estudos, visto que a importância dos itens e a direção do potencial viés associado a esses geralmente variam de acordo com o contexto em que são aplicados, mas são ignorados pela maioria dos *scores* de qualidade^{43, 44}.

Dentre as opções de ferramentas disponíveis para avaliar tais características, recomenda-se a utilização do QUADAS-2⁴⁵ – *Quality Assessment of Diagnostic Accuracy Studies*. Trata-se da versão atual da ferramenta mais comumente utilizada em revisões sistemáticas dos estudos de acurácia diagnóstica.

Essa ferramenta é recomendada por diversas organizações, entre elas a Agency for Healthcare Research and Quality, Cochrane Collaboration e o National Institute for Health and Clinical Excellence de Londres.

A ferramenta original QUADAS incluía 14 questões sequenciais que avaliavam o risco de viés, as fontes de variabilidade (agora denominada como aplicabilidade) e questões relacionadas ao quão adequado o estudo estava relatado no artigo, em termos de características metodológicas e resultados. Cada item era classificado como “sim”, “não” ou “incerto”, com o “sim” representando uma boa resposta. Essa versão original foi revista, pois apresentava algumas limitações relacionadas à sobreposição de alguns dos itens e situações em que a ferramenta era de difícil utilização, como nos casos que o padrão de referência envolvia um período de seguimento, por exemplo. Adicionalmente, o grupo de trabalho da Cochrane questionava que alguns itens eram relacionados à problemas de relato dos dados, ao invés de questões relacionadas ao risco de viés propriamente dito.

O QUADAS-2 atual está estruturado em quatro domínios-chave, que representam as principais fontes de vieses. Esses domínios referem-se à: 1) Seleção de pacientes, 2) Teste índice, 3) Padrão de referência e 4) Fluxo e tempo.

Cada domínio é avaliado em termos de risco de viés e, exceto para o domínio “fluxo e tempo”, em termos de aplicabilidade do teste. Vale destacar que a aplicabilidade refere-se ao quão os estudos primários se assemelham à questão da pesquisa da revisão, em termos de população, exames e condição-alvo.

Diversas questões norteadoras auxiliam no julgamento do risco de viés em cada domínio. Para o julgamento da aplicabilidade, apesar de estruturado de forma similar, não são incluídas questões norteadoras.

Na avaliação do risco de viés, para cada um dos quatro domínios, as questões norteadoras devem ser respondidas como “sim”, “não” ou “incerto”. Se todas as respostas forem “sim”, significa baixo risco para viés; se qualquer questão for respondida como “não”, alerta para o risco de viés. Ao final das respostas, julga-se o risco de viés como “baixo”, “alto” ou “incerto”, sem a atribuição de pontuações. A aplicabilidade também é julgada como “baixa”, “alta” ou “incerta”.

No julgamento do risco de viés é possível omitir ou incluir questões norteadoras em cada um dos domínios, a depender do desenho primário do estudo. Os critérios para omissão seriam excluir as questões que não se aplicam ao estudo e para inclusão, a necessidade de alguma questão norteadora que não está coberta no domínio. A recomendação é que não adicione questões sem necessidade, sob o risco de adicionar complexidade desnecessária. Quando a versão modificada for julgada como finalizada, dois revisores devem avaliar a concordância do julgamento para um estudo e somente se adequada, utilizar para os demais estudos incluídos na revisão. Caso contrário, a versão modificada deve ser revista.

A sumarização dos resultados pode ser representada em formas tabulares (Figura 1) ou gráficas (Figura 2) e são bastante úteis para a interpretação dos resultados do QUADAS-2.

A tradução não validada do QUADAS-2 encontra-se no Anexo F.

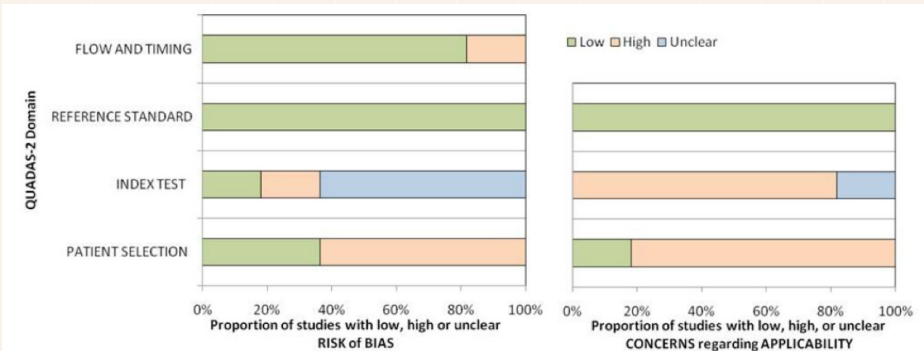
Figura 1 – Apresentação tabular dos resultados do QUADAS-2 para os estudos incluídos

Study	RISK OF BIAS				APPLICABILITY CONCERNS		
	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD	FLOW AND TIMING	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD
Study 1	😊	😊	😊	😊	😞	😊	😊
Study 2	😊	😊	😊	😊	😞	😊	😊
Study 3	😞	😞	😊	😊	😞	😊	😊
Study 4	😞	😞	😊	😊	😞	😊	😊
Study 5	😞	?	😊	😊	😞	😊	😊
Study 6	😞	?	😊	😊	😞	?	😊
Study 7	😞	?	😊	😊	😞	😊	😊
Study 8	😞	?	😊	😊	😞	?	😊
Study 9	😞	?	😊	😊	😞	😊	😊
Study 10	😞	?	😊	😞	😞	😊	😊
Study 11	😊	?	😊	😞	😊	😊	😊

😊 Low Risk 😞 High Risk ? Unclear Risk

Fonte: elaboração própria.

Figura 2 – Apresentação gráfica dos resultados do QUADAS-2 para os estudos incluídos



Fonte: QUADAS-2 – Background Document. Disponível em: <<http://www.bris.ac.uk/quadas/quadas-2/>>. Acesso em: out. 2013

Como principais vantagens da ferramenta, podemos citar a transparência no julgamento e a padronização dos critérios. Adicionalmente, a ferramenta apresenta flexibilidade, permitindo omitir ou inserir questões norteadoras nos domínios e ampliando seu contexto de aplicação. As limitações da ferramenta referem-se à natureza subjetiva do julgamento dos domínios, principalmente relacionados à aplicabilidade, onde o julgamento não está facilitado pelas questões norteadoras. Assim, o julgamento pode variar a depender do revisor que está utilizando a ferramenta.

3.8.2.3 Extração dos dados quantitativos

Os dados quantitativos são aqueles que permitirão o cálculo da metanálise, se aplicável, ou a apresentação sistemática do desempenho de um teste. Estes dados são apresentados pelas medidas de desempenho de um teste diagnóstico e permitirão compor a tabela de contingência 2x2 do estudo.

A acurácia pode ser apresentada de diferentes maneiras. As medidas de desempenho mais comumente reportadas pelos estudos são sensibilidade, especificidade, razões de verossimilhança, valores preditivos e curvas ROC. O total de participantes “doentes” e “não-doentes” é necessário para calcular as probabilidades pré-teste e pós-teste de doença. Se possível, as tabelas 2x2 devem ser construídas para todos os subgrupos relevantes.

As definições dessas medidas e as abordagens para os cálculos das metanálises são apresentados na seção de resultados a seguir.

DIRETRIZ: EXTRAÇÃO DE DADOS

- 1) Elaborar ficha clínica padrão com campos para coleta de dados descritivos (características e risco de viés dos estudos incluídos) e dados quantitativos (medidas de desempenho do teste).
- 2) Extrair os dados de cada estudo incluído na revisão sistemática.
- 3) A extração dos dados deve ser realizada por dupla de revisores, de forma independente.
- 4) Resolver as discordâncias por consenso ou por um terceiro revisor.

3.9 Resultados

Independente do número de artigos elegíveis para a revisão sistemática, nessa etapa todos os dados de interesse, contidos em ficha clínica, devem ter sido extraídos dos estudos em consenso.

De posse desses dados, podem-se compilar as informações em forma de tabelas e utilizar os dados numéricos para calcular a metanálise, se aplicável.

3.9.1 Apresentação dos dados descritivos

É importante para o leitor que sejam apresentados os dados demográficos (idade, sexo, raça, número de pacientes incluídos nos estudos), bem como dados metodológicos do estudo, como natureza prospectiva ou retrospectiva, definições de padrão de referência, período de tempo entre a aplicação do teste índice e do padrão de referência, metodologia dos testes índices e referência, dados de prevalência, etc.

Devido à natureza assimétrica da maioria dos testes diagnósticos (alguns testes são bons para excluir uma doença, outros para confirmar), é importante apresentar os resultados complementares, inclusive a sensibilidade e a especificidade de cada estudo.

Uma vez que esses dados estiverem compilados em forma de tabelas, é possível ter uma ideia qualitativa acerca da diversidade clínico-metodológica entre os estudos. Entretanto, conforme será discutido mais adiante, essa é apenas uma das fontes para heterogeneidade entre os estudos, de forma que ela deve ser investigada mais formalmente.

3.9.2 Sumário dos efeitos do tratamento nos estudos, cálculo e apresentação da metanálise

A abordagem para o cálculo da metanálise e avaliação da heterogeneidade depende do sumário estatístico que será selecionado para as análises.

As abordagens para metanálises de estudos de acurácia diagnóstica podem ser agrupadas em três categorias, de acordo com a maneira que elas tratam a natureza binária dos dados da acurácia do teste:

- Métodos que consideram independentes as análises de cada aspecto de desempenho do teste;
- Métodos que sumarizam as medidas de desempenho do teste em uma única medida estatística;
- Métodos que usam modelos estatísticos que simultaneamente consideram as duas dimensões do desempenho do teste.

A avaliação da acurácia de um teste requer conhecimento de duas grandezas, a sensibilidade e a especificidade. Os métodos para o cálculo da metanálise para acurácia diagnóstica têm que lidar com duas medidas sumárias simultaneamente, ao invés de uma (como nos casos das metanálises de estudos de intervenção).

Por conta da aplicabilidade e das limitações de cada método estatístico possível para o cálculo da metanálise (que serão discutidos a seguir), sugere-se que a decisão acerca de qual abordagem metanalítica utilizar esteja baseada na presença do efeito do ponto de corte, se existente, e da heterogeneidade entre os estudos.

Para tanto, é importante entender quais são as estatísticas que representam as medidas de desempenho do teste e como elas se relacionam, para depois entender os métodos possíveis utilizados para calcular a metanálise do desempenho entre os estudos.

3.9.2.1 Tipos de variáveis e medidas de desempenho em estudos diagnósticos

Quando um novo teste ou uma alternativa ao teste que já é utilizado na prática clínica é desenvolvido, seu desempenho deve ser testado perante o melhor teste que se tem na atualidade para diagnosticar a patologia de interesse. Esse “melhor teste disponível” é chamado de padrão de referência e é assumido como quem define a verdadeira presença ou ausência da doença. Portanto, os resultados do teste a ser testado (o teste índice) serão comparados aos resultados obtidos pelo teste padrão de referência, com o objetivo de identificar se este teste índice tem desempenho aceitável em relação ao padrão de referência para ser utilizado na prática clínica.

A acurácia de um teste diagnóstico será avaliada pela medida da habilidade de um teste em detectar a presença de uma doença. O verdadeiro *status* de doença de cada indivíduo é considerado como um tipo de dado dicotômico ou binário porque o indivíduo “tem a doença” ou “não tem a doença”. Apesar de essa suposição representar uma simplificação da realidade de um diagnóstico, ela é utilizada pela maioria dos testes disponíveis na prática clínica.

Entretanto, nem sempre os resultados são apresentados como variáveis dicotômicas. Eles podem ser apresentados como:

- Dados ordinais: os resultados dos testes de imagem geralmente são reportados em categorias, em forma de descrição verbal, como 1=definitivamente normal; 2=resumidamente normal; 3=incerto 4=presumidamente anormal; 5=definitivamente anormal;
- Dados discretos: contagem, como o número de eventos observados;
- Dados contínuos: os resultados dos testes são reportados como escala contínua, por exemplo a concentração de uma substância.

A maioria da categorização ordinal ou binária emerge pela aplicação de limiares de positividade (ou valores de corte) em intervalos dos dados contínuos. O estudo primário fornece esse valor de limiar, que foi definido de acordo com a intenção do teste (se teste de triagem, teste diagnóstico ou prognóstico) ou por algum outro parâmetro justificado. Para esse valor de corte, então, é possível compor a tabela de contingência 2x2, apresentando o número de casos enquadrado em cada situação da tabela, conforme representação abaixo.

Tabela 1 – Tabela de contingência 2 x 2.

	Doença (D+)	Não-Doença (D-)	Total
Teste índice positivo (T+)	Verdadeiros positivos (a)	Falsos positivos (b)	Testes positivos (a+b)
Teste índice negativo (T-)	Falsos negativos (c)	Verdadeiros negativos (d)	Testes negativos (c+d)
Total	Com doença (a+c)	Sem doença (b+d)	N (a+b+c+d)

Fonte: elaboração própria.

Nesse tipo de tabela, considera-se nas colunas a situação de doença como presente (D+) ou ausente (D-). A definição de presença ou ausência da doença é feita mediante aplicação do padrão de referência. Nas linhas dessa tabela estariam os casos em que a doença foi considerada positiva pelo resultado do teste índice (ou seja, os casos em que os valores do teste índice ficaram acima do ponto de corte definido pelo estudo) e os casos em que a doença foi considerada ausente, também baseado nos resultados do teste índice (que foram abaixo do valor de corte definido). Com os dados inseridos na tabela de contingência, todas as estatísticas de desempenho do teste índice podem ser estimadas.

A sensibilidade de um teste é a probabilidade condicional do teste ser positivo dada a presença da doença.

A especificidade de um teste é a probabilidade condicional do teste ser negativo dada a ausência da doença.

O valor preditivo positivo do teste índice representa a probabilidade de um indivíduo ter realmente a doença, dado que apresentou um resultado positivo do teste índice.

O valor preditivo negativo do teste índice representa a probabilidade de um indivíduo não ter realmente a doença, dado que apresentou um resultado negativo do teste índice.

A razão de verossimilhança é muito útil no processo de tomada de decisão, pois ela define o desempenho do teste diagnóstico ou regra de predição clínica para a confirmação ou afastamento de determinada suspeita diagnóstica.

A razão de verossimilhança positiva é dada pela razão entre a probabilidade de se encontrar um teste positivo em quem tem a doença sobre a probabilidade de se encontrar um teste positivo em quem não tem a doença. Portanto, a razão de verossimilhança positiva diz quantas vezes é mais provável um resultado de teste positivo em quem tem a doença do que em quem não a tem.

A razão de verossimilhança negativa é dada pela razão entre a probabilidade de se encontrar um teste negativo em quem tem a doença sobre a probabilidade de se encontrar um teste negativo em quem não tem a doença. Portanto, a razão de verossimilhança negativa diz quantas vezes é mais provável um resultado de teste negativo em quem tem a doença do que em quem não a tem.

Quando a razão de verossimilhança se aproxima de um, isso quer dizer que o teste não é capaz de mudar a chance pós-teste da doença. Razões de verossimilhança menores do que um diminuem a chance pós-teste de doença quando o resultado do teste é negativo. Razões de verossimilhança maiores do que um aumentam a chance pós-teste de doença, quando o resultado do teste é positivo.

A razão de chances diagnóstica sumariza a acurácia diagnóstica do teste índice como um único número que descreve quantas vezes maiores são as chances de se obter um resultado positivo em uma pessoa com a doença do que em uma pessoa sem a doença. Tem pouca relevância clínica direta, mas é muito importante para permitir o cálculo da metanálise dos estudos, como será discutido mais adiante.

Estudos primários que avaliam um teste em diferentes limiares de positividade, geralmente, apresentam uma curva ROC. Essa curva analisa a acurácia de um único teste em uma única população, em diferentes limiares de positividade. O gráfico plota a sensibilidade *versus* “1 – especificidade”. Desse modo, uma medida global da acurácia do teste é obtida pelo cálculo da área sob a curva ROC, em que um valor de 0,5 é obtido se o teste não tem aplicabilidade clínica (linha diagonal do gráfico) e um valor de um se o teste é perfeito. Para cada valor de corte estabelecido tem-se sensibilidade e especificidade diferentes.

A representação matemática, bem como exemplos de cálculo e um maior detalhamento desses conceitos estão apresentados no Apêndice 4.

3.9.2.2 Modelos de Análise

Assim como nas metanálises de ensaios clínicos randomizados, há dois modelos básicos que podem ser usados no sumário estatístico dos resultados de estudos individuais de acurácia diagnóstica.

O modelo de efeito fixo assume que todos os estudos seriam como certa amostra aleatória de um grande estudo comum, e que eventuais diferenças nos desfechos seriam resultado apenas de erro randômico. O sumário estatístico é mais simples, e essencialmente consiste em calcular uma média ponderada do resultado de cada estudo. O peso de cada estudo representa o inverso da variância do parâmetro estudado. Este modelo pode ser utilizado na metanálise individual das sensibilidades e especificidades e é utilizado para as curvas ROC sumária (SROC) e para um parâmetro utilizado no modelo hierárquico HSROC.

O modelo de efeito randômico assume que em adição à presença do erro randômico, eventuais diferenças entre os resultados dos estudos também podem representar diferenças reais entre as populações estudadas e os procedimentos avaliados. O fator para calcular o peso de cada estudo na metanálise é mais complexo matematicamente, e inclui a variância intra e inter-estudos. Logo, é utilizado para alguns dos parâmetros dos dois modelos hierárquicos de metanálise (bivariado e HSROC).

A heterogeneidade é esperada nos resultados dos estudos de acurácia de testes diagnósticos, portanto modelos de efeito randômico são preferidos para descrever a variabilidade da acurácia do teste entre os estudos.

3.9.2.3 Métodos Estatísticos

A escolha do método metanalítico para sumarizar os resultados dos estudos diagnósticos vai depender da variabilidade observada nos resultados dos estudos. Existem várias fontes de diversidade clínico-metodológica entre os estudos, além de variabilidades decorrentes do efeito de ponto de corte explícito, que se refere à presença de diferentes pontos de corte para positividade em estudos que avaliaram o mesmo teste índice.

Se os diferentes estudos incluídos reportaram o mesmo ponto de corte para o teste, tratar as estimativas de desempenho do teste como medidas independentes está correto. Por outro lado, se existe um efeito do ponto de corte explícito essa abordagem não é adequada, já que as estimativas de desempenho estão correlacionadas e não serão independentes. Nesses casos é preferível utilizar a curva ROC sumária, que proverá uma estimativa mais acurada.

Modelos hierárquicos são abordagens mais eficazes para estimar o desempenho do teste, pois consideram a variabilidade nos estudos e entre os estudos. Podem ou não considerar covariáveis adicionais no modelo. Entretanto, a necessidade de programas estatísticos específicos, bem como pessoas capacitadas para sua utilização, limitam seu uso na maioria das metanálises de estudos diagnósticos de acurácia.

Abaixo, seguem os fundamentos das abordagens metanalíticas mais comuns utilizadas nas revisões sistemáticas de estudos de acurácia diagnóstica.

3.9.2.3.1 Metanálises individuais de sensibilidade e especificidade (*summary operating point*)

Sensibilidade e especificidade são proporções simples (ver Apêndice 4 para definições matemáticas) e como tal podem ser metanalisadas da forma clássica, considerando suas médias ponderadas pelo inverso da variância, tanto pelo modelo

de efeito fixo (inverso da variância padrão)⁴⁶ quanto pelo modelo de efeito randômico (*DerSimonian and Laird*)⁴⁷.

Nesses casos, as medidas de sensibilidade e especificidade seriam tratadas como independentes e podem ser calculadas metanálises para cada um desses parâmetros, apresentados da forma clássica de gráfico de floresta (ou seja, seriam apresentados dois gráficos de floresta, um para sumarizar a sensibilidade e outro para sumarizar a especificidade entre os estudos).

Essa abordagem só deve ser utilizada se diferentes estudos que avaliaram o mesmo teste índice também reportaram o mesmo ponto de corte para positividade do teste (ou seja, o efeito do ponto de corte está ausente).

Variações nas estimativas de sensibilidade e especificidade são esperadas mesmo se diferentes estudos reportem o mesmo ponto de corte. Isso porque, além do erro aleatório, existem diversos outros parâmetros que podem diferir, como diferenças nos fundamentos dos métodos diagnósticos, diferenças de geração entre os equipamentos, questões relacionadas à calibração do equipamento, diferenças inter-observadores, etc. Um teste para heterogeneidade pode ser aplicado. Em caso positivo, as fontes de diversidade devem ser investigadas.

Nos casos em que diferentes estudos investigam o mesmo teste índice e reportam pontos de cortes distintos, a utilização dessa abordagem metanalítica irá produzir uma estimativa que não irá refletir a verdadeira média ponderada, produzindo estimativas errôneas acerca do desempenho do teste. Portanto, não deve ser realizada. Nos casos em que os diferentes estudos reportem pontos de corte distintos, a curva ROC sumária é o método preferido em relação a esta abordagem.

3.9.2.3.2 Curva ROC sumária (SROC)

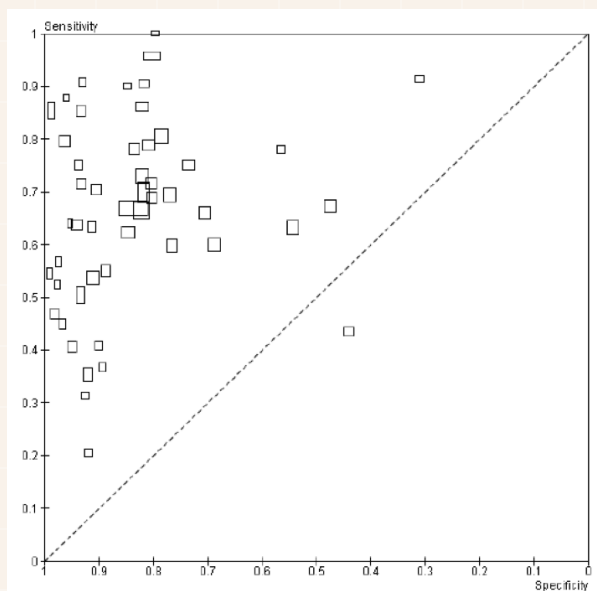
Nos estudos primários, os possíveis valores de corte para positividade de um teste índice e seus respectivos valores de sensibilidade e especificidade podem ser graficamente representados por uma curva ROC, que apresenta a sensibilidade no eixo vertical (y) e 1-especificidade no eixo horizontal (x). Via de regra, por meio dessa curva ROC, os estudos ou derivam uma medida de acurácia do teste, baseado no cálculo da área sob a curva ROC, ou definem um valor de corte para a positividade do teste, baseado em algum critério (pode ser o ponto na curva que maximiza a sensibilidade e especificidade do teste ou baseando o limiar de positividade de acordo na sua intenção, ou seja, se utilizado para triagem, diagnóstico ou prognóstico).

Se os estudos da revisão sistemática apresentarem valores diferentes para positividade do mesmo teste índice, podemos utilizar esses diferentes pontos de corte e correspondentes pares de sensibilidade e especificidade para estimar uma

nova curva ROC. Esse método de cálculo de metanálise é chamado de curva ROC sumária ou SROC.

A SROC é um modelo de regressão linear, proposto por Moses e Littenberg^{48, 49}. Inicialmente, cada estudo contribui com um ponto de corte na curva, que corresponde a um valor único de sensibilidade-especificidade. Esta disposição gráfica é chamada de ROC *plot*. O tamanho do ponto reflete a precisão das estimativas, conforme ilustração abaixo:

Figura 3 – Exemplo hipotético de um ROC plot



Fonte: elaboração própria.

Para a estimativa da curva SROC, dois parâmetros devem ser calculados para cada estudo: um representado por D e outro por S.

O parâmetro D é a variável dependente do modelo e representa o logaritmo natural da razão de chances diagnóstica (lnRCD) do estudo, demonstrado matematicamente por:

$$D = \ln[S/(1-S)] - \ln[(1-E)/E], \text{ em que } S \text{ é a sensibilidade e } E \text{ a especificidade.}$$

O parâmetro S é a variável independente do modelo e representa uma medida da proporção total de resultados positivos do teste demonstrado matematicamente por:

$$S = \ln[S/(1-S)] + \ln[(1-E)/E], \text{ em que } S \text{ é a sensibilidade e } E \text{ a especificidade.}$$

A partir do cálculo das estimativas D e S de cada estudo, um modelo de regressão linear simples é calculado, com esses parâmetros se correlacionando conforme:

$$D = \alpha + \beta S.$$

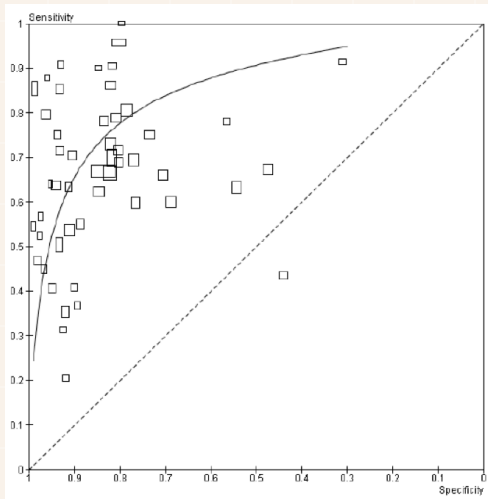
Em que α é uma constante (denominada intercepto) e β é um parâmetro de forma da curva. Os valores de α e β são calculados pelo modelo.

Esse modelo de regressão linear utiliza o modelo de efeito fixo não-ponderado ou considera cada estudo pelo inverso da variância do lnRCD.

A SROC é produzida para calcular a sensibilidade esperada ao longo de um intervalo de valores escolhidos para especificidade. O intervalo de especificidades escolhido, em que a curva será desenhada, é geralmente restrito ao intervalo observado nos dados dos estudos incluídos, a fim de evitar extrapolações.

Após a realização dos cálculos, a representação gráfica da nova curva ROC estimada pode ser ilustrada conforme:

Figura 4 – Exemplo hipotético de uma curva SROC



Fonte: elaboração própria.

Como limitações desse método, temos que ele não permite considerar a variabilidade intrínseca dos estudos e a variabilidade entre os estudos. Se a análise visual do gráfico for sugestiva de heterogeneidade, esse modelo só permite fazer análises descritivas/exploratórias preliminares. O método não permite fazer inferências estatísticas formais, que requerem análises mais complexas e o uso de modelos multiníveis (hierárquicos).

Adicionalmente, não permite determinar intervalos de confiança e valores de p das estimativas de desempenho do teste.

3.9.2.3.3 Modelos Hierárquicos de metanálise de estudos de acurácia diagnóstica

Modelos hierárquicos são preferíveis para o cálculo de metanálise de estudos de acurácia diagnóstica, pois permitem considerar a variabilidade no estudo (erro amostral) e a variabilidade entre os estudos (heterogeneidade).

Dois métodos são comumente mais utilizados: o modelo bivariado⁵⁰ e o modelo hierárquico da SROC (HSROC)⁵¹.

Os dois modelos hierárquicos envolvem distribuições estatísticas em dois níveis. No primeiro nível, modelam as contagens das células da tabela 2x2, extraídas de cada estudo utilizando distribuição binomial e transformações logarítmicas das proporções. No segundo nível, efeitos randômicos de estudos são assumidos considerando a heterogeneidade na acurácia do teste diagnóstico entre os estudos, correlacionadas com a variabilidade amostral no primeiro nível. O modelo bivariado e o modelo HSROC são matematicamente equivalentes quando não se considera ajuste por covariável, mas diferem em parametrizações. A parametrização do modelo bivariado modela sensibilidade, especificidade e a correlação direta entre elas, enquanto que a parametrização da HSROC modela funções da sensibilidade e especificidade para definir uma curva ROC sumária.

Os modelos de regressão múltipla apresentam definições matemáticas que são de difícil interpretação para o público em geral e fogem do escopo deste manual. Nesse sentido, apenas os parâmetros utilizados para sua construção serão comentados. É importante destacar que, independente do racional matemático para a construção do modelo, na prática poucos dados são requeridos para a obtenção dos resultados desses modelos. Esses dados são basicamente as contagens das células da tabela 2x2, obtidas a partir das medidas de desempenho do teste. Entretanto, é necessário utilizar programas estatísticos específicos, uma vez que o Review Manager não roda tais modelos (ver seção 3.9.2.4).

Ainda, é possível estender o modelo hierárquico para a inclusão de covariáveis (fontes de diversidade clínico-metodológica entre os estudos). Esses modelos estendidos são os considerados ideais para o cálculo de metanálise de estudos de acurácia diagnóstica. Entretanto, conforme será discutido na seção 3.9.2.4, rodam especificamente no programa estatístico SAS, por meio da publicação da macro do SAS que automatiza o modelo⁵².

a) Modelo bivariado

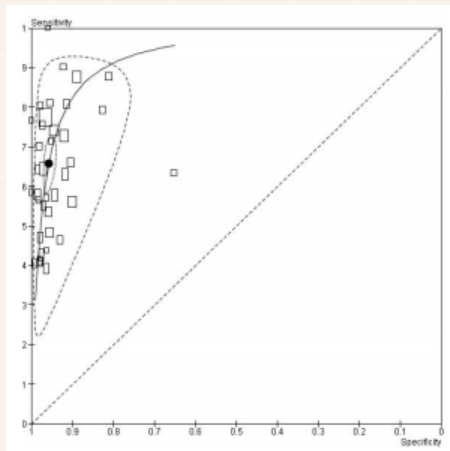
O método bivariado modela a sensibilidade e especificidade diretamente em dois níveis correspondentes às variações nos estudos e entre eles⁵⁰. No primeiro nível, assume-

se que a variabilidade da sensibilidade e especificidade para cada estudo tem uma distribuição binomial. No segundo nível, as transformações logarítmicas da sensibilidade e especificidade para cada estudo são assumidas como de distribuição normal.

O primeiro nível é então correlacionado com o segundo por meio de um modelo binomial único, que considera ambos os níveis, construído em função de cinco parâmetros (logarítmicos de sensibilidade e especificidade, das variâncias da sensibilidade e especificidade e de correlação), quando nenhuma covariável é inserida no modelo. O modelo gera estimativas numéricas para esses parâmetros que permitem, por meio de alguns cálculos adicionais, obter a metanálise da sensibilidade e especificidade. Tais cálculos envolvem transformações inversas de funções logarítmicas da sensibilidade e especificidade, utilização dos valores de erro padrão para cálculo do intervalo de confiança, entre outros.

A representação gráfica do ponto que sumariza a sensibilidade e a especificidade na curva ROC também pode ser construída e o ponto que sumariza o par sensibilidade-especificidade está destacado em preto, conforme ilustração abaixo:

Figura 5 – Exemplo hipotético de uma SROC utilizando o modelo bivariado



Fonte: elaboração própria.

b) Modelo HSROC

O método HSROC modela a acurácia do teste diagnóstico. O objetivo primário é estimar uma curva ROC sumária, porém considerando dois níveis de variabilidade, uma nos estudos e outra entre os estudos ⁵¹.

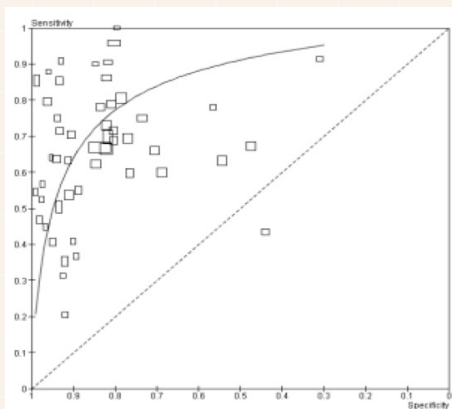
No primeiro nível, as contagens de testes positivos nos grupos doentes e não-doentes são assumidas como de distribuição binomial. Ao contrário do modelo bivariado, que

trabalha com as estimativas de sensibilidade e especificidade, este modelo trabalha no segundo nível com as medidas de razão de chances diagnósticas, por meio de sua transformação logarítmica.

O modelo resulta em parâmetros estimados para acurácia, via lnRCD (por meio do valor de α – intercepto, como descrita para a SROC) e sua variância, para o ponto de corte e sua variância, e para o parâmetro β (relacionado à inclinação da reta que permite o desenho da curva). Os primeiros dois parâmetros são colocados como modelo de efeitos randômicos de distribuição normal. Já o parâmetro de forma β só pode ser estimado como modelo de efeito fixo.

A HSROC pode então ser construída, ao definir um intervalo de valores de “1-especificidade” e utilizar a estimativa média do parâmetro de acurácia e do parâmetro β para calcular os correspondentes valores de sensibilidade, conforme ilustração abaixo:

Figura 6 – Exemplo hipotético de uma curva HSROC



Fonte: elaboração própria.

3.9.2.4 Programas para cálculo de metanálise

Assim como em metanálises de ensaios clínicos randomizados, existem programas gratuitos e comerciais que calculam metanálises de estudos de acurácia diagnóstica.

Para algumas abordagens de metanálises recomenda-se o programa Review Manager disponibilizado pela Colaboração Cochrane, pois ele tem uma interface de fácil utilização para o usuário e não se limita à realização de análises estatísticas: ele permite redigir o protocolo da revisão de acordo com o padrão de publicação da Cochrane, acessa o risco de viés dos estudos incluídos apresentando em forma gráfica e tem uma interface com o GRADE, que é o programa que permite graduar a qualidade da evidência da metanálise. Entretanto, o Review Manager não permite

utilizar todas as abordagens de metanálises possíveis discutidas neste manual. Nesse sentido, o programa só permite estimar a metanálise das sensibilidades e especificidades tratadas de forma independente (*summary operating point*) e pelo método SROC.

Ele não roda nenhum modelo hierárquico de metanálise (modelo bivariado e HSROC) e, portanto, não gera nenhuma estimativa para os parâmetros do modelo. Por outro lado, uma vez que estes modelos são rodados em programas estatísticos específicos, o Review Manager permite que os valores destas estimativas sejam utilizados para construir a forma gráfica de apresentação dos resultados dos modelos hierárquicos bivariado e HSROC, conforme foi ilustrado nas seções anteriores.

Uma outra opção bastante utilizada é o programa Meta-DiSc⁵², de *download* gratuito para uso acadêmico. Esse programa também apresenta uma interface amigável e é específico para metanálises de estudos de acurácia diagnóstica.

Está disponível no endereço eletrônico <http://www.hrc.es/investigacion/metadisc_en.htm>.

Dentre suas funcionalidades, é possível descrever os resultados primários dos estudos e explorar a heterogeneidade.

O programa calcula estimativas de acurácia e intervalos de confiança de estudos individuais e apresenta os resultados como tabulações numéricas ou gráficas, em dois formatos: gráficos de floresta (*forest plot*) para sensibilidade e especificidade, razões de verossimilhança ou razão de chances diagnóstica; e apresentação dos resultados dos estudos individuais em curvas ROC, com ou sem uma curva SROC. Por outro lado, as opções de análises utilizando os modelos hierárquicos (modelo bivariado e HSROC) ainda estão em desenvolvimento.

Em relação à exploração da heterogeneidade, com o Meta-Disc é possível explorar o efeito do ponto de corte, bem como realizar as estatísticas Cochran-Q, Chi-quadrado e o I^2 . O programa realiza análises de metaregressão univariada e multivariada, utilizando os modelos Moses & Littenberg⁴⁹ (ponderado pelo inverso da variância ou tamanho de amostra e não ponderado), de modelos de efeito fixo ou randômico.

O Meta-Disc utiliza como sistema operacional o Windows (versão 95 ou maior), de processador Pentium ou equivalente, com um mínimo de 32 MB de RAM e mínimo de 20 MB de disco rígido.

Para as análises que utilizam os modelos hierárquicos, existem como opções os pacotes para HSROC e o MADA no "R" (gratuitos) e o metandiplot no STATA, como opção comercial. Esses pacotes utilizam como covariáveis apenas as variáveis para

parametrização dos modelos e não são estendidos para inclusão de outras variáveis que representem as diversidades clínicas, por exemplo.

Pacotes comerciais permitem rodar modelos estendidos do modelo bivariado ou do HSROC, em que é possível inserir covariável(is) no modelo, para explorar mais formalmente as fontes de heterogeneidade e sua(s) influência(s) nas estimativas de desempenho do teste. A automatização desses modelos no SAS, utilizando o Proc NLMIXED, pode ser encontrado por meio da publicação dessa MACRO para o SAS pelo grupo da Cochrane de estudos diagnósticos⁵³.

3.9.3 Investigando a heterogeneidade

Conforme comentado ao longo deste manual, a diversidade clínico-metodológica entre os estudos é uma característica esperada em estudos de acurácia diagnóstica.

Essas diferenças podem ser resultantes do acaso, de erros no cálculo dos índices de acurácia ou da verdadeira diversidade clínico-metodológica⁵⁴. Estas são decorrentes das diferenças no desenho do estudo, condução, características dos participantes, intervenções, gravidade da doença, teste índice ou padrão de referência (tempo de aplicação/submissão, aspectos técnicos dos equipamentos ou materiais utilizados, variações laboratoriais ou inter/intra-observadores)^{5,55}.

Outra fonte de diversidade que é exclusiva dos testes diagnósticos é o chamado efeito do ponto de corte⁵. Ele representa as variações nos valores de corte escolhidos para a positividade do teste entre diferentes estudos que investigaram um mesmo teste índice.

Quando há suspeita de heterogeneidade entre os estudos incluídos na metanálise, as fontes e seus efeitos na estimativa da metanálise devem ser investigados.

A maneira pela qual a heterogeneidade será investigada vai depender da abordagem metanalítica adotada. Nesse sentido, para cálculos de metanálise que consideram medidas de sensibilidade e especificidade independentes, um teste I^2 de Higgs² alto aponta a presença de heterogeneidade. Nesses casos, análises exploratórias podem ser realizadas. As limitações das análises exploratórias são que elas apenas servem para fazer inferências. Ainda, quando poucos estudos são incluídos na metanálise, reestimar os parâmetros considerando subgrupos pode não ser viável, por limitação de tamanho de amostra e geração de resultados muito imprecisos.

De maneira análoga, explorar a heterogeneidade quando a abordagem metanalítica utilizada foi a SROC só é possível por meio de análises exploratórias e descritivas. A disposição dos estudos nos ROC *plots* permite inspecionar visualmente as posições dos estudos no gráfico e sugerir a presença de heterogeneidade que pode ser explorada por

análises de sensibilidade/subgrupos. Uma análise de sensibilidade bastante comum consiste em identificar os estudos *outliers*, excluí-los e recalculer as estimativas, a fim de verificar sua influência nas estimativas de desempenho do teste.

Nos modelos hierárquicos convencionais, apenas as influências dos parâmetros utilizados para o modelo são exploradas e, portanto, apenas análises exploratórias/descriptivas podem ser realizadas, conforme abordagem anterior.

O único método que permite explorar formalmente a heterogeneidade dos estudos diagnósticos de acurácia são os modelos de regressão hierárquicos estendidos, que permitem a inclusão da(s) covariável(eis) no modelo, quantificando sua influência. Esses modelos são rodados em programas estatísticos específicos. A automatização de tais modelos está disponível para o programa estatístico SAS⁵², publicada pelo grupo de trabalho em estudos diagnósticos da Cochrane.

3.10 Avaliação do risco de viés na Revisão Sistemática

Assim como vieses podem ser identificados nos estudos primários, as revisões sistemáticas também não estão isentas de risco de viés, que podem comprometer a qualidade da evidência gerada. Nas revisões sistemáticas, vieses são introduzidos de várias maneiras no processo de seleção dos estudos. Outros vieses são introduzidos no processo de condução da revisão sistemática, quando as características metodológicas adequadas não são contempladas. É importante que estes vieses sejam identificados e evitados, para que os resultados da revisão possam ser validados.

O processo de seleção dos estudos é fundamental para garantir uma revisão sistemática de qualidade, na qual toda a evidência disponível para uma questão específica de pesquisa em nível global possa estar sistematicamente apresentada e/ou sumarizada em uma única estimativa de efeito. A construção da estratégia de busca adequada, mais sensível que específica, a busca em diversas fontes de dados e literatura cinzenta e a não utilização de limites temporais ou de idioma garantem a recuperação de todas essas evidências.

Atenção especial deve ser dada a um tipo de viés chamado de viés de múltiplas publicações, em que os autores de um mesmo estudo publicam o resultado em diversos periódicos. Se estas n publicações não forem identificadas como um mesmo estudo, as populações serão somadas e os resultados computados n vezes, comprometendo a metanálise.

Ainda no processo de identificação dos estudos, o risco de viés de publicação, que seria a tendência dos estudos com resultados nulos ou negativos não serem publicados, está sempre presente e deve ser investigado. Com a ausência de dados negativos, as estimativas tendem a se apresentar superestimadas. O viés de

publicação é tão relevante em revisões sistemáticas que métodos formais utilizados para estimar sua presença serão discutidos na seção seguinte.

Durante a condução da revisão sistemática, a utilização de fichas clínicas de avaliação da elegibilidade e de extração de dados utilizadas por dupla de revisores, de forma independente, garante a obtenção mais acurada dos dados, menos susceptível a erros de interpretação de texto.

Por fim, o conhecimento e utilização dos métodos estatísticos adequados e as alternativas aplicáveis para o cálculo da metanálise contribuem para a validade interna dos achados.

3.10.1 Avaliando o viés de publicação

Estudos que evidenciam um efeito benéfico tendem a ser mais publicados do que estudos com resultados negativos⁵⁶. Há muitos fatores que predispõem a esse problema. Os autores podem desistir de escrever o artigo por acharem que não há significância nos resultados ou os editores das revistas podem não demonstrar interesse em publicar o artigo devido ao desinteresse por parte dos leitores⁵⁷. Por isso a importância de pesquisar fontes de dados não publicados e reunir a totalidade da evidência.

O gráfico de funil é o método mais conhecido para avaliar o viés de publicação⁵⁸. Trata-se de um gráfico que plota uma estimativa dos achados do estudo contra o tamanho de amostra ou alguma medida de precisão. A medida indireta do formato do gráfico é que permite inferir a presença ou ausência do viés de publicação. Assim, na ausência de viés de publicação, os estudos estarão dispersos em formato de funil invertido no gráfico. Na base, estariam os estudos com menores tamanhos de amostras, que podem apresentar variações nas tendências da estimativa do efeito por mero acaso, já que não têm poder nem precisão suficientes para demonstrar uma estimativa pontual confiável. À medida que estudos maiores são publicados, a estimativa de efeito tende a se concentrar em um valor médio, cada vez mais preciso, configurando o vértice do funil. Se na base do funil tivermos ausências de pontos (que representam os estudos), principalmente do lado direito da figura, sugere-se a presença de viés de publicação.

Entretanto, a inspeção visual do gráfico é considerada muito subjetiva, e inúmeros métodos estatísticos formais são utilizados para acessar quantitativamente essa assimetria em revisões sistemáticas de ensaios clínicos randomizados². O manual da Cochrane recomenda a utilização de tais estatísticas quando um mínimo de dez estudos estiverem incluídos na revisão, caso contrário estes testes não apresentarão poder estatístico adequado para acessar o risco de viés de publicação³².

Em revisões sistemáticas de estudos de acurácia diagnóstica, os determinantes do viés de publicação são diferentes dos aplicados para ensaios clínicos randomizados. Em contraste, não há a definição de uma hipótese nula ou o cálculo de um valor de p associado. Utilizar esses métodos para acessar o risco de viés de publicação em estudos de acurácia diagnóstica pode produzir resultados enganoso^{59,60}.

Para revisões de testes diagnósticos, gráficos de funis separados para sensibilidade e especificidade (após suas transformações logarítmicas) são improváveis de serem úteis para detectar efeito de tamanho de amostra porque tais parâmetros irão variar em função dos valores de corte e erro aleatório. A interpretação simultânea de dois gráficos de funil relacionados e dois testes para detectar a assimetria da curva também representa um desafio. A medida da razão de chances diagnóstica (RCD) sumariza a acurácia do teste e é utilizada para sumarizar a curva ROC em metanálises. Neste sentido, os gráficos de funil podem ser construídos baseados no lnRCD como parâmetro para o eixo X e com parâmetros relacionados ao tamanho da amostra no eixo Y do gráfico⁵⁸. Os testes para assimetria do funil seriam baseados em testes de regressão e de correlação de postos, sendo os de regressão aqueles que apresentam maior poder para detecção dessa assimetria. Um $p < 0,10$ é significativo para assimetria do funil e representa, conseqüentemente, a presença de publicação na revisão sistemática de estudos de acurácia diagnóstica.

DIRETRIZ: RESULTADOS

- 1) Apresentar tabela com características de base relevantes dos estudos incluídos e, pelo menos, as principais estimativas de desempenho do teste nos estudos.
- 2) Apresentar fluxo de seleção dos artigos.
- 3) Identificar a abordagem metanalítica adequada e definir o método estatístico para o cálculo.
- 4) Explorar as fontes de heterogeneidade das metanálises, de acordo com a abordagem metanalítica definida.
- 5) Avaliar o viés de publicação da revisão sistemática, se aplicável.

4 RELATO E APLICABILIDADE DOS RESULTADOS

4.1 Estruturando a discussão e conclusão

Uma discussão estruturada ajuda no relato das considerações e implicações da revisão para a prática clínica. Alguns tópicos, baseados no Handbook da Cochrane³² e literatura⁶² auxiliam nessa estruturação.

4.1.1 Estruturando a discussão

Na discussão, resumem-se os principais achados e as incertezas pendentes, procurando não repetir os principais resultados na forma quantitativa.

Outro componente importante na discussão é apresentar os pontos fortes e as limitações da revisão.

Adicionalmente, devem-se comparar os achados da revisão com os resultados de outras revisões sistemáticas publicadas previamente ou com dados de estudos importantes incluídos na revisão e discutir as principais razões que corroboram ou divergem entre os estudos.

Na tabela abaixo seguem, resumidamente, as sugestões de estrutura para discussão:

Tabela 2 – Estruturação da discussão da Revisão Sistemática

Resumir os principais resultados
Discutir a abrangência geral e aplicabilidade da evidência
Concordâncias e discordâncias com outros estudos e revisões
Destacar os pontos fortes da revisão
Potenciais vieses e limitações da revisão

Fonte: elaboração própria.

4.1.2 Estruturando a conclusão

A conclusão sumariza as ideias principais desenvolvidas na discussão, podendo ser divididas em duas seções:

4.1.2.1 Implicações para prática clínica

O significado dos achados nas implicações para a prática clínica deve ser tão inequívoco quanto possível. Não devem ir além da evidência que foi revisada e ser justificada pelos dados apresentados na revisão.

4.1.2.2 Implicações para pesquisas futuras

Se a metanálise foi julgada como conclusiva, isto deve ser relatado na conclusão. Por outro lado, metanálises que demonstram não haver diferenças entre os grupos não necessariamente significam que não haja diferenças entre eles. A depender da imprecisão da estimativa (evidenciada pelo intervalo de confiança), a metanálise pode não ter demonstrado diferença por não ter atingido um tamanho de amostra ou número de eventos suficientes que lhe confira poder estatístico adequado para evidenciar tais diferenças. Nesses casos, é importante explorar essa alternativa, enfatizando a necessidade de novos estudos para responder à questão, com metodologia e parâmetros de estimativas adequados.

5 ETAPAS OPCIONAIS

5.1 Avaliação da qualidade da evidência

Em estudos de intervenção, a qualidade da evidência proveniente da metanálise pode ser graduada de acordo com o sistema GRADE^{2,62}, proposto pelo grupo *Grades of Recommendation, Assessment, Development and Evaluation*. Tem como objetivo graduar a qualidade da evidência da metanálise por meio de critérios padronizados que permitem rebaixar e aumentar a qualidade dessa evidência. Ainda, permite inserir essa evidência no contexto da aplicabilidade clínica, através da força da recomendação. Esse sistema tem sido adotado por mais de 70 organizações envolvidas na elaboração de diretrizes e revisões sistemáticas, entre elas a Organização Mundial da Saúde, *American College of Physicians*, *American Thoracic Society*, *UpToDate* e a *Cochrane Collaboration*.

As categorias do GRADE para qualidade da evidência implicam no gradiente de confiança da estimativa do efeito do teste diagnóstico para os desfechos que são considerados importantes para o paciente⁶³.

Evidências de alta qualidade provêm de ensaios clínicos randomizados que diretamente compararam o impacto da estratégia diagnóstica alternativa em desfechos importantes para o paciente (por exemplo, estudos que avaliaram os níveis de BNP na insuficiência cardíaca), com ausência de limitações no desenho e condução do estudo, imprecisão (isto é, poder em detectar diferenças nos desfechos importantes para os pacientes), inconsistência, evidência indireta e viés de publicação.

No âmbito do diagnóstico, apesar dos estudos de acurácia também iniciarem como nos ensaios clínicos randomizados de alta qualidade, frequentemente a qualidade da evidência assim se mantém⁶⁴. Isso porque a evidência deve ser rebaixada no quesito “evidência indireta” (*indirectness*) quando os desfechos mensurados nos estudos ficam limitados aos desfechos de acurácia, que se comportam como desfechos substitutos para os desfechos que são considerados importantes para o paciente. Portanto, principalmente no contexto da força da recomendação, é necessário se fazer inferências sobre o impacto do teste em desfechos que são importantes para o paciente, o que reduz a qualidade da evidência e força da recomendação. Esse é o principal desafio em utilizar a abordagem GRADE para classificar a qualidade da evidência e força da recomendação em estudos clássicos de acurácia diagnóstica.

Para a graduação da evidência considerando os demais quesitos, os conceitos são bastante similares e podem ser resumidos de acordo com a tabela abaixo, que descreve os fatores que rebaixam a qualidade da evidência para os estudos de acurácia diagnóstica, bem como eles diferem em relação a outras intervenções:

Tabela 3 – Fatores que diminuem a qualidade da evidência de estudos de acurácia diagnóstica e como eles diferem dos demais critérios de classificação para outras intervenções.

Fatores que determinam e podem diminuir a qualidade da evidência	Explicações e diferenças da qualidade da evidência para outras intervenções
Limitações no desenho do estudo	<p>Crítérios diferentes para estudos de acurácia: estudos do tipo transversal (<i>cross-sectional</i>) ou coorte em pacientes com incerteza diagnóstica e comparações diretas dos resultados do teste com um apropriado padrão de referência são considerados como de alta qualidade e pode mover-se para moderada, baixa ou muito baixa qualidade dependendo de outros fatores. Adicionalmente, para não perder pontos nesse quesito, pacientes consecutivos devem ser recrutados como uma única coorte e não classificados pelo estado de doença; a seleção e o processo de encaminhamento devem estar claramente descritos; os testes devem ser realizados em todos os pacientes da mesma população; o novo teste e o padrão de referência devem estar bem descritos; os avaliadores dos resultados devem estar cegos em relação aos resultados do teste índice e do padrão de referência.</p>
Evidência indireta (<i>indirectness</i>)	<p>Em nível de desfecho, a limitação intrínseca ao estudo de acurácia é a ausência de evidência direta sobre o impacto do teste em desfechos importantes para o paciente. Os autores acabam tendo que fazer inferências sobre o balanço entre a influência presumida do teste nos desfechos importantes de quaisquer diferenças nos verdadeiros e falsos positivos e verdadeiros e falsos negativos em relação às complicações e custos dos testes. Por esse motivo, estudos de acurácia fornecem baixa qualidade de evidência para fazer recomendações devido à evidência indireta para os desfechos, de forma similar aos desfechos substitutos no caso de estudos de tratamento.</p>

Continua

Conclusão

Fatores que determinam e podem diminuir a qualidade da evidência	Explicações e diferenças da qualidade da evidência para outras intervenções
Evidência indireta (<i>indirectness</i>)	<p>Em relação à população, a qualidade da evidência pode ser rebaixada nesse quesito se diferenças importantes existirem entre a população estudada e a população na qual o teste seria recomendado.</p> <p>Em relação aos testes estudados, a qualidade da evidência pode ser rebaixada nesse quesito se existirem diferenças importantes nos testes investigados e na experiência diagnóstica dos responsáveis por aplicá-los nos estudos em relação aos profissionais que os utilizariam na prática clínica. Também poderia ser rebaixada se os testes que estivessem em análise não tivessem sido comparados em um mesmo estudo, mesmo se comparados com um terceiro teste comum, considerado como padrão de referência.</p>
Inconsistência	<p>Inconsistências inexplicáveis na sensibilidade, especificidade ou razões de verossimilhança podem reduzir a qualidade da evidência.</p> <p>Ainda, intervalos de confiança largos para estimativas da acurácia do teste ou proporções de verdadeiros e falsos positivo e negativo podem reduzir a qualidade da evidência nesse quesito.</p>
Viés de publicação	O risco é provável quando a fonte de evidências é apenas de estudos muito pequenos ou como resultado da assimetria do gráfico de funil

Fonte: Schönemann e cols.⁶⁴

6 CONCLUSÕES DA DIRETRIZ

Neste documento, procurou-se abordar de forma simples e prática as principais etapas necessárias para a elaboração e condução de uma revisão sistemática de estudos de acurácia diagnóstica de qualidade.

Estudos de acurácia diagnóstica apresentam características e desafios específicos, que devem ser compreendidos, a fim de estimarmos de forma adequada o desempenho de um teste. Nesse sentido, a escolha da abordagem metanalítica é fundamental. Os fundamentos matemáticos das abordagens metanalíticas são complexos e não se enquadram nos objetivos desse manual. São necessários programas estatísticos específicos para esse fim, bem como uma equipe apta à sua utilização.

Deve-se ressaltar que esta diretriz não esgota de forma alguma o assunto, de modo que o pesquisador que pretende realizar uma revisão deverá se aprofundar nos tópicos abordados por meio de leitura específica.

A programação da revisão sistemática é fundamental. A definição da equipe, das etapas que cada um irá participar, de um cronograma exequível, dos recursos necessários, da necessidade de consultoria, entre outros, são os fatores que irão garantir resultados válidos e com aplicabilidade para a prática clínica.

REFERÊNCIAS

1. THE COCHRANE COLLABORATION. Diagnostic Test Accuracy Working Group. **Handbook for DTA reviews**. Disponível em: <<http://srdta.cochrane.org/>>. Acesso em: maio 2013.
2. BRASIL. Ministério da Saúde. Secretaria de Ciência, Tecnologia e Insumos Estratégicos. **Diretrizes metodológicas: elaboração de revisão sistemática e metanálise de ensaios clínicos randomizados**. Brasília: Ministério da Saúde. 2012. 92 p.
3. COOK, D. J.; MULROW, C. D.; HAYNES, R. B. Systematic reviews: synthesis of best evidence for clinical decisions. **Annals of Internal Medicine**, Philadelphia, v. 126, n. 5, p. 376-380, 1997.
4. IRWING, L. et al. Guidelines for meta-analyses evaluating diagnostic tests. **Annals of Internal Medicine**, Philadelphia, v. 120, n. 8, p. 667-676, 1994.
5. DEEKS, J. J. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. **British Medical Journal**, London, v. 323, n. 7305, p. 157-162, 2001.
6. RANSOHOFF, D. F.; FEINSTEIN, A. R. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. **The New England Journal of Medicine**, Waltham, v. 299, n. 17, p. 926-930, 1978.
7. THE COCHRANE COLLABORATION. Disponível em: <<http://www.cochrane.org/>>. Acesso em: maio 2013.
8. THE COCHRANE COLLABORATION. **The Cochrane Library**. Disponível em: <<http://www.thecochranelibrary.com/view/0/index.html>>. Acesso em: maio 2013.
9. BIBLIOTECA VIRTUAL EM SAÚDE (BVS). Disponível em: <<http://www.bireme.br/php/index.php>>. Acesso em: maio 2013.
10. UNIVERSITY OF YORK. **Centre for Reviews and Dissemination - CRD**. Disponível em: <<http://www.york.ac.uk/inst/crd/>>. Acesso em: maio 2013.
11. PUBMED. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/>>. Acesso em: jun. 2013.
12. BRASIL. Ministério da Saúde. Secretaria de Ciência, Tecnologia e Insumos Estratégicos. **Rede Brasileira de Avaliação de Tecnologias em Saúde - Sisrebrats**. Disponível em: <<http://189.28.128.101/rebrats/visao/sociedade/estudo.cfm>>. Acesso em: jun. 2013.

13. MEDICAL Journal editors take hard line on drug research. **The Washington Post**, 10 set. 2004. Disponível em: <<http://www.smh.com.au/articles/2004/09/09/1094530773888.html>>. Acesso em: jun. 2013.
14. THE COCHRANE LIBRARY. **Cochrane Register of Diagnostic Test Accuracy Studies - CRDTAS**. Disponível em: <<http://www.thecochranelibrary.com/details/browseReviews/4920981/Diagnostic-test-accuracy.html>>. Acesso em: jun. 2013.
15. GOLDER, S. et al. Developing efficient search strategies to identify reports of adverse effects in MEDLINE and EMBASE. **Health Information and Libraries Journal**, Malden, v. 23, n. 1, p. 3-12, 2006.
16. WHITING, P. et al. Systematic reviews of test accuracy should search a range of databases to identify primary studies. **Journal of Clinical Epidemiology**, Maryland, v. 61, n. 4, p. 357-364, 2008.
17. THE MEDION DATABASE. Disponível em: <<http://www.mediondatabase.nl/>>. Acesso em: jun. 2013.
18. INTERNATIONAL FEDERATION OF CLINICAL CHEMISTRY AND LABORATORY MEDICINE. **Evidence-Based Laboratory Medicine (C-EBLM) Base**. Disponível em: <<http://www.ifcc.org/ifcc-education-division/emd-committees/c-eblm/evidence-based-laboratory-medicine-c-eblm-base>>. Acesso em: jun. 2013.
19. UNIVERSITY OF BIRMINGHAM. **ARIF databases**. Disponível em: <<http://www.birmingham.ac.uk/research/activity/mds/projects/HaPS/PHEB/ARIF/databases/index.aspx>>. Acesso em: jun. 2013.
20. UNIVERSITY OF YORK. Centre for Reviews and Dissemination - CRD. **Health Technology Assessment Database (HTA)**. Disponível em: <<http://www.crd.york.ac.uk/CRDWeb/>>. Acesso em: jun. 2013.
21. UNIVERSITY OF YORK. Centre for Reviews and Dissemination - CRD. **Database of Abstracts of Reviews of Effects (DARE)**. Disponível em: <<http://www.crd.york.ac.uk/CRDWeb/>>. Acesso em: jun. 2013.
22. SONG, F.; KHAN, K. S.; SOTTON, A. J. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic test accuracy. **International Journal of Epidemiology**, Oxford, v. 31, n. 1, p. 88-95, 2002.
23. GREENHALGH, T.; PEACOCK, R. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. **British Medical Journal**, London, v. 331, n. 1, p. 1064-1065, 2005.

24. GLANVILLE, J. Searching for diagnostic tests: which databases, which filters? In: **Annual Meeting of Health Technology Assessment International (HTAi): pushing the frontiers of information management**, Barcelona, v. 14, 2007.
25. FRASER, C. et al. **Searching for diagnostic test accuracy studies: an application to screening for open angle glaucoma (OAG)** [abstract]. XIV Cochrane Colloquium; 2006 Oct 23-26; Dublin, Ireland: 88.
26. ELSEVIER EMBASE TEAM. **Access 24 million records with Embase**. [mensagem pessoal]. Mensagem recebida por <mfigueiro@hcor.com.br>. Em 24 nov. 2010.
27. JADAD, A. R.; COOK, D. J.; BROWMAN, G. P. A guide to interpreting discordant systematic reviews. **Canadian Medical Association Journal**, Ottawa, v. 156, n. 10, p. 1411-1416, 1997.
28. BIBLIOTECA VIRTUAL EM SAÚDE (BVS). **LILACS**. Disponível em: <<http://lilacs.bvsalud.org/>>. Acesso em: jun. 2013.
29. PROQUEST. **Dissertations & Theses Database**. Disponível em: <<http://www.proquest.com/en-US/catalogs/databases/detail/pqdt.shtml>>. Acesso em: jun. 2013.
30. GOOGLE acadêmico. Disponível em: <<http://scholar.google.com.br/schhp?hl=pt-BR>>. Acesso em: nov. 2013.
31. TRIP DATABASE: clinical search engine. Disponível em: <<http://tripdatabase.com/>>. Acesso em: nov. 2013.
32. HIGGINS, J. P. T.; GREEN, S. (Ed.). **Cochrane Handbook for Systematic Reviews of Interventions**. Version 5.0.2. The Cochrane Collaboration, 2009. Disponível em: <<http://www.cochrane.org/training/cochrane-handbook>>. Acesso em: out. 2013.
33. DE VET, H. C. W. et al. Searching for Studies. In: **Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy**. Version 0.4. The Cochrane Collaboration, 2008. Disponível em: <<http://srdta.cochrane.org/sites/srdta.cochrane.org/files/uploads/Chapter07-Searching-%28September-2008%29.pdf>>. Acesso em: out. 2013.
34. BOSSUYT, P. M. et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. **Croatian Medical Journal**, Zagreb, v. 44, n. 5, p. 639-650, 2003.
35. COHEN, J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. **Psychological Bulletin**, Washington, v. 70, n. 4, p. 213-220, 1968.

36. BOSSUYT, P. M. The quality of reporting in diagnostic test research: getting better, still not optimal. **Clinical Chemistry**, Washington, v. 50, n. 3, p. 465-466, 2004.
37. SMIDT, N. et al. Quality of reporting of diagnostic accuracy studies. **Radiology**, Oak Brook, v. 235, n. 2, p. 347-353, 2005.
38. COLE, M. G. Impact of geriatric home screening services on mental state: a systematic review. **International Psychogeriatrics**, Cambridge, v. 10, n. 1, p. 97-102, 1998.
39. BREWER, D. A. et al. Should relatives of patients with colorectal cancer be screened? A critical review of the literature. **Disease of the Colon & Rectum**, New York, v. 37, n. 12, p. 1328-1338, 1994.
40. WALTER, S. D.; IRWIG, L.; GLASZIOU, P. P. Meta-analysis of diagnostic tests with imperfect reference standards. **Journal of Clinical Epidemiology**, Maryland Heights, v. 52, n. 10, p. 943-951, 1999.
41. WHITING, P.; HARBORD, R.; KLEIJNEN, J. No role for quality scores in systematic reviews of diagnostic accuracy studies. **BMC Medical Research Methodology**, London, p. 5-19, 2005.
42. MULROW, C. D. et al. Assessing quality of a diagnostic test evaluation. **Journal of General Internal Medicine**, Bethesda, v. 4, n. 4, p. 288-295, 1989.
43. JUNI, P. et al. The hazards of scoring the quality of clinical trials for meta-analysis. **The Journal of the American Medical Association**, Chicago, v. 282, n. 11, p. 1054-1060, 1999.
44. JUNI, P.; ALTMAN, D. G.; EGGER, M. Systematic reviews in health care: assessing the quality of controlled clinical trials. **British Medical Journal**, London, v. 323, n. 7303, p. 42-46, 2001.
45. WHITING, P. et al. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. **BMC Medical Research Methodology**, London, p. 3:25, 2003.
46. DEEKS, J. J.; ALTMAN, D. G.; BRADBURN, M. J. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In: EGGER, M; SMITH, G. D.; ALTMAN, D.G. (Eds.). **Systematic reviews in healthcare: meta-analysis in context**. London: BMJ Publishing Group, 2001, p. 285-312.
47. DERSIMONIAN, R.; LAIRD, N. Meta-analysis in clinical trials. **Controlled Clinical Trials**, Maryland Heights, v. 7, n. 3, p. 177-188, 1986.

48. LITTENBERG, B.; MOSES, L. E. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. **Medical Decision Making**, New Jersey, v. 13, n. 4, p. 313-321, 1993.
49. MOSES, L. E.; SHAPIRO, D.; LITTENBERG, B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. **Statistic in Medicine**, Malden, v. 12, n. 14, p. 1293-1316, 1993.
50. REITSMA, J. B. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. **Journal of Clinical Epidemiology**, Maryland Heights, v. 58, n. 10, p. 982-990, 2005.
51. RUTTER, C. M.; GATSONIS, C. A. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. **Statistic in Medicine**, Malden, v. 20, n. 19, p. 2865-2884, 2001.
52. ZAMORA, J. et al. Meta-DiSc: a software for meta-analysis of test accuracy data. **BMC Medical Research Methodology**, London, p. 6:31, 2006.
53. TAKWOINGI, Y.; DEEKS, J. J. **MetaDAS**: a SAS macro for metaanalysis of diagnostic accuracy studies. User Guide Version 1.3. 2010. Disponível em: <<http://srdata.cochrane.org/>>. Acesso em: nov. 2013.
54. LIJMER, J. G.; OSSUYT, P. M.; HEISTERKAMP, S. H. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. **Statistic in Medicine**, Malden, v. 21, n. 11, p. 1525-1537, 2002.
55. HIGGINS, J. P.; THOMPSON, S. G. Quantifying heterogeneity in a meta-analysis. **Statistic in Medicine**, Malden, v. 21, n. 11, p. 1539-1558, 2002.
56. DICKERSIN, K. et al. Publication bias and clinical trials. **Controlled Clinical Trials**, Maryland Heights, v. 8, n. 4, p. 343-353, 1987.
57. EASTERBROOK, P. J. et al. Publication bias in clinical research. **The Lancet**, London, v. 337, n. 8746, p. 867-872, 1991.
58. STERNE, J. A.; EGGER, M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. **Journal of Clinical Epidemiology**, Maryland Heights, v. 54, n. 10, p. 1046-1055, 2001.
59. DEEKS, J. J.; MACASKILL, P.; IRWIG, L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. **Journal of Clinical Epidemiology**, Maryland Heights, v. 58, n. 9, p. 882-893, 2005.

60. SONG, F. et al. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. **International Journal of Epidemiology**, Oxford, v. 31, n. 1, p. 88-95, 2002.
61. DOCHERTY, M.; SMITH, R. The case for structuring the discussion of scientific papers. **British Medical Journal**, London, v. 318, n. 7193, p. 1224-1225, 1999.
62. ATKINS, D. et al. Grading quality of evidence and strength of recommendations. **British Medical Journal**, London, v. 328, n. 7454, p. 1490, 2004.
63. SCHUNEMANN, H. J. et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. **British Medical Journal**, London, v. 336, n. 7653, p. 1106-1110, 2008.
64. GUYATT, G. H. et al. What is “quality of evidence” and why is it important to clinicians? **British Medical Journal**, London, v. 336, n. 7651, p. 995-998, 2008.
65. BERLIN, N. I. Breast cancer screening between ages 40 and 49. **The Cancer Journal from Scientific American**, v. 1, n. 3, p. 187-190, 1995.
66. KNOTTNERUS, J. A.; MURIS, J. W. Assessment of the accuracy of diagnostic tests: the cross-sectional study. **Journal of Clinical Epidemiology**, Maryland Heights, v. 56, n. 11, p. 1118-1128, 2003.
67. RUTJES, A. W. et al. Evidence of bias and variation in diagnostic accuracy studies. **Canadian Medical Association Journal**, Ottawa, v. 174, n. 4, p. 469-476, 2006.
68. GLAS, A. S. et al. Tumor markers in the diagnosis of primary bladder cancer: a systematic review. **The Journal of Urology**, Philadelphia, v. 169, n. 6, p. 1975-1982, 2003.
69. LIJMER, J. G. et al. Empirical evidence of design-related bias in studies of diagnostic tests. **The Journal of the American Medical Association**, Chicago, v. 282, n. 11, p. 1061-1066, 1999.
70. WHITING, P. et al. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. **Annals of Internal Medicine**, Philadelphia, v. 140, n. 3, p. 189-202, 2004.
71. RUTJES, A. W. et al. Case-control and two-gate designs in diagnostic accuracy studies. **Clinical Chemistry**, Washington, v. 51, n. 8, p. 1335-1341, 2005.
72. SACKETT, D. L.; HAYNES, R. B. The architecture of diagnostic research. **British Medical Journal**, London, v. 324, n. 7336, p. 539-541, 2002.

APÊNDICES

APÊNDICE 1: Construção da estratégia de busca: símbolos e operadores booleanos

A construção da estratégia de busca se dá pela definição dos termos que serão utilizados.

Alguns recursos, tais como utilização de caracteres especiais, permitem ampliar as opções dos termos. O uso de símbolos como * (asterisco), \$ (cifrão), ? (ponto de interrogação) vai depender de cada base utilizada. O símbolo de truncamento no MEDLINE/PubMed e Embase é o * (asterisco), conforme exemplificado por: diag*. Ao utilizar esta construção, estamos buscando artigos que contenham as seguintes palavras: “diagnostic”, “diagnose”, “diagnoses”, “diagnosed”.

No Embase, o uso do ? (ponto de interrogação) representa a recuperação de mais de um termo, como mostra o exemplo: wom?n, busca artigos que utilizem as palavras “woman” e “women”.

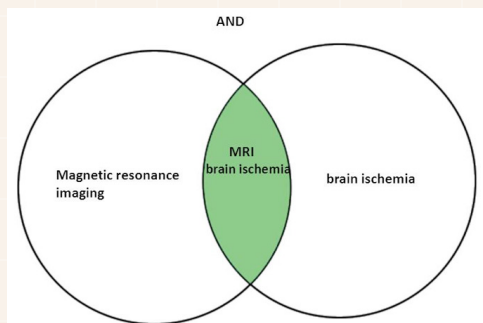
A combinação dos termos se dá através de utilização dos operadores booleanos. São eles o “AND”, “OR” e “NOT”. Eles irão fazer a ligação dos termos de busca e representam as seguintes ações:

AND: representa a intersecção dos termos. Quando utilizado, informamos à base que queremos um artigo que contenha os dois termos relacionados. Por exemplo, se queremos encontrar artigos que tenham avaliado a ressonância magnética em comparação à tomografia para identificação de isquemia cerebral, devemos utilizar o AND entre os termos que caracterizam esses dois testes.

A ordem dos termos não altera os resultados, ou seja, recuperaremos a mesma quantidade de artigos se utilizarmos: “magnetic resonance imaging” AND “brain ischemia” ou “brain ischemia” AND “magnetic resonance imaging”.

A figura 1 abaixo ilustra o conceito de busca utilizando o operador booleano AND:

Figura 1 – Operador booleano AND



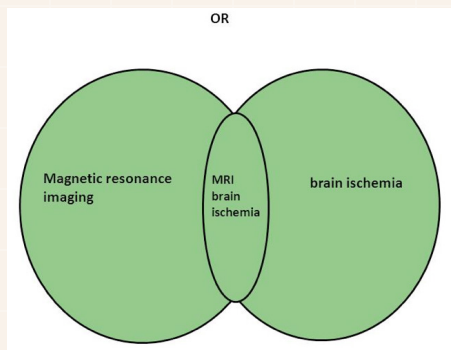
Fonte: elaboração própria.

OR: representa a soma, a união dos termos. Ele permite recuperar artigos que contenham tanto uma quanto a outra expressão, e também artigos que contenham ambos os assuntos. No exemplo ilustrado pela figura 2 abaixo, utilizarmos o operador OR “*magnetic resonance imaging*” OR “*brain ischemia*”, recupera-se todos os artigos com “*magnetic resonance imaging*” independente de seu uso para identificação de isquemia cerebral e os artigos que avaliaram qualquer intervenção na isquemia cerebral. Também espera-se recuperar os estudos que relacionaram ambos os termos.

A ordem dos termos também não altera os resultados.

A figura 2 abaixo ilustra o conceito de busca utilizando o operador booleano OR:

Figura 2 – Operador booleano OR



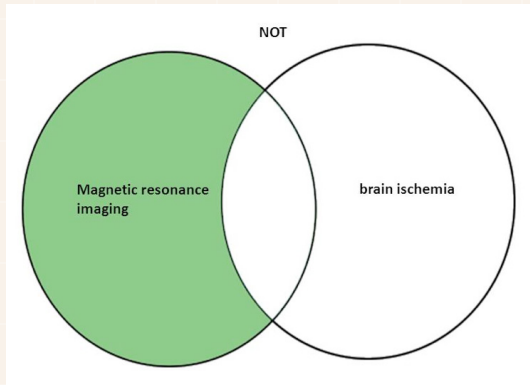
Fonte: elaboração própria.

NOT: operador booleano que exclui o termo subsequente. Pelo exemplo anterior, o arranjo “*magnetic resonance imaging*” NOT “*brain ischemia*” recupera somente artigos com o assunto “*magnetic resonance imaging*”, ou seja, excluem todos os estudos que tenham incluído a situação clínica de isquemia cerebral, inclusive os artigos que relacionem ambos os termos “*magnetic resonance imaging*” e “*brain ischemia*”.

Neste caso, a ordem dos termos altera os resultados. Assim, se utilizarmos o arranjo “*magnetic resonance imaging*” NOT “*brain ischemia*”, recuperaremos apenas artigos que contenham assuntos relacionados a “*magnetic resonance imaging*” se fizermos o arranjo “*brain ischemia*” NOT “*magnetic resonance imaging*” recuperaremos somente artigos com o assunto “*brain ischemia*”.

A figura 3 abaixo ilustra o conceito de busca utilizando o operador booleano NOT:

Figura 3 – Operador booleano NOT



Fonte: elaboração própria.

Elaborar uma única versão de uma estratégia de busca geralmente não é possível. O revisor pode se deparar com uma quantidade grande de artigos, o que pode inviabilizar a seleção de tamanho quantitativo. Neste caso, terá que lançar mão de algum termo que deixe a busca mais específica.

Assim, para otimizar o tempo, sugere-se que a pesquisa seja realizada “linha por linha”, ou seja, que para cada termo da questão de pesquisa explorado realiza-se uma pesquisa individual. Com isto, torna-se possível realizar diversas combinações utilizando os operadores booleanos. Os termos podem ser copiados em arquivo *Word* para que, sempre que necessário rodar uma nova estratégia, possam ser adicionados na caixa de pesquisa da respectiva base.

APÊNDICE 2: Gerenciadores de referências: o que são e como utilizá-los

Um gerenciador de referências apresenta inúmeras funções, inclusive, gerenciar as referências na condução de uma revisão sistemática.

Ele permite adicionar os resultados de todas as bases eletrônicas pesquisadas em um único arquivo, bem como referências advindas da busca manual, gerando uma biblioteca.

Ainda, permite organizar de forma prática todas as potenciais citações a serem incluídas na revisão sistemática, classificando-as por ordem alfabética de título, nome do autor ou número de identificação, o que é muito útil na remoção de referências duplicadas.

Várias são as opções de gerenciadores de referências disponíveis no mercado. A tabela abaixo fornece as principais opções comerciais e gratuitas, bem como a disponibilidade para os sistemas operacionais.

Tabela 1 – Gerenciadores de referências bibliográficas

Programa	Acesso	Windows	Mac OS X	Linux
Aigaion	Gratuito	Sim	Sim	Sim
Bebop	Gratuito	Sim	Sim	Sim
BibDesk	Gratuito	Não	Sim	Não
Biblioscope	Comercial	Sim	Não	Não
Bibus	Gratuito	Sim	Em teste	Sim
Bookends	Comercial	Não	Sim	Não
Citavi	Comercial	Sim	Não	Não
Connotea	Gratuito	Sim	Sim	Sim
Docear	Gratuito	Sim	Sim	Sim
EndNote	Comercial	Sim	Sim	Não
JabRef	Gratuito	Sim	Sim	Sim
Jumper 2.0	Gratuito	Sim	Sim	Sim
KBibTeX	Gratuito	Em teste	Em teste	Sim
Mendeley	Gratuito versão básica	Sim	Sim	Sim
Papers	Comercial	Sim	Sim	Não

Continuação

Pybliographer	Gratuito	Parcial	Parcial	Sim
Qiqqa	Gratuito	Sim	Não	Não
refbase	Gratuito	Sim	Sim	Sim
RefDB	Gratuito	Sim	Sim	Sim
Reference Manager	Comercial	Sim	Não	Não
Referencer	Gratuito	Não	Não	Sim
RefWorks	Comercial	Sim	Sim	N/A
Scholar's Aid	Gratuito na versão básica	Sim	Não	Não
Sente	Comercial	Não	Sim	Não
Wikindx	Gratuito	Sim	Sim	Sim
WizFolio	Gratuito versão básica	Sim	Sim	Sim
Zotero	Gratuito versão básica	Sim	Sim	Sim

Fonte: elaboração própria.

Cada gerenciador de referências tem suas peculiaridades, que devem ser exploradas na ocasião de sua utilização. Para gerenciadores como o EndNote® e o Reference Manager®, a maioria das etapas fundamentais para a criação de uma biblioteca única são de fácil entendimento e execução.

Independente do gerenciador de referências utilizado, deve-se obter o resultado da busca em arquivo texto (txt) para posterior inserção no gerenciador. Assim, utiliza-se o comando “exportar” das bases eletrônicas. Para o MEDLINE, exporta-se os resultados da busca clicando em “send to” (localizado no canto superior direito da tela de resultados) e depois seleciona-se a opção “file” e formato “MEDLINE”. Então, clica-se no botão “Creat file” para criar o arquivo de texto. A figura 4 abaixo ilustra o procedimento:

Figura 4 – Tela do PubMed para criar o arquivo da busca

The screenshot shows the PubMed interface with a search query: "("Magnetic Resonance Imaging"[Mesh]) AND "Brain Injuries"[Mesh]". The results are displayed in a list format. A dropdown menu titled "Choose Destination" is open, showing options like "File", "Clipboard", "E-mail", "Order", and "My Bibliography". The "MEDLINE" option is selected. Below the menu, there are options for "Download 2481 items.", "Format" (set to MEDLINE), and "Sort by" (set to Recently Added). The search results list two articles:

- Neuroscience. NFL kicks off brain injury rese...**
Underwood E.
Science. 2013 Mar 22;339(6126):1367. doi: 10.1126/sci... available.
PMID: 23520084 [PubMed - indexed for MEDLINE]
[Related citations](#)
- Functional magnetic resonance imaging o... after childhood brain injury: reliance on a left... network...**
Morgan AT, Masterton R, Pigdon L, Connelly...
Brain. 2013 Feb;136(Pt 2):648-57. doi: 10.1093/brain/awt355. Epub 2013 Jan 31.
PMID: 23378215 [PubMed - indexed for MEDLINE]
[Related citations](#)

Fonte: PUBMED, 2013.

Para exportar os resultados da busca no Embase, deve-se clicar em “*export*”, localizado na barra azul superior da tela de resultados. Então, escolhe-se o formato de exportação “*Plain text*” “*Full record*” e, em seguida, clica-se em “*export*”. Esse processo pode ser visualizado nas figuras 5 e 6.

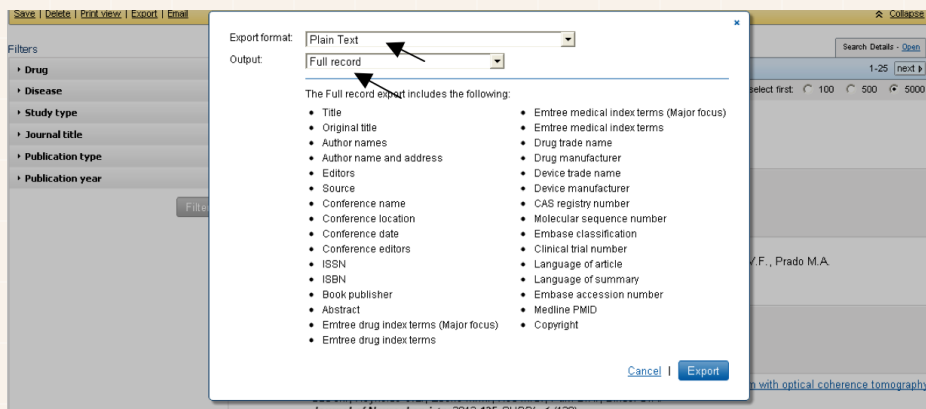
Figura 5 – Tela do Embase com resultado da busca e opção para exportar

The screenshot shows the Embase interface with a search query: "brain injury/exp AND [embase]/flm". The results are displayed in a list format. The top navigation bar has a blue bar with "Export" highlighted. The search results list four articles:

- Neuroimaging in aphasia treatment research. Quantifying brain lesions after stroke**
Crinion J, Holland A L, Copland D A, Thompson C K, Hillis A E.
NeuroImage 2013 73 (208-214)
Embase Abstract Index Terms [View Full Text](#)
- Basal ganglia lesions in children and adults**
Bekiesinska-Figatowska M, Mierzewska H, Jurkiewicz E.
European Journal of Radiology 2013 82:5 (637-649)
Embase Abstract Index Terms [View Full Text](#)
- The role of striatal acetylcholine in cerebral ischemic lesion and functional recovery**
Goncalves D F, Guzman M S, Nikolova S, Gros R, Massensini A R, Bartha R, Prado V F, Prado M A.
Journal of Neurochemistry 2013 125 SUPPL 1 (213)
Embase Abstract Index Terms [View Full Text](#)
- Proton magnetic resonance spectroscopy findings in mild traumatic brain injury**
Dabrota D, Sivak S, Bittsansky M, Grossmann J, Demkova A, Kurca E.
Journal of Neurochemistry 2013 125 SUPPL 1 (155)

Fonte: EMBASE, 2013.

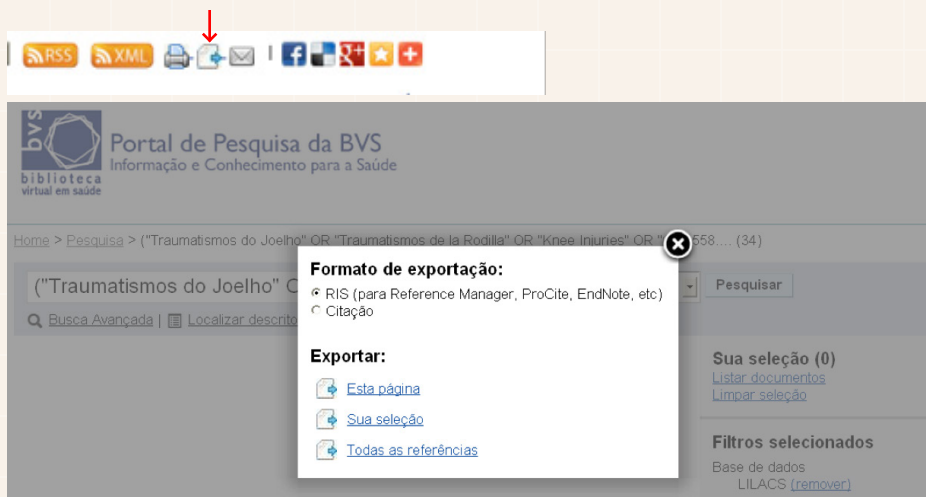
Figura 6 – Tela do Embase para exportação dos resultados da busca



Fonte: EMBASE, 2013.

Na LILACS para exportação dos resultados da busca, deve-se clicar no ícone que aparece na barra abaixo, e, se for utilizado um gerenciador de referências como o Endnote, ProCite ou Reference Manager, marcar a primeira opção da tela, “RIS”, e depois clicar em “Todas as referências” para rodar a exportação.

Figura 7 – Tela da LILACS para exportação dos resultados da busca



Fonte: BVS/LILACS, 2013.

O resultado da busca em cada base de dados deve ser exportado para um arquivo de texto (txt), que depois será importado para o gerenciador de referências. Para isto, no gerenciador, deve-se selecionar o filtro de importação correspondente à base para importar cada arquivo txt. Os gerenciadores possuem inúmeros filtros e dificilmente alguma base pesquisada não encontrará seu filtro de importação correspondente.

APÊNDICE 3 – Tipos de estudos de acurácia diagnóstica

Estudos de acurácia diagnóstica possuem diversas especificidades que os tornam bastantes diferentes dos estudos de intervenção.

Inicialmente, a intenção do teste pode ser distinta: o estudo pode tratar da avaliação de um teste com finalidade de triagem de uma doença, ou de diagnóstico propriamente dito ou mesmo de um teste com valor prognóstico. A definição e relevância dos desfechos nos estudos primários são dependentes da intenção desse teste. Por exemplo, efeitos adversos emocionais e desfechos éticos ou legais podem ser mais significativos para testes utilizados em triagem de uma patologia do que para o diagnóstico, devido à alta prevalência de falsos-positivos que são associados a esse tipo de teste. Mamografia é um bom exemplo, uma vez que a consequência emocional a um resultado falso-positivo é significativo. Adicionalmente, testes de triagem são conduzidos em indivíduos que não apresentam sintomas da doença, portanto, qualquer consequência adversa do teste pode ser mais pronunciada. Por outro lado, pode-se negligenciar a tomada de decisão médica adequada se ela estiver baseada em um resultado falso-negativo de um teste. Neste caso, as consequências legais de uma ação assim podem ser significativas. O não diagnóstico de câncer de mama é um exemplo real desse tipo de conduta equivocada⁶⁵.

Os testes diagnósticos diferem dos de triagem pelo tipo de população que estão sendo testados. Enquanto o teste diagnóstico é utilizado para confirmar ou rejeitar uma doença em pessoas que apresentam risco ou sintomas para a doença, os de triagem são utilizados em populações assintomáticas ou pré-sintomáticas. Por outro lado, testes com objetivo prognóstico são utilizados em indivíduos sabidamente portadores da doença para prever desfechos clínicos, como mortalidade, por exemplo. Adicionalmente, testes prognósticos têm a capacidade de alterar o tratamento da uma doença, a depender do seu resultado.

Portanto, dependendo da intenção de um teste, os desfechos e suas respectivas medidas de desempenho são definidas. Para estudos de acurácia diagnósticas, os desfechos são as medidas de desempenho dos testes.

Os estudos de acurácia têm natureza transversal. A abordagem típica é incluir uma série de pacientes consecutivos que apresente a suspeita da doença e submeter cada paciente ao teste índice e ao padrão de referência. Idealmente o teste índice e o padrão de referência devem ser interpretados por pessoas que estejam cegas em relação aos resultados do outro teste e o tempo decorrente entre a aplicação dos testes não pode se estender a ponto de alterar o grau de gravidade da doença. Outros termos podem ser utilizados para referenciar esse desenho de estudo, como *cross-seccional*, ou *single-gate design* (no sentido em que há apenas uma fonte de população no estudo, que cumpre os mesmos critérios de elegibilidade) ou mesmo como estudos

Diversas variações deste desenho podem ser encontradas na literatura e estas têm impacto direto das estimativas de desempenho do teste.

Todas essas variações se enquadram em um tipo de estudo denominado caso-controle diagnóstico. Esse conceito difere do conceito de estudos caso-controle em epidemiologia, onde são utilizados para responder questões sobre a etiologia de uma doença, que é o desfecho do estudo. Diferentemente dos estudos de coorte, os de caso-controle em epidemiologia revertem a ordem prospectiva da evolução da doença (exposição ao fator de estudo ao longo do tempo e consequente desenvolvimento da doença) e se iniciam com a doença, para depois se investigar o grau de exposição ao fator de estudo. Esta abordagem é bastante útil em casos de doenças raras ou com grande período de latência, que despenderiam muitos anos de seguimento e altos custos associados, se investigados de forma prospectiva. O termo “estudos de caso-controle diagnósticos” é utilizado para referir os estudos em que a situação de doença (presente ou ausente) é conhecida antes da aplicação do teste índice. Esta distinção explica o racional para a utilização dos termos “casos” e “controles”. De forma análoga, o desfecho de interesse já está detectado pelo padrão de referência e o teste índice se comporta como o fator de exposição do estudo. Infelizmente essa terminologia pode levar à confusão, dada à natureza transversal dos estudos de acurácia diagnóstica, em que os testes índice e o padrão de referência são aplicados no mesmo participante ao mesmo tempo. Assim, eles diferem dos estudos etiológicos, pois não há uma janela temporal entre a exposição e a ocorrência da doença, entre outras características⁶⁶. Esse tipo de estudo de caso-controle diagnóstico também tem uma única fonte de população no estudo (*single-gate design*), geralmente a mesma da abordagem clássica dos estudos diagnósticos.

Outras variações dos estudos de caso-controle diagnóstico são permitidas e emergem quando os casos e os controles são provenientes de distintas fontes de população (*two-gate designs*), ou seja, critérios de elegibilidade são definidos para os casos e outros critérios para os controles. Os casos estão baseados na positividade do padrão de referência e os controles podem ser participantes saudáveis ou com diagnósticos alternativos.

Existem problemas inerentes aos desenhos de estudo de caso-controle diagnóstico, que podem levar a vieses. A inclusão seletiva dos casos com doença mais avançada provavelmente pode levar a superestimação da sensibilidade⁶⁷. Por outro lado, quando os casos são derivados de indivíduos com doença moderada ou em estágio inicial, resulta na subestimação da sensibilidade⁶⁸. Ainda, a inclusão de controles saudáveis provavelmente pode levar a superestimação da especificidade^{69,70}. Quando o grupo controle é derivado de pacientes com diagnóstico alternativo, a especificidade pode ser sub ou superestimada, dependendo da alternativa diagnóstica⁷¹.

Estudos de desenho caso-controle podem produzir uma estimativa válida de desempenho do teste se os casos forem pareados com resultados positivos no padrão de referência (em termos do espectro da gravidade da doença) e os controles pareados

com resultados negativos no padrão de referência (em termos do espectro da condição alternativa). Na prática, entretanto, é difícil que a amostra seja pareada de maneira correta⁷¹. Este tipo de delineamento de estudo é mais útil nas fases precoces do desenvolvimento do teste⁷².

O verdadeiro valor de um teste diagnóstico é estabelecido apenas em estudos que incluem pacientes que se assemelham à prática clínica, sendo que o teste é útil apenas se for capaz de distinguir as condições clínicas que possam ser semelhantes e confundir-se entre si. De fato, o propósito de um teste diagnóstico é auxiliar no diagnóstico de pacientes com suspeita de doença, ou seja, pacientes sem nenhum achado óbvio que possa prontamente confirmar ou excluir o diagnóstico. Para tanto, idealmente, a amostra do estudo deve ser consecutiva ou selecionada de forma aleatória naqueles em que a condição-alvo é suspeita ou, em casos de estudos de triagem, na população alvo. Esse tipo de amostra é obtido em estudos que têm um delineamento típico.

Quando a revisão sistemática inclui ambos os delineamentos de estudo, deve-se dar atenção ao método de análise escolhido e o impacto do desenho de estudo deve ser avaliado, inclusive por análises de sensibilidade.

APÊNDICE 4 – Medidas de desempenho dos testes diagnósticos

As opções para sumarizar a acurácia diagnóstica em um estudo individual foca na avaliação do desempenho do teste índice em indivíduos com a doença e sem a doença (sensibilidade e especificidade do teste) ou nas implicações dos resultados positivos e negativos do teste índice (valores preditivos e razões de verossimilhança).

Para cada valor de corte determinado para positividade do teste índice, a população do estudo pode ser distribuída de acordo com sua condição de doença decorrente da definição deste valor de corte. Esta disposição dos casos deve estar apresentada de acordo com a tabela de contingência 2x2 conforme ilustrado abaixo:

	Doença (D+)	Não-Doença (D-)	Total
Teste índice positivo (T+)	Verdadeiros positivos (a)	Falsos positivos (b)	Testes positivos (a+b)
Teste índice negativo (T-)	Falsos negativos (c)	Verdadeiros negativos (d)	Testes negativos (c+d)
Total	Com doença (a+c)	Sem doença (b+d)	N (a+b+c+d)

A partir da disposição dos dados dessa maneira, medidas sumárias para a acurácia do teste podem ser calculadas como proporções de doenças positivas e negativas (estatísticas condicionais à situação de doença) ou proporções de resultados de teste positivos e negativos (estatísticas condicionais ao resultado do teste índice).

Sensibilidade e especificidade

São medidas condicionadas ao estado de doença, se positiva ou negativa.

A sensibilidade é definida como a probabilidade do resultado do teste índice ser positivo caso a doença esteja, de fato, presente. Matematicamente, utilizando os dados da tabela 2x2, sensibilidade = $a/(a+c)$. Essa medida representa os verdadeiros positivos da amostra. Pode ser expressa em proporção ou porcentagem e seu intervalo de confiança pode ser calculado.

A especificidade é definida como a probabilidade do resultado do teste índice ser negativo quando, de fato, a doença está ausente. Matematicamente, utilizando os dados da tabela 2x2, especificidade = $d/(b+d)$. Essa medida representa os verdadeiros negativos da amostra. Pode ser expressa em proporção ou porcentagem. Muitas vezes, os termos razão de falso positivo ou fração de falso positivo são utilizados como termos complementares à especificidade, uma vez que são calculados como “1 – especificidade” ou $b/(b+d)$.

Os valores de sensibilidade e especificidade são ocasionalmente combinados em uma medida denominada Índice de Youden, calculada como “sensibilidade + especificidade – 1”. Esse índice não fornece uma interpretação probabilística direta, mas fornece um índice geral da acurácia do teste que atribui pesos iguais aos erros do teste (falsos negativos e falsos positivos). Valores perto de 1 indicam alta acurácia; valor igual a zero indicam que o teste não tem valor diagnóstico.

Uma característica importante a ser destacada em relação à sensibilidade e à especificidade é que estas medidas não são afetadas pela prevalência da doença. Ao contrário dos valores preditivos, que serão comentados a seguir, seu intervalo de confiança pode ser calculado.

Valores preditivos

São medidas condicionadas aos resultados do teste índice, se positivo ou negativo. O valor preditivo positivo (VPP) de um teste é definido como a probabilidade de que um indivíduo tenha realmente a doença, dado que apresentou um resultado positivo do teste índice. Isto é diferente da sensibilidade, que é a probabilidade de que o paciente tenha um resultado positivo de teste índice, dado que ele apresenta a doença. Matematicamente, utilizando os dados da tabela 2x2, $VPP = a/(a+b)$. Esse resultado pode ser apresentado como proporção ou porcentagem.

O valor preditivo negativo (VPN) de um teste é definido como a probabilidade de que um indivíduo não tenha realmente a doença, dado que apresentou um resultado negativo do teste índice. Isto é diferente de especificidade, que é a probabilidade de que um indivíduo apresente um resultado negativo do teste índice, dado que ele não tem a doença. Matematicamente, utilizando os dados da tabela 2x2, $VPN = d/(c+d)$. Esse resultado também pode ser apresentado como proporção ou porcentagem.

Os valores preditivos positivo e negativo são afetados pela prevalência da doença. O termo “prevalência da doença” também pode ser interpretado no estudo como os casos de desfecho ou doença positivos (soma do número de verdadeiros positivos e falsos positivos). Neste sentido, o aumento na prevalência da doença leva a um aumento no valor preditivo positivo e a uma diminuição no negativo. Quando a prevalência é baixa, o valor preditivo positivo é baixo, independente dos valores de sensibilidade e especificidade. Uma alta prevalência sempre acarretará um aumento no VPP e uma diminuição do VPN.

Razão de verossimilhança ou *likelihood ratios*.

Razão de verossimilhança é muito útil no processo de tomada de decisão, pois ela define o desempenho do teste diagnóstico ou regra de predição clínica para confirmação ou afastamento de determinada suspeita diagnóstica. Tal razão transforma a chance

pré-teste da doença, ou seja, aquela definida intuitivamente com base nos achados de sinais, sintomas e exames iniciais, em uma chance pós-teste da doença. Nesse sentido, pode-se dizer que a razão de verossimilhança é utilizada para atualizar uma probabilidade pré-teste de doença, uma vez que o resultado de teste é conhecido.

$$\text{Chance Pós-Teste} = \text{Chance Pré-Teste} \times \text{Razão de Verossimilhança}$$

Portanto, quando a razão de verossimilhança se aproxima de um, significa dizer que o teste não é capaz de mudar a chance pós-teste da doença. Razões de verossimilhança menores que um diminuem progressivamente a chance pós-teste de doença quando o resultado do teste é negativo. Razões de verossimilhança maiores que um aumentam progressivamente a chance pós-teste de doença, quando o resultado do teste é positivo.

Razão de verossimilhança positiva

Para se entender razão de verossimilhança quando o resultado do teste é positivo, é preciso responder a duas perguntas:

- Qual é a probabilidade de se encontrar um teste (+) em quem tem a doença?
- Qual é a probabilidade de se encontrar um teste (+) em quem não tem a doença?

A razão entre essas duas probabilidades é a razão de verossimilhança (+). Portanto, a razão de verossimilhança positiva diz quantas vezes é mais provável encontrar um resultado de teste positivo em quem tem a doença do que em quem não a tem.

Matematicamente, a razão de verossimilhança positiva pode ser calculada como:

$$\text{LR} + (\textit{likelihood ratio} +) = \text{sens}/(1-\text{espec}) \text{ ou } (a/(a+c)) / (b/(b+d)).$$

Razão de verossimilhança negativa

Para se entender razão de verossimilhança quando o resultado do teste é negativo, é preciso responder a duas perguntas:

- Qual é a probabilidade de se encontrar um teste (-) em quem tem a doença?
- Qual é a probabilidade de se encontrar um teste (-) em quem não tem a doença?

A razão entre essas duas probabilidades é a razão de verossimilhança (-).

Portanto, a razão de verossimilhança negativa diz quantas vezes é mais provável encontrar um resultado de teste negativo em quem tem a doença do que em quem não a tem.

Matematicamente, a razão de verossimilhança negativa pode ser calculada como:

$$\text{LR} - (\textit{likelihood ratio} -) = (1-\text{sens})/\text{espec} \text{ ou } (c/(a+c)) / (d/(b+d)).$$

Dado que chance e probabilidades (risco) não significam a mesma coisa, é importante saber que existem nomogramas disponíveis, como no Centro de Medicina Baseada em Evidências de Oxford em que as chances pré e pós já estão automaticamente

transformadas em probabilidades, um termo mais prático para quem usa a informação para tomada de decisão.

Um alto valor de razão de verossimilhança para um resultado positivo e um baixo valor de razão de verossimilhança para resultados negativos de teste indicam que ele é útil.

Razão de chances (*Odds ratio*) diagnóstica

A razão de chances diagnóstica (RCD) sumariza a acurácia diagnóstica do teste índice como um único número que descreve quantas vezes maior é a chance de se obter um resultado positivo em uma pessoa com a doença do que em alguém sem a doença.

A descrição de acurácia em termos de razão de chances tem pouca relevância clínica direta e raramente é utilizada como sumário estatístico nos estudos primários. Geralmente, estamos mais interessados na soma do número de falsos negativos e falsos positivos, enquanto que a razão de chances diagnóstica reflete seus produtos. Entretanto, esta é uma medida muito importante na construção do modelo metanalítico e muitas vezes ela terá que ser calculada. A fórmula matemática de seu cálculo é apresentada conforme:

$$RCD = LR+/LR- = (\text{sens} \times \text{espec}) / (1-\text{sens}) \times (1-\text{espec}) = (ad) / (bc).$$

EXEMPLOS DE MEDIDAS DE DESEMPENHO DE UM TESTE DIAGNÓSTICO:

Exemplo 1:

Para exemplificar os conceitos apresentados acima, considere o exemplo 1 abaixo proveniente de dados hipotéticos.

	Doença (D+)	Não-Doença (D-)	Total
Teste índice positivo (T+)	81	591	672
Teste índice negativo (T-)	45	674	719
Total	126	1265	1391

Aplicando as fórmulas descritas, obteremos as seguintes medidas:

Sensibilidade: $81/126 = 0,64$. Interpretação: Na amostra, 64% dos pacientes que têm a doença (baseado no resultado do teste padrão de referência) foram corretamente identificados pelo teste índice.

Especificidade: $674/1265 = 0,53$. Interpretação: Na amostra, 53% dos pacientes que não têm a doença (baseado no resultado do teste padrão de referência) foram corretamente identificados pelo teste índice.

Valor preditivo positivo: $81/672 = 0,12$. Interpretação: 12% é a probabilidade de o indivíduo ter a doença se ele apresentou um resultado positivo do teste índice.

Valor preditivo negativo: $674/719 = 0,94$. Interpretação: 94% é a probabilidade de o paciente não ter a doença se ele apresentou um resultado negativo do teste índice.

LR+ = $0,64/(1-0,53) = 1,36$. Interpretação: indica quantas vezes é mais provável um resultado de teste positivo em quem tem a doença do que em quem não tem.

Probabilidade pós-teste de doença dado que o resultado do teste foi positivo = probabilidade pré-teste de doença X LR+.

A probabilidade (risco) pré-teste da doença é a prevalência desta. A chance (que é diferente da probabilidade) pré-teste pode ser utilizada para calcular a probabilidade pós-teste da doença e é dada por:

Chance pré-teste da doença: $\text{Prevalência}/1 - \text{prevalência} = (126/1391)/(1 - (126/1391)) = 0,09/(1-0,09) = 0,099$.

Assim, a chance pós-teste de doença é dada por $= 0,099 \times 1,36 = 0,135$.

Convertendo chance em probabilidade (risco), devemos aplicar a seguinte fórmula:
 $P = \text{chance}/(1 + \text{chance}) = 0,135/(1 + 0,135) = 0,12$.

Assim, 12% é a probabilidade da doença dado que o resultado foi positivo (que é o valor preditivo positivo).

LR- = $(1 - 0,64)/0,53 = 0,68$. Interpretação: indica quantas vezes é mais provável um resultado de teste negativo em quem tem a doença do que em quem não tem.

Probabilidade pós-teste de doença dado que o resultado do teste foi negativo = probabilidade pré-teste de doença X LR-.

Chance pós-teste para doença dado que o resultado do teste foi negativo $= 0,099 \times 0,68 = 0,067$.

Convertendo chance em probabilidade $= 0,067/(1 + 0,067) = 0,06$.

Esta é a probabilidade de doença dado que o resultado do teste foi negativo, o que equivale a $1 - \text{VPN}$. Portanto, $\text{VPN} = 1 - 0,06 = 0,94$, como demonstrado acima.

Razão de Chances Diagnóstica (RCD) = $LR+ / LR- = \text{sens} \times \text{espec} / (1 - \text{sens}) \times (1 - \text{espec}) = (ad) / (bc)$

- $RCD = 1,36 / 0,68 = 2,0$
- $RCD = 0,64 \times 0,53 / (1 - 0,64) \times (1 - 0,53) = 0,3392 / 0,36 \times 0,47 = 2,004$
- $RCD = 81 \times 674 / 45 \times 591 = 54.594 / 26,595 = 2,05$

Exemplo 2:

Agora, considere a outra tabela 2x2 de dados hipotéticos abaixo, que apresenta o mesmo tamanho da amostra do exemplo anterior, para demonstrarmos o efeito da prevalência da doença nos valores preditivos:

	Doença (D+)	Não-Doença (D-)	Total
Teste índice positivo (T+)	386	370	756
Teste índice negativo (T-)	214	421	635
Total	600	791	1391

Nesse caso, se calcularmos a sensibilidade e a especificidade considerando esses dados, observaremos que essas não se alteraram em relação ao exemplo anterior. Por outro lado, os valores preditivos calculados diferem totalmente dos valores do exemplo anterior, demonstrando o efeito da prevalência (ou desfecho positivo) nessas medidas.

Sensibilidade: $386/600 = 0,64$. Interpretação: Na amostra, 64% dos pacientes que têm a doença (baseado no resultado do teste padrão de referência) foram corretamente identificados pelo teste índice.

Especificidade: $421/791 = 0,53$. Interpretação: na amostra, 53% dos pacientes que não têm a doença (baseado no resultado do teste padrão de referência) foram corretamente identificados pelo teste índice.

Valor preditivo positivo: $386/756 = 0,51$. Interpretação: 51% é a probabilidade de o indivíduo ter a doença se ele apresentou um resultado positivo do teste índice.

Valor preditivo negativo: $421/635 = 0,66$. Interpretação: 66% é a probabilidade de o paciente não ter a doença se ele apresentou um resultado negativo do teste índice.

Utilizando os dados do exemplo 1, a prevalência da doença (ou desfecho positivo) mudou de $126/1391 = 9\%$ para $600/1391 = 43\%$. O aumento na prevalência levou ao aumento do valor preditivo positivo e à diminuição do valor preditivo negativo.

As demais medidas podem ser calculadas usando o mesmo racional e fórmulas do exemplo 1 acima.

Limiares de positividade do teste.

Quando o resultado do teste é uma medida contínua, define-se um limiar a partir do qual valores superiores a este limiar tornam os casos como doença positiva e valores abaixo, como não-doença. Se o limiar de positividade se alterar, o número de casos para cada medida sumária de desempenho do teste diagnóstico se altera. Esta dependência do limiar é um aspecto fundamental para a avaliação de um teste diagnóstico.

A escolha do limiar de positividade envolve uma decisão entre aumentar a sensibilidade à custa de redução da especificidade e vice-versa. A maioria dos pesquisadores deve avaliar cuidadosamente a importância relativa da sensibilidade e da especificidade do teste para estabelecer o ponto de transição diagnóstica mais adequado, cuja estratégia geral seria dependente da intenção do teste.

Assim, se a principal preocupação é evitar resultado falso-positivo (o resultado do teste pode indicar uma cirurgia arriscada para o paciente, por exemplo), então o ponto de corte deve objetivar o máximo de especificidade. Por outro lado, se a preocupação maior é evitar resultado falso-negativo (o resultado do teste em suspeito de AIDS, por exemplo), então o ponto de corte deve objetivar o máximo de sensibilidade.

Curvas ROC - (*Receive Operator Characteristic curve*)

Estudos primários que avaliam um teste em diferentes limiares de positividade geralmente apresentam uma curva ROC. A curva ROC analisa a acurácia de um único teste em uma única população. Ela compara a acurácia de um teste em diferentes limiares de positividade.

Essa curva está em um gráfico cujos valores de sensibilidade e especificidade são obtidos pela variação do limiar de positividade ao longo de todos os seus possíveis valores. Se outro limiar é definido, a sensibilidade e especificidade são calculadas novamente.

A decisão acerca do limiar de positividade, conforme dito anteriormente, dependerá da intenção do teste. Baixos valores para o limiar produzirão mais verdadeiros e falso positivos. Critérios mais rigorosos para resultados positivos produzirão menos positivos, com baixa sensibilidade e alta especificidade. Muitos autores preferem

definir este limiar baseado no ponto que maximiza a curva. Esse ponto fornecerá as maiores sensibilidades e especificidades possíveis para o teste.

O gráfico plota sensibilidade (verdadeiros positivos) *versus* “1 – especificidade” (falsos positivos). Deste modo, uma medida global da acurácia do teste é o cálculo da área sobre a curva ROC, em que um valor de 0,5 é obtido se o teste não tem aplicabilidade clínica e um valor de 1 se o teste é perfeito. A curva ROC começa na origem (0,0), vai verticalmente para acima do eixo Y (0,1) e depois horizontalmente até (1,1). Um bom teste estará mais próximo possível dessas coordenadas.



ANEXO A: Tabela das principais bases de dados e respectivos acessos eletrônicos

Bases de revisões sistemáticas e de estudos de acurácia de testes diagnósticos	Endereço	Acesso ao conteúdo
Cochrane Review: Diagnostic Test Accuracy	http://www.thecochranelibrary.com/details/browseReviews/578385/Diagnostic-test-accuracy.html	Restrito
The Cochrane Library	http://www.thecochranelibrary.com/view/0/index.html	Restrito
Biblioteca Cochrane	http://cochrane.bvsalud.org/portal/php/index.php?lang=pt	Livre
CRD (Centre for Reviews and Dissemination) DARE - Database of Abstracts of Reviews of Effects	www.york.ac.uk/inst/crd	Restrito
MEDLINE/Clinical Queries	http://www.ncbi.nlm.nih.gov/pubmed/clinical	Restrito/Livre
MEDION – Meta-analyses van Diagnostisch Onderzoek	www.mediondatabase.nl	Restrito
ARIF - Aggressive Research Intelligence Facility	www.arif.bham.ac.uk/	Restrito
C-EBLM database	http://www.ifcc.org/ifcc-education-division/emd-committees/c-eblm/evidence-based-laboratory-medicine-c-eblm-base/	Restrito
Bases essenciais para busca de Estudos Primários		
MEDLINE/PubMed	www.pubmed.gov	Restrito/Livre
Embase	www.embase.com	Restrito
LILACS	http://lilacs.bvsalud.org/	Restrito /Livre

Continua

Continuação

Bases Opcionais		
ISI of Knowledge	http://thomsonreuters.com/	Restrito *
Scopus	http://www.scopus.com/home.url	Restrito *
SCIRUS	http://www.scirus.com/	Restrito
Banco de teses da USP	http://www.teses.usp.br/	Livre
Banco de teses Portal Capes	http://www.capes.gov.br/servicos/banco-de-teses	Livre
Banco de teses IBICT	http://bdtd.ibict.br/	Livre
ProQuest Dissertations & Theses Databases	http://www.proquest.com/en-US/catalogs/databases/detail/pqdt.shtml	Restrito
Bases Especializadas		
CINAHL - Índice cumulativo em enfermagem e ciências afins	www.cinahl.com	Restrito *
PsycINFO - Behavioral & Social Sciences: Psychology	psycinfo.apa.org	Restrito *
PEDro - Physiotherapy Evidence Database	http://www.pedro.org.au/	Restrito
BBO - Bibliografia Brasileira de Odontologia	http://odontologia.bvs.br/	Livre
ADOLEC – Base de dados de adolescentes e jovens	http://www.adolesc.br/	Livre
BDENF - Base de dados de enfermagem	http://enfermagem.bvs.br	Livre
BVS-PSICO - Base de dados em psicologia	http://www.bvs-psi.org.br/php/index.php	Livre
Bases Complementares		
Bandolier Sumário de evidências	http://www.medicine.ox.ac.uk/bandolier/Universidade de Oxford	Restrito
Portal do Ministério da Saúde	Revisões Sistemáticas promovidas pelo Departamento de Ciência e Tecnologia portal.saude.gov.br/portal/saude/visualizar_texto.cfm?idtxt=25514	Livre
FDA U.S. Food and Drug Administration	http://www.fda.gov/	Restrito /Livre
Guidelines International Network	http://www.g-i-n.net/	Livre

Continua

Continuação

Buscadores na Web		
Google Acadêmico	http://scholar.google.com.br/schhp?hl=pt-BR	Restrito /Livre
TripDatabase Turning Research into Practice	http://www.tripdatabase.com/	Restrito /Livre
Busca otimizada nos sites dos membros da INAHTA #	Desenvolvida no laboratório do Google: http://by.ly/ats ou http://ats.by.ly	Restrito
Registro de Ensaio Clínicos		
Clinical trials	www.clinicaltrials.gov	Restrito
Registro Brasileiro de Ensaio Clínicos	www.ensaiosclinicos.gov.br	Livre

*Bases com acesso via portal CAPES (www.periodicos.capes.gov.br) Bases Essenciais: Na elaboração de uma RS é requisito mínimo; Bases Opcionais: Ampliar e buscar mais evidências; Bases Especializadas: Necessárias quando a pergunta envolve o assunto da base; Bases Complementares: Ampliar e buscar mais evidências e outras fontes de informação pertinente; Buscadores de busca: Otimizar a busca em várias fontes

ANEXO B – Exemplos de construções de estratégia de busca

A questão de pesquisa considerada foi Ressonância Magnética *versus* tomografia no diagnóstico de lesões nos joelhos. Como a maioria das bases de dados trabalha com termos em inglês, estes assim serão utilizados para os fins de exemplo.

Os seguintes exemplos de construções de estratégia de busca podem ser considerados:

Exemplo de construção de estratégia de busca no MEDLINE.

Quadro 1 – Exemplo de construção de estratégia de busca no MEDLINE/PubMed

MEDLINE/PubMed
Utilização do descritor de assunto (termo Mesh) sensibilizado com a utilização dos “entry terms”.
População/condição: pacientes com traumatismo nos joelhos
“Knee Injuries”[Mesh] OR (Injuries, Knee) OR (Injury, Knee) OR (Knee Injury)
AND
Intervenção: ressonância magnética “Magnetic Resonance Imaging”[Mesh] OR (imaging, Magnetic Resonance) OR (NMR Imaging) OR (Imaging, NMR) OR (Zeugmatography) OR (Tomography, MR) OR (Tomography, NMR) OR (MR Tomography) OR (NMR Tomography) OR (Tomography, Proton Spin) OR (Proton Spin Tomography) OR (Magnetization Transfer Contrast Imaging) OR (MRI Scans) OR (MRI Scan) OR (Scan, MRI) OR (Scans, MRI) OR (fMRI) OR (MRI, Functional) OR (Functional MRI) OR (Functional MRIs) OR (MRIs, Functional) OR (Magnetic Resonance Imaging, Functional)
AND
Controle: “Tomography”[Mesh] OR (tomographies) OR (tomography)

Fonte: elaboração própria.

Notar que o resultado de pesquisa #4 (2631 estudos encontrados) é o resultado final da busca pela estratégia definida na base de dados MEDLINE, que recuperou artigos potencialmente elegíveis para esta questão de pesquisa.

Figura 1 – Tela MEDLINE/history search

Search	Add to builder	Query	Items found	Time
#4	Add	Search ((#1) AND #2) AND #3	2631	09:23:23
#3	Add	Search Tomography"[Mesh] OR tomographies OR tomography	642541	09:22:31
#2	Add	Search "Magnetic Resonance Imaging"[Mesh] "Magnetic Resonance Imaging"[Mesh] OR (imaging, Magnetic Resonance) OR (NMR Imaging) OR (Imaging, NMR) OR (Zeugmatography) OR (Tomography, MR) OR (Tomography, NMR) OR (MR Tomography) OR (NMR Tomography) OR (Tomography, Proton Spin) OR (Proton Spin Tomography) OR (Magnetization Transfer Contrast Imaging) OR (MRI Scans) OR (MRI Scan) OR (Scan, MRI) OR (Scans, MRI) OR (fMRI) OR (MRI, Functional) OR (Functional MRI) OR (Functional MRIs) OR (MRIs, Functional) OR (Magnetic Resonance Imaging, Functional)	338572	09:18:13
#1	Add	Search "Knee Injuries"[Mesh] OR (Injuries, Knee) OR (Injury, Knee) OR (Knee Injury)	25790	09:17:29

Fonte: PUBMED, 2013.

Figura 2 – Tela do PubMed/MEDLINE

NCBI Resources How To Sign in to NCBI

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed ((#1) AND #2) AND #3 [Search](#) [Help](#)

[RSS](#) [Save search](#) [Advanced](#)

[Show additional filters](#) **Display Settings:** Summary, 20 per page, Sorted by Recently Added **Send to:** **Filters:** [Manage Filters](#)

Article types
Clinical Trial
Review
more ...

Text availability
Abstract available
Free full text available
Full text available

Publication dates
5 years
10 years
Custom range...

Species
Humans
Other Animals


[Clear all](#)
[Show additional filters](#)

Results: 1 to 20 of 2631 << First < Prev Page 1 of 132 Next > Last >>

[Medial Retinaculum Plasty Versus Medial Patellofemoral Ligament Reconstruction for Recurrent Patellar Instability in Adults: A Randomized Controlled Trial.](#)
Ma LF, Wang F, Chen BC, Wang CH, Zhou JW, Wang HY.
Arthroscopy. 2013 Apr 5. doi:pii: S0749-8063(13)00096-0. 10.1016/j.arthro.2013.01.030.
[Epub ahead of print]
PMID: 23568588 [PubMed - as supplied by publisher]
[Related citations](#)

[Sensitivity and specificity of magnetic resonance imaging for knee injury and clinical application for the Naresuan University Hospital.](#)
Laoruengthana A, Jarusriwanna A.
J Med Assoc Thai. 2012 Oct;95 Suppl 10:S151-7.
PMID: 23451455 [PubMed - indexed for MEDLINE]
[Related citations](#)

[\[Diagnosis of medial collateral ligament injury by stress X-ray and MRI of knee joint\].](#)
Zhang LJ, Chen JL, Xu Y, Zhu SB.
Zhongguo Gu Shang. 2012 Nov;25(11):951-3. Chinese.
PMID: 23427600 [PubMed - indexed for MEDLINE]
[Related citations](#)

Results by year

[Download CSV](#)

Find related data
Database:
[Find items](#)

Search details
((#1) AND #2) AND #3
[Search](#) [See more...](#)

Fonte: PUBMED, 2013.

Quadro 2 – Exemplo de estratégia na LILACS

População/condição: Pacientes com traumatismo nos joelhos
 (“Traumatismo do Joelho” OR “Traumatismos de La Rodilla” OR “Knee Injuries” OR “C26.558.554”)

AND

I = Intervenção: Ressonância Magnética

(“Imagem por Ressonância Magnética” OR “Imagem por Ressonância Magnética” OR “Magnetic Resonance Imaging” OR “Imageamento de Ressonância Magnética” OR “Tomografia por RM” OR “Imagem por RMN” OR “Tomografia por RMN” OR “Tomografia do Spin do Próton” OR “Varreduras por IRM” OR “Imagem Contrastada por Transferência de Magnetização” OR “IRMf” OR “Imagem por Ressonância Magnética Funcional” OR “IRM Funcional” OR “E01.370.350.500” OR “E01.370.350.825.500” OR “SP4.001.002.015.044.010.006”)

AND

C = Controle: Tomografia
 (“Tomografia” OR “Tomography” OR “E01.370.350.825”)

Fonte: elaboração própria.

Abaixo segue a construção da estratégia para a LILACS. Notar que entre os conjuntos não é usado o operador AND, ele é automático pela base.

Quadro 3 – Exemplo de Construção de estratégia na LILACS

(“Traumatismo do Joelho” OR “Traumatismos de La Rodilla” OR “Knee Injuries” OR “C26.558.554”)
 (“Imagem por Ressonância Magnética” OR “Imagem por Ressonância Magnética” OR “Magnetic Resonance Imaging” OR “Imagem de Ressonância Magnética” OR “Imageamento de Ressonância Magnética” OR “Tomografia por RM” OR “Imagem por RMN” OR “Tomografia por RMN” OR “Tomografia do Spin do Próton” OR “Varreduras por IRM” OR “Imagem Contrastada por Transferência de Magnetização” OR “IRMf” OR “Imagem por Ressonância Magnética Funcional” OR “IRM Funcional” OR “E01.370.350.500” OR “E01.370.350.825.500” OR “SP4.001.002.015.044.010.006”) (“Tomografia” OR “Tomography” OR “E01.370.350.825”)

Fonte: elaboração própria.

Figura 3 – Resultado da busca na LILACS

Portal de Pesquisa da BVS
Informação e Conhecimento para a Saúde

Home > Pesquisa > ("Traumatismo do Joelho" OR "Traumatismos de La Rodilla" OR "Knee Injuries" OR "C26.558.55... (23)

("Traumatismo do Joelho" OR "Traumatismos de La Rodilla" OR Título,resumo, assunto) Pesquisar

Busca Avançada | Localizar descritor de assunto

Curto | Ordem do resultado | 20

Resultados 1 - 20 de 23 1 2 Próxima > Última >>

1. **Accuracy of magnetic resonance in identifying traumatic intraarticular knee lesions**
Vaz, Carlos Eduardo Sanches; Camargo, Olavo Pires de; Santana, Paulo José de; Valezi, Antonio Carlos.
Clinics; 60(6): 445-450, Dec. 2005. tab.
Artigo em Inglês | LILACS | ID: 416489
[Mostrar mais](#) [Texto completo](#) [Fotocópia](#) [Documentos relacionados](#)

2. **Lesão meniscal por fadiga / Meniscal injury due to fatigue**
Camanho, Gilberto Luis.
Acta ortop. bras; 17(1): 31-34, 2009. *ilus, graf, tab*.
Artigo em Inglês, Português | LILACS | ID: 509091
[Mostrar mais](#) [Texto completo](#) [Fotocópia](#) [Documentos relacionados](#)

3. **Autoinjerto osteocondral de rodilla. Resultado clínico y radiológico a largo plazo / Osteochondral autograft of the knee. Clinical and radiological outcome**
Mastropiero, Javier; Ciccarello, Víctor Andres; Davila, Alberto.
Rev. Asoc. Argent. Ortop. Traumatol; 77(1): 57-65, mar. 2012.
Artigo em Espanhol | LILACS | ID: 649170

Sua seleção (0)
[Listar documentos](#)
[Limpar seleção](#)

Filtros selecionados
Base de dados
LILACS ([remover](#))

Filtrar

expandir todos fechar todos

Texto completo

Disponível (6)

Coleções

Bases de dados internacionais

Base de dados

LILACS (23)

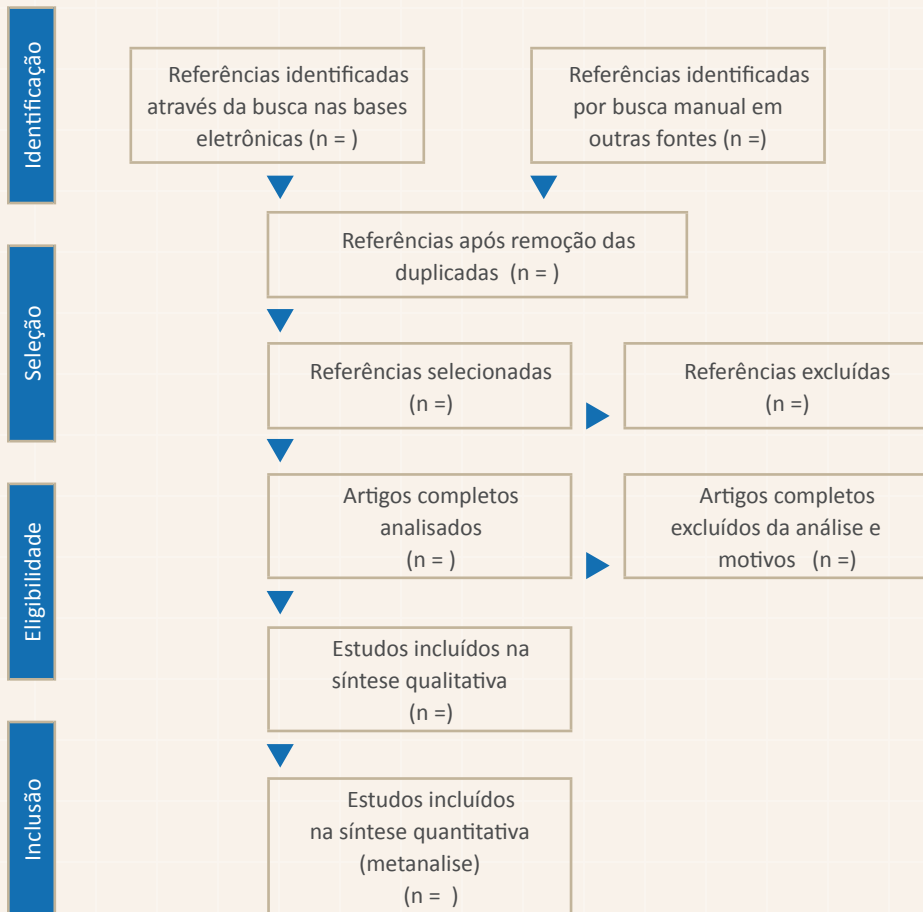
Assunto principal

Traumatismos do Joelho (23)

Fonte: BVS, 2013.

Notar que o resultado da busca recuperou 23 artigos.

ANEXO C – Fluxo de seleção dos artigos da revisão sistemática



ANEXO D – Exemplo de ficha clínica para revisões sistemáticas

REVISÃO SISTEMÁTICA DE TESTE DIAGNÓSTICO FICHA CLÍNICA

ID do estado: Dupla checagem

Autor: Ano:

Nome do revisor: _____

DADOS DO ESTUDO

Teste de diagnóstico avaliado: _____

Padrão Ouro: _____

Testado para a doença: _____

Total de pacientes no estudo:

Total de pacientes concluídos nas análises:

Crítérios de inclusão

População: _____ Teste usado Desfechos individuais reportados

Desenho do estudo: _____

Dados da tabela 2x2 sensibilidade especificidade valor preditivo positivo

Valor preditivo negativo LR+ LR-

Crítérios de inclusão (metodológicos)

Teste referência apropriado Seguimento completo Incorporação de vies

Consecutivo ou amostra aleatória Cegamento Somente estudos prospectivos

Descrição adequada da amostra

Outros critérios metodológicos

Comentários sobre os critérios de inclusão

ANEXO E – Principais fontes de viés e variabilidade nos estudos de acurácia diagnóstica.**Quadro 1– Fontes de vieses**

Fonte	Viés ou variação	Descrição
Características demográficas	Varição	Testes têm desempenhos diferentes em amostras variadas. Portanto, características demográficas podem levar a variações nas estimativas do desempenho do teste. Essas características constituem o espectro da doença.
Gravidade da doença	Varição	Diferenças na gravidade da doença entre os estudos podem levar a diferentes estimativas de desempenho do teste. Esta característica constitui o espectro da doença.
Prevalência da doença	Viés	A prevalência da condição-alvo varia de acordo com o cenário e pode afetar a estimativa de desempenho do teste. Em situações de alta prevalência de doença, os intérpretes do teste estão mais propensos a classificar o teste como anormal (viés de contexto). Adicionalmente, alteram as estimativas que levam em conta dados de prevalência, como acurácia, por exemplo.
Seleção dos pacientes	Viés	O delineamento do estudo e processo de seleção determina a composição da amostra do estudo. Se o processo de seleção não incluir pacientes com espectro similar ao da população em que o teste será utilizado na prática, os resultados estarão comprometidos ou poderão ser úteis apenas no perfil da população estudada.
Protocolo de teste: material e métodos		
Execução do teste	Varição	A descrição adequada de como foram executados os testes índice e padrão de referência é importante, pois variações nas medidas de acurácia diagnóstica podem resultar das diferenças na execução dos testes.
Tecnologia do teste	Varição	Quando as características de um teste diagnóstico mudam com o tempo como resultado do aperfeiçoamento tecnológico ou experiência do operador do teste, as estimativas do desempenho do teste podem ser afetadas.
Paradoxo de tratamento	Viés	Ocorre quando se inicia o tratamento, baseado no resultado de apenas um teste antes de submeter o paciente ao outro teste. Neste caso, o estado de doença pode ser alterado entre os testes.

Continuação

Viés de progressão da doença	Viés	Ocorre quando existe um atraso entre a aplicação do teste índice e o padrão de referência que permite que o estado de doença seja alterado.
Padrão de Referência e procedimentos de verificação		
Padrão de Referência inapropriado	Viés	O erro de diagnóstico derivado de um padrão de referência equivocado pode resultar na subestimação do desempenho do teste índice.
Viés de verificação diferencial	Viés	Ocorre quando o diagnóstico obtido pelo teste índice é verificado utilizando diferentes padrões de referência.
Interpretação (processo de leitura dos testes)		
Viés de inspeção	Viés	Interpretação do teste índice ou padrão de referência é influenciado pelo conhecimento dos resultados do outro teste. Assim, é importante que os testes sejam interpretados de forma independente, sem o conhecimento prévio do resultado de outro teste anteriormente aplicado.
Viés de inspeção clínica	Viés	A disponibilidade de informações clínicas, como idade, sexo, sintomas, comorbidades, etc., durante a interpretação do teste pode afetar a estimativa de desempenho do teste.
Viés de incorporação	Viés	Ocorre quando o resultado do teste índice é utilizado para estabelecer o diagnóstico final (como parte do padrão de referência).
Variabilidade de observação	Variação	A interpretação sobre o resultado de um teste pode variar entre os observadores e isso pode afetar a estimativa de acurácia do teste. A reprodutibilidade intrapessoal e interpessoal afeta a aplicabilidade do teste na prática.
Análises		
Manipulação de resultados indeterminados	Viés	Um teste diagnóstico pode produzir resultados que são indeterminados, com frequência variável que depende do teste. Geralmente, este problema não é reportado nos estudos e esses resultados indecifráveis simplesmente são removidos das análises.
Escolha arbitrária do valor do limiar de positividade	Variação	A seleção do valor de limiar para o teste índice que maximiza a sensibilidade e especificidade do teste pode levar a superestimação das medidas de desempenho de um teste. As estimativas calculadas nesse valor de corte em uma população independente pode não ser o mesmo que no estudo original.

Fonte: Adaptado de Whiting e col.¹⁴ (2008).

ANEXO F – QUADAS-2(Tradução não validada⁴⁵)

113

FASE 1: Formule a questão de pesquisa da revisão

<i>Pacientes</i> (cenário clínico, utilização pretendida do teste índice, apresentação, testes prévios)
<i>Teste(s) Índice</i>
<i>Padrão de Referência e condição-alvo</i>

FASE 2: Desenhe o diagrama de fluxo para o estudo primário**FASE 3: Julgamentos do risco de viés a aplicabilidade**

O QUADAS-2 é estruturado de modo que os quatro domínios-chave são cada um classificado em termos de risco de viés e preocupações relacionadas à aplicabilidade da questão de pesquisa (como definida acima). Cada domínio-chave apresenta uma série de questões norteadoras para ajudar a alcançar o julgamento sobre vieses e aplicabilidade.

DOMÍNIO 1: SELEÇÃO DOS PACIENTES**A. Risco de viés**

Descreva os pacientes incluídos (testes prévios, apresentação, uso pretendido do teste índice e cenário clínico):

- | | |
|---|-----------------|
| • Os pacientes foram recrutados de maneira consecutiva ou através de amostras aleatórias? | Sim/Não/Incerto |
| • O desenho caso-controle foi evitado? | Sim/Não/Incerto |
| • O estudo evitou exclusões inapropriadas? | Sim/Não/Incerto |

[Continua](#)

Continuação

Pode a seleção dos pacientes ter introduzido viés?

RISCO: BAIXO/ALTO/INCERTO

B. Preocupações relacionadas à aplicabilidade

Descreva os pacientes incluídos (testes prévios, apresentação, uso pretendido do teste índice e cenário clínico):

Existe uma preocupação que os pacientes incluídos não correspondem à questão de pesquisa?

PREOCUPAÇÃO: BAIXA/ALTA/INCERTA

DOMÍNIO 2: TESTE(S) ÍNDICE:

Se mais de um teste índice foi utilizado, por favor completar para cada teste.

A. Risco de viés

Descreva os pacientes incluídos (testes prévios, apresentação, uso pretendido do teste índice e cenário clínico):

- Os resultados do teste índice foram interpretados sem o conhecimento dos resultados do teste padrão de referência? Sim/Não/Incerto
- Se um limiar de positividade foi utilizado, ele foi pré-especificado? Sim/Não/Incerto

Pode a condução ou interpretação do teste índice ter introduzido vieses?

RISCO: BAIXO/ALTO/INCERTO

B. Preocupações relacionadas à aplicabilidade

Existe uma preocupação que o teste índice, sua condução ou interpretação difere da questão de pesquisa da revisão?

PREOCUPAÇÃO: BAIXA/ALTA/INCERTA

DOMÍNIO 3: PADRÃO DE REFERÊNCIA**A. Risco de viés**

Descreva os pacientes incluídos (testes prévios, apresentação, uso pretendido do teste índice e cenário clínico):

- O padrão de referência provavelmente classificou corretamente a condição-alvo? Sim/Não/Incerto
- Os resultados do padrão de referência foram interpretados sem o conhecimento dos resultados do teste índice? Sim/Não/Incerto

Pode o padrão de referência, sua condução ou interpretação ter introduzido vieses?

RISCO: BAIXO/ALTO/INCERTO

B. Preocupações relacionadas à aplicabilidade

Existe uma preocupação que a condição-alvo como definida pelo padrão de referência não corresponde à questão de pesquisa?

PREOCUPAÇÃO: BAIXA/ALTA/INCERTA

DOMÍNIO 4: FLUXO E TEMPO**A. Risco de viés**

Descreva qualquer paciente que não tenha recebido o(s) teste(s) índice e/ou o padrão de referência ou que foi excluído da tabela 2x2 (refere-se ao diagrama de fluxo):

Descreve o intervalo de tempo e qualquer intervenção entre o teste(s) índice e o padrão de referência:

- Existiu um intervalo de tempo apropriado entre a aplicação do(s) teste(s) índice e o padrão de referência? Sim/Não/Incerto
- Todos os pacientes receberam o padrão de referência? Sim/Não/Incerto
- Os pacientes receberam o mesmo padrão de referência? Sim/Não/Incerto
- Todos os pacientes foram incluídos nas análises Sim/Não/Incerto

Pode o fluxo dos pacientes ter introduzido viés?

RISCO: BAIXO/ALTO/INCERTO

Esta obra foi impressa em papel *couché* fosco 240 g/m² (capa) e papel *off set* 90 g/m² (miolo) pela Nome da Gráfica, em novembro de 2014. A Editora do Ministério da Saúde foi responsável pela normalização (OS 0211).

ISBN 978-85-334-2129-5



DISQUE SAÚDE



Disque Saúde 136
www.saude.gov.br

Biblioteca Virtual em Saúde do Ministério da Saúde
www.saude.gov.br/bvs



Ministério da
Saúde