DIRETRIZES METODOLÓGICAS

Sistema GRADE – manual de graduação da qualidade da evidência e força de recomendação para tomada de decisão em saúde

2ª edição: revista, ampliada e atualizada

TEXTO PARA A CONTRACAPA

SUMÁRIO

LISTA DE QUADROS

LISTA DE FIGURAS

LISTA DE TABELAS

Siglas e abreviaturas

AMSTAR 2	A MeaSurement Tool to Assess systematic Reviews 2
ATS	avaliação de tecnologias em saúde
BMJ	British Medical Journal
CEBM	Oxford Centre for Evidence-based Medicine
CMED	Câmara de Regulação de Medicamentos
DM2	diabetes melito tipo 2
DPP4	dipeptidil peptidase 4
ECR	ensaio clínico randomizado
EtD	*evidence to decision*
GLP1	peptídeo semelhante a glucagon 1
GRADE	Grading of Recommendations, Assessment, Development and Evaluation
HR	*hazard ratio*/razão de risco
IC95%	intervalo de confiança de 95%
NMA	*network meta-analysis*/metanálise em rede
NOS	escala de Newcastle-Ottawa
OR	*odds ratio*/razão de chances
PBE	práticas baseadas em evidência
PCDT	Protocolos Clínicos e Diretrizes Terapêuticas
PICO	população, intervenção, comparador e desfecho
REBRATS	Rede Brasileira de Avaliação de Tecnologias em Saúde
RoB 2.0	Risk of Bias 2.0
ROBINS-I	Risk Of Bias In Non-randomized Studies – of Interventions
RR	risco relativo
RS	revisão sistemática
SGLT2	cotransportador de sódio-glicose 2
SIGN	Scottish Intercollegiate Guidelines Network
SNC	sistema nervoso central
SoF	*summary of findings*/sumário dos resultados

APRESENTAÇÃO

O presente manual consiste em um documento para auxiliar pesquisadores na elaboração de recomendações com um paradigma que se denomina Saúde Baseada em Evidências e apresenta o sistema *GRADE – Grading of Recommendatons Assessment, Development and Evaluaton* como forma de avaliação da certeza da evidência e da força da recomendação.

Serão apresentados os princípios para a formulação de recomendações abordando a avaliação da certeza da evidência, síntese de evidência, construção de tabelas de evidências para a decisão, em especial sobre intervenções terapêuticas. O presente manual também apresenta novidades como a utilização do GRADE para testes e estratégias diagnósticas, abordagens para estudos de prognóstico, incidência e prevalência, e intervenções em saúde pública. Além disso, abordagens recentes como metanálise em rede e modelagem também são tópicos discutidos ao longo dos capítulos do atual manual GRADE.

O público alvo deste documento são os elaboradores de documentos relacionados a avaliações de tecnologias em saúde (ATS), revisões sistemáticas e diretrizes clínicas. A aplicação principal deste manual está na avaliação da evidência e na formulação de recomendações para diretrizes clínicas, revisões sistemáticas e ATS.

1. Saúde baseada em evidências: sistemas para avaliação da certeza da evidência e para a graduação da força da recomendação

A pesquisa científica atual produz um volume considerável de novas informações com alto grau de variabilidade metodológica, com implicações diretas para a qualidade das evidências proporcionadas pelos estudos. Um sistema eficiente de pesquisa deve abordar problemas de saúde considerados importantes para as populações, bem como as intervenções e resultados considerados importantes tanto por gestores, quanto por clínicos e pacientes (1). No entanto, o financiamento público da pesquisa está correlacionado apenas modestamente com a carga da doença, algo ainda mais evidente em país e de baixa e média renda. Assim, torna-se necessária a elaboração de materiais de síntese que facilitem o acesso a recomendações baseadas em fontes de evidências de alta qualidade e grau considerável de confiança, bem como sua interpretação, de modo a fornecer subsídios técnico-científicos para a tomada de decisão por profissionais da saúde e para os gestores (2).

É indicado que pesquisadores, usuários de diretrizes clínicas e metodologistas adotem práticas baseadas em evidência (PBE), pois elas podem contribuir para uma melhor fundamentação de decisões clínicas ou de saúde pública. A PBE é uma abordagem caracterizada pela utilização de diversas ferramentas e técnicas de epidemiologia, estatística, metodologia científica e informática para promover a integração entre a experiência clínica e as melhores evidências disponíveis, assim como identificar os valores e preferências dos pacientes (3). Nessa abordagem, também se consideram aspectos como a segurança e a efetividade das intervenções, bem como aspectos éticos na totalidade das ações individuais ou coletivas.

Tomadores de decisão e gestores devem considerar não apenas a magnitude das estimativas do efeito de potenciais vantagens e desvantagens esperadas para uma tecnologia, mas também o quão confiável aquela estimativa realmente é (4). Avaliar a confiabilidade das evidências disponíveis representa um componente-chave na PBE, e a certeza da evidência é formalmente descrita como “a extensão da nossa confiança de que as estimativas de efeito são corretas ou adequadas para apoiar uma decisão ou recomendação específica” (5) (Quadro 1). Isso envolve a probabilidade de

um evento (no caso de evidências, um desfecho) ocorrer e a confiança de que essa avaliação não é resultante de um efeito aleatório (6).

Quadro 1 – Definição de certeza da evidência adotada pelo GRADE *Working Group*
No contexto de uma revisão sistemática (RS), as classificações de certeza da evidência refletem a extensão da nossa confiança de que as estimativas de efeito estão corretas. No contexto de recomendações clínicas, as classificações de certeza da evidência refletem o grau da nossa confiança de que as estimativas de efeito são adequadas para apoiar uma decisão ou recomendação específica.
Nota: “qualidade da evidência” é o mesmo que “certeza da evidência”.

Fonte: elaboração própria.

Alguns sistemas de avaliação da qualidade da evidência foram desenvolvidos com o objetivo de embasar e instrumentalizar a PBE e graduar a força da recomendação, os quais consistem em informar a certeza das evidências apresentadas pelos ensaios clínicos e sugerir que uma recomendação de determinada conduta seja adotada ou rejeitada. Um dos primeiros esforços para caracterizar explicitamente a força de recomendações e o nível de evidência em cuidados de saúde foi publicado pelo *Canadian Task Force on the Periodic Health Examination* em 1979 (7). O nível de evidência foi baseado apenas no desenho adotado para o estudo clínico (ensaios clínicos randomizados [ECR]: nível I de evidência; estudos de coorte e caso-controle: nível II de evidência; opinião de especialistas: nível III de evidência), enquanto a força da recomendação foi baseada no grau de evidência.

Desde então, inúmeros sistemas de classificação da certeza da evidência e da força de recomendações clínicas foram criados como alternativas para avaliar as informações geradas por ensaios clínicos. Destacam-se, nesse sentido, os sistemas desenvolvidos pelo *Oxford Centre for Evidence-based Medicine* (CEBM) (8) e pela *Scottish Intercollegiate Guidelines Network* (SIGN). No entanto, muitos desses sistemas são pouco abrangentes e focados essencialmente no delineamento adotado pelos estudos avaliados, e não há padronização entre os diferentes sistemas em relação à caracterização da evidência (9). O Quadro 2 exemplifica a falta de consistência dos códigos utilizados entre os diferentes sistemas de classificação.

Quadro 2 – Níveis de evidência para recomendações adotadas por diferentes organizações sobre a utilização de anticoagulantes orais em pacientes com fibrilação atrial e doença reumática valvar mitral

Instituição	Nível de evidência	Grau de recomendação
American Heart Association (10)	C-EO	Classe I
Scottish Intercollegiate Guidelines Network (11)	2+, 4	D

Nota: apesar de as recomendações serem embasadas nas mesmas evidências, o julgamento sobre a qualidade da evidência e o grau de recomendação diferem em forma e força.
As diretrizes da *American Heart Association* indicam nível de evidência C-EO obtido de opiniões de especialistas baseadas na experiência clínica. A classe 1 de grau de recomendação indica uma forte força de recomendação onde os benefícios superam amplamente os riscos.
As diretrizes da *Scottish Intercollegiate Guidelines Network* sugerem o nível de evidência 2+ para resultados de estudos caso-controle ou de coorte bem delineados, com baixo risco de viés e probabilidade moderada de relação causal. O grau de recomendação D pertence aos níveis de evidência 3 e 4 ou consistem em extrapolação do nível de evidência 2+.

Fonte: elaboração própria.

Assim, a utilização de uma abordagem unificada e sistematizada para graduação da força de recomendação pode minimizar vieses e possíveis erros na interpretação dos estudos clínicos, auxiliando na interpretação das evidências disponíveis.

Pensando nas limitações identificadas nos sistemas anteriores, no início da década de 2000, um grupo de pesquisadores, metodologistas e especialistas em RS e em saúde pública colaboraram de modo informal para desenvolver um sistema mais abrangente de graduação da certeza da evidência e da força de recomendação. O *Grading of Recommendations, Assessment, Development and Evaluation* (GRADE) *Working Group* descreveu as primeiras abordagens do sistema em um artigo publicado no *British Medical Journal* (BMJ) em 2004 (12). Desde 2006, o BMJ solicita, nas instruções aos autores, a utilização preferencial do sistema GRADE para graduar a certeza da evidência em diretrizes clínicas. Assim, em uma série inicial de seis artigos publicados no próprio BMJ, o GRADE *Working Group* apresentou as recomendações e a importância da criação do sistema para sociedades médicas. Em sequência, uma coletânea de artigos contendo as diretrizes do sistema GRADE foi

publicada de forma sistemática no *Journal of Clinical Epidemiology* apresentando os fatores que rebaixam a certeza de uma evidência tanto em ECR quanto em estudos não randomizados. Ainda, a série inicial de artigos sobre o sistema GRADE apresentou um conjunto de considerações que podem elevar a certeza da evidência em estudos não randomizados. Após quase duas décadas de trabalho, o grupo continua a aperfeiçoar, sistematizar e expandir seus métodos por meio de seminários e eventos científicos, bem como correspondências eletrônicas, os quais se tornaram um laboratório de desenvolvimento e aprimoramento da metodologia GRADE.

1.1 O sistema GRADE e suas vantagens

O sistema GRADE pode ser definido como um sistema universal, transparente e sensível para graduação da certeza e da força da recomendação de um conjunto de evidências. O sistema propõe uma análise sistemática para definir questões de pesquisa e desfechos de interesse e avaliar a certeza do conjunto de evidências sobre uma temática de interesse em revisões sistemáticas e diretrizes clínicas, estabelecendo uma abordagem de recomendação do uso de tecnologias para gestores e profissionais da saúde (13). A avaliação do sistema GRADE não é realizada sobre estudos individuais, uma vez que o foco da sua aplicação ocorre sobre o conjunto de evidências geradas a partir de uma RS da literatura. Diversas vantagens surgiram no horizonte com essa abordagem, o que promoveu sua rápida adoção por diversas entidades e associações de saúde no mundo. Assim, conforme mencionado anteriormente, o sistema GRADE foi desenvolvido para preencher as lacunas dos sistemas de graduação de evidência anteriores e apresenta algumas vantagens (14):

- foi criado por um grupo representativo internacional de desenvolvedores de diretrizes clínicas;
- apresenta uma separação clara entre o julgamento da confiança da estimativa de efeito e da força de recomendação;
- avalia explicitamente a importância de desfechos de estratégias alternativas de gestão;
- possui critérios abrangentes de rebaixamento e elevação da certeza da evidência;

● promove um processo transparente de transformação de evidências em recomendações;

● indica explicitamente os valores e as preferências;

● inclui interpretações claras e pragmáticas sobre a força e a fraqueza de recomendações para profissionais, pacientes e formuladores de políticas;

● é útil para revisões sistemáticas e avaliações de tecnologias em saúde (ATS).

O sistema GRADE utiliza níveis de evidência para classificar a confiança da informação utilizada para apoiar uma determinada recomendação. Assim, a avaliação da certeza da evidência é realizada para cada desfecho analisado para a tecnologia de interesse, utilizando o conjunto disponível de evidências. Conforme descrito no Quadro 3, a certeza da evidência pode, então, ser classificada em quatro níveis, que representam a confiança da estimativa de efeito apresentada: alto, moderado, baixo ou muito baixo.

Quadro 3 - Níveis de certeza da evidência de acordo com o sistema GRADE.

Nível	Definição	Implicações	Fonte de informação
Alto	Há forte confiança de que o verdadeiro efeito esteja próximo do estimado.	É improvável que estudos adicionais modifiquem a confiança da estimativa de efeito.	- Ensaios clínicos bem delineados, com amostra representativa. - Em alguns casos, estudos não randomizados bem delineados, com achados consistentes*.
Moderado	A confiança da estimativa de efeito é moderada.	Futuros estudos podem modificar a confiança da estimativa de efeito, podendo, inclusive, modificar a estimativa.	- Ensaios clínicos com limitações não significativas**. - Estudos não randomizados bem delineados, com achados consistentes*.
Baixo	A confiança da estimativa de efeito é limitada.	Futuros estudos provavelmente terão um impacto importante na confiança da estimativa de efeito.	- Ensaios clínicos com limitações moderadas**. - Estudos não randomizados comparativos (coorte e caso-controle).
Muito baixo	A confiança da estimativa de efeito é muito limitada. Há um importante grau de incerteza nos achados.	Qualquer estimativa de efeito é incerta.	- Ensaios clínicos com limitações graves**. - Estudos não randomizados comparativos com presença de limitações**. - Estudos não randomizados e não comparativos***.

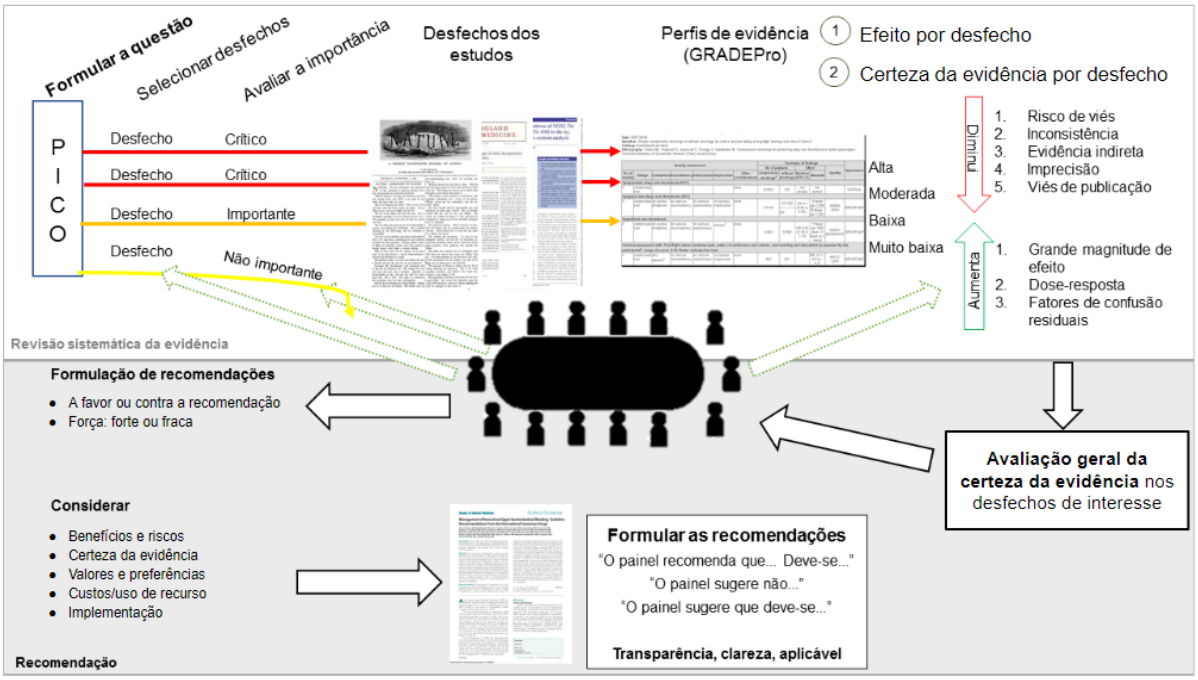
			- Opinião de especialistas.

* Estudos de coorte sem limitações metodológicas, com achados consistentes e com tamanho de efeito grande e/ou gradiente dose-resposta.
** Limitações: vieses no delineamento do estudo, inconsistência nos resultados, desfechos substitutos ou validade externa comprometida.
*** Séries e relatos de casos.
Fonte: elaboração GRADE Working Group (https://www.gradeworkinggroup.org/).

A classificação inicial da certeza é realizada a partir de critérios como delineamento do estudo e ferramenta adotada para avaliação do risco de viés. A partir da classificação inicial, diversos fatores que podem reduzir ou elevar a certeza são utilizados para analisar o conjunto final de evidências. Os fatores responsáveis pela redução no nível de certeza da evidência são os seguintes: limitações metodológicas (risco de viés), inconsistência (heterogeneidade), evidência indireta, imprecisão e viés de publicação. Adicionalmente, a certeza da evidência, caso não tenha sido reduzida devido aos fatores descritos acima, pode ser elevada por três fatores: grande magnitude de efeito, gradiente dose-resposta e fatores de confusão residuais (15).

O sistema GRADE, quando utilizado no contexto de elaboração de uma recomendação clínica, como em diretrizes ou protocolos, utiliza a força da recomendação para apoiar sua conduta. A certeza da evidência é um dos aspectos considerados durante o processo de recomendação clínica, em conjunto com o balanço entre os riscos e os benefícios do uso da tecnologia de interesse, as preferências e os valores de pacientes e familiares e os custos envolvidos na adoção da tecnologia. A força da recomendação pode ser a favor ou contra a tecnologia proposta em diretrizes clínicas e pareceres técnicos e pode ser específica para o cenário avaliado. A Figura 1 apresenta uma visão esquemática do sistema GRADE durante a elaboração de uma diretriz clínica.

Figura 1 – Visão esquemática da metodologia GRADE para síntese de evidências e desenvolvimento de recomendações



A metade superior descreve os passos relacionados a revisões sistemáticas e a recomendações para cuidados de saúde, enquanto a parte inferior descreve os passos específicos para fazer as recomendações.
Fonte: baseada na Figura 2 do *GRADE Handbook* da Cochrane *Training* (16).

O objetivo deste manual é auxiliar pesquisadores na elaboração de recomendações em tecnologias de saúde com base em análises de certeza da evidência. Nos próximos capítulos, serão apresentados os princípios para a formulação de recomendações, abordando a certeza da evidência em estudos clínicos no âmbito da saúde. O presente manual consiste em um documento revisado e ampliado das "Diretrizes metodológicas: Sistema GRADE – manual de graduação da qualidade da evidência e força de recomendação para tomada de decisão em saúde", publicado em 2014 pelo Ministério da Saúde (17), e possui como público-alvo elaboradores de revisões sistemáticas e ATS.

2. Elaboração da questão de pesquisa e escolha dos desfechos

- O GRADE requer uma especificação clara do cenário, população, intervenção, comparador(es) e desfechos relevantes;
- As perguntas devem indicar de maneira suficientemente específica: as populações, as intervenções e os desfechos;
- Os desfechos de interesse devem ser aqueles relevantes para os pacientes e priorizados na formulação da questão de pesquisa. Os desfechos substitutos (*surrogate outcomes*) geralmente proporcionam qualidade de evidência inferior devido ao seu caráter indireto, contudo podem ser usados em um contexto onde não houver um desfecho de relevância clínica, de forma justificável, devendo ser apresentado com limitações;
- No uso de desfechos substitutos, deve-se deixar explícita a sua relação com desfechos clinicamente relevantes e considerar a avaliação complementar de desfechos clínicos com a evidência de estudos observacionais.
- Em uma diretriz clínica, idealmente, a questão de pesquisa deve ser desenvolvida e/ou validada junto a um painel de especialistas com diferentes formações, sendo direcionada para a busca de respostas relevantes para a elaboração de uma recomendação clínica.
- Na construção de questões de pesquisa, utilizar o acrônimo PICO (população, intervenção, comparadores e desfechos) em vez de PICOT (adicionando T, referente a tempo) ou PICOS (adicionando S de *study desing*, referente ao desenho do estudo).

2.1. Definição da questão de pesquisa

A definição da **questão de pesquisa** faz parte do desenvolvimento de uma revisão sistemática (RS) tanto quanto em uma diretriz clínica, onde é determinada a partir de uma **dúvida clínica** a ser respondida. Mais especificamente para uma diretriz clínica, esse processo ocorre durante o momento de escopo, onde os participantes do painel definem quais as **dúvidas clínicas** são relevantes e pertinentes em um determinado contexto de saúde (14). Sobre esse aspecto, é importante esclarecer sobre as diferenças entre uma questão para uma diretriz clínica ou solicitação de avaliação de incorporação de tecnologias, e uma questão para uma síntese de evidência. Em uma diretriz clínica a pergunta direciona-se a realizar uma

recomendação sobre utilizar ou não determinada intervenção, e com que ênfase (recomendações fortes ou condicionais). Assim, para obter a resposta dessa questão, muitas vezes é necessário realizar um balanço entre as evidências para benefícios e riscos, além de avaliar custos, valores e preferências dos pacientes, viabilidade de implementação, entre outros. Essa abordagem é semelhante no processo de avaliação de tecnologias em saúde, no qual uma determinada questão é respondida por meio da incorporação ou não da tecnologia no sistema de saúde e, no caso de incorporação, quais seriam seus condicionantes (por exemplo, redução de preço, compartilhamento de risco ou desenvolvimento de novas evidências).

Por sua vez, a questão de pesquisa no processo de síntese de evidências de um método sistemático visa a responder sobre a efetividade desta (riscos e benefícios) e é parte imprescindível do processo de desenvolvimento de recomendações, seja para diretrizes, seja para decisões de incorporação de tecnologias. Nesse caso, o objetivo do processo de síntese de evidências não é responder se a intervenção deve ser utilizada na prática clínica ou se deve ser incorporada no sistema de saúde; tampouco deve se limitar à informação se a intervenção é efetiva e segura ou não. A resposta de uma questão de pesquisa em uma síntese de evidências deve nos direcionar a qual é a estimativa de efeito da intervenção para cada um dos desfechos avaliados, de forma a subsidiar o processo de tomada de decisão.

Dentro da avaliação do sistema GRADE, é importante que os principais itens de uma questão de pesquisa estejam definidos de maneira clara. Neste ponto, a estratégia PICO é o método de elaboração mais recomendado e aceito para questões de saúde em sínteses de evidência. Quando realizado de maneira adequada, instrumentaliza o pesquisador a definir a melhor estratégia de busca, seleção e avaliação dos estudos que serão a base para alcançar a resposta da sua respectiva dúvida clínica. Utilizando a estratégia PICO, consegue-se delinear de forma estruturada os principais componentes de uma questão de pesquisa: os pacientes ou população a quem devem ser investigadas e/ou aplicadas as recomendações; a intervenção terapêutica, diagnóstica ou um fator de exposição; a intervenção de controle ou alternativa; e o(s) desfecho(s) de interesse (Quadro 4).

Quadro 4 - Definição do acrônimo PICO.

P – População **(*Patient*, em inglês)**	Em qual população estamos interessados? Quais as características dessa população? Há subgrupos que precisam ser considerados?
I – Intervenção **(*Intervention*, em inglês)**	Qual a intervenção que queremos avaliar? Qual estratégia/teste diagnóstico queremos avaliar (teste índice)? Qual fator de exposição queremos investigar?
C – Comparador **(*Control*, em inglês)**	Qual é a principal alternativa já disponível para comparação com a tecnologia a ser avaliada? Por exemplo: cuidado padrão usado na prática clínica, teste referência para avaliação diagnóstica, etc. No caso de diretrizes clínicas para o SUS, o comparador deve ser uma alternativa já disponível no sistema público de saúde.
O – Desfecho **(*Outcome*, em inglês)**	O que é realmente importante para o paciente? Quais desfechos devem ser considerados? Em que situações podemos considerar desfechos substitutos?

Fonte: elaboração própria.

No contexto do sistema GRADE, o acrônimo PICO é utilizado para a definição do escopo da questão que será avaliada. Dessa forma, primeiramente é importante identificar se os resultados foram baseados em trabalhos em que a amostra estudada seja correspondente a população a qual a dúvida clínica se direciona, ou seja, apresenta os mesmos critérios de inclusão. Se uma revisão, baseada em uma população ampla, que apresenta características basais diferentes (gravidade da doença, por exemplo), gera estimativa de efeito para uma dada intervenção, essas estimativas poderão ser irreais para algumas subpopulações de pacientes. Neste contexto, corre-se o risco de a decisão clínica tomada não ser a mais adequada para uma parte dos indivíduos, como os mais vulneráveis para aquele desfecho, por exemplo.

Dessa forma, sugere-se iniciar com uma questão de pesquisa com população mais ampla e, em seguida, definir uma população mais restrita, considerando eventuais especificidades que possam explicar qualquer heterogeneidade encontrada nos futuros resultados. Em relação a subgrupos, as especificações podem incluir diferenças em pacientes, intervenções, escolha do comparador, o(s) resultado(s) ou fatores relacionados ao viés; por exemplo, estudos de alto risco de viés produzem efeitos diferentes dos estudos de baixo risco de viés.

A definição da intervenção também precisa ser clara e bem definida, e requer atenção nas suas particularidades de apresentação, como: diferentes alternativas de esquemas de tratamento (dose, via de administração, tempo de intervenção, por exemplo), diferentes estratégias ou teste diagnóstico (teste índice, por exemplo) ou formas de avaliar um fator de risco (fonte de dados, tempo de mensuração do desfecho, por exemplo). Da mesma forma, para a população, as estimativas poderão ser diferentes conforme as especificações utilizadas para avaliar a tecnologia (veja o exemplo no Quadro 5).

Um outro desafio para a definição da PICO surge quando a tecnologia estudada apresenta múltiplos comparadores. Os autores de RS devem ter clareza na escolha do comparador, principalmente quando esse comparador não é estabelecido na prática clínica. Em uma diretriz, os painelistas devem especificar o(s) comparador(es) escolhido(s), sobretudo, quando várias tecnologias estão envolvidas: eles devem especificar se a recomendação está sugerindo que todas elas são igualmente recomendadas ou, se há alguma tecnologia que será mais fortemente recomendada em detrimento de outra (para exemplificar esse ponto, veja o Quadro 6) (18). Por vezes, deverá ser utilizado o sistema GRADE para comparações múltiplas (metanálise em rede), como está especificado no capítulo 8. Sistema GRADE para metanálises em rede.

Ao eleger o último item da questão PICO, os autores precisam definir quais desfechos serão considerados para responder à dúvida clínica, assim como a importância relativa entre eles no processo de tomada de decisão. Na próxima seção, são detalhadas as características que devem ser consideradas em relação aos desfechos.

Além do tradicional acrônimo PICO, existem outros itens que por vezes são definidos no momento da estruturação da questão de pesquisa como o tempo (PICOT) e desenho do estudo (PICOS) (Quadro 7). Contudo, por padronização, recomenda-se utilizar o acrônimo PICO.

Sobre o tempo, no sistema GRADE, esse pode ser considerado como um componente do próprio desfecho (por exemplo, mortalidade em 30 dias em pacientes hospitalizados). Quando combinados desfechos com diferentes períodos avaliados, essas diferenças podem ser consideradas dentro do domínio de evidência indireta.

Sobre o desenho do estudo, é importante salientar que uma mesma questão de pesquisa pode ser respondida por meio de diferentes delineamentos. Por exemplo, uma questão pode ser respondida tanto por estudos observacionais que avaliaram determinada intervenção quanto por ensaios clínicos, apesar desse último ser o desenho preferencial. É importante ressaltar que a condução de uma RS pode envolver limitações relacionadas ao tipo de estudo (como ensaios clínicos ou estudos observacionais), ao tipo de publicação (seja um resumo, pré-print ou artigo completo), ao ano de publicação e/ou condução do estudo, bem como ao idioma. Essas limitações são parte da metodologia da RS, não sendo componentes da questão de pesquisa. Por fim, é importante salientar que o uso dessas restrições é comum em RS, sendo geralmente aplicadas para otimizar a carga de trabalho no processo de identificação e síntese de evidências; muitas vezes essas restrições resultam em uma resposta parcial à questão de pesquisa, devendo ser avaliada sua adequabilidade em cada revisão (Quadro 7).

Quadro 5 - Escolha de pacientes e intervenção – Um exemplo prático

Uma RS sobre o uso de agentes antiagregantes plaquetários apresenta como objetivo avaliar a probabilidade para ocorrência de eventos cardiovasculares em pacientes com elevado risco para essa condição. Dessa forma, a questão PICO construída inclui como população os indivíduos que podem se beneficiar de agentes antiagregantes plaquetários e a intervenção em estudo é a administração desses agentes. Na análise sobre a abrangência dos pacientes, os autores deste estudo observam que os agentes antiplaquetários podem apresentar diferentes magnitudes de efeito conforme a variabilidade de risco na linha de base (por exemplo, prevenção primária ou prevenção secundária); assim como as intervenções dos estudos

individuais descrevem o uso de diferentes esquemas de intervenção como baixa dose de ácido acetilsalicílico, todas as doses de ácido acetilsalicílico ou outras possibilidades de agentes antiagregantes plaquetários, por exemplo. Entretanto, apesar do risco absoluto diferir substancialmente, observa-se que o risco relativo dos estudos individuais são semelhantes e indicam a mesma direção de resultado. Dessa maneira, os autores são incentivados a conduzirem uma análise incluindo todos os estudos gerando um único risco relativo que se aplica ao grupo geral, mas as conclusões e/ou recomendações (no contexto de diretrizes) podem diferir entre os subgrupos de pacientes conforme as especificações consideradas como relevantes, em especial, pela diferença absoluta do risco ser diferente.

Explicação sobre diferenças entre efeitos absolutos e efeitos relativos no processo de tomada de decisão é apresentada na seção 3.3.4 (imprecisão)

Fonte: adaptado de Guyatt et al. (18).

Quadro 6 - Escolha de comparador – Um exemplo prático

Para avaliar a escolha do comparador, pode-se seguir o exemplo do uso de antiagregantes plaquetários, pensando em dois possíveis cenários de dúvida clínica: 1º) Devemos utilizar antiagregantes plaquetários na prevenção primária em pacientes com elevado risco para eventos cardiovasculares? 2º) Qual o antiplaquetário deve ser utilizado nessa população?

No primeiro cenário, sabendo que o objetivo principal é saber se é indicado o uso de antiagregantes plaquetários, é indicado que o comparador seja o placebo, ou seja, o não uso de antiplaquetário; já no segundo cenário, entendendo que o objetivo é analisar a superioridade ou não inferioridade de diferentes agentes antiagregantes plaquetários, é indicado que o comparador seja o agente utilizado na prática clínica.

Fonte: adaptado de Guyatt et al. (18).

Quadro 7 - Além do tradicional acrônimo PICO

Existem algumas derivações do acrônimo PICO como a PICOT, por exemplo, onde

a versão adiciona a letra T que representa o tempo (time) determinado para avaliação do desfecho; ou PICO, onde S representa a definição do desenho

metodológico dos estudos (study) incluídos. Tais modificações não são necessariamente úteis ou relevantes nesta fase do processo, partindo do pressuposto que o tempo necessário para a duração de um tratamento ou acompanhamento a uma exposição podem não ser estabelecidos/padronizados, nem determinados como um item de investigação; assim como a definição de qual a melhor abordagem metodológica para responder uma **dúvida clínica,** podem não ser conhecidos quando a **questão de pesquisa** é feita. Ainda em relação ao **desenho do estudo**, é importante ressaltar que em alguns contextos, não há a disponibilidade de ensaios clínicos randomizados que respondam à **dúvida clínica**, ou ainda existam estudos observacionais que fornecem evidências que estão associados a alta confiança nas estimativas. Dessa forma, muitas vezes não é sensato restringir um desenho de estudo de antemão (14). Uma vez que questões de pesquisa podem ser respondidas com diferentes delineamentos, o desenho do estudo não deve ser visto como integrante da questão de pesquisa, mas sim quanto ao processo pelo qual a questão é respondida. Assim, a sugestão é padronizar o uso do acrônimo PICO na estruturação da questão de pesquisa.

Fonte: elaboração própria.

2.2. Escolha dos desfechos e sua priorização no sistema GRADE

Na avaliação do sistema GRADE, o resultado e/ou a recomendação em relação ao uso ou incorporação de determinada tecnologia considera todo o conjunto de evidências, incluindo o balanço entre os riscos e os benefícios. Dessa forma, faz-se necessária a identificação dos desfechos, assim como a atribuição da sua respectiva importância no processo de tomada de decisão, considerando o contexto de diretrizes clínicas ou de avaliação de tecnologias em saúde.

É comum que desfechos considerados relevantes para uma tecnologia não sejam elencados, como os danos e eventos adversos da aplicação de uma intervenção, por exemplo. Os autores de uma RS podem até eleger qual desfecho focarão; contudo, no contexto das diretrizes clínicas ou de avaliação de tecnologias em saúde, os desfechos devem ser definidos conforme sua relevância ao paciente, e não apenas pela disponibilidade de evidências na literatura (18). Adicionalmente, no processo de avaliações com o objetivo de incorporação de novas tecnologias,

idealmente os desfechos de interesse devem ser previamente alinhados com o tomador de decisão (gestor).

Em diretrizes clínicas e avaliação de tecnologias em saúde, é comum serem avaliados desfechos substitutos, visto que os resultados de desfechos importantes para o paciente são relativamente infrequentes ou precisam ser observados por longos períodos. Como o uso de resultados desses desfechos mostrou-se equivocado para basear recomendações em algumas situações, o sistema GRADE orienta que sejam especificados os desfechos importantes para o paciente e, apenas se necessário, os substitutos a serem usados na sua ausência. Ressalta-se que onde faltam evidências para um desfecho considerado importante, essa lacuna precisa ser reconhecida e deve-se ponderar o uso de um desfecho substituto como uma evidência indireta, em detrimento de ignorar o desfecho, visto que essa ação pode influenciar a recomendação final (Quadro 8 e Quadro 9) (18).

Como a maioria das revisões sistemáticas não abrange as evidências para todos os desfechos elencados, a tomada de decisão geralmente baseia-se em revisões sistemáticas de diferentes fontes e/ou na elaboração de revisões sistemáticas próprias. Em resumo, os desenvolvedores de diretrizes ou ATS devem considerar desfechos substitutos quando faltam evidências de alta qualidade sobre desfechos importantes. Os desfechos devem ser indicados com suas respectivas justificativas e esse processo deve ser feito com atenção, pois o uso de um desfecho substituto pode reduzir a qualidade da evidência devido à presença de evidência indireta (18).

Quadro 8 – Escolha de desfecho – um exemplo prático

Para definição do desfecho, podemos citar a avaliação de uma nova tecnologia para o tratamento de diabetes melito. Nesse cenário hipotético, os autores elegem como desfecho importante a ocorrência de eventos macrovasculares e microvasculares, devido à sua relevância clínica para o paciente. Caso esse dado não esteja disponível e haja apenas evidências para o impacto sobre o nível de hemoglobina glicosilada, esse desfecho será considerado e descrito como desfecho substituto. Esses casos podem ser explicados pelo fato de que os resultados importantes são

relativamente infrequentes ou precisam ser observados por longos períodos de tempo, mostrando uma tendência em medir os desfechos substitutos.

Fonte: elaboração própria.

Quadro 9 – Desfechos importantes – Diretrizes

Em geral, no desenvolvimento de diretrizes é indicado que os painelistas priorizem os resultados relacionados aos eventos de morbimortalidade e efeitos adversos para as recomendações. Contudo, também devem ser elegíveis outros desfechos como hospitalização, função, incapacidade, qualidade de vida, inconveniência e uso de recursos. Veja mais detalhes nas seções 2.3. *Uso consciencioso de desfechos substitutos* e 2.4. *Hierarquização da importância dos desfechos* deste capítulo.

Fonte: elaboração própria.

2.3. Uso consciencioso de desfechos substitutos

Há condições em que desfechos finais procedentes de estudos clínicos randomizados não são disponíveis, sendo justificável o uso de desfechos substitutos, além de desejável complementação da evidência por meio de estudos observacionais. Em especial, destaca-se:

- Tecnologias para doenças raras e/ou negligenciadas;
- Novas tecnologias em oncologia;
- Tecnologias antigas, já integradas na prática clínica.

O desenvolvimento de ensaios clínicos é desafiador no contexto de doenças raras, em especial em doenças ultrarraras. As dificuldades de recrutamento, ligadas à baixa prevalência, implicam na perda relativa do poder do estudo em detectar diferenças estatisticamente significativas entre as intervenções propostas para avaliar o impacto nos desfechos avaliados. Assim, é comum utilizar desfechos contínuos e/ou desfechos compostos, para identificar potenciais diferenças com a amostra disponível, muitas vezes consistindo em desfechos substitutos. Muitas dessas doenças raras são crônico-degenerativas, que cursam sem efeitos significativos sobre a mortalidade em curto e médio prazo. A obtenção de desfechos definitivos (principalmente mortalidade) como resultado direto dos ensaios clínicos é praticamente inviável, pois exigiria não só manter os pacientes sob tratamento

durante décadas, mas também manter o grupo-controle sob os efeitos de um tratamento ineficaz durante o mesmo período.

Destaca-se que a base para o desenvolvimento de um ensaio clínico randomizado é o princípio da equipolência (*equipoise*), no qual há importante incerteza do benefício de uma alternativa em teste sobre o seu controle, o que justifica um esquema de randomização que alocará uma proporção dos pacientes no grupo controle. Com a superioridade demonstrada em desfechos intermediários, em especial em circunstâncias nas quais não há alternativas terapêuticas, muitas vezes não se pode assumir o pressuposto de equipolência (19).

Em estudos de oncologia, seguindo o princípio da equipolência, temos observado cada vez mais a substituição do desfecho de sobrevida global por sobrevida livre de progressão. Assim, em diversos estudos avaliando novos tratamentos oncológicos, no momento de progressão da doença no grupo controle, é oferecida a possibilidade de passar a utilizar a intervenção proposta. Esse cruzamento do grupo controle para o grupo intervenção dilui o potencial efeito esperado em um estudo clínico, limitando as conclusões em relação à sobrevida global (20). De forma semelhante, em doenças órfãs, pelo mesmo princípio, são definidos desfechos intermediários como os desfechos primários em estudo clínicos. A melhora nesses desfechos justifica um estudo de extensão, transferindo os pacientes do grupo-controle para o grupo de tratamento. Assim, pressupor a inexistência de efeito em desfechos finais é incorreto.

Em doenças raras crônico-degenerativas, frequente por exemplo, o teste de caminhada de 6 minutos (TC6M) é um desfecho. Com base na fisiopatologia da doença, por exemplo, espera-se que a perda progressiva da força muscular, avaliada pelo TC6M, resulte em perda de deambulação. Assim, a perda de força muscular em membros inferiores seria o desfecho que atua como mediador para perda da deambulação, sendo avaliado pelo TC6M (21).

A dosagem de biomarcadores, apesar de menos comum, por vezes também pode ser utilizada como desfechos substitutos: por exemplo, em acromegalia a evolução da doença é decorrente diretamente do excesso de GH e de IGF-1 ao longo do tempo, com as manifestações ocorrendo tardiamente; sua normalização é considerado como desfecho de interesse tanto em pesquisa como alvo terapêutico

na prática clínica (22). É importante lembrar que em algumas doenças comuns os desfechos substitutos também podem ser utilizados, sendo amplamente aceitos em alguns casos. Por exemplo, em HIV a contagem de linfócitos T CD4 possui uma associação com infecções oportunísticas (em especial quando inferior a 200 células/mm^3), e em osteoporose, o escore T em densitometria óssea é um preditor importante de fratura. Apesar do impacto em desfechos clínicos ser sempre desejável, o seu uso não implica necessariamente em redução importante da certeza da evidência, sendo usualmente utilizado para tomada de decisão, tanto do ponto de vista regulatório quanto na incorporação de tecnologias.

De forma semelhante, para tecnologias antigas e doenças negligenciadas, muitas vezes sem interesse comercial corrente e/ou sem financiamento para desenvolvimento de estudos clínicos definitivos, as decisões precisam ser tomadas a luz da evidência disponível, muitas vezes baseada em desfechos substitutos.

É importante salientar que a impossibilidade de obter desfechos finalísticos em determinadas condições não exime as incertezas inerentes a essa opção. Contudo também é incorreto assumir que não há efeito clínico, sendo menosprezados os achados referentes aos desfechos substitutos, em especial quando os estudos existentes não foram delineados para avaliar desfechos finais. Assim, é importante que sejam considerados alguns aspectos adicionais no processo de tomada de decisão, entre eles, a definição de quanto o desfecho substituto é acurado em representar desfechos clinicamente relevantes e, muitas vezes, buscar evidências complementares para a efetividade em desfechos clínicos a partir de estudos observacionais, em especial em doenças raras, doenças negligenciadas e intervenções antigas.

2.4. Hierarquização da importância dos desfechos

A hierarquização da importância dos desfechos é um processo essencial para a tomada de decisão clínica, que deve ser realizada, idealmente, antes de iniciar a busca pelas evidências. Essa definição deve ser baseada nos desfechos relevantes para os pacientes, preferencialmente incluindo-os neste processo. Para definir a ordem de importância, pode-se utilizar um sistema de votação ou métodos de

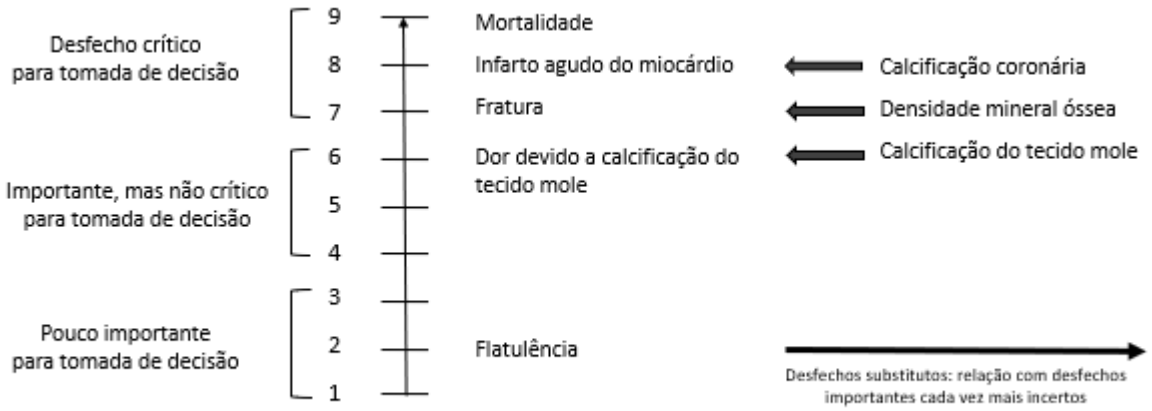
consenso (Delphi, por exemplo). No sistema GRADE, a importância dos desfechos é classificada em uma escala de 1 a 9 dividida em três categorias (Figura 2):

- 1-3: desfechos pouco importantes para o processo de decisão;
- 4-6: desfechos importantes para o processo de decisão;
- 7-9: desfechos críticos para o processo de decisão.

Em uma diretriz clínico-assistencial, a definição dos desfechos conforme seu nível de importância (pouco importante, importante ou crítico) direciona o processo de desenvolvimento de recomendações, assim como uma avaliação mais pragmática do balanço entre riscos e benefícios. Em especial, as decisões sobre a qualidade geral das evidências que suportam uma recomendação podem depender de quais resultados são designados como críticos para a tomada de decisão (por exemplo, aqueles classificados com 7, 8 ou 9, na escala mencionada) e quais não são. Por exemplo, um painel de diretrizes indica que evidências de alta qualidade suportam todos os resultados, exceto um, que, por sua vez, apresenta apenas evidências de baixa qualidade disponíveis. Se esse resultado em questão for crítico, a qualidade geral da evidência será designada como de baixa qualidade; mas se os painelistas considerarem que o resultado restante é importante, mas não crítico, a classificação geral da qualidade da evidência será de alta qualidade para a recomendação associada (18).

Como acontece para os outros componentes, não se pode oferecer regras rígidas para o julgamento da importância dos desfechos. O que o sistema GRADE propõe é identificar as questões incluídas e permitir uma avaliação transparente e compreensível dos julgamentos envolvidos. Dessa forma, os painelistas de diretrizes podem ter subsídios para debater as questões, e os usuários das diretrizes fazem sua própria avaliação da adequação das conclusões geradas (18).

Figura 2 - Importância relativa dos desfechos



Fonte: adaptado de Guyatt et al. (18).

2.5. Definindo questões no contexto de tomada de decisão

Como descrito anteriormente, definir uma questão de pesquisa em saúde inclui especificar todos os desfechos de interesse. Os autores de RS ou desenvolvedores de diretrizes clínicas ou ATS devem considerar todos os resultados relevantes para avaliar o uso ou não de uma determinada intervenção (terapêutica ou diagnóstica).

Para facilitar o processo, é disponibilizada o GRADEpro (www.gradepro.com), que consiste em uma ferramenta para auxílio no desenvolvimento de diretrizes, facilitando a implementação do sistema GRADE. O GRADEpro permite a seleção de dois modelos diferentes para formular uma questão de pesquisa, conforme proposto abaixo (14):

• Deve-se usar [intervenção] vs. [comparação] para [problema de saúde]?

• Deve-se usar [intervenção] vs. [comparação] em [população]?

Além de um formato para perguntas sobre diagnóstico:

• A [intervenção] vs. [comparação] deve ser usada para diagnosticar [condição-alvo] em [problema de saúde e/ou população]?

Perguntas de exemplo (14):

1. As escovas de dentes manuais *versus* escovas de dentes elétricas devem ser usadas para a saúde bucal?

2. Os corticoides nasais tópicos devem ser usados em crianças com rinite alérgica persistente?3. 'O oseltamivir versus nenhum tratamento antiviral deve ser usado para tratar pessoas com gripe?

4. A mensuração do nível sérico de troponina I seguida de estratégias apropriadas de manejo deve ser usada no tratamento do infarto agudo do miocárdio?

Note que, no contexto de tomada de decisão (como temos em diretrizes clínico-assistenciais e pedidos de incorporação), a questão de interesse difere da pergunta de uma síntese de evidência (PICO) por dois fatores principais. Primeiro, diferente de uma questão de RS, onde usualmente se questiona os benefícios e riscos de uma determinada terapia, é questionada se a mesma deve ser utilizada ou não (ou incorporada, no contexto de avaliação de novas tecnologias), para a qual serão levados outros fatores na tomada de decisão, conforme será discutido no capítulo 5. Segundo, não são estabelecidos os desfechos específicos na questão de pesquisa, pois o conjunto dos resultados desses desfechos (e não eles isoladamente) é que determina a recomendação.

3. Avaliação da certeza da evidência

3.1 Níveis de certeza para o conjunto final de evidências

O nível de evidência representa a confiança da informação utilizada para apoiar uma determinada recomendação. No sistema GRADE, a avaliação da qualidade da evidência é realizada para cada desfecho analisado para a tecnologia de interesse por meio do conjunto disponível de evidências. Sugere-se que a certeza da evidência seja interpretada de forma diferente na elaboração de revisões sistemáticas e de recomendações em saúde:

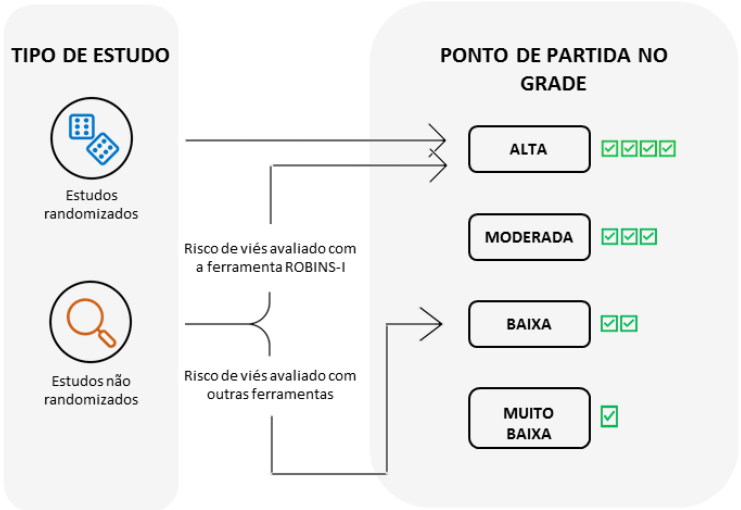
- na elaboração de revisões sistemáticas, a certeza da evidência reflete o nível de confiança de que uma estimativa de efeito está correta;
- na elaboração de recomendações para uma diretriz clínica, a certeza da evidência reflete o nível de confiança de que uma estimativa de efeito é adequada para apoiar uma recomendação específica, levando em consideração o contexto específico da evidência.

O GRADE tem quatro níveis de evidência, também conhecidos como certeza da evidência ou qualidade da evidência: muito baixo, baixo, moderado e alto Quadro 3.

3.2 Como avaliar a certeza do conjunto final de evidências

A classificação inicial da certeza de evidência é definida a partir do delineamento do estudo (Figura 3). O ECR é o delineamento de estudo mais adequado para questões relacionadas a intervenções e, quando considerado, inicia com uma alta certeza de evidência. Estudos não randomizados estão intrinsicamente associados a viés de confusão, o que pode influenciar nos resultados observados para uma dada intervenção. Assim, há duas possibilidades: a) iniciam com baixa certeza de evidência na avaliação do GRADE, ou; b) caso o instrumento ROBINS-I seja adotado para avaliação do risco de viés do conjunto de estudos avaliado, a certeza da evidência é iniciada como alta; contudo, uma vez que o ROBINS-I aplica métricas mais rigorosas na avaliação da evidência de estudos observacionais (comparando-os a um estudo randomizado teórico), a evidência é rebaixada de 1 a 3 níveis devido a confusão (ou ao menos, confusão residual) (23). Apesar de uma abordagem baseada no uso do ROBINS-I ser mais abrangente, as duas alternativas permanecem válidas e tendem a resultar no mesmo nível de evidência na maioria das situações.

Figura 3 – Identificação do ponto de partida da avaliação da certeza da evidência no sistema GRADE conforme o delineamento do estudo



ROBINS-I = *Risk of Bias in Non-randomized Studies – of Interventions*.
Fonte: elaboração própria.

Há cinco fatores que podem reduzir a certeza da evidência, descritos no Quadro 10. Em contrapartida, caso estejam presentes fatores que aumentem a confiança dos resultados apresentados, há a possibilidade de aumentar o nível de certeza da evidência (Quadro 10). Dessa forma, uma determinada evidência pode ser proveniente de limitações em mais de um dos fatores descritos acima: quanto mais sérias forem essas limitações, menor será a certeza da evidência.

Quadro 10 – Fatores que reduzem ou aumentam a certeza da evidência

Fator	**Consequência**
Fatores que reduzem a qualidade da evidência	
Limitações metodológicas (risco de viés)*	↓ 1 ou 3 níveis
Inconsistência	↓ 1 ou 2 níveis
Evidência indireta	↓ 1 ou 2 níveis
Imprecisão	↓ 1 ou 3 níveis
Viés de publicação	↓ 1 ou 2 níveis
Fatores que aumentam a qualidade da evidência	
Magnitude de efeito elevada	↑ 1 ou 2 níveis
Fatores de confusão residuais que aumentam a confiança da estimativa	↑ 1 nível
Gradiente dose-resposta	↑ 1 nível
* O rebaixamento do risco de viés em até três níveis pode ocorrer pela utilização da ferramenta ROBINS-I em estudos observacionais.	

Fonte: elaboração GRADE Working Group (https://www.gradeworkinggroup.org/).

Geralmente, o Grupo GRADE recomenda não aumentar o nível de certeza da evidência caso ele já tenha sido reduzido anteriormente, especialmente na presença de limitações metodológicas, uma vez que o rebaixamento da certeza da evidência indica que resultados dos estudos devem ser interpretados com cautela. Dessa forma,

os critérios para aumentar a certeza da evidência aplicam-se principalmente a estudos observacionais bem delineados.

Por fim, opiniões de especialista não são formalmente caracterizadas como evidência, devendo-se, preferencialmente, buscar outras fontes de informação, como estudos observacionais não comparados (séries e relatos de casos). Assim, recomendações oriundas de opiniões de especialistas são classificadas como nível de evidência “muito baixo”.

3.3 Domínios que podem reduzir a certeza no conjunto final de evidências

A seguir são discutidos detalhadamente os cinco fatores que podem resultar na redução da certeza da evidência de resultados específicos e, assim, reduzir a confiança da estimativa de efeito.

3.3.1 Risco de viés

- O risco de viés de cada estudo incluído no corpo de evidências de um desfecho deve ser analisado individualmente por meio de ferramentas específicas para cada tipo de delineamento.

- Todas as avaliações individuais devem ser consideradas no julgamento do risco de viés no âmbito do sistema GRADE. O julgamento é específico para cada desfecho de interesse de uma questão PICO (população, intervenção, comparador e desfecho) e deve considerar todo o corpo de evidências identificado para cada desfecho.

- O julgamento do domínio risco de viés no sistema GRADE não deve ser obtido a partir de uma simples média das avaliações de risco de viés dos estudos individuais; deve ser acompanhado de informações explícitas e transparentes, com esclarecimentos sobre as razões que motivaram eventuais penalizações.

Terminologia

As terminologias “limitações dos estudos” e “risco de viés” podem ser utilizadas para fazer referência a diferentes situações e remeter a significados distintos. Um aspecto

básico nesta distinção é que as limitações dos estudos são fatores que podem afetar a validade ou a confiabilidade dos resultados, independentemente da presença de vieses. Já o risco de viés é uma probabilidade de que os resultados sejam enviesados, o que pode ser causado por vários fatores, incluindo as limitações dos estudos. Outra diferença é que as limitações dos estudos podem ser de natureza mais geral, enquanto o risco de viés geralmente se refere a fatores específicos que podem afetar os resultados de um estudo. Desse modo, é importante esclarecer e diferenciar os dois contextos em que a expressão “risco de viés” será utilizada.

- Risco de viés de estudos individuais: refere-se à avaliação das limitações e da qualidade metodológica de cada um dos estudos incluídos em uma RS. Em geral, é realizada por meio do uso de ferramentas de avaliação do risco de viés, que são específicas para cada tipo de delineamento.

- Risco de viés do corpo de evidências: refere-se à avaliação do domínio risco de viés no sistema GRADE, um dos cinco domínios que podem reduzir a confiança nos resultados encontrados para uma determinada questão PICO. Nesse sentido, essa avaliação é mais ampla e, para realizá-la, consideram-se todas as avaliações de risco de viés dos estudos individuais que contribuem diretamente para os resultados do desfecho de interesse.

Avaliação

Limitações no delineamento e na execução dos estudos podem enviesar as estimativas de efeito de um tratamento e levar a conclusões equivocadas. A confiança dos resultados, da estimativa de efeito e das recomendações subsequentes diminui se os estudos apresentarem limitações metodológicas consideráveis, ou seja, limitações que podem influenciar nos resultados da intervenção de interesse. Sendo assim, esse domínio do sistema GRADE avalia a presença dessas limitações ou de vieses no corpo de evidências considerado para cada desfecho de interesse de uma questão PICO. Quanto mais graves forem as limitações, maior será a probabilidade de a qualidade da evidência ser reduzida.

As limitações ou os vieses dos estudos clínicos podem variar de acordo com o contexto ou tipo de estudo. É recomendado que ferramentas apropriadas para avaliação do risco de viés sejam utilizadas *a priori*, as quais devem ser escolhidas de acordo com o delineamento de cada estudo incluído na revisão (Quadro 11). Somente

após o julgamento individual com as ferramentas de risco de viés, é possível avaliar a certeza da evidência no âmbito do domínio risco de viés do sistema GRADE (Figura 4).

Quadro 11 – Ferramentas para avaliação do risco de viés ou das limitações dos estudos de acordo com o delineamento

É importante que cada estudo seja avaliado de acordo com a ferramenta apropriada para o seu delineamento. De acordo com o Manual de Revisões Sistemáticas da Cochrane, a ferramenta *Risk of Bias 2.0* (RoB 2.0) deve ser utilizada para avaliar o risco de viés de estudos randomizados e a ferramenta *Risk of Bias in Non-randomized Studies of Interventions* (ROBINS-I) deve ser utilizada para avaliar o risco de viés de estudos comparativos não randomizados que tenham por objetivo avaliar a eficácia e a segurança de alguma intervenção.

Algumas ferramentas disponíveis na literatura estão elencadas abaixo.

- Ensaios clínicos randomizados: RoB 2.0,[1] escala de Jadad, *JBI Checklist for Randomized Controlled Trials*.
- Ensaios clínicos não randomizados: ROBINS-I[1] (24).
- Estudos de coorte: escala de Newcastle-Ottawa (NOS) para estudos de coorte; *JBI Checklist for Cohort Studies* (25, 26); ROBINS-E.
- Estudos caso-controle: NOS para estudos caso-controle; *JBI Checklist for Case Control Studies* (25, 27), ROBINS-E.
- Estudos transversais: *JBI Checklist for Analytical Cross Sectional Studies* (28).
- Séries de caso: *JBI Checklist for Case Series* (29).
- Relatos de caso: *JBI Checklist for Case Reports* (30).
- Estudos de prevalência: *JBI Checklist for Prevalence Studies* (31).
- Estudos de acurácia diagnóstica: *Quality Assessment of Diagnostic Accuracy Studies 2* (QUADAS-2); *JBI Checklist for Diagnostic Test Accuracy Studies* (32, 33).
- Revisões sistemáticas: *A Measurement Tool to Assess Systematic Reviews 2* (AMSTAR 2); ROBIS; *JBI Checklist for Systematic Reviews* (34-36).

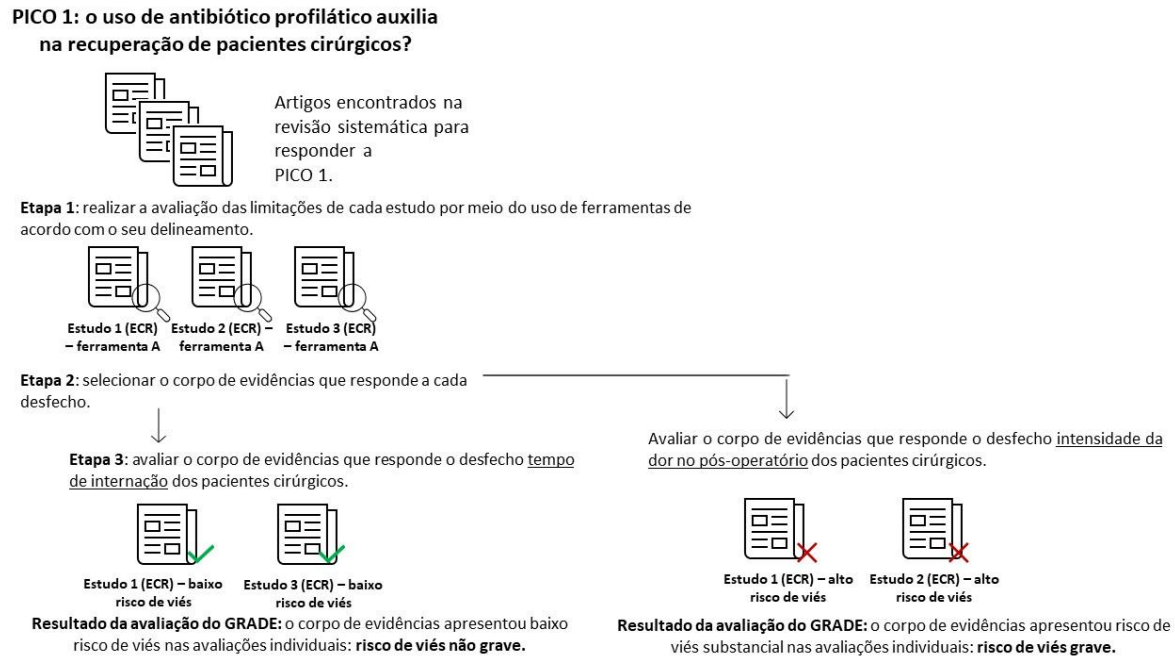
[1] Ferramentas baseadas em domínios cuja avaliação deve ser realizada para cada um dos desfechos de interesse do estudo.

É importante salientar que essas ferramentas são apenas alguns exemplos dos materiais disponíveis na literatura e podem ser atualizadas ou modificadas a qualquer momento.

Fonte: elaboração própria.

Mais detalhes sobre os possíveis vieses a serem encontrados em ECR e ensaios clínicos não randomizados são abordados ao longo deste capítulo.

Figura 4 – Exemplo ilustrando as etapas de avaliação do domínio risco de viés no sistema GRADE



ECR = ensaio clínico randomizado; PICO = população, intervenção, comparador e desfecho.
Fonte: elaboração própria.

De modo geral, a primeira etapa da avaliação do domínio risco viés é identificar os delineamentos dos estudos incluídos na revisão e usar ferramentas apropriadas para avaliar o risco de viés de cada estudo individual. Posteriormente, o julgamento do domínio risco de viés no sistema GRADE é realizado a partir de uma análise cuidadosa das limitações do corpo de evidências de cada desfecho. Para isso, é necessário considerar em qual grau as limitações dos estudos avaliados podem enviesar os resultados do desfecho. Se houver muitas limitações ou limitações graves, a certeza da evidência pode ser reduzida em um ou até dois níveis nesse domínio, partindo de uma avaliação “não grave” para “grave” ou “muito grave”.

Alguns princípios podem ser considerados na avaliação do corpo de evidências de cada desfecho no GRADE:

- uma avaliação criteriosa deve considerar o quanto cada estudo contribui para a estimativa de efeito. Essa contribuição é representada pelo peso do estudo

em uma metanálise, sendo usualmente um reflexo do tamanho da amostra e da incidência do desfecho (número de eventos). Estudos com tamanho de amostra e número de eventos maiores, em geral, podem contribuir mais para a avaliação geral do risco de viés nos respectivos desfechos;

- não é recomendado definir a avaliação geral do domínio risco de viés no sistema GRADE a partir da média das avaliações individuais de cada estudo. Deve-se realizar uma avaliação criteriosa da contribuição de cada estudo e, como guia geral, concentrar-se nos estudos de melhor qualidade;
- é necessário ser conservador na penalização e redução da certeza da evidência. Para reduzir a certeza da evidência, deve-se ter confiança de que há risco de viés substancial no corpo de evidências de cada desfecho;
- em caso de dúvidas e incertezas, deve-se reconhecer a dificuldade da situação e esclarecer detalhadamente os motivos das decisões tomadas para os leitores da revisão ou da diretriz (para mais informações, ver o capítulo 4. Síntese de evidências);
- por fim, todo julgamento da certeza da evidência no sistema GRADE deve ser acompanhado de informações detalhadas e transparentes que esclareçam as razões que motivaram eventuais penalizações (para mais informações, ver o capítulo 4. Síntese de evidências).

A Quadro 12 sumariza as etapas para realizar a avaliação do domínio risco de viés no âmbito do sistema GRADE a partir da avaliação de estudos individuais.

Quadro 12 – Julgamento do domínio risco de viés no sistema GRADE		
Risco de viés intra-estudo (estudos individuais)	**Risco de viés inter-estudos (corpo de evidências)**	**Julgamento do risco de viés**
Baixo risco de viés para todos os critérios-chave.	A maior parte das informações/estimativas para um determinado desfecho advém de estudos com **baixo risco de viés.**	Sem limitações graves – **não penalizar.**
Limitações críticas para um critério ou algumas limitações para múltiplos critérios de avaliação.	A maior parte das informações/estimativas advém de estudos com risco de **viés moderado ou com algumas preocupações.**	Com limitações graves: rebaixar um nível – **grave.**
Limitações críticas para um ou mais critérios, suficiente para reduzir substancialmente a confiança nos resultados.	A maior parte das informações/estimativas advém de estudos com **alto risco de viés.**	Com limitações muito graves: rebaixar dois níveis – **muito grave.**

Fonte: adaptado de Guyatt et al. (37).

É importante observar que a Quadro 12 ilustra a avaliação do domínio risco de viés a partir do pressuposto de que os estudos iniciam com certeza da evidência ALTA. Esse pode ou não ser o ponto de partida, dependendo de algumas questões. A Figura 3 demonstra os fatores que determinam o ponto de partida de uma avaliação no âmbito do sistema GRADE.

Os principais vieses que devem ser considerados no âmbito do sistema GRADE para estudos randomizados (falta de sigilo de alocação, falta de

cegamento/mascaramento, avaliação incompleta de pacientes e desfechos, relato seletivo de desfechos e uso de medidas de desfecho inadequadas para ensaios randomizados) e não randomizados (confusão, viés na seleção de participantes do estudo, viés na classificação das intervenções, viés por desvio das intervenções pretendidas, viés por dados ausentes, viés na mensuração dos desfechos e viés na seleção dos desfechos reportados para estudos não randomizados). Todos estão descritos a seguir.

Principais riscos de viés em ECR

Falta de sigilo de alocação

Nos ensaios randomizados, a falta de sigilo de alocação ocorre quando os pesquisadores que estão recrutando pacientes para participar do estudo sabem ou podem prever para qual grupo o próximo paciente será alocado. Ou seja, a sequência aleatória gerada não é sigilosa ou é previsível (por exemplo, alocação pelo dia da semana, data de nascimento, número do prontuário). A falta de sigilo de alocação permite a quebra do processo de randomização, elemento essencial dos ECR.

Falta de cegamento (mascaramento)

A falta de cegamento ou mascaramento ocorre quando os pacientes, cuidadores, pesquisadores que selecionam e/ou aferem o desfecho e os analistas de dados sabem em qual braço do estudo os pacientes foram alocados (ou a possibilidade de identificar qual é o medicamento que é administrado em cada uma das etapas de um estudo cruzado). O conhecimento dessa informação pode afetar a estimativa de efeito encontrada pelo estudo, pois os pacientes podem se sentir mais motivados ou não a realizar o tratamento. Ainda, pesquisadores ou pacientes que aferem o desfecho podem ser mais meticulosos ou não em relação a certos cuidados prestados, entre outros potenciais impactos. A ausência de cegamento pode ou não influenciar o risco de viés de um estudo ou do corpo de evidências no GRADE, dependendo principalmente do tipo de desfecho sendo avaliado. Desfechos com componentes subjetivos, como aplicação de escalas, redação de laudos ou interpretação de exames, são mais suscetíveis a variações por falta de cegamento do que desfechos objetivos, como óbito, por exemplo.

Avaliação incompleta de pacientes e desfechos

A avaliação incompleta de pacientes e desfechos ocorre quando há perda de seguimento e falta de aderência ao princípio de intenção de tratar em estudos de superioridade; ou perda de seguimento e falha na condução de ambas as análises, considerando apenas aqueles que aderiram ao tratamento e para os quais havia dados disponíveis, em estudos de não inferioridade.

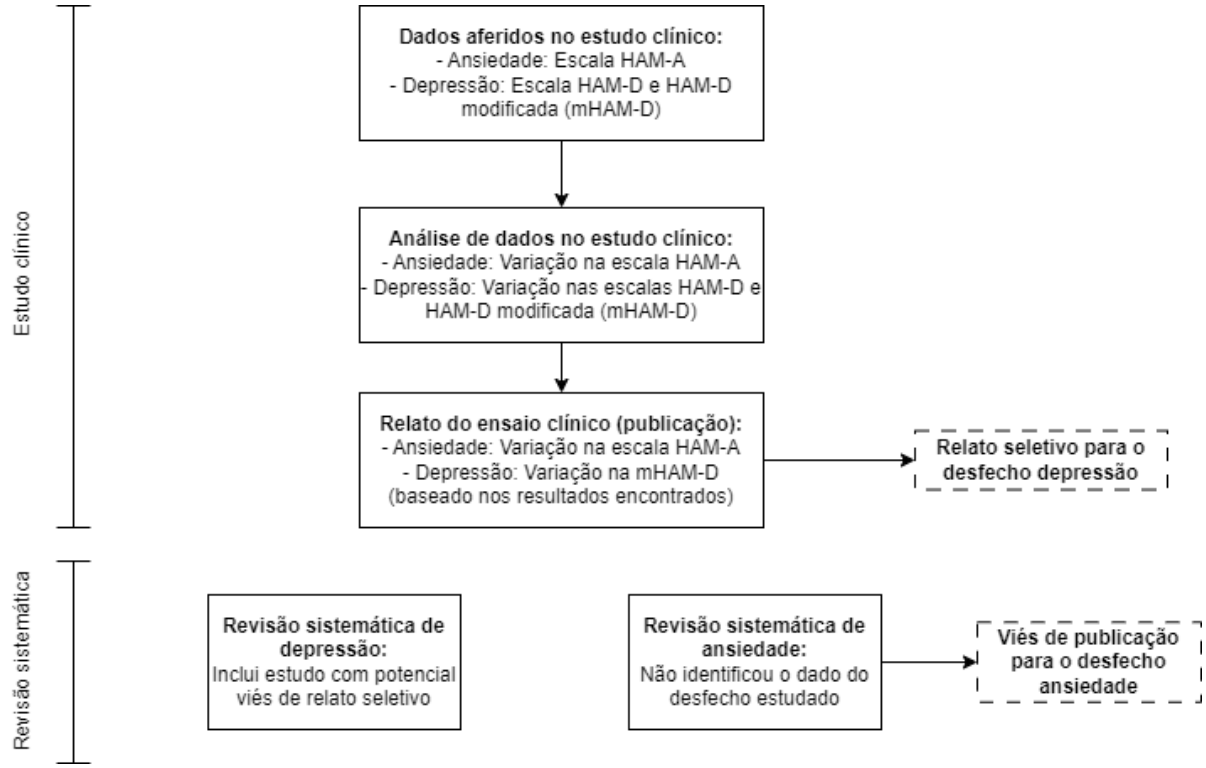
No entanto, a importância de taxas específicas de perda de seguimento varia amplamente e depende da relação entre a perda de seguimento e o número de eventos. Em geral, quanto maior for a proporção de perda de seguimento em relação ao número de eventos nos grupos controle e intervenção e quanto maior for a diferença de perdas entre esses grupos, maior será o risco de viés relacionado a dados ausentes. Por exemplo, uma perda de seguimento de 5% em ambos os grupos, com taxas de evento de 20% no grupo intervenção e 40% no grupo controle, dificilmente resultaria em alto risco de viés; entretanto, se as taxas de evento girassem em torno de 2% e 4%, respectivamente, perdas de 5% em cada grupo poderiam comprometer consideravelmente os resultados do estudo.

É importante observar que dados ausentes podem ou não introduzir viés nas análises de um estudo. Em geral, alto risco de viés é considerado quando a ausência do dado está relacionada ao seu valor real dentro de cada grupo de intervenção do estudo e difere entre os grupos.

Relato seletivo de desfechos

O relato seletivo de desfechos ocorre se houver relato incompleto ou ausente de alguns desfechos de acordo com os resultados encontrados. É importante diferenciar essa categoria do risco de viés por perda de seguimento/dados ausentes. Também é comum a confusão entre relato seletivo de desfechos e viés de publicação, contudo consistem em vieses distintos. Enquanto no viés de publicação, não há dado sobre o desfecho de interesse, em relato seletivo de desfecho, os pesquisadores selecionam intencionalmente resultados que favorecem sua intervenção ou hipótese. Um exemplo é escolher relatar a eficácia de um tratamento com base na escala que apresenta os resultados mais favoráveis. Na Figura 5, exemplificamos a ocorrência de relato seletivo de desfechos em contraste com viés de publicação.

Figura 5 – Exemplos de viés de relato seletivo de desfechos e viés de publicação.



Fonte: Adaptado de Higgins et al. (38)

Para que haja evidência de que não ocorreu relato seletivo de desfechos, é necessário ter acesso ao protocolo desenvolvido antes da condução do estudo; no caso de ensaios clínicos randomizados, uma boa prática é avaliar se os desfechos apresentados nas publicações são os mesmos descritos no registro conforme a base do *clinicaltrials.gov.*

Uso de medidas de desfecho inadequadas

O uso de medidas de desfecho inadequadas ocorre quando o método para mensuração do desfecho é inapropriado ou quando a mensuração do desfecho pode ter ocorrido de maneira diferente entre os grupos de intervenção de um estudo.

Outros fatores também podem ser considerados em estudos com delineamentos mais específicos, como efeitos de transporte (*carryover effect*) em ensaios clínicos cruzados e viés de recrutamento em estudos randomizados em *cluster*. Por fim, é importante identificar se houve interrupção precoce do estudo por

benefício, fator que pode resultar em uma superestimação substancial da estimativa de efeito. O risco de superestimação é maior em estudos com pequeno número de eventos (menos de 500 eventos, porém ainda mais preocupante em estudos com menos de 200 eventos). Evidências empíricas sugerem que critérios formais para interrupção não reduzem esse viés.

Principais riscos de viés em EC não randomizado

Viés de confusão

Nos ensaios não randomizados, o viés de confusão ocorre quando não há mensuração ou há mensuração inadequada dos principais fatores prognósticos que influenciam a propensão de um indivíduo receber ou não a intervenção de interesse. Entre os exemplos mais comuns, estão a gravidade da doença, a presença de comorbidades, o nível de acesso e utilização dos serviços de saúde e a condição socioeconômica. Os fatores relevantes podem variar de acordo com a condição clínica em questão. Ainda, mesmo que mensurados, esses fatores podem introduzir vieses quando não são utilizados adequadamente na seleção ou na análise do estudo (24) .

Existem diferentes tipos de confusão que podem ser introduzidos em estudos não randomizados, como confusão devido a diferenças basais entre os grupos, confusão tempo-dependente e confusão residual. É importante que os revisores considerem todos os tipos durante a avaliação de risco de viés (39).

Viés na seleção de participantes do estudo

O viés na seleção de participantes do estudo ocorre quando estes são escolhidos de modo a excluir indivíduos com base em características, tempo de seguimento ou desfechos que se relacionam tanto com a intervenção quanto com o desfecho de interesse do estudo, podendo introduzir viés. Em estudos de coorte, por exemplo, a seleção de coortes diferentes de pacientes que receberam ou não uma intervenção constitui um viés de seleção. Em estudos caso-controle, um possível exemplo é o pareamento exagerado ou insuficiente entre os dois grupos de participantes. É importante observar que o viés de seleção independe da validade externa do estudo, como possibilidade de generalização dos resultados, aplicabilidade ou transferibilidade para outras populações (39).

Viés na classificação das intervenções

O viés na classificação das intervenções ocorre quando há erro de classificação diferencial ou não diferencial das intervenções. Erro não diferencial significa que o erro não está associado ao desfecho e, usualmente, ocasiona uma estimativa de efeito enviesada na direção da nulidade. Já o erro de classificação diferencial está relacionado ao desfecho ou ao risco de desfecho, sendo o tipo que mais pode enviesar os achados de um estudo (24). Um exemplo é o viés de recordação em estudos do tipo caso-controle, em que o conhecimento do *status* de caso ou controle pode afetar a lembrança da intervenção anterior ou dos desfechos de interesse.

Viés por desvio das intervenções pretendidas

O viés por desvio das intervenções pretendidas ocorre quando há diferenças sistemáticas entre os braços experimental e controle em termos de cuidado dispensado aos participantes, podendo representar desvio da intervenção pretendida. Um exemplo é o uso de co-intervenções pelos participantes de forma desbalanceada entre os grupos (39). Avaliar o desfecho neste item da ferramenta depende do efeito de interesse: efeito de ser designado a uma intervenção ou efeito de iniciar e aderir a uma intervenção (24).

Viés por dados ausentes

O viés por dados ausentes ocorre quando o seguimento dos participantes incluídos no estudo não é concluído, acarretando ausência de dados. A perda de seguimento é um problema maior quando há indícios de variações de acordo com o grupo alocado ou com fatores prognósticos basais. Além disso, outra prática que resulta em viés por dados ausentes é a exclusão de participantes das análises devido à ausência de dados sobre o desfecho ou variáveis confundidores (24).

Viés na mensuração dos desfechos

O viés na mensuração dos desfechos ocorre quando há erros diferenciais ou não diferenciais na mensuração dos desfechos de interesse. Esse tipo de viés pode ser introduzido quando avaliadores ou pacientes estão cientes do *status* da intervenção, quando métodos diferentes de avaliação de desfecho são utilizados de acordo com o grupo de participantes ou quando há erros de medida possivelmente relacionados ao *status* da intervenção e seus efeitos (24).

Viés na seleção dos desfechos relatados

O viés na seleção dos desfechos relatados ocorre quando a seleção dos resultados a serem relatados é feita de acordo com os achados do estudo, de modo a impedir que a estimativa seja combinada em metanálises ou a priorizar desfechos que tiveram resultados mais favoráveis (24).

Considerações adicionais

Como considerações adicionais sobre o domínio risco de viés no sistema GRADE, é importante reiterar que a avaliação de risco de viés deve ser feita por meio de ferramentas estruturadas e validadas, considerando os critérios nelas elencados e, preferencialmente, as particularidades de cada desfecho de interesse. Somente após a avaliação com uso da ferramenta escolhida e apropriada, conforme um julgamento no âmbito do sistema GRADE deve ser realizado, considerando todo o corpo de evidências reunido para cada desfecho.

Os vieses explorados até aqui não representam necessariamente a totalidade dos aspectos avaliados pelas ferramentas ou dos fatores a serem considerados em cada avaliação. Além disso, o peso de cada critério ou da avaliação individual de cada estudo é incerto. Não há um escore quantitativo para avaliação do domínio risco de viés no sistema GRADE. Os autores devem ponderar a influência de cada estudo e considerá-la na sua avaliação. Certo grau de subjetividade e variação é esperado, desde que acompanhado de explicações transparentes para o julgamento realizado.

Outro aspecto importante é que o risco de viés deve ser considerado no contexto de outras limitações avaliadas pelo GRADE. Se, por exemplo, houver dúvida em relação a dois problemas de qualidade, um relacionado ao risco de viés e outro à precisão, sugere-se penalizar pelo menos um dos dois domínios do sistema GRADE.

Por fim, é necessário esclarecer que, em situações nas quais as evidências disponíveis são oriundas de um único ECR, não é adequado reduzir a certeza da evidência apenas por esse motivo. É possível que um único ECR robusto, de tamanho amostral adequado e bem conduzido forneça evidências de alta qualidade.

3.3.2. Inconsistência

- 'Inconsistência refere-se a uma heterogeneidade estatística não explicada entre os resultados do estudos que contribuem com a estimativa do efeito da intervenção que serão utilizados na tabela GRADE'O julgamento da

inconsistência é baseado na similaridade entre as estimativas de efeito, na sobreposição dos intervalos de confiança (IC) e em critérios estatísticos, como I^2; contudo, parâmetros estatísticos não devem ser utilizados isoladamente na avaliação da inconsistência.

- O domínio de inconsistência pode depender do alvo da avaliação da certeza da evidência, sendo impactado pela avaliação em relação a efeitos nulos ou ao limiar de diferença minimamente importante (do inglês *minimal important difference* - MID).
- É recomendada a utilização de questões de pesquisa amplas para diretrizes clínicas e revisões sistemáticas, explorando os resultados inconsistentes entre os estudos através de um pequeno número de hipóteses geradas a priori para possíveis explicações (incluindo a direção do efeito).
- O sistema GRADE sugere reduzir a certeza de evidência caso haja inconsistência importante nos resultados, mesmo nos casos em que o estudo desenvolveu análises de sensibilidade *a priori* em paralelo.

Avaliação

A avaliação de inconsistência no sistema GRADE é um critério que pode ser considerado no rebaixamento do nível de certeza, tendo suas recomendações iniciais publicadas em 2011 com predominante foco sobre desfechos binários. Estudos envolvendo uma mesma questão de pesquisa podem, usualmente, apresentar diferenças clínicas e metodológicas na sua execução (populações incluídas, tempo de seguimento, entre outros). Como consequência, pode haver variabilidade do efeito nos resultados dos estudos de forma individual, e a medida sumária de uma metanálise pode não representar adequadamente as especificidades do conjunto dos resultados. Os avaliadores devem explorar as possíveis explicações para a heterogeneidade e, nos casos em que não for identificada uma explicação plausível, a classificação da certeza da evidência deve ser reduzida. A escolha sobre reduzir em um ou dois níveis depende do julgamento quanto à magnitude da inconsistência dos resultados (40). A certeza da evidência pode ser reduzida por inconsistência, mas não deve ser elevada por ser muito consistente (41). Em 2023, o sistema GRADE divulgou atualizações no processo de avaliação de inconsistência, abordando questões gerais sobre aspectos do processo de avaliação e recomendações sobre o

processo de identificação da credibilidade de subgrupos para a análises de sensibilidade (42).

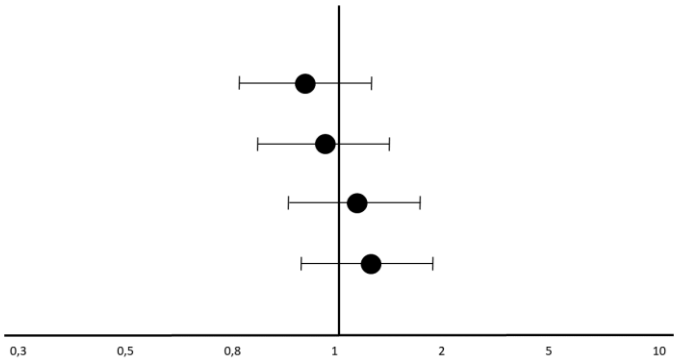
Embora existam métodos estatísticos para mensurar a heterogeneidade, há quatro critérios que devem ser utilizados para avaliar se o nível de certeza de uma evidência deve ser reduzido (40):

1. Grande variação nas estimativas pontuais dos estudos, sendo que a direção dos efeitos não deve ser utilizada como um critério isolado para inconsistência (Quadro 13);
2. Sobreposição mínima ou ausência de sobreposição dos IC, sugerindo que a variação é maior do que se esperaria ao acaso;
3. Testes de heterogeneidade com um valor abaixo do valor de p, verificando se todos os estudos apresentam a mesma magnitude de efeito;
4. Estatística I^2 com valores elevados (Quadro 14).

Quadro 13 – O impacto da direção do efeito nas decisões sobre inconsistência

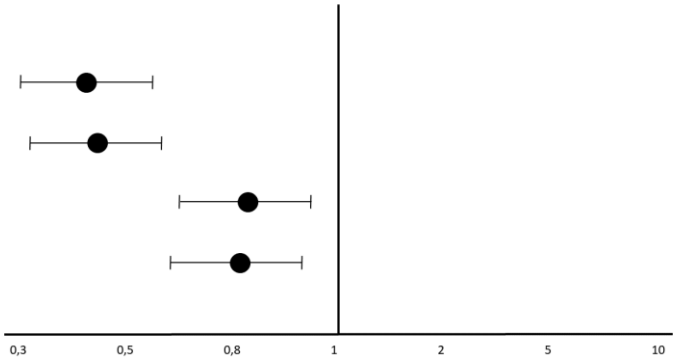
Na Figura 6, o nível de certeza da evidência não deve ser reduzido por inconsistência, pois diferenças na direção não constituem um critério para variabilidade caso a magnitude das diferenças entre as estimativas pontuais seja pequena.

Figura 6 – Diferenças na direção, mas heterogeneidade mínima



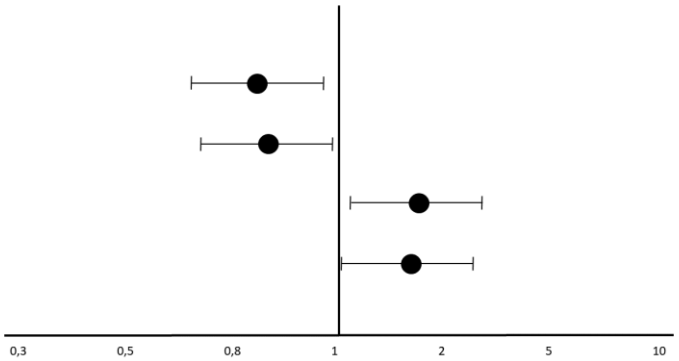
Uma inconsistência é considerada importante quando ela reduz a certeza do resultado em relação a uma decisão particular. Veja que, no exemplo da Figura 7, há uma variabilidade substancial, mas as diferenças envolvem a magnitude e não a direção do efeito do tratamento. Nesse cenário, o autor pode ou não considerar essas diferenças como algo importante para rebaixar a certeza da evidência. Além disso, questões como o tipo de desfecho (contínuo ou binário), a medida de efeito adotada (absoluta ou relativa) e características do estudo (como o tamanho amostral, e relevância clínica dos desfechos) também impactam no IC95% e, consequentemente, na interpretação da inconsistência.

Figura 7 – Heterogeneidade substancial, mas importância questionável



A Figura 8 apresenta um cenário no qual a magnitude da variabilidade dos resultados é idêntica ao da Figura 7. Entretanto, alguns dos estudos sugerem benefício e outros sugerem malefício, indicando a necessidade de rebaixar a certeza da evidência.

Figura 8 – Heterogeneidade e importância substanciais



Fonte: adaptado de Guyatt et al. (41).

Quadro 14 – Possibilidade de classificação da estatística I^2

O valor de I^2 quantifica a proporção de variação das estimativas pontuais que ocorreu devido a diferenças entre os estudos (41). Embora a determinação do que seria um valor de I^2 alto seja subjetiva, pode-se utilizar a seguinte classificação:

< 40%: possivelmente baixo

30-60%: possivelmente moderado

50-90%: possivelmente substancial

75-100%: possivelmente elevado

As sobreposições nos intervalos classificatórios e o uso do termo “possivelmente” demonstram a incerteza que envolve esse julgamento, sendo importante destacar as limitações dessa estatística (43): a) quando a amostra de cada estudo for pequena, as estimativas de pontos podem variar substancialmente, por questões aleatórias, e o valor de I^2 pode ser baixo; b) por outro lado, quando o tamanho das amostras for grande, uma pequena diferença na estimativa de ponto pode gerar um valor de I^2 alto. Ainda, a estatística τ^2 (*tau square*) é uma medida de variabilidade que apresenta a vantagem de não ser dependente do tamanho da amostra.

Lembrete: todas as abordagens estatísticas apresentam limitações, e os seus resultados devem ser interpretados em conjunto com uma análise subjetiva das estimativas pontuais e das sobreposições dos intervalos de confiança.

Fonte: adaptado de Guyatt et al. (41).

Na presença de inconsistência durante a realização de uma metanálise, devem ser realizadas análises de sensibilidade (análises de subgrupo e/ou metarregressão) para identificar os fatores clínicos e metodológicos responsáveis por essas diferenças. Recomenda-se que a metanálise inclua testes formais para identificar se as hipóteses *a priori* explicam as possíveis inconsistências entre subgrupos, mesmo quando a variabilidade existente possa ser explicada ao acaso (40). Se o tamanho do

efeito diferir entre os estudos, as explicações sobre a inconsistência possivelmente se devem aos seguintes itens (40):

- população (por exemplo, as intervenções apresentam maior efeito relativo na população com maior carga de doença);
- intervenção (por exemplo, um maior efeito, ao qual esteja associado doses mais elevadas do medicamento);
- desfechos (por exemplo, tempo de duração do seguimento do estudo);
- metodologia do estudo (por exemplo, ECR com altos ou baixos riscos de viés).

Desta forma, variações no delineamento do estudo, seja por critérios de elegibilidade da população, diferentes estratégias de um mesmo tratamento (doses ou frequência, por exemplo), comparadores e até desfechos (por exemplo, duração do acompanhamento), caracterizam a variabilidade na questão PICO que pode impactar diretamente na avaliação de inconsistência. O grupo GRADE considera que o domínio de inconsistência aborda, sobretudo, a variabilidade nos resultados, e não necessariamente no desenho do estudo. Desta forma, a variabilidade de resultados (inconsistência) resultante da variabilidade do desenho do estudo é considerada uma oportunidade para explorar os seus motivos (42).

Se a RS indicar que os resultados são similares entre os estudos avaliados, a variabilidade na PICO do estudo aumenta a generalização destes resultados. Por outro lado, se os resultados diferem substancialmente entre os estudos incluídos na RS, diferenças no delineamento do estudo (questão PICO) oferecem uma oportunidade a ser explorada. Se uma inconsistência puder ser explicada por diferenças na população, na intervenção ou nos desfechos, os autores da metanálise podem oferecer diferentes estimativas entre os grupos de participantes, intervenções ou desfechos. Painelistas de diretrizes estariam, então, propensos a apresentar diferentes recomendações para diferentes grupos de pacientes e intervenções. Se uma alta variabilidade para o tamanho do efeito permanecer inexplicada, o nível de certeza da evidência deve ser rebaixado, considerando que não foi possível realizar uma análise estratificada. Tanto autores de metanálises quanto painelistas de diretrizes devem considerar até que ponto estão incertos sobre o efeito, considerando a inconsistência, e o quão relevantes essas inconsistências são para a confiança no resultado (40).

O alvo da certeza da evidência

Em 2017, o grupo GRADE discutiu a necessidade de avaliar o impacto da estimativa pontual em relação a um determinado limiar (ou em um intervalo deles) para determinar a certeza da evidência, em especial no contexto de tomada de decisão (diretrizes e avaliação de tecnologias em saúde) (44). O estabelecimento deste alvo requer a contextualização da avaliação, proporcionando limiares de efeito nulo, MID ou uma rede de limiares (que podem ter intervalos de efeitos triviais, pequenos, moderados ou grandes).

A partir das recomendações publicadas em 2023, a metodologia GRADE indicou que a adoção destes limiares como alvo da certeza da evidência possui implicações diretas para o domínio de inconsistência. Os resultados podem diferir entre os estudos, mas se todas as estimativas pontuais estiverem acima do limiar adotado, o rebaixamento por inconsistência será inapropriado. Por outro lado, recomenda-se o rebaixamento por inconsistência caso o mesmo grau de inconsistência seja observado (isto é, os resultados diferindo entre os estudos) mas as estimativas pontuais estando substancialmente divididas entre ambos os lados do limiar adotado. Desta forma, a contextualização da análise a partir da adoção de um limiar de efeito acarretará em modificações no grau de inconsistência avaliada, sendo condicionada, em especial, se os resultados são consistentes ou não para suportar dada recomendação (42).

Contextualização do I^2

Adicionalmente, a metodologia GRADE atualizou direcionamentos sobre a adoção da avaliação estatística pelo I^2 para avaliação de heterogeneidade. Como mencionado anteriormente, estudos pequenos podem apresentar uma ampla variação nas estimativas pontuais, mas ainda assim os seus intervalos de confiança estarem sobrepostos de modo a gerar um baixo I^2. Por outro lado, estudos com grande tamanho amostral, que usualmente possuem importante e considerável peso na síntese de evidências, podem apresentar intervalos de confiança estreitos de modo a ter pouca sobreposição e um alto I^2. Assim, é importante destacar que a medida de I^2 representa uma medida estatística relativa de heterogeneidade, de modo a questionar qual a proporção da variabilidade entre as estimativas é atribuível a

diferenças genuínas em oposição à variabilidade atribuível a erro de amostragem (45).

De fato, dissonâncias entre o I^2 alto e intervalos de confiança estreitos são preocupações abordadas também para a avaliação da certeza da evidência em metanálises de proporções, como as que avaliam prognóstico, incidência e prevalência, conforme detalhado no capítulo 7. Sistema GRADE para prognóstico, incidência e prevalência (46). De fato, no contexto de estudos em que resultados consistentes podem, no entanto, ser associados com um elevado I^2 , o GRADE sugere ignorar critérios baseados em estatísticas e adotar a inspeção visual de similaridade no ponto estimativas (47, 48).

Metanálises de desfechos contínuos também necessitam de contextualização na utilização do I^2, uma vez que o intervalo de confiança geralmente é mais estreito quando comparado a estudos envolvendo desfechos binários de modo a gerar um alto valor de I^2 mesmo quando os resultados são similares entre os estudos. Particularmente, este evento é observado em metanálises de desfechos contínuos devido a utilização de medidas de efeito absolutas, como diferença entre médias, mais suscetíveis a variações basais do que medidas de efeito relativos de desfechos binários. Mais uma vez, a utilização do I^2 pode induzir em erro: a interpretação de medidas estatísticas para avaliação de inconsistência deve considerar o contexto. Portanto, um alto valor de I^2 não justificaria necessariamente o rebaixamento por inconsistência ao avaliar a certeza em uma estimativa de efeito não-nula (48).

Decidindo sobre quando utilizar análises de subgrupos

Em relação à decisão de quando utilizar análises de subgrupos, o grupo GRADE inicialmente indicou que, sempre que possível, deve-se apresentar uma explicação para a inconsistência (41). A explicação pode ser baseada em diferenças na população, na intervenção ou nos desfechos, relatando duas ou mais estimativas de efeito e gerando recomendações de acordo com os subgrupos (40). Entretanto, é necessário atentar para a possibilidade de associações espúrias, que não explicam a variabilidade existente para toda a extensão da inconsistência (Quadro 15) (49).

Quadro 15 – Análises de subgrupos e suas apresentações

Autores de revisões e elaboradores de diretrizes devem exercitar o ceticismo diante de explicações baseadas em efeitos de subgrupos. Devem atentar aos critérios a seguir para julgar se a análise atingirá critérios suficientes para ser considerada convincente.

1. A característica que determina os subgrupos foi especificada entre as variáveis que seriam analisadas na linha de base ou após a randomização? (As hipóteses de subgrupo devem ser desenvolvidas *a priori*.)
2. A diferença entre os subgrupos surgiu a partir de comparações intra ou entre os estudos?
3. As estatísticas sugerem que o acaso é uma explicação improvável para a diferença entre os subgrupos?
4. A hipótese para a análise do subgrupo se deu *a priori* e incluiu a direção do resultado confirmada posteriormente?
5. A hipótese do subgrupo foi testada em uma amostra muito menor?
6. As diferenças encontradas nas análises de subgrupos foram consistentes entre os estudos e com os resultados principais?
7. As evidências externas (racional biológico ou sociológico) corroboram a hipótese de diferença entre os subgrupos?

A credibilidade dos efeitos de subgrupos deve ser discutida além de “sim” ou “não”, com análises individualizadas, sintetizando as evidências de forma a corroborar com as demais análises.

Fonte: adaptado de Sun et al. (49).

As recomendações iniciais do sistema GRADE sugerem que os autores devem sempre considerar a possibilidade de inconsistência (heterogeneidade) nos resultados com base em diferenças no delineamento dos estudos e criarem hipóteses a priori que possam explicar esta heterogeneidade. É recomendado que os pesquisadores testem as hipóteses identificadas para análise de subgrupos de modo independente ao grau de heterogeneidade. De fato, em uma RS pode ser necessário avaliar o quanto um efeito modificador, como a idade, severidade da doença ou ainda ano de publicação dos estudos, pode impactar no efeito estimado da intervenção de interesse. No entanto, a utilização de diversos subgrupos para explicar inconsistência

pode levar a avaliações espúrias ou ainda enganosas, podendo comprometer o cuidado clínico e as condutas terapêuticas para o cuidado em saúde. Desta forma, um grupo de investigadores produziu uma ferramenta para avaliar a credibilidade das análises de subgrupo, identificando o quanto um efeito de subgrupo verdadeiro existe. A ferramenta *Instrument for assessing the Credibility of Effect Modification Analyses* (ICEMAN) foi desenvolvida para identificar a plausibilidade da credibilidade de possíveis efeitos modificadores do tratamento em ECRs e RS de intervenção (50). O Quadro 16 descreve os critérios elencados no ICEMAN para avaliação da credibilidade de subgrupos em RS. O ICEMAN é um questionário contendo questões chaves, opções de respostas e uma avaliação geral da credibilidade. Apesar de seu uso no sistema GRADE ser opcional, seu uso pode ser útil nesse propósito, facilitando a avaliação da inconsistência.

Quadro 16 – Instrumento ICEMAN para avaliação da credibilidade de subgrupos em revisões sistemáticas

Instruções rápidas
Sinônimos de efeito modificador incluem efeito de subgrupos, interações, e moderação;
O instrumento é aplicável para um efeito modificador proposto no período; complete um questionário para cada desfecho, ponto de tempo, medida de efeito, e efeito modificador;
As opções de resposta a esquerda indicam definitivamente ou provavelmente reduzem, enquanto as opções de resposta a direita indicam definitivamente ou provavelmente aumentam a credibilidade;
Incertezas provavelmente reduzem a credibilidade;
Realizar comentários ou indicações auxilia a interpretação de cada questão;
O quanto um efeito modificador é importante para o paciente não é parte da avaliação de credibilidade;
O manual possibilita instruções mais detalhadas e exemplos.
Considerações preliminares
Referência do estudo:
Se disponível, referência do protocolo:
Indique o desfecho de interesse e, se aplicável, o período de tempo:
Indique a medida de efeito do desfecho de interesse (diferença relativa ou absoluta):
Indique o potencial efeito modificador de interesse:
O efeito modificador em potencial foi avaliado antes da randomização? [ ] sim, continue [ ] não, pare aqui e leia o manual de instruções (https://www.iceman.help/overview).

1. A análise da modificação do efeito é baseada na comparação dentro e não entre os ensaios?

[ ] Completamente entre Análises de subgrupos ou metaregressões comparando efeitos gerais de cada estudo individual. Isto é feito tipicamente para agregar dados de metanálise.	[ ] A maioria entre ou incerto Análise de subgrupo ou metarregressão com mais informações provenientes de efeitos gerais, mas	[ ] A maioria dentro A maioria dos ensaios fornecendo dentro do ensaio informações de subgrupo; ou análise individual de dados do participante que	[ ] Completamente dentro Todos os estudos fornecem informações dentro (subgrupos) ou dados individuais dos participantes; e análises separadas

	alguns testes fornecendo dentro do teste informações do subgrupo.	combina informações dentro e entre o estudo.	para informações entre grupos.
2. Para comparações dentro dos grupos, os efeitos de modificação foram similares entre os estudos? [ ] Não aplicável: nenhum ou uma comparação dentro-ECR			
[ ] Definitivamente não são similares O efeito modificador relatado por dois ou mais estudos e com direções de efeito claramente diferentes.	[ ] Provavelmente não são similares ou incerto O efeito modificador não é relatado por estudos individuais ou muito impreciso para indicar algo.	[ ] A maioria é similar O efeito modificador relatado por dois ou mais estudos apresentando similaridade na direção mas com diferenças consideráveis na magnitude.	[ ] Definitivamente similar O efeito modificador relatado por dois ou mais estudos apresentam direção de efeito similar e pouca diferença na magnitude.
3. Para comparações entre-grupos, o número de estudos é grande? [ ] Não aplicável? não há comparações entre RCT			
[ ] Muito pequeno Um ou dois ou no menor subgrupo; 5 ou menos para meta-regressão contínua.	[ ] Muito pequeno ou incerto 3 a 4 em subgrupos menores; 6 a 10 em análises de meta-regressão contínua	[ ] Consideravelmente grande 5 a 9 em subgrupos menores; 11 a 15 em meta-regressão contínua	[ ] Grande 10 ou mais em subgrupos menores; mais de 15 em meta-regressões contínuas.
4. A direção do efeito modificador foi corretamente apresentada em hipótese *a priori*?			
[ ] Definitivamente não Os resultados e as análises de pós-teste são inconsistentes com a direção da hipótese ou biologicamente implausíveis.	[ ] Provavelmente não ou incerto Hipótese ou direção da hipótese são vagos ou incertos.	[ ] Provavelmente sim Não há protocolo prévio disponível, mas uma declaração inequívoca de hipótese com a direção correta do efeito modificador.	[ ] Definitivamente sim Protocolo prévio disponível que inclui especificação correta da direção do efeito modificador e/ou base em racional biológico.
5. O teste de interação sugere que há uma chance de uma explicação improvável do aparente efeito modificador? (considerar o número de efeitos modificadores)			
[ ] Explicação pelo acaso é muito provável Interação ou meta-regressão com p-valor > 0,05.	[ ] O acaso é uma provável explicação ou incerteza Interação ou meta-regressão com p-valor ≤0,05 e ≥ 0,01, ou nenhum teste de interação relatado e não computado.	[ ] O acaso pode não explicar Interação ou meta-regressão com p-valor ≤0,01 e ≥ 0,005.	[ ] É improvável que o acaso possa explicar Interação ou meta-regressão com p-valor ≤0,005.
6. Os autores testaram apenas um pequeno número de efeitos modificadores ou consideraram o número de efeitos modificadores na análise estatística?			
[ ] Definitivamente não Análises explicitamente exploratórias ou um grande número de efeitos modificadores testados (mais do que 10), sendo que a multiplicidade não foi considerada nas análises.	[ ] Provavelmente não ou incerto Nenhuma menção sobre o número ou 4 a 10 efeitos modificadores testados, e o número não foi	[ ] Provavelmente sim Nenhum protocolo disponível mas declarações claras de 3 ou menos efeitos modificadores testados	[ ] Definitivamente sim Protocolo disponível e 3 ou menos efeitos modificadores testados ou o mesmo número considerado nas análises.

	considerado nas análises.		
7. Os autores adotaram modelos de efeitos aleatórios?			
[ ] Definitivamente não Utilização de modelo de efeitos fixos adotado ou explicitamente declarado.	[ ] Provavelmente não ou incerto Provável utilização de modelos de efeitos fixos	[ ] Provavelmente sim Provável utilização de modelo de efeitos aleatórios	[ ] Definitivamente sim Utilização de modelo de efeitos aleatórios ou declaração explícita de utilização.
8. Se o efeito modificador é uma variável contínua, um ponto de corte arbitrário foi evitado? [ ] Não aplicável? não é contínuo.			
[ ] Definitivamente não Análise baseada em ponto de corte exploratório.	[ ] Provavelmente não ou incerto Análise baseada em ponto de corte de origem incerta.	[ ] Provavelmente sim Análise baseada em ponto de corte pré-especificado.	[ ] Definitivamente sim Análise baseada em um continuum completo
9. Opcional: Há alguma consideração adicional que possa aumentar ou reduzir a credibilidade? [ ] Não aplicável.			
[ ] Sim, provavelmente reduz.		[ ] Sim, provavelmente aumenta.	
10. Como você classificaria a credibilidade geral do efeito modificador proposto?			
A avaliação geral deve ser guiada por itens que reduzem a credibilidade. Uma estratégia sensível é disponibilizada a seguir: • Todas as respostas indicam definitiva ou provável redução da credibilidade ou incerteza → muito baixa • Duas ou mais respostas definitivamente diminuem a credibilidade → máximo geralmente baixo, mesmo que todas as outras respostas satisfaçam os critérios de credibilidade. • Uma resposta definitivamente reduz a credibilidade → máximo geralmente moderado, mesmo que todas as outras respostas satisfaçam os critérios de credibilidade. • Duas respostas provavelmente reduzem a credibilidade → máximo geralmente moderado, mesmo que todas as outras respostas satisfaçam os critérios de credibilidade • Nenhuma resposta de opções “definitivamente” ou “provavelmente” reduz a credibilidade → probabilidade muito alta.			
Credibilidade muito baixa Nenhum efeito modificador provável. Recomenda-se utilizar o efeito geral para cada subgrupo.	**Baixa credibilidade** Nenhum efeito modificador provável Recomenda-se utilizar o efeito geral para cada subgrupo, mas note que há incerteza remanescente.	**Credibilidade moderada** Efeito modificador provável Utilizar efeito modificador em separado para cada subgrupo, mas note que há incertezas remanescentes.	**Credibilidade alta** Efeito modificador muito provável. Utilizar efeito modificador em separado para cada subgrupo.

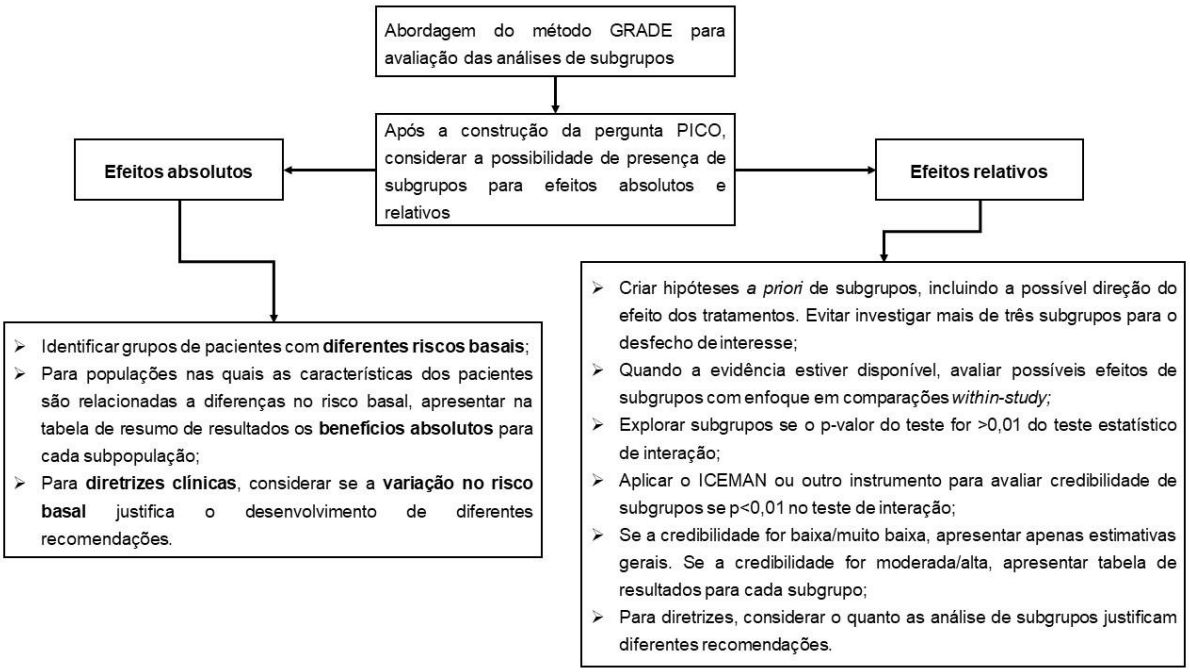
Fonte: Adaptado de Schandelmaier et al. (50).

A avaliação final de credibilidade do efeito de subgrupo através do ICEMAN possui certas similaridades com a avaliação da certeza da evidência pelo GRADE: a)

a avaliação do ICEMAN apresenta um continuum similar ao GRADE (de muito alta até muito baixa credibilidade); b) a avaliação de cada tópico contribui para a construção da avaliação geral; c) a avaliação geral possui base no julgamento prévio de cada critério individual. Por outro lado, o ICEMAN destaca a natureza contínua do julgamento da credibilidade através de uma escala visual analógica (50).

Para a avaliação geral, o ICEMAN sugere adotar a estimativa global nos resultados da RS caso o efeito de subgrupos seja muito baixa ou baixa em credibilidade, evitando assim a realização de análises de subgrupos, e a adoção de estimativas separadas para subgrupos caso a credibilidade seja alta ou moderada. A Figura 9 apresenta uma estratégia para análise da credibilidade de subgrupos.

Figura 9 – Fluxograma para realização de análises de subgrupos



Fonte: Guyatt et al. (42).

Por fim, conforme demonstrado anteriormente, avaliar inconsistência considerando a importância de cada um dos desfechos é um desafio por diversas questões (51). Primeiro, o conjunto de evidências nem sempre fornece definições claras quanto à importância dos resultados ou valores e preferências ou sobre como a priorização dos desfechos foi definida e, por vezes, inclui um conjunto diversificado

de métodos ou instrumentos para avaliação dos desfechos (52-54). Assim, pode ser difícil interpretar se as diferenças são decorrentes da variabilidade entre os instrumentos ou de outros fatores potenciais. Segundo, em alguns cenários, os autores hesitam em agrupar estimativas obtidas com diferentes instrumentos, criando um dilema para a interpretação qualitativa e não quantitativa de revisões sistemáticas, o que pode dificultar a avaliação do domínio inconsistência. Nas situações em que apenas um estudo relata os resultados de interesse, a avaliação de inconsistência é realizada de forma direta, sendo considerada inexistente (a certeza da evidência baseada em apenas um estudo provavelmente será rebaixada em outros domínios). Embora os autores sejam incentivados, sempre que apropriado, a agruparem em uma única análise os resultados de um mesmo construto/desfecho, quando isso não for possível (por exemplo, conjuntos de evidências com apresentações narrativas de resultados), a avaliação da inconsistência deve seguir os mesmos passos citados anteriormente (51).

3.3.3 Evidência indireta

- É considerada evidência direta toda evidência oriunda de pesquisas que comparam intervenções de interesse, na população de interesse, com os comparadores de interesse e que apresentam resultados de desfechos importantes para os pacientes no contexto de uma determinada questão PICO definida pelo grupo de elaboradores do painel de diretriz clínica ou RS.
- Caso desvios substanciais da PICO sejam identificados no corpo de evidências, a certeza da evidência deve ser rebaixada por evidência indireta.
- As fontes de evidência indireta são as seguintes: diferenças na população de interesse, diferenças na intervenção de interesse, diferenças nos desfechos de interesse e diferenças nos comparadores de interesse.

Avaliação

No domínio sobre evidência indireta, avalia-se se os participantes, as intervenções, os desfechos e os comparadores dos estudos são substancialmente diferentes daqueles considerados na questão de pesquisa da RS ou diretriz clínica. Para que o julgamento desse domínio seja feito de forma adequada, é fundamental

que todos os itens da questão PICO estejam claros e explícitos *a priori*, visto que esses elementos serão os principais guias para identificar se há evidência indireta no corpo de evidências.

Nesse contexto, evidência direta consiste em pesquisas que comparam diretamente as intervenções de interesse, administradas na população-alvo, e que mensuram desfechos importantes para os pacientes. De modo geral, a estimativa é mais confiável quando os resultados advêm de evidências diretas. Em situações em que há diferenças entre a questão de pesquisa dos estudos incluídos e a questão clínica de interesse, é importante que os autores de revisões sistemáticas e de diretrizes clínicas considerem incertezas em relação à aplicabilidade das evidências para a questão de pesquisa de interesse e reduzam a certeza da evidência em um ou dois níveis, conforme necessário.

A evidência pode ser considerada indireta conforme ocorrerem variações da questão PICO de interesse em maior ou menor grau. No âmbito de uma RS, as evidências encontradas que forem consideradas apropriadas e suficientemente diretas não resultarão em penalização nesse domínio do GRADE. No entanto, autores de diretrizes clínicas que eventualmente farão uso da RS em questão podem estar interessados em uma questão PICO um pouco diferente e, nesse caso, julgarem que a evidência obtida na RS é indireta para a questão PICO da diretriz.

Fontes de evidência indireta

Na sequência, são apresentadas as quatro fontes de evidência indireta — diferenças na população de interesse e na intervenção de interesse (aplicabilidade) diferenças nos desfechos de interesse (desfechos substitutos) e comparações indiretas. As diferenças pontuadas se referem sempre ao que foi definido originalmente na questão PICO de interesse.

Diferenças na população de interesse (aplicabilidade)

Diferenças na população dos estudos em relação à questão PICO de interesse são um problema comum para os autores de revisões sistemáticas e painelistas de diretrizes, podendo resultar em penalização por evidência indireta (Quadro 17, exemplo I). O efeito na certeza da evidência geral varia de acordo com o quão diferente são as populações dos estudos encontrados em relação à população de interesse.

Os ECR que avaliam o efeito de intervenções tendem a excluir pacientes com comorbidades. Nesse sentido, painelistas de uma diretriz, ao fazerem recomendações relacionadas ao uso de algum medicamento para uma população com múltiplas doenças, podem considerar reduzir a evidência em um nível nesse domínio. Em alguns casos, as diferenças podem ser verificadas por meio de análises de subgrupo ou análises combinadas de diversos estudos que incluíram diferentes populações, podendo, assim, eliminar a necessidade de penalização.

Essa discussão, em geral, se refere a populações humanas diferentes, mas há situações em que as únicas evidências disponíveis são de estudos com animais, como roedores e primatas. Há, ainda, situações em que outros tipos de estudos não desenvolvidos em seres humanos, como evidências oriundas de estudos em laboratório, podem gerar evidências bastante uteis para a tomada de decisão (Quadro 17).

Quadro 17 – Exemplos de evidência indireta oriunda de diferenças na população de interesse

I. Uma diretriz que se propõe a fazer recomendações sobre o uso de broncodilatadores em pacientes adultos e pediátricos com diagnóstico de asma pode não identificar evidências para pacientes pediátricos, visto que essa população pode não ter sido incluída nos ensaios clínicos randomizados. Sendo assim, os painelistas da diretriz podem considerar as evidências dos pacientes adultos ao fazer recomendações para pacientes pediátricos e penalizar por evidência indireta nessa população.

II. Estudos conduzidos in vitro para avaliar mudanças nos padrões de resistência e susceptibilidade das bactérias a diferentes tipos de agentes antimicrobianos podem gerar evidência indireta para avaliar a superioridade de alguns antibióticos no tratamento de infecções bacterianas em seres humanos. Uma vez que exames realizados in vitro para avaliar ação antibacteriana costumam possuir alta correlação com desfechos clínicos, em certos casos, esse tipo de evidência pode ser considerado até mesmo de alta qualidade, dependendo de adequada contextualização com especialistas na área.

Fonte: adaptado de Guyatt, et al. (41).

Diferenças na intervenção de interesse (aplicabilidade)

É importante garantir que apenas estudos com intervenções diretamente relacionadas a uma questão PICO sejam incluídos em uma RS. No entanto, exceções podem ocorrer. Em geral, quando intervenções com evidências indiretas relacionadas à questão PICO de interesse são incluídas em revisões sistemáticas ou no desenvolvimento de diretrizes, a certeza da evidência deve ser reduzida (Quadro 18, exemplo I). Em alguns casos, a intervenção utilizada será a mesma, mas pode ser administrada de maneira diferente, podendo acarretar a penalização nesse domínio (Quadro 18, exemplo II).

Outro fator a ser considerado durante o julgamento da evidência indireta é o contexto no qual uma recomendação será implementada. Painelistas de diretrizes devem considerar rebaixar a certeza da evidência nesse domínio quando concluir-se que uma determinada intervenção não pode ser implementada com o mesmo rigor ou sofisticação técnica previstos no estudo que avaliou sua eficácia (Quadro 18, exemplo III).

É importante mencionar que existe um *continuum* de similaridades entre intervenções e formas de administração que requer um julgamento cuidadoso. Em geral, é raro e desnecessário que a população pretendida e a intervenção de interesse a ser administrada sejam completamente idênticas aos estudos encontrados. Só é necessário rebaixar a certeza da evidência por evidência indireta quando essas diferenças têm potencial para modificar os resultados encontrados para os desfechos de interesse.

Quadro 18 – Exemplos de evidência indireta oriunda de diferenças na intervenção de interesse

I. Uma RS conduzida durante a elaboração de uma diretriz clínica para guiar o manejo de pacientes com COVID-19 teve por objetivo verificar se anticoagulantes devem ser utilizados no tratamento de pacientes ambulatoriais com suspeita ou diagnóstico de COVID-19. Apenas um estudo relevante foi identificado na busca, o qual avaliava o uso de sulodexida em comparação ao tratamento usual em pacientes ambulatoriais com COVID-19. Sulodexida é um anticoagulante que não possui autorização de uso no Brasil. Sendo assim, no contexto de uma diretriz clínica que visa a fazer recomendações para a população brasileira, a certeza da evidência deve ser penalizada por evidência indireta, visto que o medicamento não está entre os anticoagulantes disponíveis no contexto de interesse.

II. A intervenção de interesse de uma RS é a realização de um procedimento cirúrgico. Os autores da revisão encontraram apenas estudos conduzidos em centros clínicos de referência, nos quais as cirurgias foram realizadas por cirurgiões especializados. Não foram encontrados estudos conduzidos com cirurgiões localizados em hospitais de menor complexidade. Os autores da revisão podem concluir que os resultados de cirurgias realizadas em locais menos especializados podem diferir dos resultados apresentados nos estudos e, portanto, penalizar a avaliação da certeza da evidência por evidência indireta.

III. O impacto da implementação de um programa de exercícios físicos e acompanhamento nutricional em moradores de São Paulo foi avaliado pelos painelistas de uma diretriz que visa a fazer recomendações para a prevenção de doenças cardiovasculares na população geral. Considerou-se que os resultados obtidos nos estudos identificados provavelmente diferirão consideravelmente do impacto que será observado com a implementação da intervenção em larga escala, pois o contexto, a infraestrutura e os recursos de cada cidade são diferentes, tornando-se necessário rebaixar a certeza da evidência por evidência indireta.

Fonte: adaptado de Guyatt et al. (55).

Diferenças nos desfechos de interesse (desfechos substitutos)

O sistema GRADE especifica que tanto os elaboradores de revisões sistemáticas quanto os painelistas de diretrizes clínicas devem iniciar o processo especificando todos os desfechos importantes de interesse. Os estudos disponíveis podem ter mensurado o impacto da intervenção de interesse em desfechos relacionados, porém diferentes dos desfechos definidos como importantes para os pacientes (Quadro 19).

A diferença entre desfechos desejados e mensurados pode estar relacionada ao período de interesse (por exemplo, um desfecho medido aos 3 meses em vez de aos 12 meses). Nesse caso, dependendo do quão próximo é o período observado do período desejado ou do quão relevante é a avaliação em um período mais prolongado, os autores de revisões devem penalizar a evidência em um ou dois níveis.

Outra fonte de evidência indireta relacionada à mensuração dos desfechos é o uso de desfechos substitutos em detrimento de desfechos importantes para os pacientes. A Tabela 1 apresenta alguns exemplos de desfechos substitutos observados na literatura.

Tabela 1 – Desfechos substitutos comumente encontrados na literatura e desfechos importantes para os pacientes correspondentes

Condição clínica	Desfechos substitutos	Desfechos de relevância clínica (preferíveis)
Diabetes melito	Redução da glicemia; hemoglobina glicosilada	Sintomas diabéticos específicos; hospitalização; complicações cardiovasculares, oculares, renais e/ou neuropáticas
Hipertensão	Redução da pressão arterial	Óbito de causa cardiovascular; infarto agudo do miocárdio; acidente vascular cerebral

Condição clínica	Desfechos substitutos	Desfechos de relevância clínica (preferíveis)
Demência	Função cognitiva; marcadores biológicos	Funcionalidade, comportamento, qualidade de vida, sobrecarga dos cuidadores
Osteoporose	Cálcio e fósforo séricos; densidade óssea	Incidência de fraturas
Doença cardiovascular	Lipídios séricos; calcificação coronariana	Infarto agudo do miocárdio, eventos vasculares, óbito
Doença respiratória crônica	Parâmetros de função pulmonar	Qualidade de vida, incidência de exacerbações, mortalidade
Trombose venosa	Trombose venosa assintomática	Trombose venosa sintomática

Fonte: adaptado de Guyatt et al. (55).

Em geral, o uso de desfechos substitutos requer o rebaixamento da certeza da evidência em pelo menos um nível, às vezes até mesmo dois. Considerações referentes à biologia, ao mecanismo e à história natural da doença podem ser úteis na decisão sobre a evidência indireta. Também deve ser considerada a proximidade entre o desfecho substituto e o caminho causal do desfecho de interesse importante para o paciente. Por exemplo, para a incidência de fraturas, a densidade óssea é um desfecho substituto mais próximo do que a simples mensuração de cálcio e fósforo séricos. Para substitutos que estão muito distantes dos desfechos importantes para os pacientes, é importante reduzir a certeza da evidência em dois níveis.

Em alguns casos, é considerado justificável não rebaixar a evidência em pelo menos um nível na presença de desfechos substitutos. Uma possibilidade é quando as mudanças nesses desfechos estão altamente correlacionadas com mudanças em desfechos importantes para os pacientes nos ECR de diferentes medicamentos da

mesma classe (por exemplo, betabloqueadores, antagonistas do cálcio, bisfosfonatos). Entretanto, mesmo nesses casos, os revisores ou painelistas de uma diretriz devem ser cautelosos e justificar adequadamente a decisão de não penalização. Aconselha-se nesse caso que a correlação clínica entre o desfecho de interesse e o desfecho substituto seja adequadamente demonstrada por evidência empírica.

Algumas ferramentas para avaliar a validade de desfechos substitutos foram desenvolvidas e podem auxiliar autores de revisões e diretrizes durante a ponderação do julgamento de evidência indireta.

Quadro 19 – Exemplo de evidência indireta oriunda de diferenças nos desfechos de interesse

O corpo de evidências identificado para responder à questão PICO “O uso de vacinas para combater o vírus da *influenza* reduz a incidência de complicações respiratórias em pacientes adultos?” apresentou apenas desfechos de imunogenicidade, como produção de anticorpos contra o vírus e quantificação da carga viral dos pacientes infectados. Não foram encontrados estudos que apresentassem o efeito do uso de vacinas sobre o percentual de pacientes que fizeram uso de suplementação de oxigênio, progressão da doença, necessidade de internação em unidade de terapia intensiva e uso de ventilação mecânica. Dessa forma, a certeza da evidência foi rebaixada por evidência indireta.

Fonte: elaboração própria.

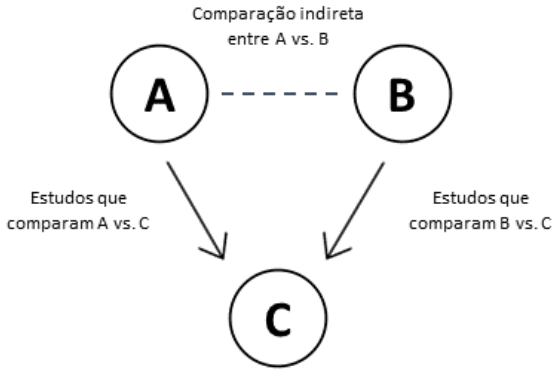
Comparações indiretas

As comparações indiretas ocorrem quando uma comparação entre a intervenção A e B não está disponível, mas A foi comparada com C em um estudo e B foi comparada com C em outro estudo. Desse modo, tais estudos permitem que comparações indiretas da magnitude do efeito de A *versus* B sejam realizadas (Figura 10).

No início do desenvolvimento do sistema GRADE, considerava-se que comparações indiretas em geral possuíam uma certeza de evidência inferior às evidências diretas. Ao longo do tempo, foram desenvolvidas novas abordagens

estatísticas, como metanálise em rede, que possui uma metodologia específica para avaliação da certeza da evidência, disponível no capítulo 8. Apesar dessa possibilidade, ainda é comum encontrar comparações indiretas de evidências que não fazem uso de métodos estatísticos adequados. A simples comparação dos resultados entre dois grupos com intervenções ativas é geralmente considerada insuficiente e, como resultado, a qualidade da evidência pode ser penalizada em até dois níveis devido à falta de abordagem apropriada para lidar com evidências indiretas.

Figura 10 – Ilustração de uma comparação indireta entre as intervenções A e B



Fonte: adaptado de Song et al. (56).

Considerações adicionais sobre o domínio evidência indireta

Algumas considerações adicionais devem ser mencionadas sobre o domínio evidência indireta, por exemplo, o mecanismo de ação de uma intervenção. No âmbito do sistema GRADE, o mecanismo de ação ou o embasamento fisiopatológico de uma intervenção, como um medicamento, não deve contribuir para aumentar ou diminuir a certeza da evidência. No entanto, o mecanismo de ação pode ter outras funções durante o processo de avaliação das evidências, como na seleção dos estudos para uma RS, na aplicabilidade das evidências para diferentes intervenções ou populações e na necessidade de análises de subgrupo.

É importante que, no momento de realizar o julgamento geral sobre o domínio evidência indireta, todas as quatro possíveis fontes sejam consideradas e seus potenciais efeitos sejam ponderados de forma combinada. Se necessário, a certeza

da evidência deve ser rebaixada em um ou dois níveis, de acordo com as fontes e o tamanho do desvio da questão PICO de interesse. Como regra geral, a evidência indireta oriunda de desfechos substitutos deve ser penalizada em pelo menos um nível; já a evidência indireta oriunda de outras fontes requer um pouco mais de cuidado e ponderação no julgamento.

Por fim, algumas fontes de evidência indireta são problemas comuns, tanto no contexto de revisões sistemáticas quanto no contexto de diretrizes clínicas (diferenças no desfecho e nos comparadores), enquanto outras são mais comuns apenas no contexto das diretrizes clínicas (diferenças na população e na intervenção).

3.3.4 Imprecisão

- O GRADE sugere avaliar a imprecisão baseado no IC do efeito absoluto como critério primário, e considerando limiares de efeito absoluto definidos de acordo com o nível de contextualização da abordagem adotada.
- As abordagens baseiam-se em níveis de contextualização, que podem ser: a) minimamente contextualizada (quando há contraste do efeito absoluto de um desfecho com pelo menos uma diferença clinicamente importante mínima); b) parcialmente contextualizada (quando há definição de diferentes faixas de efeito para avaliar o efeito absoluto de um desfecho), e; c) completamente contextualizada (quando se permite a combinação do efeito absoluto de diferentes desfechos para avaliar o efeito da intervenção frente a diferentes faixas de efeito absoluto esperado).
- A aplicação do GRADE no processo de tomada de decisão (diretrizes clínicas e incorporação de tecnologias) deve aplicar, como padrão mínimo, a abordagem minimamente contextualizada.
- O uso de abordagens parcialmente e completamente contextualizadas deve ser estimulada, contudo entende-se a dificuldade em estabelecer parâmetros para o tamanho de efeito absoluto de cada desfecho, e em especial, a ponderação entre desfechos, que limita de forma importante o seu uso.
- Quando se encontra um efeito absoluto muito grande, recomenda-se avaliar se o tamanho ótimo da informação (*optimal information size* [OIS]) foi atendido.

- Não adotar uma abordagem contextualizada limita a conclusão dos resultados alcançados e mostra-se insuficiente na tomada de decisão.

A avaliação de imprecisão no sistema GRADE é o quarto critério indicado que pode ser considerado na redução do nível da certeza de evidência, podendo-se penalizar por alta probabilidade de erro aleatório (51). Para estudos de RS, o sistema GRADE define precisão como o grau de confiança de que a estimativa de efeito de uma intervenção está próxima do verdadeiro efeito; em diretrizes clínicas, a precisão da estimativa de efeito está relacionada à confiança de que a estimativa é adequada para corroborar uma recomendação (57). Assim, o GRADE *Working Group* indica que o critério para reduzir o nível de evidência por imprecisão deve diferir entre a elaboração de diretrizes e a formulação de revisões sistemáticas. Em diretrizes, deve-se considerar um contexto mais completo para a realização do julgamento e a formulação de uma recomendação, incluindo todos os desfechos priorizados na avaliação, enquanto o julgamento para revisões sistemáticas pode incluir um contexto mínimo para avaliação, em que cada desfecho é avaliado isoladamente (57).

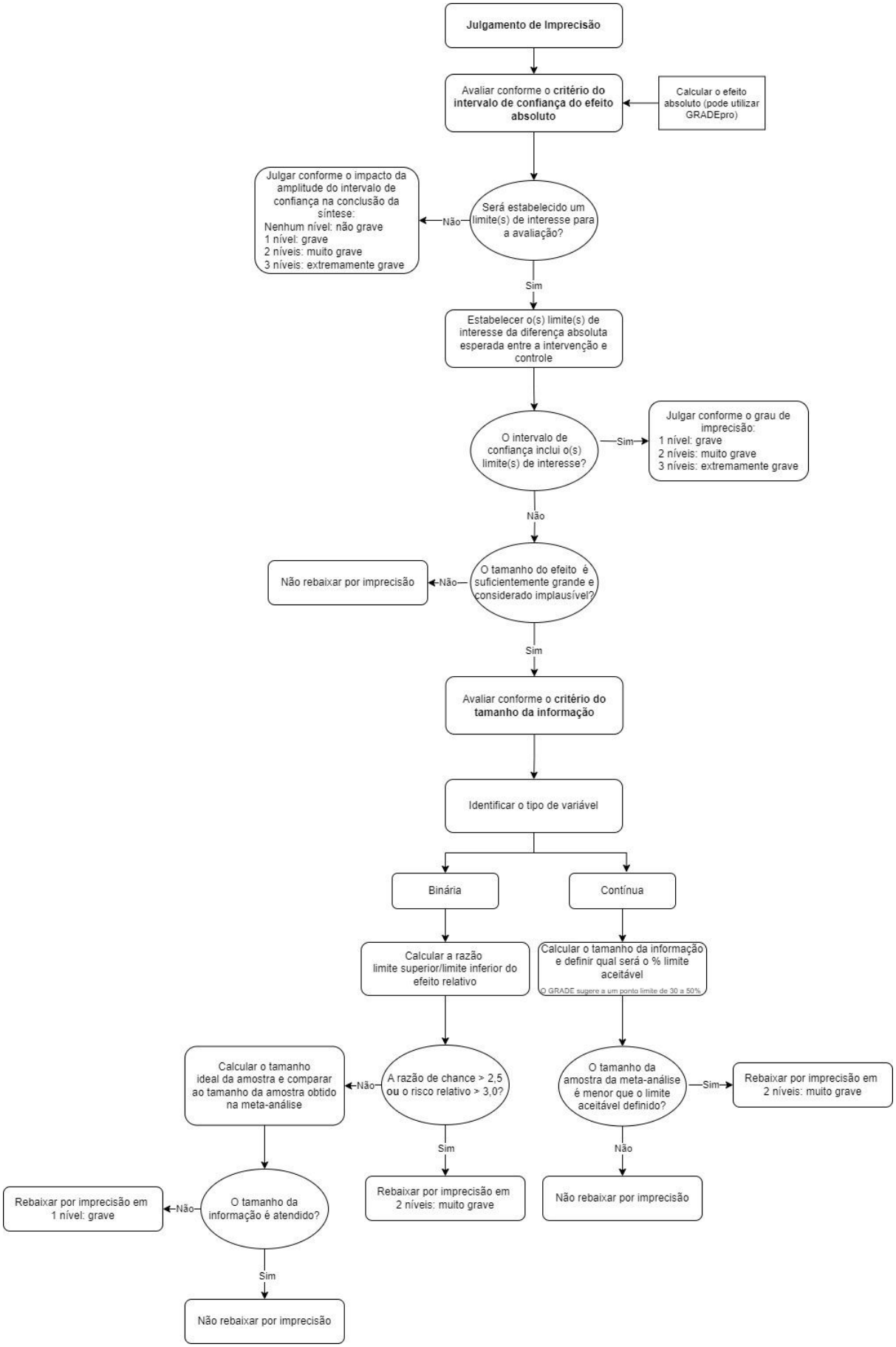
As primeiras orientações do GRADE *Working Group* indicavam que a classificação da imprecisão deveria incluir uma avaliação do IC da medida pontual e do tamanho da amostra obtidos para a síntese das evidências, sendo possível o rebaixamento da certeza de evidência em até dois níveis (51). Com a evolução do método, em 2022, o GRADE *Working Group* publicou novas orientações sobre como avaliar a imprecisão. A partir de então, orienta-se utilizar como critério primário para o julgamento o IC da diferença absoluta entre a intervenção e o controle, a partir de limite(s) de interesse estabelecido(s) conforme o nível de contextualização – minimamente, parcialmente e completamente contextualizado (Tabela 2) (58, 59). Se o IC ultrapassar o(s) limite(s), o domínio deve ser penalizado; se não ultrapassar, não deve ser penalizado (Figura 11). Em casos nos quais o IC for estreito, o efeito relativo for grande e o tamanho da amostra e o número de eventos forem menores do que o esperado, os resultados podem ser frágeis/questionáveis em relação à imprecisão. Em tal situação, pode ser apropriado utilizar como critério para julgamento o OIS, também chamado de tamanho da informação de revisão (*review information size* [RIS]) (58) .

Tabela 2 – Critérios de julgamento para o domínio imprecisão

<table>
<tr><th>Minimamente
(geralmente em revisões sistemáticas)</th><th>Parcialmente</th><th>Completamente
(geralmente em diretrizes)</th></tr>
<tr><td colspan="2">Deve-se realizar o julgamento de cada um dos desfechos de forma individual.</td><td>Deve-se realizar o julgamento dos diversos desfechos de forma simultânea.</td></tr>
<tr><td>Efeito encontrado
versus
efeito nulo ou com uma diferença minimamente importante.</td><td>Efeito encontrado
versus
faixa que representa efeitos triviais, pequenos, moderados ou grandes.</td><td>Intervenção
(conjunto de efeitos desejáveis e indesejáveis)
versus
faixa que representa efeitos triviais, pequenos, moderados ou grandes.</td></tr>
<tr><td colspan="3">É preciso explicitar o alvo e o cenário considerados antes da avaliação.</td></tr>
</table>

Fonte: elaboração própria.

Figura 11 - Fluxograma do processo de avaliação de imprecisão


Julgamento de Imprecisão
Avaliar conforme o critério do intervalo de confiança do efeito absoluto
Calcular o efeito absoluto (pode utilizar GRADEpro)
Será estabelecido um limite(s) de interesse para a avaliação?
Não
Julgar conforme o impacto da amplitude do intervalo de confiança na conclusão da síntese:
Nenhum nível: não grave
1 nível: grave
2 níveis: muito grave
3 níveis: extremamente grave
Sim
Estabelecer o(s) limite(s) de interesse da diferença absoluta esperada entre a intervenção e controle
O intervalo de confiança inclui o(s) limite(s) de interesse?
Sim
Julgar conforme o grau de imprecisão:
1 nível: grave
2 níveis: muito grave
3 níveis: extremamente grave
Não
O tamanho do efeito é suficientemente grande e considerado implausível?
Não
Não rebaixar por imprecisão
Sim
Avaliar conforme o critério do tamanho da informação
Identificar o tipo de variável
Binária
Contínua
Calcular a razão limite superior/limite inferior do efeito relativo
Calcular o tamanho da informação e definir qual será o % limite aceitável
O GRADE sugere a um ponto limite de 30 a 50%
A razão de chance > 2,5 ou o risco relativo > 3,0?
Não
Calcular o tamanho ideal da amostra e comparar ao tamanho da amostra obtido na meta-análise
Sim
Rebaixar por imprecisão em 2 níveis: muito grave
O tamanho da amostra da meta-análise é menor que o limite aceitável definido?
Sim
Rebaixar por imprecisão em 2 níveis: muito grave
Não
Não rebaixar por imprecisão
O tamanho da informação é atendido?
Não
Rebaixar por imprecisão em 1 nível: grave
Sim
Não rebaixar por imprecisão

O fluxograma do processo de avaliação de imprecisão segue os direcionamentos indicados por Zeng et al. (58)., para um cenário de avaliação minimamente contextualizado, iniciando a avaliação com o critério primário do intervalo de confiança (IC) de 95% do risco absoluto. Para o cálculo do risco absoluto e seu IC 95%, pode ser utilizada a ferramenta GRADEpro, na área de tabela de evidência (https://gradepro.org/). O tamanho ideal da amostra refere-se ao cálculo do tamanho amostral planejado para um estudo primário. Para esse cálculo, pode-se utilizar calculadoras como a disponível de forma on-line pelo link de acesso https://www.stat.ubc.ca/~rollin/stats/ssize/b2.html.

Fonte: elaboração própria.

De forma geral, os resultados de um estudo serão imprecisos se forem incluídos poucos pacientes e ocorrerem poucos eventos; com isso, o IC tende a ser amplo ao redor da estimativa pontual do efeito (Figura 12, Quadro 20) (51). Nesses casos, devido à incerteza sobre os resultados, pode ser necessário classificar a certeza das evidências em níveis mais baixos do que os usuais (57). A partir do critério primário do IC, as novas recomendações indicam que a certeza de evidência pode ser rebaixada por imprecisão em até três níveis, ou seja, grave, muito grave e extremamente grave. Já pelo critério de OIS, pode-se considerar rebaixar em até dois níveis por imprecisão. A seguir, é detalhada a avaliação a partir das duas abordagens (58, 59) — a partir do IC e a partir do tamanho ótimo da informação ou tamanho de informação da revisão.

Abordagem a partir do intervalo de confiança

O IC é uma medida de precisão estimada a partir da amostra do estudo e pode ser descrito como um intervalo de valores que provavelmente contém o parâmetro de interesse (valor verdadeiro) com um nível especificado de confiança. Como critério para avaliação em uma abordagem minimamente contextualizada, o IC deve ser utilizado pelos autores para julgar se o verdadeiro efeito está presente em relação ao efeito nulo ou se um efeito importante está presente em relação a uma diferença minimamente importante (*minimal important difference* [MID]) determinante para a tomada de decisão (Quadro 21). Nesse cenário, o grau de penalização acontecerá na proporção de imprecisão em relação à referência estabelecida, seja o efeito nulo ou ao(s) MID(s) (Quadro 22). Se a referência for apenas o efeito nulo, então a avaliação

será baseada na identificação se há um benefício ou um dano. Assim, o conjunto de evidência que apresentar um IC que englobe um efeito nulo será penalizado. Nesse caso, o grau de penalização dependerá da possibilidade de interpretação muito distintas do efeito relacionado aos limites do IC.

Em uma abordagem parcialmente contextualizada, os autores julgam se o verdadeiro efeito está localizado em uma faixa que representa um efeito trivial, pequeno, moderado ou grande (Figura 16, Quadro 23) (58). Por fim, em uma abordagem completamente contextualizada (indicada para diretrizes), os autores consideram todos os desfechos críticos ou importantes de forma conjunta (ou seja, trocando efeitos desejáveis *versus* indesejáveis de uma intervenção em saúde) e definem um limite de decisão acima do qual eles recomendariam a favor de uma intervenção e abaixo do qual eles recomendariam contra a recomendação (59).

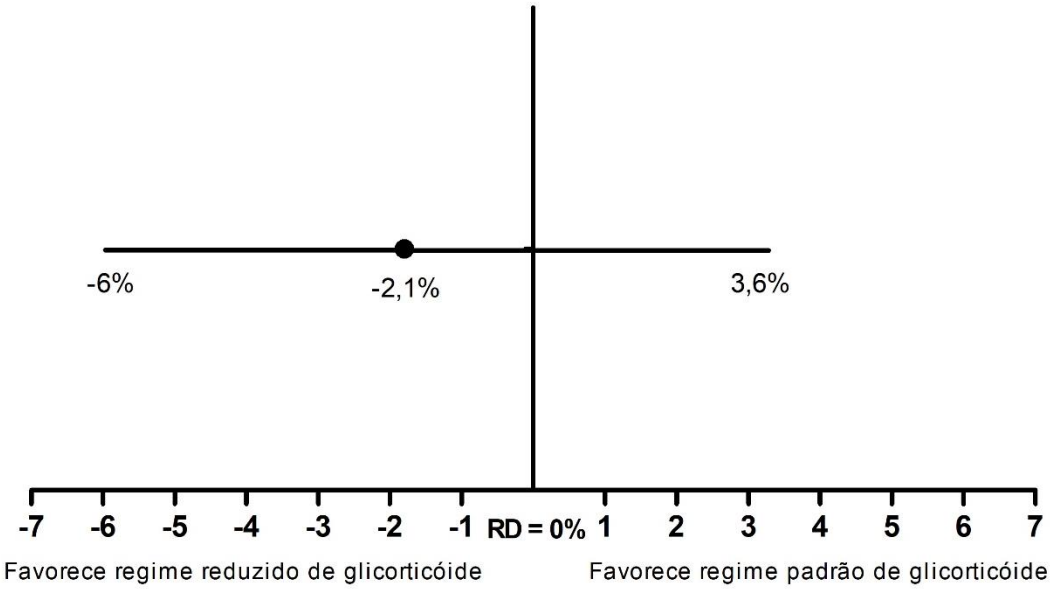
Atualmente, consideramos que a aplicação do GRADE no processo de tomada de decisão (diretrizes clínicas e incorporação de tecnologias) deve aplicar, como padrão, a abordagem minimamente contextualizada. O uso de abordagens parcialmente e completamente contextualizadas deve ser estimulada, contudo entendemos que a dificuldade em estabelecer parâmetros para o tamanho de efeito de cada desfecho, e em especial, a ponderação entre desfechos, limita bastante seu uso. A abordagem não-contextualizada se mostra insuficiente nesse contexto.

Em linhas gerais, a escolha do(s) limiar(es) para julgamento isolado ou em faixas de efeito depende da perspectiva adotada (clínica ou de saúde pública), do contexto de execução (ambiente com recursos de saúde mais ou menos desenvolvidos) e dos valores e das preferências dos pacientes, entre outros fatores (57). A definição de um limiar parte da necessidade de os autores de revisões sistemáticas identificarem o que será considerada uma diferença importante do ponto de vista clínico entre as intervenções testadas, no caso de desfechos contínuos, e o que será considerado como um efeito benéfico ou danoso relevante, em desfechos binários. Quando a medida de efeito apresenta um IC em que os dois limites sugerem inferências muito diferentes, há, consequentemente, uma alta incerteza em relação ao verdadeiro efeito. Nesse cenário, pode-se avaliar a penalização por imprecisão em três níveis (58).

Na ausência de limiar de interesse, em que a avaliação da imprecisão é baseada na relação com o efeito nulo absoluto (ou seja, diferença de risco igual a 0), os autores devem observar a representação dos limites do IC na conclusão das evidências (58). Por exemplo, em uma RS sobre regime de dose de glicocorticoides (dose reduzida *versus* dose padrão) em pacientes com vasculite, a metanálise de ECR para o desfecho mortalidade relata que o regime de dose reduzida resultou em 2,1 mortes a menos por 100 pacientes, com um IC de 6 a menos a 3,6 a mais (Figura 12) (60). Em relação ao efeito nulo, pode-se concluir que o regime de dose reduzida de glicocorticoides reduz a mortalidade. Se os autores tivessem considerado que o limite superior do IC de 3,6 a mais por 100 pacientes indica um dano importante, poderiam concluir que, embora a estimativa pontual sugira um benefício, permanece plausível a existência de um risco importante (58). Conforme recomendação do sistema GRADE, os autores podem optar pela conclusão que se sentem mais confortáveis: 1) penalizar por imprecisão em um nível e, consequentemente, concluir que “o uso de glicocorticoides em dose reduzida em comparação com a dose padrão provavelmente resulta na redução da mortalidade”, 2) penalizar por imprecisão em dois níveis e, consequentemente, concluir que “o uso de glicocorticoides em dose reduzida em comparação com a dose padrão pode resultar na redução da mortalidade”, permanecendo possível a ideia de que o esquema de dose reduzida pode impactar em um importante aumento de óbitos, 3) ou, em um último cenário, considerar que ambos os limites do IC apresentam grandes efeitos (ou seja, o IC inclui grandes benefícios e grandes danos), penalizar em três níveis e, consequentemente, concluir que há uma incerteza em relação aos efeitos das intervenções (58).

Por fim, na utilização do sistema GRADE, deve-se atentar à possibilidade de penalização dupla por imprecisão e inconsistência ao se realizar uma metanálise com um modelo de efeitos aleatórios (58).

Figura 12 – Efeito do regime de dose reduzida *versus* regime de dose padrão de glicocorticoides na mortalidade de pacientes com vasculite



Fonte: Adaptado de Zeng et al. (58).

Quadro 20 – Considerações sobre IC, significâncias estatística e clínica do resultado e confiança da estimativa

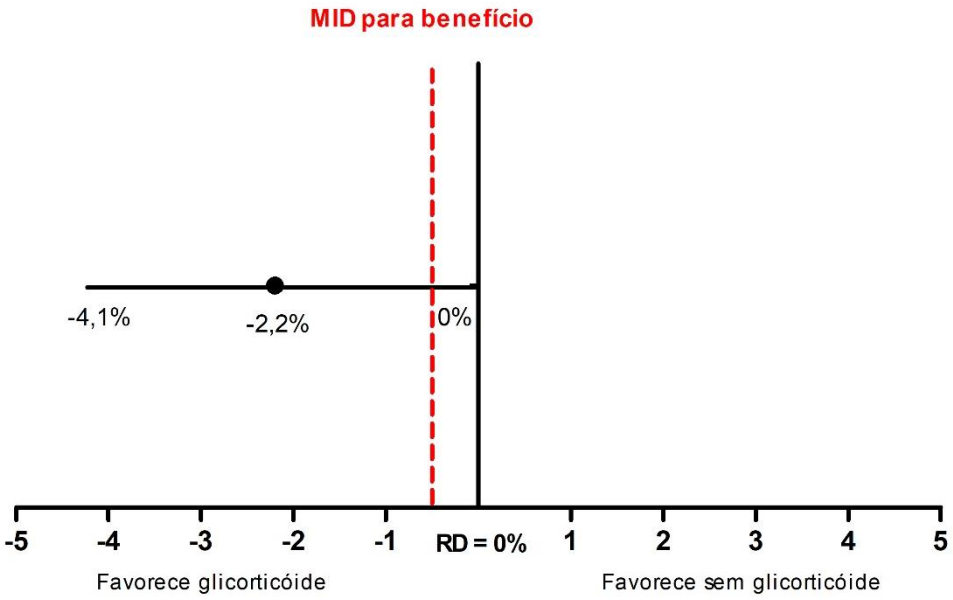
A estimativa de efeito é a estimativa pontual, como um risco relativo ou um risco absoluto, que representa a mais provável medida de efeito existente para um determinado medicamento. Em relação a isso, há o IC nessa estimativa (geralmente definido em 95%), baseado em uma medida estatística (quanto maior for o número de eventos do estudo, maior tende a ser a precisão). Por exemplo, um risco absoluto apresentou um IC95% de 0,94 a 0,97 para mortalidade, representando os valores dos melhores (redução de 6%, ou seja, 60 mortes a cada 1.000 pacientes) e piores (redução de 3%, ou seja, 30 a cada 1.000 pacientes) efeitos plausíveis dentro do IC95%. Nota-se que tanto o pior quanto o melhor cenário reduzem a mortalidade e, quando ambos apresentam a mesma direção, há significância estatística. Além disso, é necessário considerar se ambos os limites conseguem representar um resultado que seja clinicamente relevante. Previamente à obtenção dos resultados, os autores poderiam ter estabelecido uma diferença minimamente importante (ou seja, o limiar de interesse) em uma redução de 5% (ou seja, 50 óbitos por 1.000 pacientes). Nesse cenário, por mais que haja uma diferença considerada significativa do ponto de vista estatístico (pois a nulidade não está incluída no IC), não é possível alcançar uma diferença clinicamente relevante. Para a interpretação dos resultados obtidos de revisões sistemáticas e o desenvolvimento de recomendações, a confiança da estimativa do efeito está relacionada ao grau de certeza em que ela apoia uma determinada decisão (por exemplo, adotar ou não uma intervenção).

IC95% = intervalo de confiança de 95%.
Fonte: Adaptado de Zeng et al. (58).

Quadro 21 – Exemplos de aplicação da abordagem considerando o IC para o julgamento da imprecisão

Exemplo 1. Em uma RS sobre corticoides em pacientes com sepse (61), a metanálise de ensaios controlados randomizados para o desfecho mortalidade relata que o uso de corticoides resultou em uma redução de 2,2 mortes por 100 pacientes, com um IC de 4,1 a menos para 0 a menos (Figura 13). Considerando a importância do desfecho e usando uma abordagem minimamente contextualizada, poderia ser estabelecida como uma diferença minimamente importante (ou seja, o limiar de interesse) a redução de 0,5 óbitos por 100 pacientes (ou seja, 5 óbitos por 1.000 pacientes). Como a estimativa pontual está acima do MID, conclui-se que os corticoides resultam em uma redução importante na mortalidade. Como o IC cruza o MID de 0,5% (ou seja, para alguns pacientes, o efeito dos corticoides pode ser trivial), pode haver penalização por imprecisão em pelo menos um nível. Nesse exemplo, provavelmente não deveria haver penalização em dois níveis, pois o IC ultrapassa de forma discreta o limiar de interesse (-0,5 por 100). Também se destaca que o IC não inclui um aumento nas mortes com corticoides, ou seja, não apresenta um risco atribuído à intervenção. Uma conclusão equivocada de benefício não colocaria, portanto, pacientes e médicos em risco de administração de uma intervenção letal.

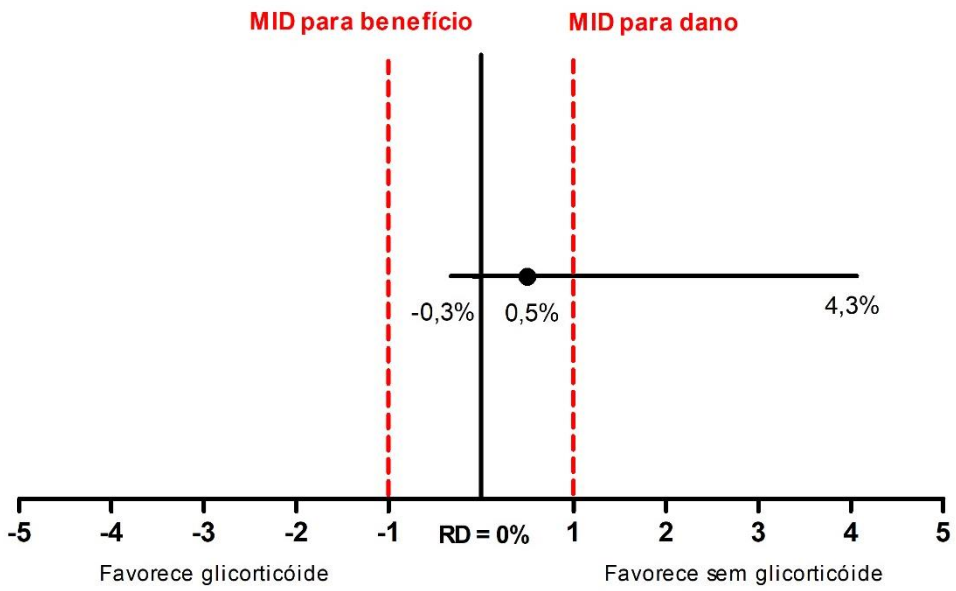
Figura 13 – Efeitos da utilização vs. não utilização de corticoides na morte de pacientes com sepse



Exemplo 2. Na mesma RS (61), a metanálise para o desfecho incidência de acidente vascular cerebral (AVC) relata que o uso de corticoides resultou em um efeito absoluto de aumento de 0,5 eventos de AVC em 100 pacientes, com um intervalo de confiança de 0,3 a menos a 4,3 a mais (Figura 14). Considerando a importância do desfecho e usando uma abordagem minimamente contextualizada, poderia ser estabelecido um limiar de MID de benefício ou dano de 1 evento por 100 pacientes. Nesse cenário, a estimativa pontual estaria na faixa de sem efeito ou efeito trivial (isto é, abaixo do limiar de MID de dano e acima de um efeito nulo), concluindo que os corticoides não apresentam efeito ou apresentam efeito insignificante sobre o AVC. Como o IC cruza o MID de um efeito de risco, pode haver penalização por imprecisão em um nível. Nesse exemplo, pode haver dúvida se o IC é grande o suficiente para justificar a penalização em dois níveis, sugerindo dano máximo de 4,3% Para essa análise, é indicado avaliar o grau de importância dada pelos pacientes e clínicos ao desfecho, em que quanto maior for a importância do desfecho, maior será a probabilidade de penalização em dois níveis Também se pode considerar o julgamento da magnitude do aumento da diferença de risco absoluto, classificando em um efeito moderado ou grande, em que quanto maior for o alcance do IC, maior será a probabilidade de penalização por imprecisão em dois níveis. Outra abordagem para a escolha do grau de penalização é considerar o impacto na formulação de uma declaração que embasará uma decisão clínica.

Nesse sentido, interpreta-se que “o uso de corticoides **provavelmente** não tenha efeito ou tenha um efeito trivial sobre a incidência de AVC” quando há redução em um nível da certeza de evidência por imprecisão, o que pode não refletir a possibilidade de que os corticoides resultem em um aumento de 4,3% eventos prejudiciais à saúde do paciente. Quando a certeza de evidência é rebaixada por imprecisão em dois níveis, pode-se concluir que “o uso de corticoides **pode** resultar em nenhum efeito ou efeito trivial sobre a incidência de AVC”, não excluindo a possibilidade de dano.

Figura 14 – Efeito da utilização vs. não utilização de corticoides em acidente vascular cerebral em pacientes com sepse

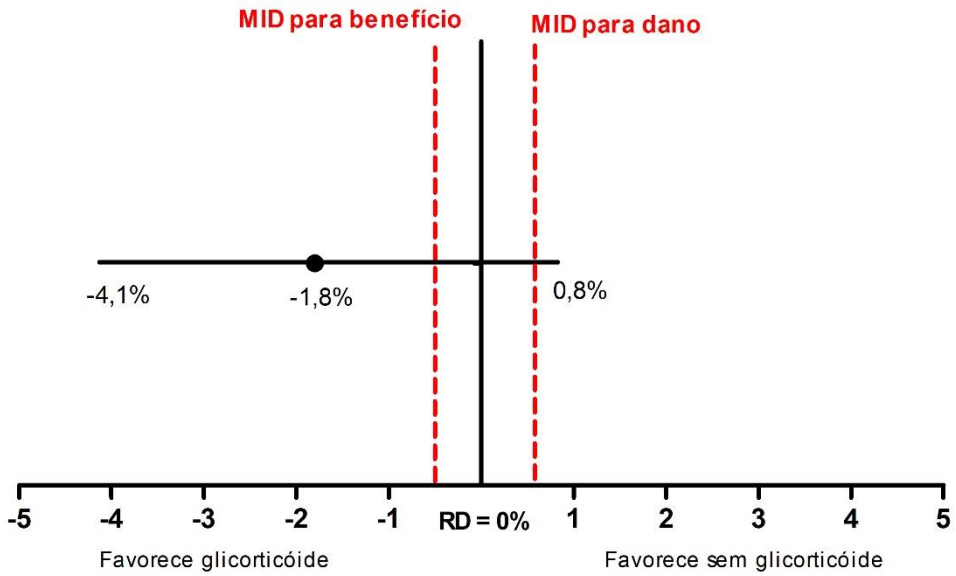


Exemplo 3. Outro desfecho avaliado na RS foi a mortalidade a curto prazo (28 e 31 dias), em que a metanálise indica que o uso de corticoides resultou em uma redução de 1,8 mortes por 100 pacientes, com um intervalo de confiança de 4,1 a menos a 0,8 a mais (

Figura 15). Considerando a importância do desfecho e usando uma abordagem minimamente contextualizada, poderia ser estabelecida como uma diferença minimamente importante (ou seja, o limiar de interesse) a redução ou o aumento de 0,5 óbitos por 100 pacientes (ou seja, 5 óbitos a menos ou a mais por 1.000 pacientes). Como a estimativa pontual está acima do MID de benefício, conclui-se

que os corticoides resultam em uma redução importante na mortalidade. Contudo, o IC inclui simultaneamente benefícios e danos importantes atribuídos à intervenção, podendo a imprecisão ser penalizada em pelo menos dois níveis. Nesse exemplo, é importante uma penalização mínima de dois níveis, pois o IC ultrapassa os dois limiares de interesse (ou seja, benefício e dano). Dessa forma, a conclusão seria que “o uso de corticoides pode ter uma redução importante na mortalidade”, mantendo a possibilidade de dano em aberto.

Figura 15 – Efeito da utilização vs. não utilização de corticosteroides na morte a curto prazo de pacientes com sepse



IC = intervalo de confiança; MID = diferença minimamente importante.
Fonte: Adaptado de Zeng et al. (58).

Abordagem a partir do tamanho ótimo da informação ou tamanho de informação da revisão

Em relação à abordagem a partir do tamanho ótimo de informação — *optimal intervention size,* OIS — trata-se de um valor único que representa o número total de participantes ou eventos necessários em uma metanálise para que ela corresponda a uma amostra relevante de um único estudo de intervenção (6, 58). Já o tamanho de informação da revisão (*review information size* [RIS]) é baseado em uma abordagem

matemática similar, sendo um conceito alternativo para atender à avaliação de imprecisão em uma abordagem parcialmente e completamente contextualizada, ou seja, em que é necessária a utilização de mais de um limiar de interesse (o GRADE *Working Group* preparou uma calculadora *on-line* para a identificação do RIS, disponível em: https://www.gradepro.org/calc/reviewinformationsize) (59). Dessa forma, o RIS é utilizado quando há possibilidade de um efeito grande não plausível (ou seja, a estimativa pontual encontra-se acima do limiar atribuído ao efeito grande) baseado em um aparente baixo número de participantes ou eventos (passo 7 do Quadro 23).

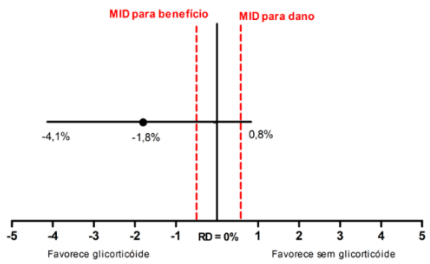
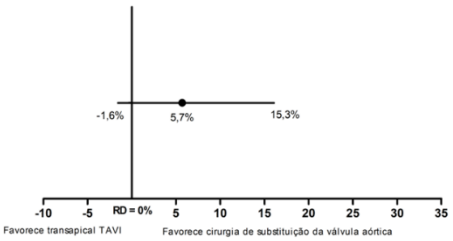
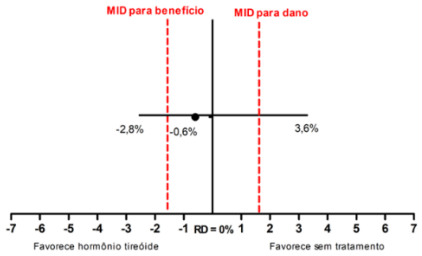
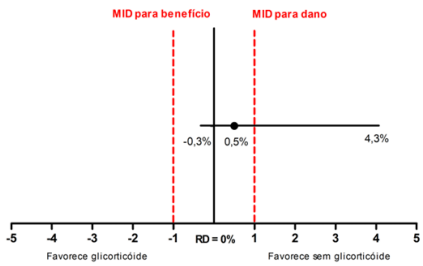
Quando o IC inclui o(s) limite(s) de interesse, deve-se classificar a imprecisão conforme descrito acima, não sendo necessário considerar o tamanho da informação. Quando o IC não inclui o(s) limite(s) de interesse e a medida de efeito tem um valor expressivo (por exemplo, redução ou aumento do RR maior que 30%) a ponto de os resultados serem considerados não plausíveis, os autores devem considerar a abordagem de avaliação de imprecisão a partir do tamanho da informação (6). Se o tamanho da informação for atendido, ou seja, a síntese das evidências é baseada em uma amostra representativa, não é necessário penalizar a certeza de evidência por imprecisão.

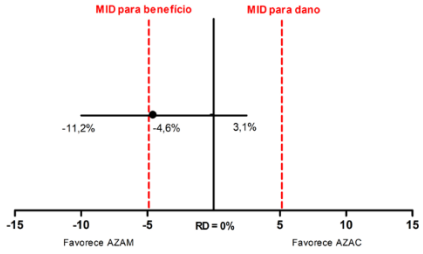
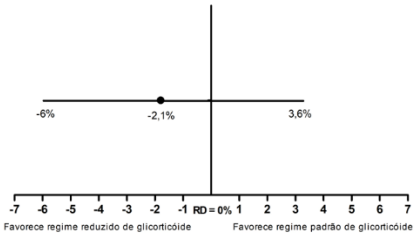
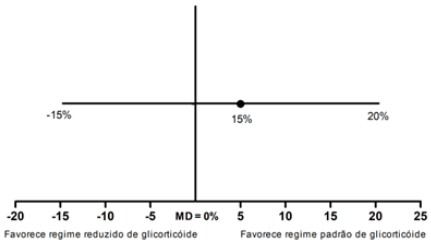
Para desfechos dicotômicos, o sistema GRADE indica que o tamanho da amostra está distante de atender ao tamanho da informação quando a proporção do limite superior para o inferior do IC é superior a 2,5 para OR e 3,0 para RR (58). Dessa forma, pode-se penalizar a certeza de evidência em dois níveis. Caso o efeito seja grande e a razão for menor do que os limites citados, deve-se calcular o tamanho da informação e compará-lo com o tamanho da amostra disponível na metanálise. Se o critério de tamanho da informação não for atendido, deve-se penalizar por imprecisão em um nível.

Para desfechos contínuos, o sistema GRADE recomenda que o cálculo do tamanho da informação seja a partir do(s) MID ou do desvio padrão (DP) de interesse; quando não houver confiança nesses valores, pode-se usar um tamanho de efeito de 0,2 DP para representar um efeito pequeno (62). Isso resulta em um tamanho amostral total de aproximadamente 800 (400 por grupo) (6, 58). O sistema GRADE sugere, com base na abordagem OIS, reduzir em dois níveis por imprecisão se o tamanho total da amostra da metanálise for menor do que o limite arbitrário de 30% a 50% do OIS. Se os autores optarem por serem mais conservadores, podem escolher

50% de OIS como limite (ou seja, 400 no total); se optarem por serem menos conservadores, podem usar 30% de OIS como limite (ou seja, 240 no total) (58).

Quadro 22 – Sumário de situações para considerar o rebaixamento por imprecisão em dois níveis aplicando a abordagem de IC em cenário minimamente contextualizado

Situação	Explicação	Representação
1	Quando a estimativa pontual reflete um benefício importante, porém o limite inferior do IC inclui a possibilidade de dano importante.	
2	Quando a estimativa pontual reflete um dano importante, porém o limite superior do IC inclui a possibilidade de benefício importante.	
3	Quando a estimativa pontual reflete um efeito trivial ou sem efeito (ou seja, sem benefício e sem dano) e o IC inclui a possibilidade tanto de um benefício importante quanto de um dano importante.	
4	Quando a estimativa pontual reflete um efeito trivial ou sem efeito (ou seja, sem benefício e sem dano) e o IC inclui a possibilidade de um dano substancialmente (ou seja, possivelmente grande) importante.	

Situação	Explicação	Representação
5	Quando a estimativa pontual reflete um efeito trivial ou sem efeito (ou seja, sem benefício e sem dano) e o IC inclui a possibilidade de um benefício substancialmente (ou seja, possivelmente grande) importante.	
6	Quando a estimativa pontual sugere um benefício e o IC inclui uma possibilidade de dano importante.	
7	Quando a estimativa pontual sugere dano e o IC inclui uma possibilidade de benefício importante.	

Nota: As circunstâncias 1, 2, 3, 4 e 5 refletem cenários em que pelo menos um MID é indicado na avaliação. As circunstâncias 6 e 7 refletem cenários em que pelo menos um MID não é estabelecido.
IC = intervalo de confiança; MID = diferença minimamente importante.

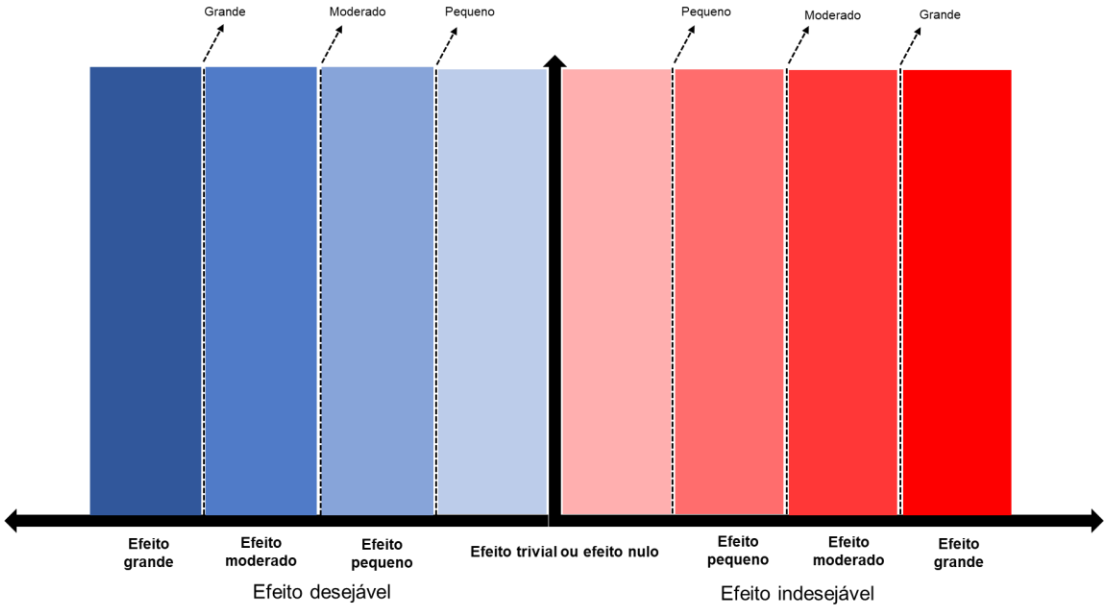
Fonte: Adaptado de Zeng et al. (58).

Quadro 23 – Passos para avaliar imprecisão utilizando uma abordagem parcialmente contextualizada	
Passo 1	Definir o desfecho como dicotômico ou contínuo.
Passo 2	Definir os limiares de interesse absolutos considerados pequenos/triviais, moderados ou grandes para cada desfecho em saúde (atentar que desfechos como mortalidade podem apresentar efeitos desejáveis caso reduzam e indesejáveis caso aumentem). Para a identificação desses limiares, recomenda-se: avaliação por outros tomadores de decisão, consenso pelas partes interessadas relevantes, evidência empírica que integre o tamanho de efeito absoluto e a importância relativa dos resultados ou, se nada mais estiver disponível, consideração da melhor opinião de especialistas no assunto. O método escolhido para a definição dos limiares deve ser mencionado.
Passo 3	Definir qual será o alvo desejado para basear a avaliação em relação aos limiares identificados no passo 2. Ou seja, decidir qual será a posição de referência para a avaliação do grau de imprecisão da estimativa pontual: entre dois limiares (por exemplo, se o objetivo a ser alcançado deve estar entre o limiar de efeito pequeno desejável e efeito pequeno indesejável ou entre o limiar de efeito pequeno e efeito moderado), acima de algum dos limiares de efeito grande ou acima ou abaixo de um limiar específico.
Passo 4	Calcular a estimativa do efeito absoluto para o conjunto de evidências do desfecho de interesse, além do IC desse efeito com base no risco basal e o efeito relativo ou, se relevante, a estimativa do risco diferencial entre os estudos da metanálise.
Passo 5	Determinar quantos limiares são cruzados pelo IC, considerando que a linha de efeito nulo não é considerada um limiar.

Passo 6	Avaliar o rebaixamento da certeza de evidência por imprecisão conforme quantos limiares forem incluídos no intervalo de confiança. A cada limiar incluído, um nível da certeza de evidência deve ser rebaixado.
Passo 7 (opcional)	Quando o efeito é grande (isto é, a estimativa pontual encontra-se acima do limiar definido como efeito grande) e baseado em um número aparentemente pequeno de eventos ou participantes, deve-se considerar o uso da abordagem RIS. A avaliação, de acordo com o RIS, verifica em que medida o tamanho da amostra da RS corresponde ao tamanho amostral necessário, penalizando por imprecisão à medida que o RIS não é alcançado.Caso contrário, deve-se finalizar a avaliação no passo 6. Quando o efeito for nulo ou trivial (ou seja, a estimativa pontual está dentro dos limiares de efeito pequeno ou nenhum efeito), verifique se o RIS para efeito nulo ou trivial foi atingido pela RS para, por exemplo, avaliar uma equivalência entre as intervenções ou considerar se um novo rebaixamento da certeza é necessário. Caso contrário, finalizar a avaliação no passo 5.
IC = intervalo de confiança; RIS = tamanho da informação de revisão; RS = revisão sistemática.	

Fonte: adaptado de Schunemann et al. (59).

Figura 16 – Estrutura ilustrativa de descrição dos limiares e intervalos referente aos efeitos triviais, pequenos, moderados e grandes



Fonte: Adaptado de Schünemann et al. (59).

3.3.5 Viés de publicação

- Evidências empíricas sugerem que, em geral, estudos estatisticamente significativos têm maior probabilidade de serem publicados do que estudos sem significância estatística (chamados de estudos “negativos”).
- Revisões sistemáticas precoces, realizadas quando poucos estudos iniciais estão disponíveis, podem superestimar a estimativa de efeito, uma vez que estudos “negativos” geralmente levam mais tempo para serem publicados (*lag-time bias*). Estudos iniciais com resultados positivos, especialmente com pequeno tamanho amostral, devem ser considerados suspeitos.
- A não publicação de estudos com resultados “negativos” parece ser uma prática frequente. O pesquisador deve suspeitar de viés de publicação se os estudos forem uniformemente pequenos, principalmente se a maioria possuir conflitos de interesse importantes (por exemplo, patrocínio da indústria farmacêutica).
- A avaliação empírica do padrão de resultados (por exemplo, gráfico de funil) pode sugerir viés de publicação, mas deve ser interpretada com cautela.

O viés de publicação ocorre quando os efeitos de uma determinada intervenção sobre um desfecho específico são sistematicamente superestimados ou subestimados devido à publicação seletiva dos estudos incluídos na síntese de evidências (63) ou não publicação de estudos completos (64). Do ponto de vista estatístico, estudos que relatam resultados significativos são mais propensos a serem aceitos para publicação nos periódicos do que os estudos que relatam resultados não significativos (65-68). Uma RS do grupo Cochrane demonstrou que ECR com resultados nulos ou negativos levaram, em média, cerca de 1 ano a mais para serem publicados do que ECR com resultados positivos (69).

Estudos com resultados negativos realizados em países onde a língua nativa não é o inglês tendem a ser submetidos a periódicos locais, não sendo facilmente localizados nas estratégias de buscas (70, 71). Também tendem a ser publicados em documentos associados à “literatura cinza” (teses, capítulos de livros, resumos), que normalmente é negligenciada em revisões sistemáticas que não realizam uma pesquisa abrangente (72). Sendo assim, se os autores não implementarem técnicas de busca rigorosas, será difícil julgar o domínio viés de publicação, uma vez que os estudos podem não ser identificados devido ao viés de publicação ou ao esforço insuficiente dos autores para identificá-los.

A confiança das estimativas pode ser reduzida em casos em que há suspeita considerável de viés de publicação (63). Ver Quadro 24 para maiores detalhes.

Quadro 24 – Possíveis fatores relacionados à existência de viés de publicação ao longo dos processos de publicação (41, 63)

Estudos-piloto e preliminares: pequenos estudos com maior probabilidade de serem "negativos" (por exemplo, aqueles com hipóteses descartadas ou fracassadas) permanecem não publicados; algumas empresas os classificam como informações privadas.

Conclusão do relatório: os autores decidem que relatar um estudo "negativo" não é interessante e não investem tempo e esforços necessários para a submissão.

Seleção do periódico para publicação: os autores decidem submeter o relatório "negativo" para avaliação em periódico não indexado, de língua não inglesa ou de circulação limitada.

Consideração editorial: o editor decide que o estudo "negativo" não justifica a revisão por pares para uma possível publicação e rejeita o manuscrito na avaliação preliminar.

Revisão por pares: os revisores concluem que o estudo "negativo" não contribui para o campo e recomendam a rejeição do manuscrito pela revista. O autor desiste ou muda para um periódico de menor impacto. Esse processo também contribui para um retardo nas publicações dos relatos negativos.

Revisão e reenvio do autor: os autores de um manuscrito com uma ou mais rejeições decidem renunciar à submissão do estudo "negativo" e não tentam submetê-lo novamente a outro periódico.

Publicação do relatório: o periódico retarda a publicação do estudo pelo fato de ser "negativo".

Fonte: Guyatt et al.(73).

O risco de viés de publicação pode ser maior nas revisões sistemáticas que se propõem a incluir estudos observacionais, inclusive estudos de prognóstico, do que nas revisões de ECR (74, 75). Isso pode ocorrer, principalmente nos cenários em que os estudos observacionais forem conduzidos automaticamente a partir de registros de pacientes ou prontuários médicos locais; nesses casos, os estudos normalmente são publicados em periódicos locais (71). Assim, será mais difícil julgar se os estudos

observacionais que apareceram na literatura representam a totalidade ou uma fração dos estudos realizados. Usualmente, essa fração ainda poderá representar os estudos com resultados mais “interessantes” ou significativos (63).

Estudos com tamanhos de amostra pequenos usualmente estão mais propensos a não serem publicados, serem publicados mais tardiamente ou a serem publicados em periódicos não indexados (67). Podem ocorrer discrepâncias entre os resultados das metanálises com pequenos estudos e os grandes ECR subsequentes em quase 20% das vezes, e o viés de publicação pode ser um dos principais contribuintes para essa discrepância (76). Assim, deve-se suspeitar de viés de publicação quando a síntese de evidências for composta por um pequeno número de estudos baseados em amostras menores, em sua maioria positivos, especialmente se forem patrocinados pela indústria ou se os pesquisadores apresentarem conflitos de interesse (77, 78).

Existem recursos que podem ser utilizados como suporte metodológico para avaliar o viés de publicação em revisões sistemáticas, incluindo inspeção visual e teste de assimetria no gráfico de funil (70, 79). Como regra geral, tanto a inspeção visual quanto o gráfico de funil devem ser utilizados para avaliar o domínio viés de publicação em metanálises que incluíram pelo menos 10 estudos (em alguns casos, pode-se utilizar a partir de cinco) (80). O *trim and fill* é outro conjunto de testes estatísticos que pode ser utilizado para julgar o viés de publicação, pois avalia por meio da remoção e imputação, o efeito que pequenos estudos positivos teriam em contrapartida a um estudo negativo (81). As mesmas explicações alternativas para a assimetria observadas no gráfico de funil se aplicariam para a avaliação do resultado desses testes (41). É importante destacar que, mesmo que seja detectada uma assimetria, ela pode não ser resultado de um viés de publicação (79, 82, 83). Em estudos menores, efeitos superestimados podem gerar um gráfico de funil assimétrico, que poderia ser explicado por outras limitações além do viés de publicação, como população do estudo ou baixa qualidade metodológica, por exemplo. Recomenda-se o uso de mais de uma ferramenta para auxiliar a tomada de decisão sobre a certeza da evidência no domínio viés de publicação (63).

Independentemente de quais testes os elaboradores de revisões sistemáticas ou diretrizes clínicas utilizem como suporte para o julgamento, eles devem estar cientes de que os testes podem estar propensos a erros e que os resultados devem

ser interpretados com cautela. Tendo em vista as dificuldades de interpretação e certeza sobre a presença de viés de publicação, o GRADE sugere que a certeza da evidência seja rebaixada em, no máximo, um nível considerando o desenho do estudo, tamanho amostral do estudo, viés de retardo nas publicações, abrangência na estratégia de busca e assimetria no gráfico de funil.

3.4 Domínios que podem elevar a certeza no conjunto final de evidências

- Três fatores podem elevar a certeza da evidência em um ou dois níveis (grande magnitude de efeito, gradiente dose-resposta e fatores de confusão residuais), os quais geralmente se aplicam a estudos observacionais.
- O sistema GRADE considera magnitude de efeito elevada quando for observado um risco relativo (RR) >2 ou <0,5 e muito grande quando for observado um RR >5 ou <0,2.
- A presença de gradiente dose-resposta é um achado que reforça a probabilidade da ocorrência de relação causa-efeito, aumentando, assim, a confiança da estimativa.
- A conclusão de que fatores de confusão residuais apoiam ainda mais as inferências sobre o efeito do tratamento também pode elevar a certeza da evidência.

O sistema GRADE possui três critérios que podem aumentar a certeza da evidência: magnitude de efeito, gradiente dose-resposta e fatores de confusão residuais que aumentam a confiança da estimativa. Circunstâncias nas quais a certeza da evidência pode ser elevada ocorrem com pouca frequência e são principalmente relevantes para estudos observacionais (inclusive estudos de coorte, caso-controle, estudos antes e depois e estudos de séries temporais) e para estudos experimentais ou de intervenção não randomizados (por exemplo, fornecendo tratamento a um dos dois grupos pareados). Embora seja teoricamente possível avaliar os resultados de ECRs, ainda não foi encontrado um exemplo adequado de tal caso. As seções a seguir discutem em detalhes os três fatores que permitem aumentar a confiança em uma estimativa de efeito.

3.4.1 Grande magnitude de efeito

Estudos observacionais metodologicamente bem executados que produzem estimativas grandes ou muito grandes e apresentam uma magnitude de efeito consistente permitem maior confiança nos resultados. Em situações como essa, mesmo que o desenho do estudo superestime os efeitos da intervenção, é improvável que explique todo o benefício ou dano aparente (84). Dessa maneira, por meio do sistema GRADE, a certeza da evidência pode ser elevada quando estudos com poucos vieses demonstram um efeito grande, já que a certeza do efeito observado é maior.

A decisão sobre elevar a certeza de uma evidência devido a efeitos grandes ou muito grandes deve considerar não apenas a estimativa pontual, mas também a precisão (amplitude do IC) em torno desse efeito. Por exemplo, um pequeno número de eventos (2/100 *versus* 10/100) para um determinado desfecho levaria a uma grande redução na estimativa de efeito, porém com um IC amplo, o que poderia ter ocorrido pelo acaso, de modo que um ou outro evento poderia modificar essa estimativa de efeito.

Com base em estudos de modelagem que fornecem estimativas de magnitudes de efeito muito improváveis de serem explicada por viés (85), o sistema GRADE define grande efeito quando for observado um RR >2 ou RR <0,5, o que pode elevar a evidência em um nível, e considera magnitude de efeito muito grande quando for observado um RR >5 ou RR <0,2 (86), o que pode elevar a evidência em dois níveis (Quadro 25). Essas estimativas são aplicáveis para RR e razão de risco *(hazard ratio,* HR). Para a razão de chances (*odds ratio*, OR), é necessário fazer uma conversão para identificar o RR correspondente; para dados contínuos, podem ser utilizadas as definições de tamanho de efeito de Cohen para considerar o que seria um tamanho grande de efeito (62).

Quadro 25 – Consequências da elevada magnitude do efeito na avaliação da qualidade da evidência

Tamanho do efeito	Interpretação	Consequência
RR ≥ 2 e ≤ 5	Magnitude de efeito grande	↑ 1 nível

ou RR ≥ 0,2 e ≤ 0,5 DMP > 0,8 ou DMP < -0,8		
RR > 5 ou RR < 0,2	Magnitude de efeito muito grande	↑ 2 níveis

RR = risco relativo; DMP: diferença de médias padronizada.
Fonte: adaptado de Guyatt et al. (87).

Os elementos relacionados à elevação da certeza da evidência devido à magnitude do efeito incluem a rapidez da resposta ao tratamento e o conhecimento da trajetória da condição (88). Por exemplo, há confiança de que a cirurgia de artroplastia do quadril tem um grande efeito não apenas por causa do tamanho da resposta ao tratamento, mas também porque a história natural da osteoartrite do quadril é uma deterioração progressiva que a cirurgia reverte rápida e uniformemente. A rapidez da resposta, comparada com a trajetória conhecida da condição, também pode ser considerada (e calculada) (88) como uma grande magnitude de efeito.

Adicionalmente, evidências indiretas podem dar suporte para grandes efeitos do tratamento. Por exemplo, não há evidência, por meio de ECR, da necessidade de utilizar anticoagulação oral em pacientes que fazem uso de válvulas cardíacas mecânicas, mas evidências de estudos observacionais sugerem um grande efeito da anticoagulação oral na diminuição de eventos tromboembólicos (89, 90). Evidências indiretas suplementares de ECR que demonstraram grandes reduções no risco de trombose com anticoagulação em condições análogas, como fibrilação atrial, aumentam ainda mais a confiança no efeito benéfico da anticoagulação (91).

Outro exemplo que permite inferir uma associação forte sem um estudo comparativo é descrito a seguir. Na avaliação do impacto da realização de colonoscopia de rotina ou nenhum rastreamento para câncer de cólon sobre a taxa

de perfuração associada a colonoscopia, uma grande série de pacientes submetidos à colonoscopia fornecerá evidências de alta qualidade sobre o risco de perfuração associado à colonoscopia. Quando as taxas de controle estão próximas de 0 (ou seja, há certeza de que a incidência de perfuração espontânea do cólon em pacientes não submetidos à colonoscopia é muito baixa), séries de casos com um número grande de pacientes podem fornecer evidências de alta qualidade, permitindo, assim, inferir uma forte associação, mesmo com um número limitado de eventos.

Porém, quando os resultados são subjetivos, é importante ter cautela ao considerar a elevação da certeza da evidência devido aos grandes efeitos observados. Isso é especialmente importante quando o conjunto de evidências apresenta risco de viés (por exemplo, estudos não cegos), imprecisão e viés de publicação.

3.4.2 Gradiente dose-resposta

A presença de gradiente dose-resposta é um achado que reforça a probabilidade da ocorrência de relação de causa e efeito. Nos casos em que há um consistente aumento do efeito em associação a um aumento na exposição, a evidência do efeito torna-se mais robusta.

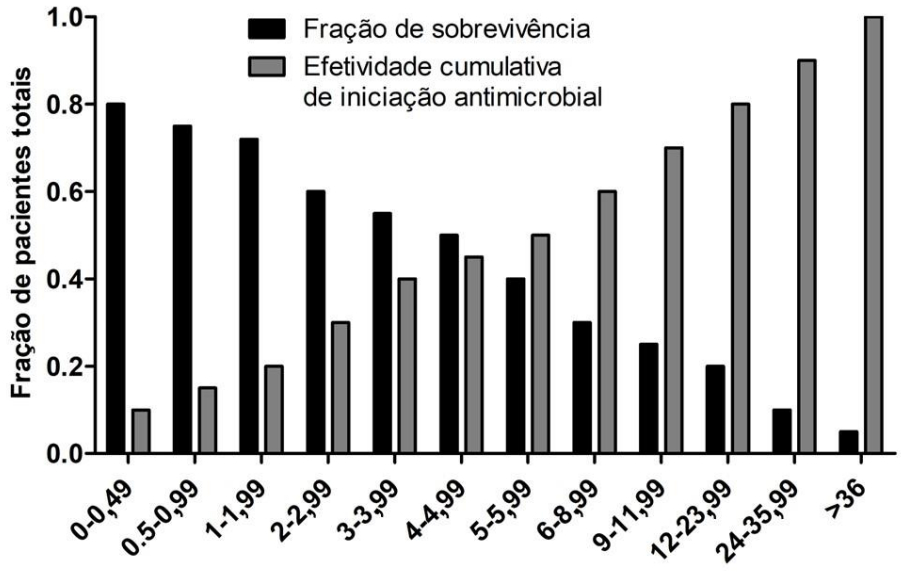
O aumento ou a diminuição de determinada exposição (por exemplo, dose, frequência ou consumo) leva também a variações nos desfechos avaliados, o que aumenta a confiança de que existe uma relação causal entre os fatores. Nesse caso, o GRADE sugere elevar a certeza da evidência em um nível.

Por exemplo, a maioria das evidências de gradiente dose-resposta no tratamento de doenças alérgicas provém de ensaios randomizados bem realizados que não requerem atualização. No entanto, não há estudos de intervenções destinadas à redução da exposição do tabagismo ativo que avaliaram o desenvolvimento de asma ou sibilos em crianças. Por outro lado, estudos observacionais encontraram um risco aumentado de desenvolvimento de doenças sibilantes na primeira infância em crianças expostas a tabagismo passivo oriundo dos pais. O gradiente dose-resposta observado justificaria a elevação da certeza da evidência (92).

Outro exemplo é a administração de antibióticos em pacientes com sepse e hipotensão (Figura 17). Há um grande aumento absoluto na mortalidade a cada hora

de atraso na administração do antibiótico. Essa relação dose-resposta aumenta a confiança de que o efeito sobre a mortalidade é real e substancial, elevando a certeza da evidência (93).

Figura 17 – Gradiente dose-resposta associado ao tempo até a administração de antibióticos em pacientes com choque séptico



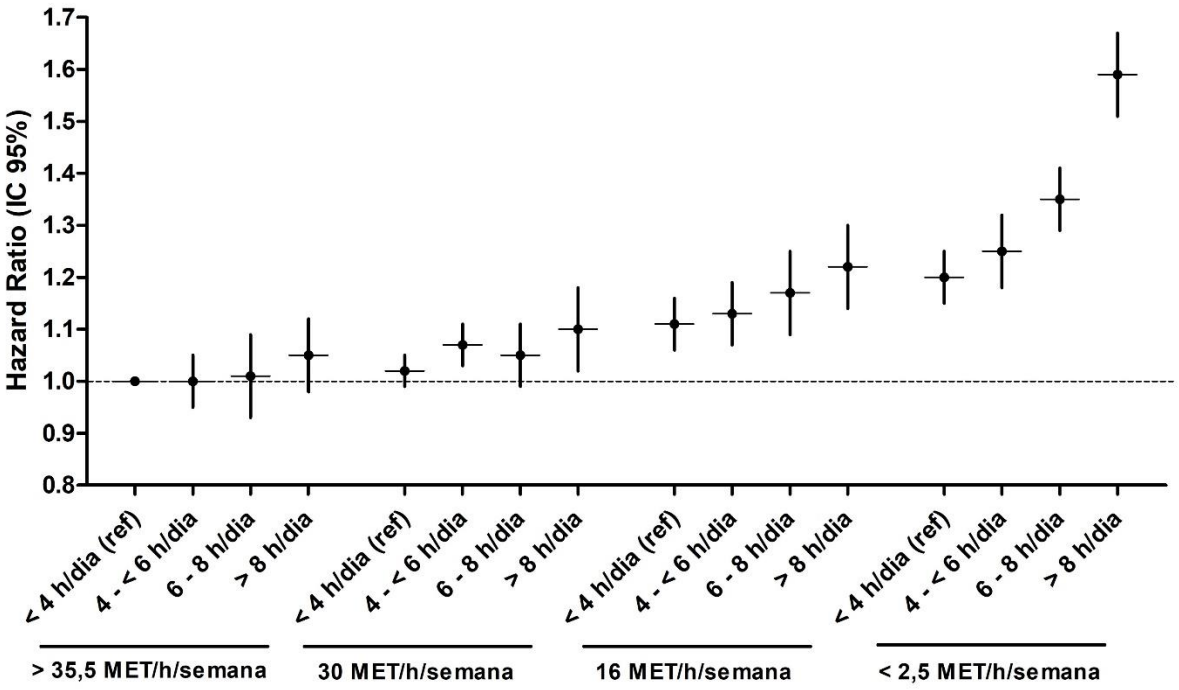
Nota: iniciação antimicrobiana efetiva cumulativa após o início de hipotensão associada a choque séptico e sobrevida. O eixo X representa o tempo em horas após a primeira documentação de hipotensão associada a choque séptico. As barras pretas representam a fração de pacientes que sobreviveram à alta hospitalar para terapia efetiva iniciada dentro do intervalo de tempo determinado. As barras cinzas representam a fração cumulativa de pacientes que receberam antimicrobianos eficazes em um determinado momento.
Fonte: Adaptada de Kumar et al. (93).

Adicionalmente, outro exemplo é o de estudos observacionais avaliando o efeito da radioterapia no desfecho desenvolvimento de neoplasias do sistema nervoso central (SNC) em crianças com leucemia linfoblástica aguda. Observou-se que pacientes que não haviam realizado radioterapia tinham uma incidência de 1% de neoplasia do SNC. Pacientes que realizaram radioterapia e receberam uma dose de radiação de 12 Gy tinham uma incidência de 1,6%, enquanto aqueles que foram expostos a uma radiação de 18 Gy tinham uma incidência de 3,3%. Dessa maneira, observou-se que, quanto maior o nível de radiação, maior a incidência de neoplasias

do SNC, aumentando a certeza de que radioterapia para leucemia linfoblástica aguda leva ao desenvolvimento de neoplasias do SNC.

Por fim, temos o exemplo da associação entre tempo diário sentado e atividade física em relação ao desfecho mortalidade por todas as causas. Uma RS com mais de 1 milhão de homens e mulheres observou uma clara associação entre mortalidade por todas as causas e aumento do tempo sentado combinado com níveis mais baixos de atividade física (Figura 18) (94).

Figura 18 – Gradiente dose-resposta da associação entre tempo diário sentado e atividade física em relação ao desfecho mortalidade por todas as causas



Fonte: Adaptada de Ekelund et al. (94).

Antes de se considerar aumentar a certeza da evidência baseando-se na existência de um gradiente dose-resposta, é importante verificar a credibilidade do gradiente por meio de cinco critérios descritos a seguir (95). Os critérios 1 e 2 são considerados os mais críticos, enquanto que as informações necessárias para os critérios 3, 4 e 5 já foram coletadas nos demais domínios do GRADE.

1) Adequação da análise: o gradiente dose-resposta pode ser identificado por meio de metanálise de dose-resposta (Figura 19), ainda pouco comuns na literatura, ou por análise de subgrupo (Figura 20). Análises com três ou mais subgrupos (Figura

20A) fornecem informações mais confiáveis que análises com apenas dois subgrupos (Figura 20B), uma vez que a probabilidade de a relação linear estar relacionada ao acaso é menor na existência de mais subgrupos. Em geral, a certeza da evidência não deve ser elevada quando o gradiente dose-resposta é fornecido a partir de apenas dois subgrupos (95).

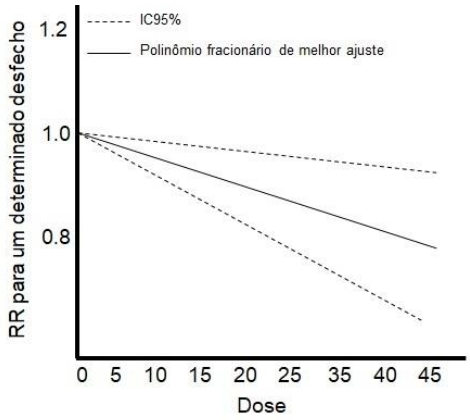
2) Existência de confundidores: a existência de confundidores pode causar um gradiente dose-resposta. Um exemplo muito conhecido envolve estudos realizados nos anos 1980 sobre gradiente dose-resposta entre o consumo diário de café e o risco de câncer de pâncreas. Esses estudos tinham o tabagismo como confundidor, que também apresenta um gradiente dose-resposta com risco de câncer de pâncreas, além de existir uma correlação positiva entre o consumo diário de cigarro e de café (95).

3) Viés ecológico: viés ecológico é caracterizado quando encontramos um gradiente dose-resposta entre os estudos, porém esse achado não se confirma nos estudos em si (Figura 21) (95).

4) Consistência entre os estudos: as informações para avaliar esse critério já devem ter sido coletadas para a avaliação do domínio inconsistência. O julgamento da concordância entre os estudos para avaliar a credibilidade do gradiente dose-resposta é intuitivo e não há um teste estatístico recomendado para isso (95).

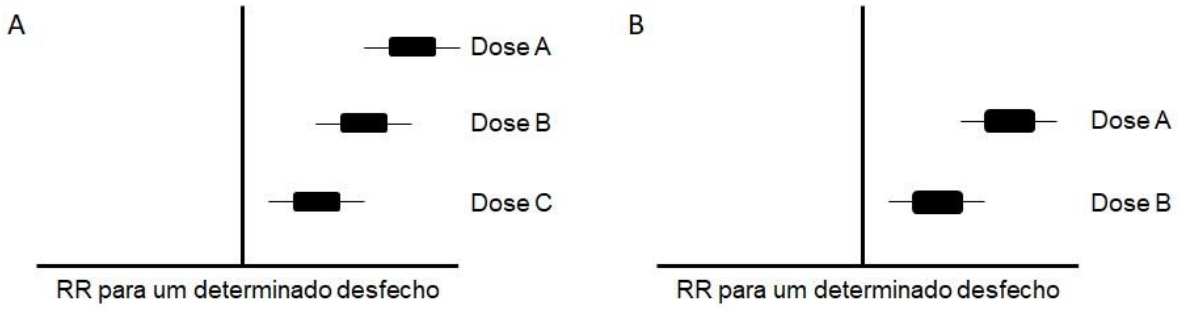
5) Apoio por evidência indireta: a existência de plausibilidade biológica para e de evidência externa aumentam a credibilidade do gradiente dose-resposta. Idealmente, esses itens devem ser considerados a priori (95).

Figura 19 – Metanálise de dose-resposta



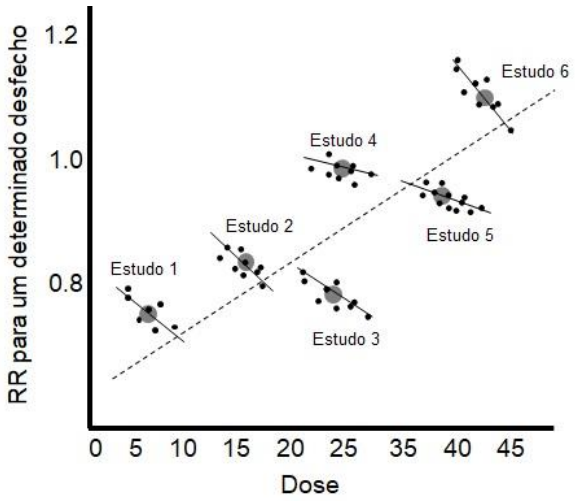
RR = risco relativo; IC95% = intervalo de confiança de 95%.
Fonte: elaboração própria.

Figura 20 – Análise de subgrupos para avaliação da existência de gradiente dose-resposta



RR = risco relativo.
Fonte: elaboração própria.

Figura 21 – Viés ecológico



RR = risco relativo.
Fonte: elaboração própria.

3.4.3 Fatores de confusão residuais em direção oposta

Os confundidores residuais podem se comportar de forma diferente em relação aos desfechos avaliados. Em sua maioria, eles se comportam de forma neutra ou favorecendo os achados encontrados. Quando são identificados confundidores residuais que, em hipótese, deveriam favorecer alguma intervenção, mas não o fazem

(por exemplo, quando uma intervenção se mostra efetiva mesmo em pacientes com pior prognóstico), a confiança da estimativa aumenta. É preciso se atentar a esse fator em dois casos: quando esses fatores subestimam a estimativa de efeito, mas é observada uma importante associação, ou quando os fatores de confusão residuais superestimaram a estimativa de efeito, mas não é encontrada associação.

Em raras ocasiões, todos os vieses plausíveis podem estar trabalhando para subestimar o verdadeiro efeito do tratamento. Por exemplo, uma RS de estudos observacionais que incluíram 38 milhões de pacientes demonstrou maiores taxas de mortalidade em hospitais privados com fins lucrativos em comparação a hospitais privados sem fins lucrativos (96). Uma fonte potencial de viés é que hospitais sem fins lucrativos tendem a atender pacientes mais graves, o que colabora para uma taxa de mortalidade maior. Entretanto, observou-se o contrário. Isso significa que o confusão residual pode estar reduzindo o tamanho do efeito encontrado, aumentando a confiança na menor mortalidade, já que o estudo foi realizado em hospitais privados e sem fins lucrativos.

Existe uma situação paralela, em que estudos observacionais não conseguem demonstrar uma associação de que todos os vieses plausíveis teriam aumentado o efeito da intervenção. Essa situação geralmente surgirá na exploração de efeitos nocivos aparentes. Por exemplo, a fenformina é um medicamento hipoglicemiante com potencial para causar acidose láctica. Por esse motivo, suspeitava-se que a metformina, hipoglicemiante oral da mesma classe farmacológica, pudesse causar efeito semelhante. Devido ao alerta desse potencial efeito adverso, os médicos eram provavelmente mais propensos a monitorar e reportar acidose láctica associada à metformina do que a outros medicamentos, causando potencial viés de aferição. Contudo, estudos observacionais com forte rigor metodológico não demonstraram associação entre esse efeito adverso e doses terapêuticas do fármaco (97). O fato do potencial viés de aferição não mostrar uma associação positiva reforça as conclusões de que não há risco aumentado de acidose láctica com o uso de metformina em doses terapêuticas, elevando a certeza da evidência.

3.4.4 Considerações sobre o aumento do nível de evidência

A utilização desses fatores em situações nas quais a certeza da evidência já havia sido reduzida previamente precisa ser avaliada com cautela. Em geral, não é recomendado elevar a certeza da evidência nessas situações, sendo esses critérios aplicáveis principalmente para o conjunto de evidências oriundo de estudos observacionais com resultados substanciais.

Em especial, não se deve elevar a certeza da evidência na presença de limitações metodológicas, uma vez que a força da associação de um efeito incerto estaria sendo elevada devido à alta probabilidade de viés. Além disso, é desaconselhado elevar a certeza da evidência por magnitude de efeito grande quando há medidas de efeito imprecisas, como, por exemplo, IC amplos.

Assim, o uso desses critérios em ensaios clínicos é limitado. Como exceção da regra acima, os critérios apresentados podem ser aplicados a estudos randomizados na presença de limitações metodológicas quando essas se referem a problemas relacionados à randomização e/ou alocação dos participantes. Nesse caso, o nível de evidência iniciaria como baixo, podendo o conjunto de evidências ser avaliado como estudos observacionais.

4. Síntese de evidências

- Recomenda-se a apresentação da síntese de evidências no formato de uma tabela,
- É importante que a síntese da evidência inclua elementos básicos, como desfechos, estudos incluídos, estimativas de efeito, efeitos absolutos potenciais e a graduação da certeza de evidência.
- Apenas os desfechos críticos e importantes devem ser incluídos na tabela, principalmente no processo de desenvolvimento de recomendações.

Após a avaliação dos domínios propostos pelo sistema GRADE, recomenda-se a realização da síntese de evidências, por ser uma forma transparente de apresentação da graduação da evidência, e da indicação dos motivos que levaram a determinado nível da certeza de evidência. Para isso, foram propostos dois formatos padronizados de tabela para a apresentação das evidências: sumário de resultados e perfil de evidências.

Na tabela do perfil de evidências, são apresentados os oito domínios da avaliação da certeza de evidência, associados ao sumário de resultados. Como exemplo, a Figura 22 apresenta uma das versões mais utilizadas. Sempre que possível, devem ser apresentadas medidas de efeitos absolutos e relativos para auxiliar na tomada de decisão. Quando as medidas não estiverem disponíveis, o resultado deve ser apresentado em forma descritiva. Esse formato é indicado na elaboração de diretrizes clínicas, pois permite uma visualização mais fácil do processo de avaliação.

Já a tabela de sumário dos resultados (*summary of findings* [SoF]) apresenta uma avaliação geral da certeza de evidência de forma concisa. Como exemplo, a Figura 23 apresenta uma das versões mais utilizadas. Devido à evolução do método, diferentes formatos de tabela SoF e perfil de evidências estão disponíveis, ficando ao critério do grupo elaborador a escolha da versão mais adequada para a compreensão e interpretação dos resultados (98). Em ambas as tabelas, o julgamento sobre os determinantes apresentados deve ser realizado de forma transparente, com indicação dos motivos que levaram a um determinado nível de evidência e os comentários adicionais, na nota de rodapé.

De modo geral, os dois formatos apresentam sete elementos básicos (99):

1. desfechos;

2. estudos incluídos;

3. estimativa relativa de efeito;

4. estimativa do efeito absoluto basal;

5. estimativa do efeito absoluto da intervenção (ou diferença de risco em relação ao grupo controle);

6. graduação da certeza de evidência;

7. explicações (explicações e informações adicionais) – apresentadas em notas de rodapé.

Figura 22 – Exemplo de apresentação detalhada dos resultados utilizando uma tabela de perfil de evidências através da ferramenta GRADEpro

Autor(es): Ministério da Saúde
Pergunta: Antagonistas da aldosterona comparado a placebo em pacientes com insuficiência cardíaca?
Contexto: Sistema Único de Saúde.
Bibliografia:

Avaliação da Certeza							Nº de pacientes		Efeito		Certeza da Evidência
Nº dos estudos	Delineamento do estudo	Risco de viés	Inconsistência	Evidência indireta	Imprecisão	Outras considerações	Antagonistas da aldosterona	Placebo	Relativo (95% CI)	Absoluto (95% CI)	
Mortalidade (seguimento: média 18 meses)											
4	ensaios clínicos randomizados	não grave[a]	não grave	não grave[b]	não grave	nenhuma	939/5730 (16.4%)	1165/5769 (20.2%)	**RR 0.81** (0.74 para 0.88)	**38 menos por 1.000** (de 53 menos para 24 menos)	⊕⊕⊕⊕ Alta

IC: Intervalo de confiança; **RR:** Risco relativo
Explicações
a. Dois estudos (AREA IN-CHF, 2000 e RALES, 1999) não reportam o modo de cegamento dos pacientes e como foi realizada a alocação sigilosa.
b. Em um estudo (RALES 1999), 29% da amostra apresentava classe IV (NYHA).

Fonte: Ferramenta *online* GRADEpro, disponível em https://www.gradepro.org/.

Figura 23 – Exemplo de apresentação dos resultados resumidos utilizando a ferramenta GRADEpro

Antagonistas da aldosterona comparado a placebo em pacientes com insuficiência cardíaca?

Paciente ou população: Pacientes com insuficiência cardíaca.
Contexto: Sistema Único de Saúde.
Intervenção: Antagonistas da aldosterona.
Comparação: Placebo.

Desfechos	Nº de participantes (estudos) Seguimento	Certeza da evidência (GRADE)	Efeito relativo (IC 95%)	Efeitos absolutos	
				Risco com placebo	Diferença de risco com antagonistas da aldosterona
Mortalidade	11499 (4 ECRs) média 18 meses	⊕⊕⊕⊕ Alta[a,b]	**RR 0.81** (0.74 para 0.88)	202 por 1.000	**38 menos por 1.000** (53 menos a 24 menos)

* **O risco no grupo de intervenção** (e seu intervalo de confiança de 95%) é baseado no risco assumido do grupo comparador e o **efeito relativo** da intervenção (e seu IC 95%).
IC: Intervalo de confiança; **RR:** Risco relativo

Nível da graduação da evidência conforme o GRADE
Certeza alta: estamos muito confiantes de que o verdadeiro efeito se aproxima da estimativa de efeito.
Certeza moderada: estamos moderadamente confiantes na estimativa de efeito: é provável que o efeito verdadeiro seja próximo do efeito estimado, mas existe a possibilidade de que seja substancialmente diferente.
Certeza baixa: nossa confiança na estimativa de efeito é limitada: o efeito verdadeiro pode ser substancialmente diferente do efeito estimado.
Certeza muito baixa: temos muito pouca confiança na estimativa de efeito: o verdadeiro efeito provavelmente será substancialmente diferente do efeito estimado.

Explicações

a. Dois estudos (AREA IN-CHF, 2000 e RALES, 1999) não reportam o modo de cegamento dos pacientes e como foi realizada a alocação sigilosa.
b. Em um estudo (RALES 1999), 29% da amostra apresentava classe IV (NYHA).

Fonte: Ferramenta *online* GRADEpro, disponível em https://www.gradepro.org/.

Desfechos

Apenas os desfechos críticos e importantes devem ser incluídos na tabela, principalmente no processo de desenvolvimento de recomendações. Conforme definido pelo GRADE *Working Group*, é aconselhado apresentar até sete desfechos para diretrizes clínicas, escolhidos de acordo com os mais importantes para a tomada de decisão.

Estudos incluídos

Para cada desfecho incluído na tabela, devem ser apresentados o delineamento, o número de estudos incluídos e o número total de participantes, bem como as referências correspondentes.

Estimativas de efeito

A estimativa de efeito é o resultado final da síntese de evidências, idealmente obtida por meio de metanálise de uma RS. Para desfechos binários, sugere-se a apresentação do risco relativo (*relative risk* [RR]), mas também podem ser apresentadas a razão de chances (*odds ratio* [OR]), a taxa de risco (*hazard ratio* [HR]) ou a razão da taxa (*rate ratio*); para desfechos contínuos, as medidas mais utilizadas são a diferença de médias (*mean difference* [MD]) e a diferença padronizada de médias (*standardized mean difference* [SMD]). Quando a estimativa sumária não estiver disponível, a informação pode ser descrita de forma narrativa e deve incluir a estimativa de efeito dos estudos individuais.

Efeitos absolutos potenciais

O efeito absoluto representa em números absolutos o impacto do risco em determinada escala populacional (efeito do risco para cada 1.000 pacientes, por exemplo). A medida do efeito absoluto é calculada a partir da estimativa do efeito relativo e é afetada, principalmente, pelo risco basal da condição de saúde avaliada. Esse processo de obtenção do efeito absoluto pode ser realizado a partir do programa online GRADEpro (<https://www.gradepro.org/>), desenvolvido pelo GRADE *Working Group* e disponibilizado com orientação em português gratuitamente.

Para a tomada de decisão, a medida do impacto absoluto é mais importante do que a do efeito relativo, pois um mesmo efeito relativo pode influenciar de forma diferente no número de casos, conforme o risco basal da condição. A medida do efeito absoluto deve ser relatada para o grupo controle (basal) e para o grupo intervenção, conforme descrito nos próximos itens.

Estimativa do efeito absoluto basal

A estimativa do risco basal deve ser, idealmente, realizada para a população para a qual a avaliação foi proposta, podendo ser obtida a partir de estudos de prognóstico e de incidência. Na ausência dessa informação, pode-se utilizar como

estimativa a taxa de eventos no grupo controle dos ensaios clínicos randomizados utilizados na síntese de evidências.

Caso a intervenção tenha sido proposta para diferentes cenários, podem ser apresentadas diferentes estimativas de risco basal e pode ser calculado o impacto para cada um dos cenários. Por exemplo, uma diretriz de insuficiência cardíaca (IC) (100) que avalia a incorporação de uma nova estratégia diagnóstica pode considerar a probabilidade pós-teste do método já disponibilizado no sistema (nesse caso, o ecocardiograma), associado ou não a um novo método (nesse caso, a porção N-terminal do peptídeo natriurético tipo B [NT-proBNP]), em três cenários distintos de risco basal conforme a probabilidade pré-teste de um diagnóstico positivo de IC: baixa (20%), média (50%) e alta probabilidade (90%) (Quadro 26). Nesse caso, o valor absoluto das medidas de cada teste variará conforme cada cenário. Nesse processo de incorporação, os dados podem auxiliar na avaliação do efeito de cada estratégia diagnóstica de cada cenário sobre a redução da fila de espera por exames mais complexos e o tempo até início do tratamento da condição.

Quadro 26 – Exemplo de apresentação da síntese de evidências

Considerando os dados apresentados abaixo, pode-se observar que, entre aqueles que possuem baixa probabilidade de insuficiência cardíaca (IC), 120 em cada 1.000 pacientes que realizam o exame de ecocardiografia seriam diagnosticados equivocadamente e poderiam receber tratamento desnecessário (100). Além disso, 38 pacientes com IC não seriam diagnosticados e não receberiam o tratamento necessário. Por outro lado, na realização de ecocardiografia apenas em pacientes com níveis de NT-proBNP elevados (≥ 125 pg/mL), 61 em cada 1.000 pacientes investigados seriam erroneamente diagnosticados com IC (59 casos a menos do que na simulação anterior) e 51 pacientes não seriam diagnosticados (13 a mais do que na simulação anterior). Em relação ao uso de recursos, essa estratégia evitaria a realização de 408 exames de ecocardiografia a cada 1.000 pacientes em investigação.

Probabilidade pós-teste de insuficiência cardíaca de acordo com o risco basal do paciente

	Probabilidade clínica		
	Baixa	**Média**	**Alta**
Probabilidade pré-teste	20%	50%	90%
NT-proBNP < 125 pg/mL	4%	14%	60%
NT-proBNP ≥ 125 pg/mL	31%	64%	94%
NT-proBNP ≥ 400 pg/mL	44%	76%	97%
Ecocardiografia negativa	5%	18%	67%
Ecocardiografia positiva	57%	84%	98%
NT-proBNP ≥ 125 pg/mL e eco negativa	9%	28%	78%
NT-proBNP ≥ 125 pg/mL e eco positiva	71%	91%	99%

Fonte: Brasil, Ministério da Saúde (100).

Impacto de diferentes exames na acurácia diagnóstica e realização de ecocardiografia de acordo com o risco basal do paciente

		Efeito esperado por 1.000 pacientes					
		Paciente com IC	**VP**	**FP**	**FN**	**VN**	**Ecocardiografias necessárias**
Baixa probabilidade (20%)	**NT-proBNP** (**125** pg/mL)	200	184	408	16	392	0
	NT-proBNP (**400** pg/mL)	200	170	216	30	584	0
	Ecocardiografia	200	162	120	38	680	1.000
	Descartar dx se NT-proBNP < 125 pg/mL **Ecocardiografia se NT-proBNP ≥ 125** pg/mL	200	149	61	51	739	592
	NT-proBNP (**125** pg/mL)	500	460	255	40	245	0

Média probabilidade (50%)	**NT-proBNP** (**400** pg/mL)	500	425	135	75	365	0
	Ecocardiografia	500	405	75	95	425	1.000
	Descartar dx se NT-proBNP < 125 pg/mL **Ecocardiografia se NT-proBNP ≥ 125** pg/mL	500	373	38	127	462	715
Alta probabilidade (90%)	**NT-proBNP** (**125** pg/mL)	900	828	51	72	49	0
	NT-proBNP (**400** pg/mL)	900	765	27	135	73	0
	Ecocardiografia	900	729	15	171	85	1.000

Legenda: NT-proBNP = porção N-terminal do peptídeo natriurético tipo B; VP = verdadeiro-positivo; VN = verdadeiro-negativo; FN = falso-negativo; FP = falso-positivo.
Fonte: Brasil, Ministério da Saúde (100).

Fonte: elaboração própria.

Estimativa do efeito absoluto da intervenção

Para permitir uma comparação com a estimativa do efeito absoluto basal, ou seja, sem o uso da intervenção, o efeito absoluto da intervenção deve ser apresentado a partir da taxa de eventos com o uso da intervenção ou da redução absoluta de eventos com o uso da intervenção. A taxa de eventos é o produto da estimativa do risco basal (ou taxa de eventos no grupo controle) e do risco relativo. Os valores absolutos são apresentados para cada 100 ou cada 1.000 pacientes (multiplicando o valor encontrado por 10), a fim de facilitar o processo de formulação de recomendações. A partir dessas informações, é possível calcular a diferença de risco com a intervenção, que é o resultado da subtração do risco basal absoluto e da taxa absoluta de eventos da intervenção, desde que apresentem o mesmo denominador.

Seguindo outro exemplo da diretriz de IC (Anexo I. Exemplo do processo de obtenção da estimativa do efeito absoluto da intervenção), o uso de mineralocorticoides reduziu em 19% (RR = 0,81) a mortalidade em um seguimento médio de 18 meses, sendo 20,2% a taxa de eventos absolutos no grupo controle da população do estudo. Com o uso de mineralocorticoides, a taxa de eventos seria de aproximadamente 16,4% (20,2% x 0,81), representando uma redução absoluta de óbitos de 3,8% (ou seja, cerca de 38 eventos a cada 1.000 pacientes tratados ou 1 evento a cada 26 pacientes tratados). Em um outro cenário, com uma estimativa de risco basal para a população brasileira de 11% (taxa anual do Brasil, fonte DATASUS, 2016), a taxa de eventos com o uso do mineralocorticoides seria de 8,9% (11,0% x

0,81), representando uma redução de óbitos de 2,1% (cerca de 21 eventos a cada 1.000 pacientes tratados ou 1 evento a cada 48 pacientes tratados). É importante ressaltar que, nesse exemplo, a interpretação dos efeitos também deve considerar o tempo de avaliação do desfecho, pois as estimativas de risco utilizadas para comparação são obtidas em diferentes períodos de seguimento (18 meses e 1 ano, respectivamente).

Graduação da certeza de evidência

No contexto de tabelas de perfil de evidência, o julgamento sobre cada domínio do GRADE deve ser apresentado em colunas específicas, com uma legenda indicando os domínios que levaram ao rebaixamento ou aumento da certeza de evidência e o resultado geral da certeza de evidência. No contexto de tabelas SoF, deve ser apresentado o resultado final do nível de graduação da certeza de evidência para cada desfecho, com uma legenda explicando o motivo que levou ao rebaixamento ou aumento da certeza de evidência na avaliação dos domínios.

Em um cenário no qual não há consenso na avaliação de algum domínio do GRADE dentro do grupo elaborador e/ou entre os especialistas, é indicado que ambas as avaliações sejam descritas, apresentando o motivo que levou ao julgamento distinto (99).

Explicações

As explicações são apresentadas em notas de rodapé, na forma de comentários concisos, claros, precisos e relevantes (101). Elas são utilizadas para esclarecer as informações contidas na tabela e explicar as escolhas e os julgamentos realizados na avaliação e classificação da certeza de evidência (justificativa para rebaixar ou não o nível de certeza da evidência, assim como razões para elevar o nível de certeza da evidência). Além disso, podem ser acrescentadas informações que facilitem a interpretação dos resultados, com considerações referentes às estimativas do risco basal e aos achados de estudos individuais.

Dessa forma, esse item tem um papel fundamental na compreensão e interpretação das evidências. Contudo, identificar quais características básicas devem ser incluídas em uma explicação é um desafio, visto que há uma grande variação no nível de detalhamento, o qual nem sempre está alinhado ao domínio (102). Ao formular a

explicação, é importante mencionar o elemento-chave responsável pelo julgamento, de acordo com os conceitos do GRADE. Na Tabela 3 são apresentados alguns exemplos de explicações apresentadas em nota de rodapé, derivados de uma revisão sobre o tema (103).

Informações adicionais

Como citado, o GRADE *Working Group* desenvolveu e disponibilizou de forma gratuita o programa *on-line* GRADEpro (<https://www.gradepro.org/>), além de uma lista de verificação para a elaboração de diretrizes clínicas(<https://cebgrade.mcmaster.ca/guidecheck.html>). O GRADEpro auxilia na estruturação da tabela de apresentação, seja no formato SoF ou perfil de evidência. O programa também fornece estrutura e funções de suporte para os usuários, permitindo a gestão do processo de desenvolvimento de diretrizes clínicas, inclusive o gerenciamento de grupos de trabalho, a declaração de conflitos de interesse, a formulação de recomendações e a redação do texto.

Recentemente, um conjunto de declarações foi desenvolvido para interpretar os resultados de revisões sistemáticas de intervenções e comunicá-los a pacientes, ao público e a profissionais de saúde, seguindo a abordagem GRADE para avaliar as evidências (104). No entanto, esse material está em língua inglesa, e o grupo de trabalho responsável pelo desenvolvimento desta diretriz metodológica, em colaboração com especialistas no assunto, está atualmente realizando a adaptação transcultural desse material para a língua portuguesa, a fim de torná-lo acessível ao público deste documento.

Tabela 3 – Indicações desejáveis de explicações em tabelas de sumário de resultados e perfil de evidência

Domínio do GRADE	Conteúdo relevante	Exemplo
Critérios que reduzem a certeza de evidência		
Risco de viés (também conhecido como limitações no delineamento e na execução do estudo)	- Se há problemas sérios ou muito sérios de risco de viés em estudos randomizados ou problemas sérios, muito sérios e extremamente sérios de risco de viés de acordo com a ferramenta ROBINS-I; - Proporção/número de estudos que apresentam falhas que podem acarretar algum viés conforme a avaliação de uma ferramenta apropriada e adequada para o delineamento; - Contribuição dos estudos para as estimativas agrupadas (peso dos estudos na análise).	Risco de viés: estudos que tiveram grande peso para a estimativa de efeito foram classificados como alto risco de viés devido à falta de sigilo de alocação da randomização e à falta de cegamento. Estudos que tiveram grande peso para a estimativa do efeito geral foram classificados como alto risco de viés devido à falta de [ocultação] em [3] de [5] estudos e à [falta de cegamento] em [2] de [5] estudos.
Imprecisão	- Se há problemas sérios, muito sérios ou extremamente sérios de imprecisão; - Interpretação dos limites do intervalo de confiança (IC); - Determinar se o tamanho ideal da informação foi atendido, quando utilizado.	Imprecisão grave (muito grave ou extremamente grave). O IC95% é consistente com a possibilidade de benefício importante e grande dano, excedendo uma diferença mínima importante e incluindo apenas [2] eventos.
Evidência indireta	- Se há problemas sérios ou muito sérios em relação à evidência indireta relacionada à questão PICO (população, intervenção, comparação e desfecho); - Qualquer diferença substancial entre as evidências identificadas e a pergunta original da revisão em relação a pacientes, intervenções, comparações e desfechos avaliados de forma diferente a ponto de questionar a estimativa de efeito obtida;	Evidência indireta grave (ou muito grave). Os pacientes incluídos nos estudos têm [a condição diferente da proposta] e diferem de forma importante da [questão PICO]. Os estudos também usaram [doses do medicamento] diferentes em relação à [pergunta].

Domínio do GRADE	Conteúdo relevante	Exemplo
	- Os autores devem preencher a tabela de evidência indireta conforme a avaliação detalhada da evidência indireta relacionada à revisão ou questão de diretriz (<www.gradepro.org>).	
Viés de publicação	- Se o viés de publicação não foi detectado ou é suspeito; - Interpretação do gráfico de funil; - Abrangência das estratégias e dos métodos de busca para identificar todas as evidências disponíveis; - Presença de pequenos estudos (muitas vezes positivos) com fins lucrativos.	Há suspeita de viés de publicação porque os estudos incluídos são pequenos e o gráfico de funil mostra assimetria.
Critérios que aumentam a certeza de evidência		
Grande efeito	- Se um grande efeito ou associação está presente e o período de tempo de exposição necessário para atingir o efeito; - Descrição explícita da magnitude de efeito considerado grande.	Grande efeito baseado em estudos observacionais bem conduzidos e sem risco importante de viés ou outras limitações, apresentando um [OR] _____ (IC 95%: ____ a ____).
Gradiente dose-resposta	- Se os estudos fornecem evidências de um gradiente dose-resposta entre intervenção ou exposição e desfecho; - Descrição explícita dos limites de intervenção ou exposição relacionados a mudanças no resultado (melhoria ou redução).	Gradiente de dose-resposta evidente. [RR] com a intervenção _____ (IC95%: ____ a ____) com doses menores que _____ e [RR] de ____ (IC 95%: ____ a _____) com doses maiores que _____.

Domínio do GRADE	Conteúdo relevante	Exemplo
Fatores de confusão residuais em direção oposta	- Se os estudos fornecem evidências de todos os fatores de confusão ou vieses plausíveis contra o efeito/associação detectado, quando um efeito/associação é detectado; - Se os estudos fornecem evidências de todos os fatores de confusão ou vieses plausíveis em favor da detecção de um efeito/associação, quando um efeito/associação não é detectado; - Descrição explícita do mecanismo pelo qual fatores de confusão ou vieses podem ter reduzido ou aumentado o efeito/associação observado.	Confundidor e/ou viés contra o efeito/associação detectado (deve ser fornecida uma descrição do mecanismo). Confundidor e/ou viés a favor da detecção de um efeito ou associação não encontrada (deve ser fornecida uma descrição do mecanismo).

Observação: o conteúdo inserido entre colchetes ([]) pode ser substituído de acordo com o contexto.
Fonte: Adaptado de Santesso et al. (102).

5. Uso do GRADE para o desenvolvimento de recomendações

- Recomenda-se o uso da voz ativa ao redigir recomendações, empregando termos como "recomendamos" e "sugerimos" para indicar força ou fraqueza das recomendações, evitando a voz passiva para manter clareza na comunicação.

- As recomendações devem especificar a população-alvo, o comparador e, quando necessário, o cenário relevante, a fim de garantir uma interpretação correta e precisa das recomendações, especialmente quando a confiança nas estimativas de efeito varia.

- É preferível apresentar recomendações a favor de uma intervenção em vez de recomendações contra uma alternativa, salvo quando a terapia não possui eficácia comprovada, sendo amplamente utilizada. Isso proporciona uma clareza maior na orientação.

A tomada de decisão em saúde é um processo complexo que deve levar em consideração diferentes fatores. O equilíbrio entre os desfechos desejáveis e indesejáveis e a aplicação dos valores e das preferências dos pacientes determinam a direção da recomendação, e esses fatores, juntamente com a certeza da evidência, determinam a força da recomendação. Tanto a direção quanto a força podem ser modificadas após considerar não somente a evidência científica, mas também fatores como valores e preferências, custo, equidade, aceitabilidade e viabilidade. O sistema GRADE oferece uma metodologia abrangente e transparente para tomada de decisão estruturada, a qual é utilizada em diversas instituições, como o Ministério da Saúde do Brasil. Por vezes, essa metodologia pode ser combinada com outros métodos, como a técnica Delphi, a técnica Delphi modificada, a técnica de Grupo Nominal, conferências de consenso, sistemas de voto e a análise de decisão multicritérios (105). Para esse propósito, o GRADE utiliza as tabelas de evidência para decisão (evidence to decision, EtD), como apresentado a seguir.

5.1 Conceitos de força de recomendação

A força da recomendação reflete a importância de adotar uma determinada conduta (ou rejeitar a conduta, dependendo da direção da recomendação). Assim como a confiança das estimativas de efeito (certeza da evidência), a força de uma recomendação pode ser conceituada nas seguintes categorias: fraca (também chamadas de condicional) ou forte e a favor ou contra a intervenção. Uma recomendação forte significa uma conduta que, salvo exceções, deve ser seguida; uma recomendação fraca/condicional é uma sugestão de conduta que deve ser seguida como rotina, mas a conduta oposta é justificável (106) .

Um resumo de como as implicações dos graus de recomendações podem ser interpretadas na perspectiva de pacientes, médicos e gestores/tomadores de decisão é apresentado no Quadro 27.

Quadro 27 – Implicações dos graus de recomendações conforme o sistema GRADE

Público-alvo	Recomendação forte	Recomendação fraca (condicional)
Gestores	A recomendação deve ser adotada como política de saúde na maioria das situações.	É necessário debate substancial e envolvimento das partes interessadas.
Pacientes	A maioria dos indivíduos desejaria que a intervenção fosse indicada; no entanto, um pequeno número não aceitaria essa recomendação.	Grande parte dos indivíduos desejaria que a intervenção fosse indicada; no entanto, alguns indivíduos não aceitariam essa recomendação.
Profissionais de saúde	A maioria dos pacientes deve receber a intervenção recomendada.	O profissional deve reconhecer que diferentes escolhas serão apropriadas para cada paciente para definir uma decisão

		consistente com os seus valores e preferências.

Fonte: adaptado de Andrews et al. (107).

Para que os painelistas realizem uma recomendação forte, eles devem levar em conta os vários fatores que influenciam a força de uma recomendação, como informações relevantes que apoiem um equilíbrio claro entre as consequências desejáveis (recomendar uma ação) e indesejáveis de uma intervenção (recomendar contra uma ação). Quando há menos confiança sobre esse equilíbrio ou quando as informações relevantes não estão disponíveis, o painel de recomendação deve ser mais conservador e, na maioria dos casos, optar por fazer uma recomendação fraca (condicional) (107).

Com o avanço da ciência, é possível que painelistas de uma diretriz clínica enfrentem decisões sobre intervenções promissoras associadas a danos ou custos apreciáveis e com evidências insuficientes de benefícios para apoiar seu uso. Nesses casos, o painel pode elaborar recomendações sobre o uso de uma intervenção apenas no contexto de pesquisa, fornecendo um estímulo importante para responder questões de pesquisa emergentes. No Quadro 28, são indicadas as condições que devem estar presentes para recomendações no contexto de pesquisa (14).

Quadro 28 – Recomendações no contexto de pesquisa

Recomendações no contexto de pesquisa são apropriadas na presença de três condições:

1) há evidências insuficientes até o momento para apoiar uma decisão a favor ou contra uma intervenção;
2) futuras pesquisas têm grande potencial para reduzir a incerteza sobre os efeitos da intervenção;
3) futuras pesquisas são de grande valia para estimar os custos previstos.

Fonte: adaptado de Andrews et al. (107).

5.2 Determinantes da direção e força da recomendação

Em uma reunião de recomendação, os painelistas fazem recomendações a favor (quando os efeitos desejáveis superam os efeitos indesejáveis) ou contra (quando o oposto é verdadeiro) uma determinada estratégia, em relação a um comparador.

No sistema GRADE, os desfechos classificados como críticos e importantes são avaliados nos fatores efeitos desejáveis e indesejáveis. Esses desfechos são baseados na RS previamente conduzida para responder à questão, que é apresentada por meio das tabelas GRADE (tabelas perfil de evidências e sumário dos resultados). A Tabela 4 apresenta categorias frequentes de consequências desejáveis e indesejáveis de uma estratégia de gestão (107).

Tabela 4 – Efeitos desejados e indesejados de uma intervenção em relação à estratégia alternativa

Efeitos desejáveis	Efeitos indesejáveis
● Aumento da sobrevida ● Redução da morbidade ● Alívio dos sintomas ● Melhora da qualidade de vida ● Redução do consumo de recursos	● Diminuição da sobrevida ● Complicações graves imediatas ● Eventos adversos ● Diminuição da qualidade de vida ● Aumento do consumo de recursos

Fonte: adaptado de Andrews et al. (107).

De acordo com o sistema GRADE, os fatores determinantes da força e da direção da recomendação são os seguintes: importância do problema, magnitude dos efeitos desejáveis e indesejáveis (e o balanço entre esses efeitos), certeza da evidência, recursos necessários, valores e preferências dos pacientes, impactos em equidade, aceitabilidade da intervenção e viabilidade de implementação (Quadro 29). O grupo elaborador da diretriz deve apresentar aos painelistas evidências relacionadas a esses fatores para embasar o julgamento e a tomada de decisões. Uma maneira de apresentar essas evidências é utilizando as tabelas EtD, que permitem uma apresentação estruturada e transparente das evidências. Além de

fornecerem um resumo conciso das melhores evidências disponíveis, as tabelas EtD organizam a discussão e identificam os motivos dos desacordos, ajudando os painelistas a decidirem se as recomendações podem e devem ser implementadas (107).

Quadro 29 – Fatores determinantes da recomendação

Domínio	Considerações
Importância do problema	Ao se iniciar o desenvolvimento de uma diretriz, deve-se analisar a magnitude e a transcendência do problema. Os painelistas devem julgar a importância e prioridade do problema. Evidências para esse domínio incluem a descrição epidemiológica da condição de interesse (como prevalência, incidência e taxa de mortalidade).
Efeitos desejáveis e efeitos indesejáveis	Normalmente, a evidência apresentada para esse domínio é baseada na RS previamente conduzida para responder à questão, que é apresentada por meio das tabelas GRADE (perfil de evidência e sumário dos resultados [*summary of findings*, SoF]). Quanto mais substanciais forem os efeitos desejáveis, mais provável será que uma intervenção seja recomendada. Por outro lado, quanto mais substanciais forem os efeitos indesejáveis, menos provável será que uma intervenção seja recomendada. Os julgamentos sobre o quão substanciais são os efeitos devem levar em conta a magnitude absoluta do efeito (proporção de pessoas que se beneficiaram) e a importância do desfecho (valorização atribuída pelas pessoas acometidas).

Quadro 29 – Fatores determinantes da recomendação	
Domínio	**Considerações**
Certeza geral da evidência	A avaliação da qualidade da evidência geral deve levar em consideração os desfechos que foram classificados como críticos. A certeza da evidência é definida como o menor nível de evidência entre os desfechos classificados como críticos. Entretanto, se todos os desfechos críticos apresentarem efeito na mesma direção (por exemplo, todos os desfechos apontam que a intervenção é superior ao comparador), a certeza da evidência pode ser definida como o maior nível de evidência entre os domínios críticos. Para esse domínio, não é necessário apresentar evidências adicionais aos painelistas.
Balanço entre riscos e benefícios	O balanço entre o risco e o benefício clínico será julgado pelo grupo, considerando todos os desfechos avaliados para a tomada de decisão. Por consenso, será definido se e quanto os benefícios se sobrepõem aos riscos para o conjunto de desfechos da intervenção avaliada. As situações nas quais os benefícios claramente se sobrepõem aos riscos, e nas quais os riscos claramente se sobrepõem aos benefícios, geralmente geram recomendações fortes. As recomendações fracas são procedentes de situações nas quais há certo equilíbrio entre os riscos e os benefícios.
Valores e preferências	Pode haver variabilidade na importância que diferentes pacientes atribuem a diferentes desfechos. Isso deve ser levado em consideração durante a definição do balanço entre riscos e benefícios; nesse caso, os desfechos não

Quadro 29 – Fatores determinantes da recomendação	
Domínio	**Considerações**
	apresentarão peso equivalente. Dados de valores e preferências dos pacientes podem ser obtidos a partir de RS da literatura, entrevista com pacientes e representantes e experiência de especialistas/percepções baseadas na prática clínica.
Certeza da evidência sobre recursos necessários (custo)	Nesses domínios, os painelistas devem avaliar a magnitude dos recursos necessários para implementar a intervenção em avaliação. Considerando a perspectiva do Ministério da Saúde do Brasil, os custos das intervenções de interesse podem ser obtidos das fontes a seguir. - Banco de Preços em Saúde (http://bps.saude.gov.br/login.jsf): sistema criado pelo Ministério da Saúde que fornece informações sobre compras públicas e privadas de medicamentos e produtos para a saúde. - Painel de Preços (http://paineldeprecos.planejamento.gov.br/): fornece informações sobre compras públicas homologadas no Sistema de Compras do Governo Federal (COMPRASNET). - Listas de preços máximos da Câmara de Regulação de Medicamentos (CMED) (http://portal.anvisa.gov.br/listas-de-precos): fornece o preço máximo de venda de medicamentos a consumidores e ao governo.

Quadro 29 – Fatores determinantes da recomendação

Domínio	Considerações
Custo-efetividade	Análises de custo-efetividade, inclusive custo-utilidade. Detalhes sobre a condução de avaliações econômicas em saúde podem ser obtidos na Diretriz de Avaliação Econômica da Rede Brasileira de Avaliação de Tecnologias em Saúde (REBRATS) (108).
Equidade	Deve-se avaliar o potencial impacto das recomendações no combate às iniquidades sociais, sendo estas mais propensas a gerarem recomendações fortes. As intervenções que reduzem as desigualdades são mais propensas a serem recomendadas do que as que não reduzem (ou as que aumentam as desigualdades) (109, 110).
Aceitabilidade	Nesse domínio, são avaliadas as preferências dos principais atores (gestores, profissionais de saúde e pacientes) em relação à intervenção, assim como às suas alternativas.
Viabilidade	É preciso identificar os recursos e a estrutura necessários para implementar a intervenção de interesse e avaliar sua disponibilidade. Quanto menos viável for uma intervenção, menos provável será que ela seja recomendada. As barreiras à implementação de uma intervenção também podem modificar a força de uma recomendação. Os profissionais da saúde podem achar inútil receber recomendações fortes se as intervenções não forem implementáveis em seus ambientes.

Quadro 29 – Fatores determinantes da recomendação	
Domínio	**Considerações**
	No entanto, se o público-alvo for formuladores de políticas, um painel pode querer fazer uma recomendação forte apesar das barreiras que atualmente dificultam ou impossibilitam a adesão dos profissionais à recomendação. Os painéis também podem incorporar a consideração de barreiras críticas, como a disponibilidade da intervenção, diretamente em suas recomendações. Mais comumente, os painéis podem ajudar os responsáveis pela implementação das recomendações ao abordar as principais barreiras à implementação de suas recomendações nas conclusões.

Fonte: adaptado de Ministério da Saúde (105).

Após a apresentação das evidências para cada um dos domínios da Tabela 5, o grupo elaborador deve firmar seu julgamento, durante os painéis, para consenso e formulação das recomendações. O sistema GRADE propõe questões que podem guiar o julgamento de cada critério (Tabela 5).

Tabela 5 – Domínios avaliados durante a tomada de decisão de acordo com o sistema GRADE

Domínios	Questão de interesse	Opções de resposta
Importância do problema	O problema é prioritário?	Não Provavelmente não Provavelmente sim Sim Varia Não se sabe
Efeitos desejáveis	Qual a magnitude dos efeitos desejáveis?	Trivial Pequena Moderada Grande Varia Não se sabe
Efeitos indesejáveis	Qual a magnitude dos efeitos indesejáveis?	Trivial Pequena Moderada Grande Varia Não se sabe
Certeza geral da evidência	Qual a certeza da evidência (nível de evidência para o conjunto de evidências)?	Muito baixa Baixa Moderada Alta
Balanço entre riscos e benefícios	O balanço entre os riscos e os benefícios favorece a intervenção ou o comparador?	Favorece o comparador Provavelmente favorece o comparador Não favorece o comparador nem a intervenção Provavelmente favorece a intervenção Favorece a intervenção Varia Não se sabe
Valores e preferências	Há incerteza ou variabilidade significativas em relação à importância que os pacientes atribuem aos principais desfechos?	Incerteza ou variabilidade importante Incerteza ou variabilidade possivelmente importante

Domínios	Questão de interesse	Opções de resposta
		Incerteza ou variabilidade provavelmente não importante Sem incerteza ou variabilidade importante
Certeza da evidência sobre recursos necessários	Qual a magnitude dos recursos necessários (custos)?	Custo grande Custo moderado Custo e economia negligenciáveis Economia moderada Economia grande Varia Não se sabe
Custo-efetividade	A custo-efetividade da intervenção favorece a intervenção ou o comparador?	Favorece o comparador Provavelmente favorece o comparador Não favorece o comparador nem a intervenção Provavelmente favorece a intervenção Favorece a intervenção Varia Não se sabe
Equidade	Quais são os impactos referentes à equidade em saúde?	Reduz Provavelmente reduz Provavelmente sem impacto Provavelmente aumenta Aumenta Varia Não se sabe
Aceitabilidade	A opção é aceitável para os principais atores interessados?	Não Provavelmente não Provavelmente sim Sim Varia Não se sabe
Viabilidade	A implementação da intervenção é viável?	Não Provavelmente não Provavelmente sim Sim

Domínios	Questão de interesse	Opções de resposta
		Varia Não se sabe

Fonte: Ministério da Saúde (105).

5.3 Qual a perspectiva adotada?

A perspectiva adotada pelo painel determina quais consequências econômicas de uma intervenção devem ser consideradas na realização de uma recomendação ou decisão. Os painéis das diretrizes clínicas devem ser explícitos sobre isso (111). A instituição demandante da diretriz geralmente é quem determina a perspectiva a ser adotada pelo painel, seja ela restrita, como de uma farmácia hospitalar, ou até mais ampla, como de um governo ou uma sociedade (112). O sistema GRADE orienta que a perspectiva seja a mais ampla possível (112).

Em uma perspectiva mais restrita, como uma farmácia hospitalar, os custos resultantes de eventos adversos seriam ignorados pela economia ocasionada pelo uso de um medicamento que reduz o risco de acidente vascular cerebral ou infarto do miocárdio, por exemplo. Já em uma perspectiva mais ampla, como da sociedade civil, serão considerados os custos diretos, bem como os indiretos (por exemplo, afastamento do trabalho, custo de transporte para receber o tratamento etc.).

Na perspectiva de um paciente usuário de um sistema de saúde, os recursos considerados são apenas os que afetam diretamente o paciente (por exemplo, custos diretos); a maioria dos custos gerados pode ser ignorada (por exemplo, custos arcados pelo governo) (112). Os médicos que atendem pacientes que não possuem acesso a um sistema de saúde público, suplementar ou privado precisam ajudá-los a tomar decisões considerando os custos que serão desembolsados. Isso é particularmente importante quando as vantagens e desvantagens clínicas são bem equilibradas e há custos substanciais envolvidos. Nessas circunstâncias, se os elaboradores de diretrizes utilizarem o sistema GRADE e disponibilizarem perfis de evidências para os usuários das diretrizes, os médicos podem revisar o sumário das evidências e garantir que a decisão do paciente de aceitar a estratégia de tratamento recomendada seja consistente com seus valores e preferências (112).

Diferentes tipos de decisões e diferentes perspectivas requerem diferentes considerações. Consequentemente, o sistema GRADE sugere conjuntos específicos

de critérios para recomendações clínicas em relação à perspectiva individual do paciente, à perspectiva populacional, a decisões de incorporação, a recomendações sobre testes diagnósticos e a recomendações sobre o sistema de saúde pública (Quadro 30) (111).

Embora existam diferenças na sua operacionalização para os diferentes tipos de decisões, a maioria dos critérios é semelhante, como pode ser observado no Quadro 30. Todos os cinco conjuntos de critérios incluem questões sobre se o problema é uma prioridade, a magnitude dos efeitos desejáveis e indesejáveis, a certeza da evidência, consideração de como os pacientes (ou outros acometidos, como cuidadores) valorizam os principais resultados, o equilíbrio entre efeitos desejáveis e indesejáveis, uso de recursos, aceitabilidade e viabilidade. Todas as estruturas que adotam uma perspectiva populacional também incluem a consideração de impactos sobre a equidade (111).

Quadro 30 – Estrutura de itens da evidência para decisão para cinco tipos diferentes de decisões

Domínio	Recomendações clínicas – perspectiva individual	Recomendações clínicas – perspectiva da população	Decisões de incorporação	Sistema de saúde e recomendações de saúde pública	Diagnóstico, triagem e outros testes (recomendações clínicas e de saúde pública – perspectivas individual e populacional)
Importância do problema	O problema é uma prioridade?				
Acurácia do teste	Não aplicável				O quão preciso o teste é?
Efeitos desejáveis e efeitos indesejáveis	Quão substanciais são os efeitos esperados desejáveis? Quão substanciais são os efeitos indesejáveis esperados?				
Certeza geral da evidência	Qual é a certeza geral da evidência dos efeitos?				Qual é a certeza da evidência da precisão do teste? Qual é a certeza da evidência para quaisquer benefícios diretos críticos ou importantes, efeitos adversos ou ônus do teste?

Domínio	Recomendações clínicas – perspectiva individual	Recomendações clínicas – perspectiva da população	Decisões de incorporação	Sistema de saúde e recomendações de saúde pública	Diagnóstico, triagem e outros testes (recomendações clínicas e de saúde pública – perspectivas individual e populacional)
					Qual é a certeza da evidência dos efeitos do manejo que se orienta pelos resultados dos testes? O quão certo é o vínculo entre os resultados dos testes e as decisões de gerenciamento? Qual é a certeza geral da evidência dos efeitos do teste?
Valores e preferências	Há incerteza importante ou variabilidade no valor que as pessoas atribuem aos principais desfechos?				Há incerteza importante ou variabilidade no valor que as pessoas atribuem

Domínio	Recomendações clínicas – perspectiva individual	Recomendações clínicas – perspectiva da população	Decisões de incorporação	Sistema de saúde e recomendações de saúde pública	Diagnóstico, triagem e outros testes (recomendações clínicas e de saúde pública – perspectivas individual e populacional)
					aos principais desfechos, inclusive a efeitos adversos e carga do teste e a desfechos no decorrer do manejo clínico guiado pelos resultados do teste?
Balanço entre riscos e benefícios	O balanço entre os efeitos desejáveis e indesejáveis favorece a intervenção ou a comparação?				O balanço entre os efeitos desejáveis e indesejáveis favorece o teste ou a comparação?
Certeza da evidência sobre recursos		Quão grandes são os recursos (custos)?			
		Qual é a certeza da evidência dos recursos (custos)?			

Domínio	Recomendações clínicas – perspectiva individual	Recomendações clínicas – perspectiva da população	Decisões de incorporação	Sistema de saúde e recomendações de saúde pública	Diagnóstico, triagem e outros testes (recomendações clínicas e de saúde pública – perspectivas individual e populacional)
necessários (custo)	A relação de custo-benefício da intervenção (o custo desembolsado em relação aos benefícios líquidos) favorece a intervenção ou a comparação?	O custo-benefício da intervenção favorece a intervenção ou a comparação?			A relação custo-benefício do teste favorece o teste ou a comparação?
Equidade	-	Qual seria o impacto na equidade em saúde?			
Aceitabilidade	A intervenção é aceitável para os pacientes, seus cuidadores e os profissionais de saúde?	A intervenção é aceitável para as principais partes interessadas?			
Viabilidade	A intervenção é viável para os	A implementação da intervenção é viável?			A implementação do teste é viável?

Domínio	Recomendações clínicas – perspectiva individual	Recomendações clínicas – perspectiva da população	Decisões de incorporação	Sistema de saúde e recomendações de saúde pública	Diagnóstico, triagem e outros testes (recomendações clínicas e de saúde pública – perspectivas individual e populacional)
	pacientes, seus cuidadores e os profissionais de saúde?				

Fonte: adaptato de Alonso-Coello et al. (111).

5.4 Redação de recomendações em saúde

Várias características devem ser observadas na redação de recomendações. É indicado que os desenvolvedores de diretrizes apresentem as recomendações em voz ativa. Por exemplo, devem-se utilizar os termos ‘‘recomendamos’’ e ‘‘sugerimos’’ para recomendações fortes e fracas, respectivamente. Recomendações na voz passiva podem não apresentar clareza (107).

Ainda, as recomendações devem sempre especificar a população e o comparador, como no seguinte exemplo: “em pacientes com insuficiência renal aguda, recomendamos a mensuração do volume da urina em hora em hora por pelo menos 24 horas”. A força dessa recomendação pode diferir se a alternativa for a cada 2 horas ou uma vez por dia. Assim, é necessária especificação adicional: ‘‘quando comparada com a medição diária do volume urinário’’ (107).

Às vezes, a recomendação pode indicar o cenário, principalmente quando a confiança das estimativas de efeito varia de acordo com ele. Por exemplo, uma recomendação sobre endarterectomia carotídea pode variar dependendo da extensão do atraso entre a apresentação de um paciente com sintomas sugestivos de estenose carotídea e a realização da cirurgia (113). Outro exemplo em que o cenário pode ser importante é uma intervenção de alto custo em países de alta e baixa renda (107).

Ademais, é preferível que sejam apresentadas recomendações a favor de uma intervenção em vez de recomendações contra uma alternativa. Por exemplo, em relação à adição de ácido acetilsalicílico ao clopidogrel em pacientes que sofreram um acidente vascular cerebral, é preferível afirmar “em pacientes que tiveram acidente vascular cerebral, sugerimos clopidogrel em monoterapia em comparação ao uso de ácido acetilsalicílico associado ao clopidogrel” em vez de “em pacientes que tiveram acidente vascular cerebral e estão utilizando clopidogrel, sugerimos não adicionar ácido acetilsalicílico”. No entanto, quando uma terapia não apresenta eficácia, mas é amplamente utilizada, as recomendações contra essa abordagem são apropriadas. Por exemplo, “em pacientes submetidos à cirurgia cardíaca que não estavam recebendo betabloqueadores anteriormente, sugerimos não iniciar terapia com betabloqueador perioperatório” (107). Outro exemplo são recomendações sobre tecnologias que não apresentaram evidência de benefício: “recomendamos não

utilizar [tecnologia] em pacientes com [condição] (recomendação forte, certeza da evidência [nível])".

Para evitar a má interpretação de uma recomendação, uma alternativa sugerida pelo sistema GRADE é o uso de símbolos (que podem ser menos confusos do que números ou letras) (114) ou termos para expressar a força das recomendações. Sugere-se o uso de "↑↑" para recomendações fortes e de "↑" para recomendações fracas/condicionais. Para desenvolvedores de diretrizes que preferem números ou letras, sugere-se o uso de ‘‘1’’ para recomendações fortes e ‘‘2’’ para recomendações fracas/condicionais. Quaisquer que sejam os termos escolhidos pelos desenvolvedores de diretrizes (por exemplo, fraco ou condicional), devem ser usados de forma sistemática em diferentes diretrizes. Ainda, as explicações sobre o significado e as implicações das recomendações fortes e condicionais devem estar facilmente acessíveis, a fim de facilitar a interpretação correta (por exemplo, na seção de métodos da diretriz ou utilizando *hiperlinks* em publicações eletrônicas) (107).

Além disso, o painel deve fornecer uma justificativa para a sua recomendação ou decisão. As conclusões também devem incluir considerações relevantes sobre subgrupos de pacientes, implementação, monitoramento e avaliação de prioridades em pesquisa. A justificativa para uma recomendação ou decisão deve considerar os julgamentos feitos pelo painel em relação aos critérios utilizados na avaliação, pensando nos principais fatores que conduziram a recomendação ou decisão. As conclusões do painel sobre as considerações de subgrupo devem especificar quais subgrupos foram considerados e como essas considerações afetaram a recomendação. Se os julgamentos do painel (e as evidências de pesquisa ou considerações adicionais que informaram esses julgamentos) e suas conclusões para um subgrupo forem muito diferentes da avaliação geral, o painel pode optar por apresentar uma estrutura EtD separada para o subgrupo (111).

As conclusões sobre as considerações de implementação devem especificar as principais preocupações sobre a viabilidade e aceitabilidade da intervenção e as estratégias para lidar com essas preocupações, bem como qualquer informação importante sobre como implementar a intervenção, particularmente para intervenções complexas e com custo elevado para o sistema de saúde.

As conclusões sobre monitoramento e avaliação devem incluir sugestões sobre quais indicadores, se houver, devem ser monitorados e qualquer avaliação necessária em relação à implementação da recomendação ou decisão. Isso é especialmente relevante para as decisões e recomendações do sistema de saúde público e suplementar. Finalmente, depois de revisar e avaliar as evidências, os painéis devem identificar prioridades de pesquisa para abordar quaisquer incertezas ou lacunas importantes nas evidências de pesquisa que informaram seus julgamentos (111).

5.5. Exemplo de uma tabela EtD e julgamento

O **Quadro 31** é um exemplo de tabela EtD de uma questão clínica e os julgamentos realizados para uma determinada diretriz elaborada para o SUS. Pode-se observar que os fatores determinantes discutidos na realização da recomendação dessa temática foram benefícios, riscos, balanço entre risco e benefícios, certeza da evidência, custos e viabilidade de implementação. Dessa maneira, salientamos que adaptações dos fatores determinantes que compõem a EtD podem ocorrer e serem determinadas pelo grupo elaborador da diretriz.

Quadro 31 – Processo de tomada de decisão referente ao uso de uma determinada tecnologia no tratamento de pacientes com uma determinada condição

Item da tabela EtD	Julgamento dos painelistas	Justificativa
Benefícios	Sem relevância clínica	A maioria dos painelistas julgou que o benefício não apresenta relevância clínica, com alguns tendendo para benefício pequeno. Não houve diferença na mortalidade por todas as causas, hospitalização ou deterioração clínica (desfechos importantes).
Riscos	Sem relevância clínica	A maioria dos painelistas considera o medicamento bem tolerado, com os riscos sendo relativamente baixos, e a relevância clínica ausente ou pequena.

Quadro 31 – Processo de tomada de decisão referente ao uso de uma determinada tecnologia no tratamento de pacientes com uma determinada condição		
Item da tabela EtD	Julgamento dos painelistas	Justificativa
Balanço dos riscos e benefícios	Equilibrado	Não há benefício claro a favor ou contra a intervenção ou o comparador. A maioria dos painelistas julgou que há um equilíbrio entre os benefícios e riscos, com leve inclinação para o comparador.
Certeza da evidência	Muito baixa	Há um pequeno número de eventos nos desfechos clinicamente relevantes.
Custos	Sem impacto importante	Estimou-se que o valor do comprimido para o SUS é, em média, R$ 6,33. De acordo com dados da CMED, com um PMVG de 17%, o valor do tratamento para um paciente por 5 dias é R$ 63,30. O painel entendeu que o custo individual é baixo, mas pode ser moderado caso haja uso em maior escala.
Viabilidade de implementação	Provavelmente sim	Atualmente, o medicamento não está disponível no SUS (Rename), mas municípios podem disponibilizar o medicamento nos hospitais. A incorporação no SUS seria factível.
Outras considerações	-	-
Recomendação: *Sugerimos não utilizar [tecnologia] em pacientes com [condição] (recomendação condicional, certeza da evidência [nível]).* Considerações gerais sobre o uso de [tecnologia] em pacientes com [condição]: ● o painel de recomendações considerou que, apesar de não poder descartar benefício e o medicamento ser relativamente seguro, no momento não há evidências suficientes para indicar o seu uso de rotina; ● o uso de [tecnologia] deve ser limitado a estudos clínicos.		

CMED = Câmara de Regulação do Mercado de Medicamentos; EtD = *evidence to decision*; PMVG = preço máximo de venda ao governo; SUS = Sistema Único de Saúde.
Fonte: Elaboração própria.

6. Sistema GRADE para testes e estratégias diagnósticos

A classificação da certeza das evidências de estudos sobre a acurácia de um determinado teste ou estratégia diagnóstica difere conceitualmente, mas compartilha a lógica fundamental dos domínios de risco de viés e de evidência indireta do sistema GRADE para intervenção, prognóstico ou outros estudos.

As recomendações relativas aos testes diagnósticos também compartilham alguns desafios das recomendações para intervenções terapêuticas, inclusive alguns desafios de acordo com a sua especificidade. Enquanto alguns testes diagnósticos relatam resultados positivos e negativos (por exemplo, testes de gravidez ou de infecção por HIV), outros testes relatam os resultados de forma ordinal (por exemplo, escala de Glasgow ou Mini Mental) ou contínua (por exemplo, medidas metabólicas). Geralmente, os testes apresentam uma associação entre a direção (aumento/redução) dos resultados e a direção da probabilidade de desenvolvimento da doença ou dos eventos adversos à medida que o resultado do teste se torna mais extremo.

Para simplificar o uso desses testes ordinais ou contínuos, muitas vezes se assumem abordagens que, em última análise, categorizam os resultados dos testes de forma binária (positivos ou negativos; presença ou ausência de doença) e, consequentemente, direcionam para a escolha de tratar ou não tratar os pacientes com base nesses resultados.

No cenário clínico, muitas vezes é necessário mais de um teste diagnóstico para a tomada de decisão junto ao paciente. Isso configura uma estratégia diagnóstica que consiste na utilização de um teste inicial mais sensível (mas não específico) que, quando positivo, é seguido por um teste mais específico (por exemplo, o teste para HIV inclui o uso de um teste ELISA que, quando positivo, é seguido pela determinação quantitativa de RNA HIV). Dessa forma, muitas vezes se pode avaliar ou recomendar o uso de uma estratégia de testes em vez de um único teste; essa recomendação geralmente é baseada na comparação com uma estratégia alternativa de testes. Independentemente do tipo de estratégia ou apresentação, os resultados geralmente apresentam uma direção na qual espera-se que o aumento/redução da escala/valores aumente/reduza a probabilidade dos eventos associados a ela (115).

Na avaliação da certeza de evidência, deve estar explícito qual o objetivo dos testes ou estratégias diagnósticos, como identificação de distúrbios fisiológicos, estabelecimento de prognóstico, monitoramento de doenças ou de resposta ao tratamento, triagem e diagnóstico (116).

Painelistas de diretrizes ou desenvolvedores de revisões sistemáticas também devem estabelecer claramente a finalidade do teste ou estratégia diagnóstica dentro do contexto da questão clínica. O Quadro 32 descreve algumas das possíveis finalidades que os testes diagnósticos podem ter.

Quadro 32 – Possíveis finalidades dos testes e estratégias diagnósticos

Finalidades	Descrição
Substituição	Um novo teste pode ser avaliado com a função de substituir um teste que se utilizava anteriormente devido a várias razões: mais preciso, menos invasivo, menos arriscado ou desconfortável, apresenta menos desafios organizacionais ou técnicos, é mais rápido ou mais fácil para fazer o laudo, menos custoso, etc.
Triagem	Um novo teste pode ser avaliado para que seja incluído antes da realização dos procedimentos diagnósticos existentes, em que apenas pacientes com resultados específicos seguirão para a etapa de testes diagnósticos usuais. Esses testes não são necessariamente mais acurados, mas geralmente são mais simples e menos onerosos.
Adição	Um novo teste pode ser avaliado para inclusão após a via usual diagnóstica existente, podendo ser utilizado para reduzir a chance de resultados falso-positivos ou falso-negativos. Os testes complementares são geralmente mais acurados, mas menos utilizados do que os testes preexistentes.
Combinação	Um novo teste pode ser utilizado simultaneamente a um teste já existente. Os resultados da utilização conjunta (teste preexistente e teste adicionado) são voltados para fazer um diagnóstico e determinar o gerenciamento do fluxo do indivíduo.

Fonte: Adaptado de Bossuyt et al. (117), Mustafa et al. (118), Schunemann and Mustafa (119).

Neste capítulo, assume-se uma abordagem diagnóstica (testes e estratégias diagnósticos) que, ao final, categoriza os resultados de forma dicotômica (quando se

deve tratar ou não um paciente). As principais estimativas de acurácia obtidas em testes são calculadas através de uma tabela de contingência do tipo 2x2, na qual os participantes dos estudos são classificados de forma cruzada a partir do resultado do teste índice e os resultados da presença ou ausência da condição de interesse são identificados através do teste padrão de referência (Tabela 6). Assim, a análise da tabela de contingência identifica o número de indivíduos com diagnóstico verdadeiro-positivo, verdadeiro-negativo, falso-positivo e falso-negativo, ou seja, os principais desfechos que serão avaliados. A partir desses índices iniciais, é possível estimar a sensibilidade, especificidade e acurácia diagnóstica do teste índice (Quadro 33).

Tabela 6 – Tabela de contingência 2x2 para estudos de testes e estratégias diagnósticos

		Resultado do teste padrão de referência	
		Presença da condição de interesse	**Ausência** da condição de interesse
Resultado do teste índice	Positivo (+)	Verdadeiro-positivos (a)	Falso-positivos (b)
	Negativo (-)	Falso-negativos (c)	Verdadeiro-negativos (d)
	Total	**Com condição de interesse (a+c)**	**Sem condição de interesse (b+d)**

Fonte: elaboração própria.

Quadro 33 – Estimativas obtidas a partir da avaliação dos resultados de testes índices em tabelas de contingência 2 x 2	
Verdadeiro-positivo	O resultado é positivo, e a condição de interesse **está presente**
Verdadeiro-negativo	O resultado é negativo, e a condição de interesse **não está presente**
Falso-positivo	O resultado é positivo, e a condição de interesse **não está presente**
Falso-negativo	O resultado é negativo, e a condição de interesse **está presente**
Sensibilidade	A probabilidade de os casos verdadeiros serem identificados corretamente pelo teste índice (**a/(a+c))**
Especificidade	A probabilidade de pessoas sem a condição de interesse serem corretamente identificadas pelo novo teste (**d/(b+d)**)
Acurácia	A probabilidade de o teste fornecer resultados corretos (**(a+d)/(a+b+c+d)**)

Fonte: elaboração própria.

6.1. Elaborando as questões que envolvem testes e estratégias diagnósticos

De forma semelhante ao que foi descrito no capítulo 2, que trata sobre a questão de pesquisa, devem-se estabelecer claramente o propósito e o papel do teste ou estratégia diagnóstica. O formato da pergunta feita pelos autores de revisões sistemáticas ou desenvolvedores de diretrizes seguem os mesmos princípios que o formato para as demais questões: “Deve-se utilizar o ‘*teste/estratégia A*’ para ‘*diagnóstico/avaliação de xxx*’ em pacientes ‘*xxx*’?”. Por exemplo: “*Deve-se utilizar*

tomografia computadorizada espiral multislice de artérias coronárias (exame em avaliação ou teste ínidice) em substituição à angiografia coronária invasiva convencional (exame em comparação ou teste de referência padrão) para reduzir os eventos coronarianos (complicações) associados a falso-negativos e, assim, também reduzir o número de tratamentos desnecessários associados aos falso-positivos?". O exemplo ilustra um raciocínio usual para a avaliação de um novo teste com a função de substituição, que busca evitar complicações associadas a uma alternativa mais invasiva e cara para a avaliação de uma condição que poderá ser tratada com eficácia. Nesse cenário, o novo teste precisaria apenas replicar os resultados do teste vigente e demonstrar maiores benefícios para a população-alvo, pressupondo-se que o novo teste categorize de forma similar os pacientes em um mesmo estágio da doença e que as consequências do manejo dos pacientes sejam similares (115).

Também se devem definir os quatro componentes da questão PIRO – população, teste índice (testes/estratégias diagnósticos em avaliação), testes/estratégias de comparação (teste de referência padrão [Quadro 35]) e os desfechos de interesse) (116, 120). Por vezes, referir-se aos testes/estratégias como "intervenção" traz o reconhecimento do princípio fundamental de que os resultados dos testes levam a decisões clínicas ou gerenciais (Quadro 34).

Quadro 34 – Exemplos de questões envolvendo testes ou estratégias diagnósticas

Abaixo, são descritos dois exemplos de perguntas de pesquisa organizadas na estrutura PIRO para avaliação de um teste diagnóstico:

Questão de pesquisa 1:

Qual o impacto do teste na detecção de papilomavírus humanos (HPV) em comparação com a inspeção visual do colo uterino (após aplicação de ácido acético) em mulheres com risco de neoplasia intraepitelial cervical (NIC), em países de média ou baixa renda, nos desfechos importantes, como recidiva da doença e sobrevida, para os pacientes?

População: mulheres em risco de câncer do colo uterino em países de baixa ou média renda.

Teste índice: rastreamento único com teste para presença de HPV e, consequentemente, tratamento para neoplasia NIC.

Teste referência: inspeção visual do colo uterino (após aplicação de ácido acético) e, consequentemente, tratamento para NIC.

Desfechos clínicos importantes: óbito por câncer de colo uterino, incidência de câncer de colo uterino, recorrência de NIC, sangramento maior, parto prematuro, infertilidade, infecções maiores e menores, tratamentos desnecessários e detecção de câncer do colo uterino durante a triagem.

Questão de pesquisa 2:

Qual a acurácia de um teste de amplificação do ácido nucleico (Xpert) para o diagnóstico de meningite tuberculosa em pacientes com suspeita de meningite tuberculosa?

População: pacientes com suspeita de meningite tuberculosa.

Testes prévios: os pacientes podem ter sido submetidos a um exame geral de saúde (histórico e exames) e, possivelmente, a radiografia de tórax.

Teste índice (novo): Xpert.

Teste de referência (comparador): cultura.

Desfechos: sensibilidade, especificidade, verdadeiro-positivo, verdadeiro-negativo,

falso-positivo e falso-negativo.

Fonte: Adaptado de Schünemann et al. (115).

Quadro 35 – Testes de referência padrão

O conceito de acurácia diagnóstica baseia-se na presença de um teste que seja considerado o “padrão-ouro”, ou seja, um teste que consegue definir com alto índice de certeza quando uma determinada doença ou fator está ou não presente nos pacientes. Neste capítulo, o termo “padrão-ouro” será utilizado para representar uma abordagem “perfeita” para definir ou diagnosticar uma doença ou condição de interesse, mesmo que a abordagem seja teórica ou hipotética. A partir dessa definição, o termo “padrão de referência” ou “teste de referência padrão” é utilizado para se referir ao exame ou ao teste/estratégia de diagnóstico que, no momento da avaliação, é considerado a melhor abordagem e o mais aceito para comparar o teste/estratégia que está em avaliação.

Fonte: Adaptado de Schünemann et al. (115).

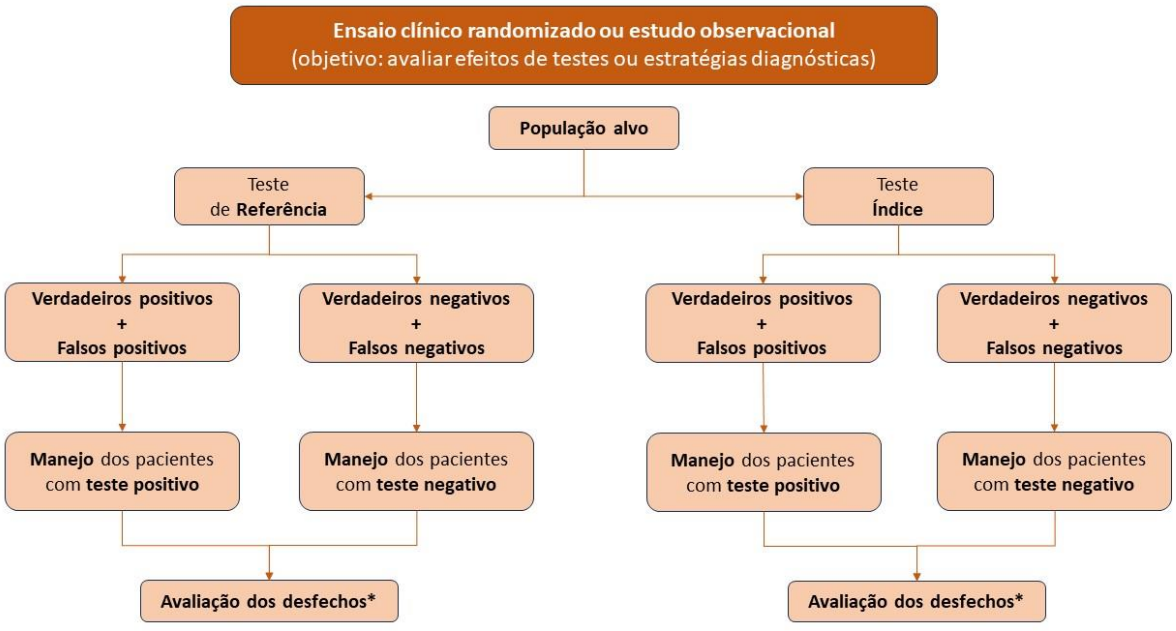
A melhor maneira de avaliar um teste ou estratégia diagnóstica é por meio de um ECR em que os pesquisadores aloquem os pacientes em uma abordagem estratégica diagnóstica de controle ou experimental que meça desfechos importantes para os pacientes (mortalidade, morbidade, sintomas, qualidade de via ou uso de recursos). A Figura 24 apresenta um fluxograma teórico demonstrando possíveis desenhos de ECR ou de estudos observacionais que se propõem a avaliar os efeitos de testes ou estratégias diagnósticas e que poderiam ser utilizados durante a avaliação da certeza de evidência dessas questões. A Figura 25 demonstra um fluxo teórico para os estudos de acurácia, que também podem ser utilizados para responder a essas questões. As duas abordagens genéricas se diferenciam conforme descrito a seguir.

- Figura 24: os pacientes são randomizados para o teste ou estratégia em investigação ou, alternativamente, para um teste/estratégia que está sendo utilizado como comparador (deve-se, idealmente, utilizar o teste vigente preconizado para a avaliação dos pacientes). Aqueles cujo teste for positivo

(casos detectados) receberão o tratamento disponível para esses pacientes (independentemente de ser um falso-positivo ou um verdadeiro-positivo). Os investigadores devem avaliar os desfechos em cada um dos grupos e comparar os resultados; sugere-se que sejam priorizados os desfechos importantes para os pacientes.

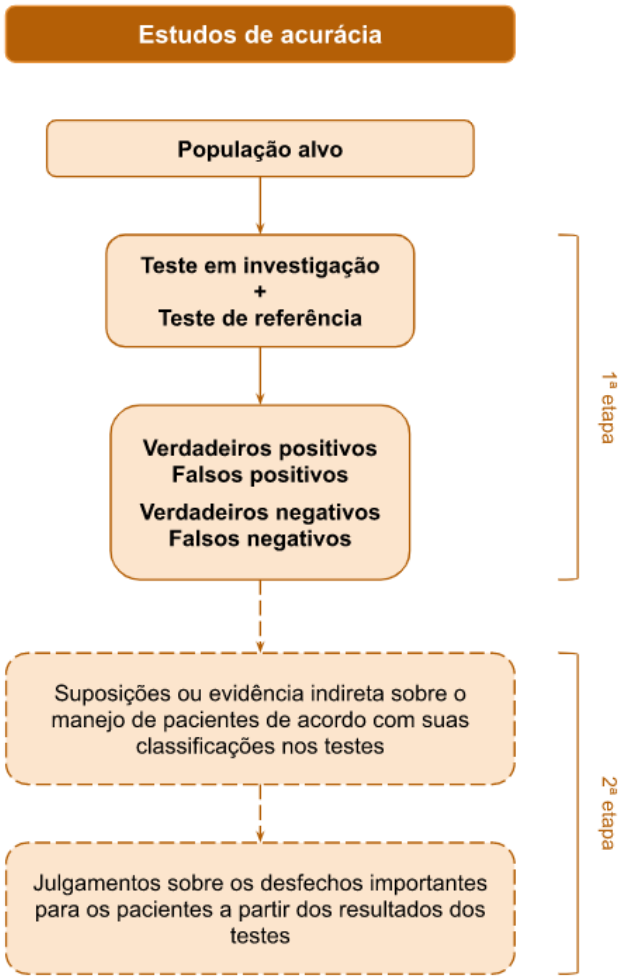
- Figura 25: a primeira etapa, os mesmos pacientes serão avaliados tanto com os testes ou estratégias diagnósticas quanto com o teste de referência. Os investigadores calcularão a acurácia do teste em investigação em comparação ao teste de referência. Em uma segunda etapa, para realizar os julgamentos sobre a importância desses resultados para os pacientes, os pacientes com testes ou estratégias de diagnóstico positivos serão (ou foram, em outros estudos) submetidos a tratamentos ou a nenhum tratamento; deve-se, então, avaliar e comparar os resultados importantes para os pacientes nos grupos.
-

Figura 24 – Fluxo esquemático de ensaios clínicos randomizados ou estudos observacionais que se propõem a avaliar os efeitos de testes ou estratégias diagnósticas



*Desfechos clínicos, preferencialmente desfechos importantes para os pacientes.
Fonte: Adaptado de Schünemann et al. (115).

Figura 25 – Fluxo esquemático dos estudos de acurácia e subsequentes processos de tomada de decisão em saúde



* Desfechos clínicos, preferencialmente desfechos importantes para os pacientes.
Fonte: Adaptado de Schünemann et al. (115).

6.2. Julgamento sobre a certeza de evidência

Em um estudo típico de acurácia, uma série consecutiva de pacientes sob suspeita de determinado diagnóstico/característica seria submetida ao teste índice (teste em avaliação) e, em seguida, todos os pacientes seriam avaliados com o teste de referência padrão (melhor método disponível para indicar a presença da condição-alvo). Quando os estudos incluídos na síntese de evidências apresentarem um desenho de estudo semelhante ao descrito na Figura 24, a avaliação da certeza das estimativas de efeito deve incluir os mesmos critérios/etapas descritos nos capítulos anteriores, para avaliação de ensaios clínicos, e partir de uma classificação alta (3.

Avaliação da certeza da evidência). Contudo, os estudos geralmente apresentam limitações que acarretam na redução da avaliação. Um dos principais motivos que ocasiona rebaixamento das evidências é o fato de que os desfechos que avaliam a acurácia são apenas desfechos substitutos dos desfechos de relevância clínica para os pacientes (115), como por exemplo, a utilização da densitometria óssea para avaliação da densidade mineral óssea e não a constatação de fratura (desfecho de relevância clínica).

O conteúdo do presente capítulo é destinado para situações em que os resultados diretos sobre os desfechos de relevância clínica para os pacientes não estão disponíveis e a síntese de evidências é baseada em estudos de acurácia diagnóstica.

A seguir, são descritos de forma geral os fatores que podem aumentar ou reduzir a certeza de evidência da síntese baseada em estudos de acurácia diagnóstica e como eles se diferem das evidências avaliadas em outros desenhos de estudo. Ao longo deste capítulo, cada domínio é explicado em maior detalhe.

Risco de viés

Estão disponíveis vários instrumentos para avaliação do risco de viés em estudos de acurácia (32, 115). Um exemplo é a ferramenta *Quality Assessment of Diagnostic Accuracy* (atualmente na sua segunda versão, QUADAS-2), a qual permite uma avaliação transparente do risco de viés com base nos seguintes domínios: seleção de pacientes, teste índice, padrão de referência, fluxo e tempo, os quais podem ser classificados como alto risco, risco incerto ou baixo risco de viés. Nessa ferramenta, se as respostas a todas as perguntas de sinalização para um domínio forem “sim”, o risco de viés pode ser considerado baixo. Se um ou mais domínios forem julgados como alto risco ou risco incerto de viés, o estudo seria classificado como contendo risco de viés (Quadro 36). A ferramenta QUADAS-2 também avalia a aplicabilidade da seleção de pacientes, o teste índice e o padrão de referência, porém a aplicabilidade deve ser avaliada junto ao domínio evidência indireta.

Na avaliação do risco de viés, devem ser consideradas possíveis limitações no delineamento e na execução dos estudos. Deve-se avaliar se a amostra incluída é representativa da população de interesse; se a comparação foi realizada de forma

independente com o teste/estratégia de referência; se todos os pacientes incluídos foram avaliados com ambos os testes (índice e referência); se foi relatada a incerteza diagnóstica; e se o teste/estratégia de referência realmente tinha condições de classificar a condição-alvo, entre outros (115).

Estudos apropriados de teste de acurácia incluem pacientes com diagnóstico incerto que são representativos da população-alvo e com diferentes espectros de gravidade da doença. Esses estudos devem incluir, preferencialmente, pacientes consecutivos ou selecionados aleatoriamente nos quais existe incerteza diagnóstica. Se os estudos falharem nesse critério e, por exemplo, incluírem pacientes gravemente afetados e controles saudáveis, a acurácia aparente de um teste provavelmente será enganosamente alta e, nesse caso, geraria risco de viés (121, 122).

Quadro 36 - Critérios de risco de viés de estudos diagnósticos da ferramenta QUADAS-2

Critérios	Seleção de pacientes	Teste índice	Teste de referência	Fluxo e tempo
Descrição	Houve descrição dos métodos de seleção de pacientes? Houve descrição dos pacientes incluídos?	Houve descrição do teste índice e sobre como ele foi conduzido e interpretado?	Houve descrição do padrão de referência e sobre como ele foi conduzido e interpretado?	Houve descrição de todos os pacientes que não receberam os testes índice ou padrão de referência ou que foram excluídos da tabela 2 x 2? Houve

Quadro 36 - Critérios de risco de viés de estudos diagnósticos da ferramenta QUADAS-2

Critérios	Seleção de pacientes	Teste índice	Teste de referência	Fluxo e tempo
				descrição do intervalo e de quaisquer intervenções entre os testes índices e o padrão de referência?
Questões sinalizadoras (sim, não ou incerto)	Questão 1: foi incluída uma amostra de pacientes consecutiva ou aleatória? Questão 2: foi evitado um desenho caso-controle? Questão 3: o estudo evitou exclusões inapropriadas?	Questão 1: os resultados do teste índice foram interpretados sem conhecimento dos resultados do padrão de referência? Questão 2: se um limiar foi usado, ele havia sido pré-especificado?	Questão 1: o teste de referência provavelmente classifica corretamente a condição clínica de interesse? Questão 2: os resultados do padrão de referência foram interpretados sem conhecimento	Questão 1: houve um intervalo adequado entre os testes índices e o teste de referência? Questão 2: todos os pacientes receberam o mesmo teste de referência? Questão 3: todos os pacientes

Quadro 36 - Critérios de risco de viés de estudos diagnósticos da ferramenta QUADAS-2

Critérios	Seleção de pacientes	Teste índice	Teste de referência	Fluxo e tempo
			dos resultados do teste índice?	foram incluídos na análise?
Risco de viés (alto, baixo ou incerto	A seleção de pacientes pode ter introduzido viés?	A conduta ou interpretação do teste de índice pode ter introduzido viés?	O padrão de referência, sua conduta ou sua interpretação pode ter introduzido viés?	O fluxo de pacientes pode ter introduzido viés?

Fonte: Adaptado de Whiting et al. (32) e Schünemann et al. (115).

Evidência indireta

A evidência direta se refere a estudos que avaliam diretamente a população e os testes/estratégias de interesse e mensuram os resultados importantes para os pacientes[1]. Já a evidência indireta é proveniente de estudos com populações, testes/estratégias, comparadores e/ou desfechos que não são exatamente iguais aos de interesse estabelecidos na questão de pesquisa. Assim como na avaliação de intervenções terapêuticas, a evidência indireta deve ser avaliada a partir de características da população, cenário, intervenção (teste novo ou índice) e comparador (padrão de referência) (115).

Em relação à população, a certeza do conjunto final de evidências pode ser rebaixada se houver diferenças importantes entre as populações estudadas e aquelas para as quais a recomendação se destina (em testes anteriores, no espectro da

doença ou comorbidade), no contexto de diretrizes. Deve-se verificar, também, se há diferenças importantes no cenário em que se aplicará a recomendação/evidência no que se relaciona às características entre os testes/estratégias aplicados ou na experiência que os profissionais poderiam ter no momento de realização do teste, como, por exemplo, pacientes atendidos em um serviço de emergência podem ser diferentes de pacientes atendidos em um consultório médico Além disso, a prevalência ou probabilidade pré-teste pode ser um guia para julgar se há evidência indireta na população, uma vez que podem haver diferenças entre prevalência média na evidência disponível e o que é encontrado na prática (115).

Em relação à intervenção ou ao teste índice, pode ocorrer evidência indireta quando os testes nos estudos/revisões encontrados foram implementados com padrões ligeiramente diferentes dos padrões usados na prática ou no contexto específico para o qual a orientação se destina (por exemplo, país, especialidade ou plano de saúde). Diferentes valores de corte ou limiares entre as configurações também podem levar à evidência indireta (e, muitas vezes, explicam a inconsistência nas análises de sensibilidade) (115).

Em relação ao comparador, pode ocorrer evidência indireta quando estudos de testes diagnósticos comparam o teste de interesse com outro teste que não tem relação com a questão de pesquisa. Ainda, se a questão clínica for sobre a escolha entre dois testes, nenhum dos quais é um padrão de referência, ambos os testes podem ser comparados diretamente com o teste de referência no mesmo estudo (115).

Por fim, o desfecho quase sempre será penalizado por evidência indireta no contexto de diretrizes clínicas, pois, embora os desfechos devam ser baseados nos desfechos de estudos de intervenção, as evidências disponíveis geralmente incluem resultados de acurácia como desfecho. Desenvolvedores de diretrizes clínicas geralmente enfrentam um cenário de evidências no qual não se tem resultados diretos sobre o impacto do uso dos testes nos desfechos críticos ou importantes para os pacientes. As recomendações e decisões devem ser baseadas nos resultados da intervenção que seguem os resultados do teste, quando a evidência disponível geralmente inclui apenas a acurácia do teste como resultado. Assim, é necessário fazer inferências sobre as possíveis influências dos falso-positivos ou verdadeiro-positivos e dos falso-negativos ou verdadeiro-negativos sobre os desfechos

importantes para os pacientes, complicações e custos relacionados aos testes. Portanto, de forma geral, os estudos de acurácia apresentam baixa certeza de evidência nos cenários de realização de recomendações em diretrizes clínicas por apresentarem resultados baseados em medidas indiretas em vez de nos desfechos importantes para os pacientes. Já na avaliação da certeza de evidência em uma RS, a evidência indireta pode não ser penalizada, pois a questão de pesquisa se concentra no item diagnóstico do teste de acurácia (115).

Inconsistência

A avaliação do domínio inconsistência parte da identificação de similaridade ou da heterogeneidade dos resultados relatados pelos estudos de testes e estratégias diagnósticos que buscaram responder à mesma pergunta PIRO. Recomenda-se que a avaliação das estimativas de sensibilidade (verdadeiro-positivo e falso-negativo) e especificidade (falso-positivo e verdadeiro-negativo) identificadas nos estudos seja realizada de forma separada, analisando tanto as similaridades das estimativas pontuais quanto a sobreposição dos IC, as variâncias das estimativas obtidas por modelos de efeitos aleatórios nas metanálises (quando disponíveis) e as análises obtidas dos testes de heterogeneidade (como a avaliação do qui-quadrado e I^2). No entanto, é importante compreender que os testes estatísticos de heterogeneidade não são considerados medidas adequadas para a utilização isolada na avaliação de inconsistência em estudos de acurácia diagnóstica.

Quando diferenças em algum componente da pergunta PIRO são identificadas, a avaliação dos resultados em subgrupos pode ser necessária para evitar o rebaixamento da evidência por inconsistência. O grupo GRADE indica que variabilidade na escolha dos pontos de corte dos testes talvez explique algumas heterogeneidades identificadas durante a avaliação de inconsistência, e a análise da curva característica de operação do receptor (*receiver operating characteristic*, ROC) poderia contribuir na elucidação da heterogeneidade. Uma inconsistência que apresente explicação em análises de subgrupos para os desfechos de sensibilidade, especificidade ou razões de proporções do teste índice entre os resultados dos estudos incluídos nas evidências pode acarretar uma redução da certeza de evidência. As orientações presentes no capítulo sobre avaliação da inconsistência em estudos de intervenção (capítulo 3) também se aplicam à avaliação das evidências

de testes de acurácia e estratégias diagnósticas, podendo ser acessadas para maiores detalhes na avaliação de cada um dos itens citados neste parágrafo (123).

Imprecisão

O domínio imprecisão, assim como a inconsistência, deve ser avaliado separadamente para os desfechos de sensibilidade e especificidade do teste índice. Recomenda-se a utilização de critérios semelhantes aos utilizados para a avaliação do domínio imprecisão em perguntas que envolvem sínteses de resultados de estudos de intervenção (3.3.4 Imprecisão). Assim, IC amplos em estimativas do teste de acurácia ou em estimativas de taxas (inclusive verdadeiro-positivos, verdadeiro-negativos, falso-positivos e falso-negativos) podem reduzir a certeza de evidência por imprecisão. No entanto, o grupo GRADE indica que a extensão da amplitude do IC para a redução da certeza de evidência é uma questão de julgamento e pode variar de acordo com o contexto de análise do teste índice.

A avaliação de imprecisão em revisões sistemáticas com foco em testes e estratégias diagnósticos ocorre a partir da inspeção da amplitude do IC no número de participantes dos estudos. O IC depende do número de eventos observados nos estudos: na avaliação de sensibilidade, isso corresponde ao número de pessoas com a doença e ao número de testes positivos; já em especificidade, corresponde ao número de pessoas sem a doença e ao número de testes negativos. Em um exemplo hipotético, se o limiar de menor sensibilidade aceitável for 0,8, uma estimativa pontual de 0,86 junto a um IC95% de 0,75 a 0,95 incluiria um intervalo abaixo do limiar de interesse e, portanto, não indicaria uma avaliação imprecisa sobre os potenciais benefícios e prejuízos do teste.

Já na avaliação de imprecisão por meio de abordagens contextualizadas, voltadas sobretudo para a realização de diretrizes, as estimativas de sensibilidade e especificidade devem ser transformadas em números absolutos de verdadeiro-positivos, falso-positivos, falso-negativos e verdadeiro-negativos de acordo com a prevalência da condição de saúde estudada. A identificação de limiares que reflitam implicações clínicas para o tratamento de pacientes e da condição estudada é recomendada para a avaliação de imprecisão em abordagens contextualizadas. Quando os limites dos IC incluem valores que podem resultar em diferentes conclusões sobre os valores do teste índice, a certeza da evidência deve ser

rebaixada por imprecisão. Assim, um IC relativamente estreito ainda pode ser amplo o suficiente para rebaixar a evidência. Por exemplo, se a prevalência for de 1% em uma determinada condição testada em 1.000 pessoas/ano, espera-se que 10 pessoas tenham a doença. Nesse caso, um IC mais amplo em torno da sensibilidade pode levar a uma menor preocupação com a imprecisão em comparação a um IC de menor amplitude em um cenário de prevalência de aproximadamente 40% (indicando que 400 pessoas apresentariam a doença a cada 1.000 testadas). A imprecisão, dessa forma, pode ser considerada não grave quando a amplitude do IC for estreita, de modo que seu limite inferior ou superior não alteraria a recomendação do teste índice.

Viés de publicação

Um alto risco na avaliação do viés de publicação (por exemplo, evidência com base em estudos pequenos ou assimetria no gráfico de funil) podem reduzir a certeza da evidência no cenário da avaliação de testes de acurácia e estratégias diagnósticas (123). Devem-se utilizar critérios semelhantes aos utilizados para avaliar o viés de publicação em perguntas sobre tratamentos, detalhados no capítulo 3.

Fatores que podem aumentar a certeza da evidência (gradiente dose-resposta, grande magnitude de efeito e fatores de confusão)

Para esses itens, o grupo GRADE considera que os métodos ainda não foram completamente desenvolvidos e, portanto, o julgamento deve ser baseado em critérios semelhantes aos utilizados para avaliar evidências nas perguntas de tratamentos. No entanto, determinar a presença de gradiente dose-resposta ou de alta probabilidade de a doença/característica estar corretamente associada aos resultados pode ser importante para aumentar a certeza da evidência. Porém, existem discordâncias sobre se ou como esses efeitos devem ter um papel na avaliação da certeza da evidência em estudos diagnósticos (123).

A certeza das evidências em testes de acurácia pode aumentar se os resultados da curva ROC mostrarem uma relação clara e consistente entre sensibilidade e especificidade (o equivalente para avaliação em estudos de tratamento seria o gradiente dose-resposta) (123). Uma acurácia muito alta em teste e uma mínima presença de confusão residual oposto também podem aumentar a certeza da evidência nos resultados dos testes (87). Porém, deve-se ter parcimônia

nas tentativas de aumentar a certeza da evidência com base no tamanho de efeito dos resultados, pois ainda não existe um consenso, inclusive entre o GRADE *Working Group,* (123) que indique essa ação (Quadro 37).

Quadro 37 – Avaliação da certeza da evidência de testes ou estratégias diagnósticas em cenários complexos e desfechos importantes para os pacientes

Para mais detalhes sobre o pensamento e as orientações do GRADE *Working Group* sobre a avaliação da certeza da evidência em cenários complexos (avaliação de desfechos importantes para os pacientes e recomendações), envolvendo a utilização de diversos testes e estratégias diagnósticas, sugere-se a leitura do artigo GRADE *Guidelines*: 22 (124), em que os autores apresentam detalhes sobre o processo de tomada de decisões.

Fonte: adaptado de Schünemann et al. (124).

6.3. Síntese das evidências para testes/estratégias diagnósticos utilizando o sistema GRADE

O julgamento sobre os desfechos de testes ou estratégias diagnósticas deve ser baseado em evidências obtidas por meio de uma RS, preferencialmente com metanálise (125). É recomendado que a apresentação da síntese de evidências para cada desfecho seja realizada através de uma tabela de perfil de evidências (Figura 26) ou SoF (Figura 27), pois ajuda a garantir a transparência para o processo de tomada de decisão (123). As tabelas utilizadas para apresentar os desfechos de testes ou estratégias diagnósticas têm um formato de apresentação específico para esse delineamento, que difere do formato utilizado nas tabelas de apresentação de perguntas sobre terapias ou intervenções (123). Entre as diferenças, está a configuração padrão dos quatro desfechos definidos para uma questão de diagnóstico: verdadeiro-positivos, falso-negativos, verdadeiro-negativos e falso-positivos.

Os formatos propostos para a apresentação estão disponíveis por meio da ferramenta GRADEpro (<https://www.gradepro.org/>). Deve-se sempre priorizar um formato de apresentação que permita a visualização mais fácil do processo de avaliação, de forma transparente e objetiva (123).

Figura 26 – Exemplo de estrutura para apresentação detalhada dos resultados de teste/estratégia diagnóstica utilizando uma tabela de perfil de evidências através da ferramenta GRADEpro

Pergunta: deve-se usar [teste em estudo] para diagnosticar [condição de interesse] em [problema de saúde e/ou população]?

Sensibilidade	0.00 (IC95% 0,00 a 0,00)
Especificidade	0.00 (IC95% 0,00 a 0,00)

Prevalências	[número] %	[número] %	[número] %

Desfecho	Nº. de estudos (Nº. de pacientes)	Delineamento do estudo	Fatores que podem reduzir a certeza da evidência					Efeito para 100 pacientes testados	Acurácia do teste
			Risco de viés	Evidência indireta	Inconsistência	Imprecisão	Viés de publicação	Probabilidade pré-teste de [número]%	
Verdadeiro-positivos (pacientes com [condição de interesse])	[número] estudos [número] pacientes							0 (0 para 0)	
Falso-negativos (pacientes incorretamente classificados como não tendo [condição de interesse])								0 (0 para 0)	

Verdadeiro-negativos (pacientes sem [condição de interesse])	[número] estudos [número] pacientes							0 (0 para 0)	
Falso-positivos (Pacientes com [condição de interesse] incorretamente classificados)								0 (0 para 0)	

Fonte: Elaboração própria. Realizada na ferramenta *online* GRADEpro, disponível em https://www.gradepro.org/.

Figura 27 – Exemplo de estrutura para apresentação de resultados de teste/estratégia diagnóstica resumidos e utilizando a ferramenta GRADEpro

Deve-se usar [teste em estudo] para diagnosticar [condição de interesse] em [problema de saúde e/ou população]?

Paciente ou população: [problema de saúde e/ou população]

Contexto: teste

Teste novo: [teste comparador] | **Ponto de corte**:

Teste de referência: teste | **Limiar**: teste

Sensibilidade combinada: 0,70 (IC95% 0,60 a 0,80)| **Especificidade combinada**: 0,80 (IC95% 0,70 a 0,90)

Resultado do teste	Número para 100 pacientes testados (IC95%) **Prevalência [número]%** Comumente visto em	Número de participantes (estudos)	Certeza da evidência (GRADE)
Verdadeiro-positivos	0 (0 a 0)	([número])	-

Falso-negativos	0 (0 a 0)			
Verdadeiro-negativos	0 (0 a 0)	([número])	-	
Falso-positivos	0 (0 a 0)			

IC = intervalo de confiança.

Fonte: Elaboração própria. Realizada na ferramenta *online* GRADEpro, disponível em https://www.gradepro.org/.

6.4. Elaboração de tabela de evidência para decisão de testes/estratégias diagnósticos utilizando o sistema GRADE

Os critérios para a avaliação da EtD em testes e estratégias diagnósticos são os mesmos utilizados em recomendações clínicas, saúde pública e decisões de incorporação sobre intervenções de saúde (5. Uso do GRADE para o desenvolvimento de recomendações). Contudo, há a distinção do item “acurácia do teste”, em que é questionada a acurácia do teste ou estratégia diagnóstica. Devido aos desfechos específicos relacionados à acurácia, também são propostos direcionamentos específicos para o processo de avaliação em alguns itens, que são voltados para a precisão e acurácia do teste ou estratégia diagnóstica (125).

Deve-se realizar uma avaliação detalhada de cada item que compõe a avaliação da EtD, principalmente no impacto da acurácia do teste para os desfechos importantes para o paciente (125). Ademais, caso seja observada uma baixa acurácia na avaliação do item “acurácia do teste” para um novo teste em comparação a um teste ou uma estratégia já existente, é improvável que seja necessário considerar detalhadamente os outros itens da tabela EtD, pois uma baixa acurácia geral já é suficiente para direcionar uma decisão contra o seu uso.

Na seção de material suplementar há um exemplo de avaliação da EtD para uma questão PIRO (Anexo II).

7. Sistema GRADE para prognóstico, incidência e prevalência

Prognóstico, incidência e prevalência são métricas fundamentais para a tomada de decisões em saúde. O prognóstico refere-se à estimativa do risco de ocorrência de eventos futuros; a incidência quantifica novos casos de determinada condição clínica em uma população específica; e a prevalência indica a proporção de indivíduos que apresentam uma dada condição em um momento ou período específico.

Considerando a relevância destas estimativas na área da saúde, torna-se essencial avaliar a certeza da evidência, garantindo decisões bem fundamentadas. Apesar disso, tal prática ainda pode ser considerada incipiente. Por exemplo, um estudo que avaliou 235 RS de prevalência identificou que apenas 9, ou seja, 3,8% das revisões, incluíram algum tipo de avaliação formal da certeza da evidência, sendo que apenas 4 (1,7%) seguiu alguma recomendação do grupo GRADE (126, 127).

Na última década, o GRADE *Working Group* publicou recomendações sobre a avaliação da certeza das estimativas de prognóstico (47, 127, 128). Isso inclui o prognóstico geral, que reflete o risco de um determinado evento ou condição clínica surgir em uma população - conceito que pode ser comparado à incidência. Detalhes sobre essas recomendações serão abordados nas próximas seções. No entanto, até a data de elaboração deste manual, não há recomendações formais do grupo GRADE para avaliar a certeza da evidência das estimativas de prevalência. Por isso, sugere-se adaptar a metodologia utilizada para avaliar estimativas de prognóstico, conforme observado em algumas RS publicadas (129, 130). Ao longo do texto, serão dados exemplos de como tal prática pode ser realizada.

7.1 Prognóstico em Ciências da Saúde

De maneira geral, o termo prognóstico descreve a probabilidade de um evento futuro em uma população de interesse. Entretanto, estudos de prognóstico podem apresentar diferentes objetivos, como estabelecer o risco de determinado desfecho em uma população de interesse, avaliar o efeito de uma característica do indivíduo sobre o risco de um desfecho ou desenvolver um modelo preditivo do desfecho de interesse, conforme apresentado no Quadro 38.

Quadro 38 – Principais tipos e objetivos de estudos de prognósticos.

Tipo de estudo	Objetivo de estudo	Exemplo
Prognóstico geral	Estabelecer a probabilidade de um desfecho de interesse em uma população.	Risco de sangramento em pacientes com fibrilação atrial em uso de antagonistas da vitamina K.
Fatores prognósticos	Estabelecer o impacto de características do indivíduo sobre o risco de desenvolver o desfecho de interesse.	Influência da idade no risco de sangramento em pacientes com fibrilação atrial.
Modelo preditivo de desfecho (ou de risco)	Desenvolvimento de um modelo prognóstico que considera simultaneamente uma série de fatores prognósticos e classifica os indivíduos em vários níveis de risco	Escores CHAD2 e CHADS-VASC para avaliação do risco de eventos cerebrovasculares.

Fonte: adaptado de Iorio et al. (48).

Os critérios para avaliar a certeza da evidência variam conforme o tipo de estudo prognóstico. Discutiremos a aplicação da metodologia GRADE para determinar a certeza da evidência em estudos focados no prognóstico geral (48). Esta abordagem pode ser diretamente aplicada a estimativas de incidência e adaptada para estimativas de prevalência. Informações sobre a avaliação da evidência de estudos que avaliam estudos prognósticos podem ser obtidas no artigo GRADE número 28 (47).

7.2 Avaliação da certeza da evidência em estudos sobre prognóstico

A avaliação da certeza da evidência para os desfechos obtidos de estudos de prognósticos ocorrerá por meio dos cinco domínios que podem rebaixar a certeza da evidência: risco de viés, inconsistência, evidência indireta, imprecisão, e viés de publicação. Adicionalmente, dois dos três domínios que podem elevar a certeza da evidência parecem ser aplicáveis a estudos de prognóstico: magnitude de efeito e efeito dose-resposta (131). Em estudos de prevalência, entende-se que não há aplicabilidade dos domínios que podem elevar a certeza da evidência, sendo essas estimativas avaliadas apenas por meio dos cinco domínios que podem rebaixar a certeza da evidência.

A interpretação do julgamento final em quatro níveis (alto, moderado, baixo ou muito baixo) da certeza da evidência em estudos de prognóstico está apresentada no Quadro 39.

Quadro 39 - Significado dos níveis de evidência para desfechos de prognóstico.

Nível de certeza	**Definição**
Alta	Estamos muito confiantes de que o verdadeiro prognóstico (probabilidade de eventos futuros) está próximo ao estimado.
Moderada	Estamos moderadamente confiantes de que o verdadeiro prognóstico (probabilidade de eventos futuros) provavelmente está próximo ao estimado, mas existe a possibilidade de que seja substancialmente diferente.
Baixa	Nossa confiança na estimativa é limitada: o verdadeiro prognóstico (probabilidade de eventos futuros) pode ser substancialmente diferente do estimado.
Muito baixa	Nós temos muito pouca certeza na estimativa: o verdadeiro prognóstico (probabilidade de eventos futuros) provavelmente será substancialmente diferente da estimativa.

Fonte: adaptado de Iorio et al. (48).

7.3. Delineamento do estudo

Estimativas de prognóstico são originadas a partir de estudos longitudinais observacionais, como estudos de coorte, e até mesmo ensaios clínicos, nos quais cada braço pode ser considerado um estudo observacional. A avaliação da certeza da evidência em desfechos de prognósticos considera que estudos de coorte iniciam com alta confiança, enquanto estimativas provenientes de estudos experimentais, como ECRs, iniciam com baixa certeza na evidência. Isso ocorre porque os critérios de elegibilidade de ECRs geralmente são restritos e menos generalizáveis, excluindo indivíduos relevantes para a questão de interesse em prognóstico (127). Cabe notar que ECRs pragmáticos de considerável tamanho amostral e com critérios amplos para a elegibilidade de pacientes podem gerar estimativas confiáveis em prognósticos, e tornam-se uma exceção à regra.

Dados de prevalência são tradicionalmente obtidos a partir de estudos transversais, mas também podem ser obtidos a partir da linha de base de estudos longitudinais, sejam observacionais ou experimentais. A certeza da evidência inicia como alta ao se tratar de estimativas obtidas a partir de estudos observacionais, seguindo o mesmo racional apresentado para estimativas de prognóstico.

7.*4 Risco de viés*

As limitações no delineamento e na execução dos estudos individuais, que podem levar à super- ou subestimação das medidas de interesse, devem ser observadas na avaliação do risco de viés.

Para estudos de prognóstico geral, devemos avaliar: se a amostra avaliada foi representativa e incluiu um grupo homogêneo de indivíduos que não apresentavam o desfecho de interesse no início do período de observação; se houve seguimento completo e suficiente para o desenvolvimento do desfecho de interesse; se a mensuração do desfecho foi realizada de maneira não enviesada e objetiva; e se houve registro e ajuste para os principais fatores confundidores.

Tal avaliação pode ser realizada com o auxílio de ferramentas desenvolvidas para esse objetivo. O *QUality In Prognosis Studies* (QUIPS), detalhado no **Quadro 40**, é uma ferramenta desenvolvida para a avaliação do risco de viés em estudos com foco

em desfechos prognósticos, especificamente prognóstico geral (132). Digno de nota, essa ferramenta pode não ser indicada para os demais delineamentos de estudos de prognóstico (como fatores prognósticos e modelos prognósticos); para esses estudos, outras ferramentas podem ser utilizadas, como o instrumento PROBAST (*Prediction model Risk Of Bias ASsessment Tool*) com foco na avaliação do risco de viés de modelos preditivos de desfecho (133).

Quadro 40 – Domínios do QUIPS para avaliação do risco de viés em estudos com foco em desfechos prognósticos

Domínios	Descrição	Exemplo de alto risco de viés
Participação no estudo	Julga a representatividade da amostra do estudo.	Baixa taxa de participação no estudo; Diferentes taxas de distribuição de idade e sexo em comparação à população de origem; Alta elegibilidade amostral.
Atrito	Verifica se há tendência na associação relatada entre o fator prognóstico e o desfecho por uma avaliação de desfechos em participantes que completaram o estudo.	A diferença entre os participantes que completaram e não completaram o estudo distorce a associação entre o fator prognóstico e o desfecho.
Mensuração do fator prognóstico	Verifica o processo de mensuração do fator prognóstico, permitindo julgar se o estudo o mediu de maneira semelhante, válida e	Pesquisas que adotam método não confiável para avaliar o fator prognóstico ou utilizam abordagens diferentes entre os participantes.

Quadro 40 – Domínios do QUIPS para avaliação do risco de viés em estudos com foco em desfechos prognósticos

Domínios	Descrição	Exemplo de alto risco de viés
	confiável entre todos os participantes.	
Mensuração do desfecho	Verifica como foi mensurado o desfecho, permitindo o julgamento de semelhança entre os estudos para a avaliação do desfecho de modo confiável e válido entre todos os participantes.	Observada quando há mensuração de diferentes formas para o desfecho em relação à exposição ao fator prognóstico.
Fatores de confusão	Verifica potenciais variáveis de confusão importantes e inclusão de variáveis em análises multivariadas pré-especificadas.	Sobreposição de impacto de fatores potencialmente prognósticos que podem alterar o efeito do fator prognóstico de estudo, e ausência de ajustes em análise multivariada.
Análise estatística e relatos seletivos	Analisa a adequação da análise estatística do estudo e a integridade dos relatos.	Análise estatística equivocada e com suspeita de relato apenas de fatores associados aos resultados de forma positiva foram reletados.

Adaptado de: Hayden et al. (132).

Uma alta variação no risco de viés pode ser identificada entre os estudos de estimativas de prognóstico, e o peso dos estudos no conjunto de evidências deve ser considerado no processo de avaliação. Desta forma, a inclusão de um ou mais estudos

com alto risco de viés não necessariamente reduz a certeza da evidência caso contribuam apenas com uma pequena proporção de eventos na estimativa agrupada. O GRADE *Working Group* recomenda que análises de sensibilidade sejam realizadas durante a avaliação da certeza da evidência para identificar se os resultados de estimativas são semelhantes em estudos com menor e maior risco de viés, de modo a identificar o impacto dos estudos de alto risco de viés na certeza (127). Caso uma grande discrepância entre as estimativas de prognóstico seja identificada, o grupo GRADE recomenda a adoção apenas de estimativas dos estudos de menor risco de viés, sem o rebaixamento da certeza da evidência para o domínio de risco de viés (47).

Para a avaliação do risco de viés em estudos de prevalência, estudo recente identificou 30 ferramentas designadas para o julgamento deste domínio (134). Todas as ferramentas identificadas apresentaram limitações, sendo o instrumento *Joanna Briggs Institute Checklist for Prevalence Studies* considerado o mais adequado entre os atualmente disponíveis (119). Os principais fatores a serem considerados na avaliação do risco de viés de estudos de prevalência são a representatividade da amostra avaliada e a mensuração adequada da condição de interesse.

7.5 Inconsistência

A avaliação de inconsistência em estimativas de estudos de prognóstico é semelhante ao recomendado para avaliação em estudos de intervenção (41). Ou seja, os critérios de avaliação de inconsistência incluem a avaliação da variabilidade nas estimativas pontuais, presença e extensão da sobreposição dos ICs, e onde as estimativas pontos se situam em relação aos limiares de decisão.

A utilização de abordagens estatísticas, como o I^2, possui menor utilidade em estudos de prognóstico e prevalência (127). Frequentemente, os estudos incluídos possuem grande tamanho amostral, o que pode levar a ICs estreitos, resultando em alto I^2 e indicação de heterogeneidade estatística mesmo na ausência de uma inconsistência importante. Estudo que avaliou 134 metanálises de prevalência identificou uma mediana de I^2 de 96,9% (intervalo interquartil de 90,5 - 98,7%), sendo que 125 (93,3%) dos estudos apresentou $I^2 \geq 70\%$ (46).

Assim, ao avaliar a inconsistência de evidências de prognóstico e prevalência, a extensão da variação das estimativas pontos é o fator de maior relevância, e estatísticas como o I^2 devem ser interpretadas com cautela (46, 127). Além disso, os

autores de RS devem gerar hipóteses *a priori* que possam explicar uma alta inconsistência, caso essa seja observada. Análises de subgrupos podem ser realizadas e, se houver diferença entre os grupos, cada estimativa pode ser considerada separadamente, dispensando a necessidade de penalizar a evidência por inconsistência (46, 127).

7.6 Imprecisão

A avaliação do domínio de imprecisão em estudos de prognóstico e prevalência deve ser realizada considerando a largura do IC 95% em torno da estimativa pontual.

Em uma abordagem contextualizada, que deve ser realizada sobretudo no contexto do desenvolvimento de recomendações para diretrizes clínicas, deve-se considerar penalizar a certeza na evidência quando o efeito no paciente ou a ação clínica derivada da estimativa em avaliação seria diferente a depender se o limite superior ou o limite inferior do intervalo representasse a verdade. Caso se considere que a tomada de decisão clínica não mude devido à amplitude do IC, o julgamento será de uma estimativa precisa e a certeza da evidência não será rebaixada. A decisão final sobre limiares de risco para a tomada de decisão pode diferir para os autores de revisões sistemáticas e painéis de diretrizes clínicas, resultando em diferentes julgamentos por imprecisão. Idealmente os autores de revisões sistemáticas devem apresentar seus resultados de forma a permitir que os usuários examinem as implicações considerando diferentes limiares.

7.7 Evidência indireta

A avaliação da evidência indireta em estudos de prognóstico e prevalência deve ser realizada em abordagem semelhante ao indicado pelo GRADE *Working Group* para estudos de intervenção (55). Isto é, os autores devem considerar o quanto a população e os desfechos dos estudos correspondem à população e aos desfechos de interesse. Quando as estimativas de efeito são consideradas de evidência indireta, a certeza da evidência pode ser rebaixada em um ou dois níveis a depender do grau do comprometimento que a avaliação poderá apresentar.

7.8 Viés de publicação

Viés de publicação ocorre quando decisões sobre a publicação de resultados de estudos são influenciadas pelo resultado de testes estatísticos ou pela magnitude ou direção do efeito observado. Como consequência, os resultados dos estudos

identificados e incluídos em uma síntese são sistematicamente diferentes dos resultados dos estudos não identificados, resultando em estimativas enviesadas.

Tal conceito é facilmente aplicável em estudos que avaliam a efetividade e segurança de intervenções, porém o mesmo não ocorre para estudos de prognóstico geral ou prevalência. Considerando que tais estimativas não são avaliadas para significância estatística, e que são estimativas não comparadas, fatores como o resultado de testes estatísticos ou a direção do efeito não se aplicam. A influência da magnitude da estimativa sobre a decisão de publicação é incerta e pode variar a depender do desfecho em estudo. Assim, para avaliar tal domínio, os revisores devem definir se, no contexto de interesse, a evidência observada está sujeita a viés de publicação.

Cabe ressaltar que testes estatísticos comumente utilizados, como o teste de Egger, são aplicáveis quando os dados apresentam distribuição normal e há baixa heterogeneidade, pressupostos que não são atendidos por estimativas de prognóstico e prevalência, tornando tais testes inapropriados.

7.9 Domínios que podem elevar a certeza da evidência

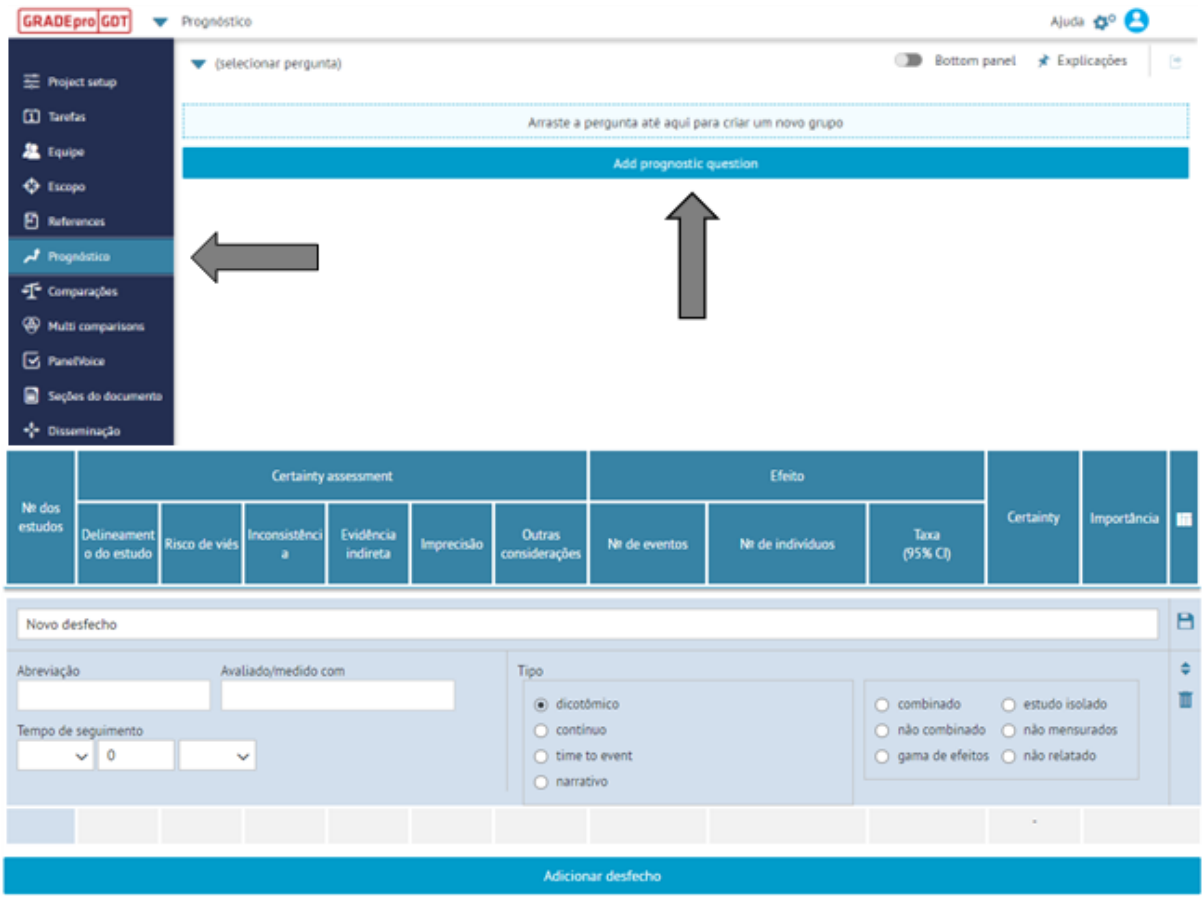
Os critérios GRADE para aumentar a confiança em estudos de tratamento incluem grande magnitude efeito, gradiente de dose-resposta e confundidores residuais na direção oposta. Em geral, não se observa situações claras para aumento do nível de evidência, em especial nessas situações onde não temos medidas comparadas de efeito. Do ponto de vista teórico, situações análogas podem existir para os dois primeiros critérios em revisões sistemáticas de prognóstico. Por exemplo, um aumento no número de eventos durante o período de seguimento do estudo em um padrão bem definido, linear ou não, pode aumentar a confiança em qualquer um dos pontos de dados que contribuem para o padrão, evidenciando um gradiente dose-resposta. Entendemos que as situações nas quais há aumento na certeza da evidência são pouco comuns; caso ocorra, a mesma deve ser adequadamente explicada e justificada.

7.10. *Construção da tabela sumária de evidências*

A criação da tabela sumária de evidências para a questão de prognóstico pode ser realizada diretamente no site do GRADEpro (disponível em https://www.gradepro.org/) na opção “adicionar uma questão de prognóstico”, conforme Figura 28. Para a construção da tabela, deve-se identificar:

- o tempo de seguimento para mensuração do desfecho,
- o desfecho mensurado, o número de estudos,
- o delineamento dos estudos,
- a avaliação dos cinco domínios que podem diminuir a certeza da evidência,
- outras considerações relevantes, incluindo a avaliação dos domínios que podem elevar a certeza da evidência,
- a descrição do efeito em termos relativos, absolutos ou narrativos.

Figura 28 - Construção da tabela sumária de evidências em estudos de prognóstico através da ferramenta GRADEpro



Fonte: Ferramenta *online* GRADEpro, disponível em https://www.gradepro.org/.

A avaliação da certeza da evidência em estudos de prognóstico, prevalência e incidência pelas recomendações do GRADE *Working Group* segue abordagem similar àquelas voltadas aos estudos de intervenção, porém com importantes adaptações necessárias para incorporar as características dessas estimativas. A adoção da orientação do GRADE e documentação da sua sequência lógica de aplicação garantem a compreensão da evidência disponível e os motivos para a classificação da certeza da evidência.

8. Sistema GRADE para metanálises em rede

abordagem sistema GRADE para metanálise em rede (*network meta-analysis*, NMA) tem avançado nos últimos anos com o fortalecimento dos conceitos de NMA, tornando o processo de avaliação da certeza da evidência mais eficiente. Em 2014, foi publicado pelo GRADE *Working Group* o primeiro estudo sobre como avaliar a qualidade das evidências que apoiam as estimativas do efeito de um tratamento obtidas por meio de NMA. Em 2018, o grupo publicou um novo artigo com o objetivo de esclarecer o processo (135). Desde então, outros artigos mais específicos foram publicados com o intuito de aprofundar a avaliação da certeza da evidência para NMA, os quais serão abordados ao longo deste capítulo (136-139).

8.1 Conceitos de metanálise em rede

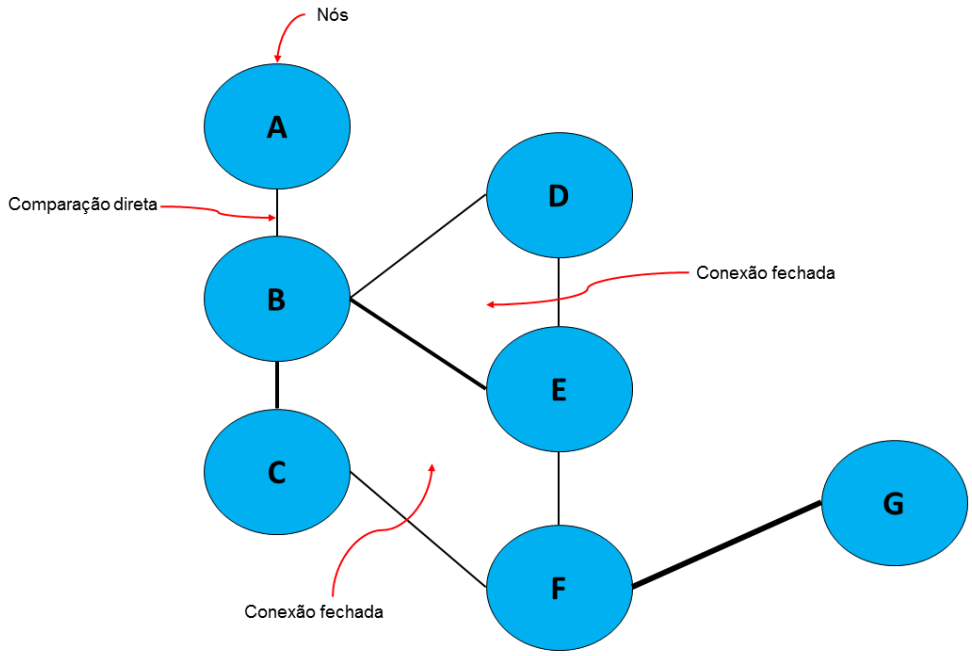
A NMA, ou metanálise de comparações múltiplas, é um método estatístico para comparar múltiplas intervenções (três ou mais) simultaneamente em uma única análise, combinando evidências diretas e indiretas em uma rede de vários estudos que abordam a mesma questão de pesquisa (140).

As metanálises convencionais foram um avanço para a PBE, permitindo a reunião de informações com minimização de vieses. Embora esse tipo de análise seja restrito a comparações entre dois tratamentos (141), são frequentes os cenários clínicos com diversas opções terapêuticas disponíveis, não apenas duas. Nesses casos, metanálises convencionais de ensaios clínicos não fornecem informações completas sobre qual seria a melhor opção de tratamento. Essa situação é especialmente prevalente nos contextos de classes terapêuticas inovadoras e doenças raras ou negligenciadas, uma vez que a condução de um ensaio clínico requer grandes amostras, altos investimentos e tempo (142-144). Nesse sentido, as NMA podem ser uma alternativa promissora para processos de tomada de decisão (145).

As NMA podem ser realizadas por meio de diferentes técnicas. Há modelos disponíveis para todos os tipos de dados brutos, que produzem diferentes medidas de efeito agrupado, utilizando o método Bucher (146) ou abordagens frequentistas (147) e bayesianas (148), com diferentes programas. Em relação à estrutura das NMA, pode apresentar diferentes configurações, as quais são representadas por gráficos ou diagramas de rede. Os círculos, ou nós, representam a intervenção ou tecnologia de interesse. As linhas que conectam os nós representam as comparações diretas

disponíveis na literatura, como demonstra a Figura 29. Ainda, a representação gráfica da rede pode incluir a quantidade de evidências diretas disponíveis por meio do alargamento das linhas entre os nós e o número de estudos para cada intervenção (tamanho de nós) (149, 150). A presença de conexões fechadas (*closed loops)* indica que existe tanto evidência direta quanto indireta para determinada comparação na rede, e as conexões abertas apresentam resultados que dependem extensivamente das comparações indiretas estimadas (151, 152) .

Figura 29 – Componentes básicos do diagrama de redes



Uma rede é composta por pelo menos três nós (intervenções e comparadores) conectados por linhas (comparações diretas). Neste diagrama, a largura das linhas é proporcional ao número de evidências diretas disponíveis na literatura. Conexões fechadas podem ser formadas de acordo com a disponibilidade de evidências diretas e indiretas na literatura (por exemplo, B vs. C vs. E vs. F representa uma conexão fechada; B vs. D vs. E é outra conexão fechada). A evidência indireta é calculada utilizando um comparador em comum (por exemplo, estimativas entre A e D são feitas através de B; estimativas entre E e G são feitas através de F).
Fonte: adaptado de Tonin et al. (153).

O Quadro 41 apresenta conceitos e definições de NMA.

Quadro 41 - Conceitos e definições de NMA

Conceito	Definição
Comparador comum	Tratamento comum entre as intervenções que se deseja comparar. Se uma rede tem três tratamentos (A, B e C) e A está diretamente ligado a B, enquanto C também está diretamente ligado a B, o comparador comum dessa rede será B.
Comparação direta	Comparação entre duas intervenções por meio de estudos que comparam diretamente tecnologias ativas (*head to head*) ou comparação com placebo.
Comparação indireta	Estimativa de efeito fornecida por duas ou mais comparações diretas que compartilham um comparador comum (como estudos de A *versus* C e estudos de B *versus* C quando A *versus* B é a comparação de interesse).
Rede	Um conjunto de estudos de intervenções para uma condição clínica que permite, por meio de comparações diretas e indiretas, calcular os efeitos relativos de todos os tratamentos em comparação ao placebo ou tratamento padrão e entre si, em um desfecho específico.
Diagrama e geometria da rede	A base dessa análise de rede é um diagrama de rede (gráfico) em que cada nó representa uma intervenção e as linhas de conexão entre os nós representam um ou mais estudos nos quais as intervenções foram comparadas diretamente. A descrição das características da rede de intervenções, que pode incluir o uso de

Quadro 41 - Conceitos e definições de NMA	
Conceito	**Definição**
	estatísticas numéricas resumidas, é considerada uma avaliação da geometria da rede.
Loops	Duas ou mais comparações diretas que contribuem para uma estimativa indireta. *Loops* de primeira ordem são aqueles que envolvem apenas uma única intervenção adicional. Por exemplo, se a intervenção de interesse for A *versus* B, as estimativas diretas de A *versus* C e B *versus* C constituem um *loop* de primeira ordem. Um *loop* de segunda ordem envolveria duas outras intervenções (por exemplo: A *versus* C, C *versus* D e D *versus* B).
Ordem de classificação (*rank*)	Cálculos de probabilidade para ranqueamento das alternativas, em termos da probabilidade de ser o melhor tratamento.
Intransitividade	Diferenças nas características do estudo que podem modificar o efeito do tratamento nas comparações diretas (como A *versus* C e B *versus* C), que formam a base para a estimativa indireta do efeito da comparação de interesse (A *versus* B), criam viés na avaliação indireta de A *versus* B. Os fatores que podem modificar os efeitos do tratamento incluem diferentes características do paciente, diferentes cointervenções, extensão diferente em que as intervenções de interesse são administradas de forma otimizada, comparadores diferentes e diferenças na mensuração do resultado.

Quadro 41 - Conceitos e definições de NMA	
Conceito	**Definição**
Heterogeneidade	Diferenças nas estimativas de efeito entre estudos que avaliaram a mesma comparação.
Incoerência	Diferenças entre estimativas diretas e indiretas.

Fonte: adaptado de Tonin et al. (153) e Puhan et al. (154).

8.2 Uso do sistema GRADE para NMA

A avaliação da certeza da evidência para NMA utilizando o sistema GRADE é realizada por meio de uma abordagem em quatro etapas (154):

1. apresentação das estimativas diretas e indiretas do tratamento para cada comparação da rede de evidências. A estimativa direta é fornecida por uma comparação direta (estudos de A *versus* B), e a estimativa indireta é fornecida por duas ou mais comparações diretas que compartilham um comparador em comum (por exemplo, foram inferidos os efeitos dos estudos de A *versus* B, A *versus* C e B *versus* C);
2. classificação da certeza de cada estimativa de efeito direto e indireto;
3. apresentação da estimativa da NMA para cada comparação da rede de evidências;
4. classificação da certeza da evidência de cada estimativa de efeito da NMA.

Objetivando fortalecer a base conceitual para a avaliação da certeza da evidência de NMA e tornar o processo mais eficiente, foram descritos quatro avanços (135):

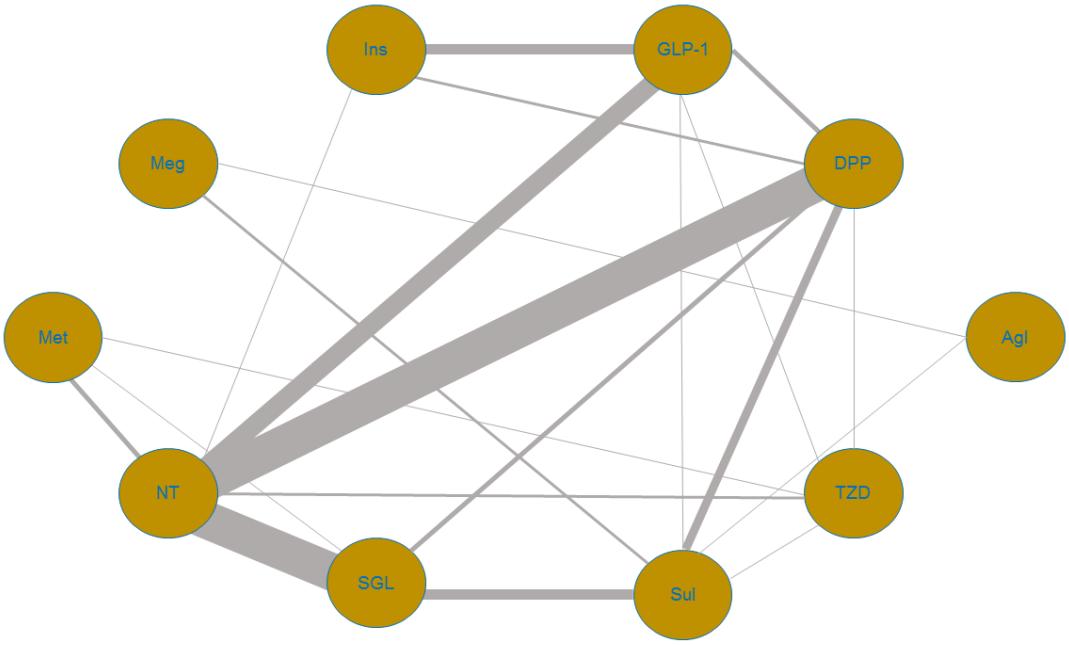
1. não é necessário considerar a imprecisão ao classificar as estimativas diretas e indiretas para informar a classificação das estimativas da NMA;

2. não é necessário avaliar/classificar a evidência indireta quando a certeza da evidência direta for alta e a contribuição da evidência direta para a estimativa da rede for, pelo menos, tão grande quanto a da evidência indireta;
3. não se deve confiar em um teste estatístico de incoerência global da rede para avaliar a incoerência no nível de comparação aos pares;
4. na presença de incoerência entre evidências direta e indireta, a certeza da evidência de cada estimativa pode ajudar a decidir em qual estimativa acreditar.

8.3 Exemplo da avaliação do sistema GRADE para NMA

Para ilustrar cada etapa da avaliação do sistema GRADE para NMA, será utilizada uma questão da diretriz para o controle da glicemia em pacientes com diabetes melito tipo 2 (DM2) no Sistema Único de Saúde (155). Para essa diretriz, a evidência disponível na literatura sobre o uso de agentes antidiabéticos em pacientes com DM2 foi sumarizada pela realização de NMA. Foram realizadas NMA distintas para avaliar o uso de hipoglicemiantes como monoterapia e também como terapia de intensificação. A Figura 30 mostra a rede de evidências formada pelos tratamentos disponíveis.

Figura 30 – Exemplo de rede formada pelos tratamentos medicamentosos para DM2



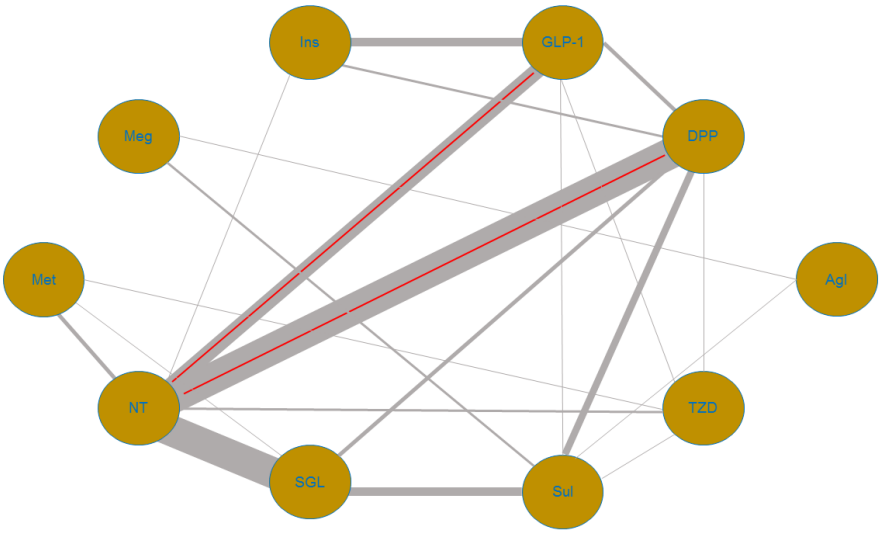
Fonte: Adaptado de Brasil (155).

Etapa 1. Apresentar estimativas de efeito direto e indireto para cada comparação de interesse da NMA

Antes de iniciar a avaliação da certeza da evidência, é necessário identificar quais evidências diretas e indiretas contribuíram para a rede de comparações de interesse. Para a escolha da evidência indireta, deve-se considerar o número de pacientes e o número de estudos incluídos (156).

Após essa identificação, os revisores devem apresentar as estimativas de efeito direto e indireto para cada desfecho, em cada comparação de interesse. Existem diversas abordagens para o cálculo de estimativas indiretas e, para o exemplo apresentado aqui, é utilizado o método *split node*, que separa a evidência em uma comparação específica (um "nó") de estimativas diretas e indiretas do efeito do tratamento (157). Por exemplo, a evidência direta para a comparação entre inibidores da dipeptidil peptidase 4 (DPP4) e agonista do receptor do peptídeo semelhante ao glucagon tipo-1 (GLP1) sobre o uso de agentes antidiabéticos em pacientes com DM2 apresentou um RR de 0,76 (IC95% 0,31 a 1,88) (Figura 31). A evidência indireta inclui um *loop* de primeira ordem, com um RR de 0,54 (IC95% 0,25 a 1,18).

Figura 31 – Exemplo de rede formada pelos tratamentos medicamentosos inibidores de DPP4 *versus* GLP1para DM2



Fonte: Adaptado de Brasil (155).

Etapa 2. Avaliação da certeza da evidência de estimativas dos efeitos diretos e indiretos

A avaliação da certeza da evidência deve ser realizada separadamente para evidências diretas e indiretas. A avaliação de estimativas de efeitos diretos abrange a avaliação do sistema GRADE de cada comparação para a qual estão disponíveis comparações diretas *(head to head*). Para a avaliação preliminar da evidência direta, devem ser avaliados os domínios de risco de viés, inconsistência, evidência indireta e viés de publicação, conforme orientação do sistema GRADE indicado na Figura 32 (5, 154). O domínio imprecisão deve ser considerado somente para obtenção da avaliação final das estimativas dos efeitos diretos e indiretos (Quadro 42).

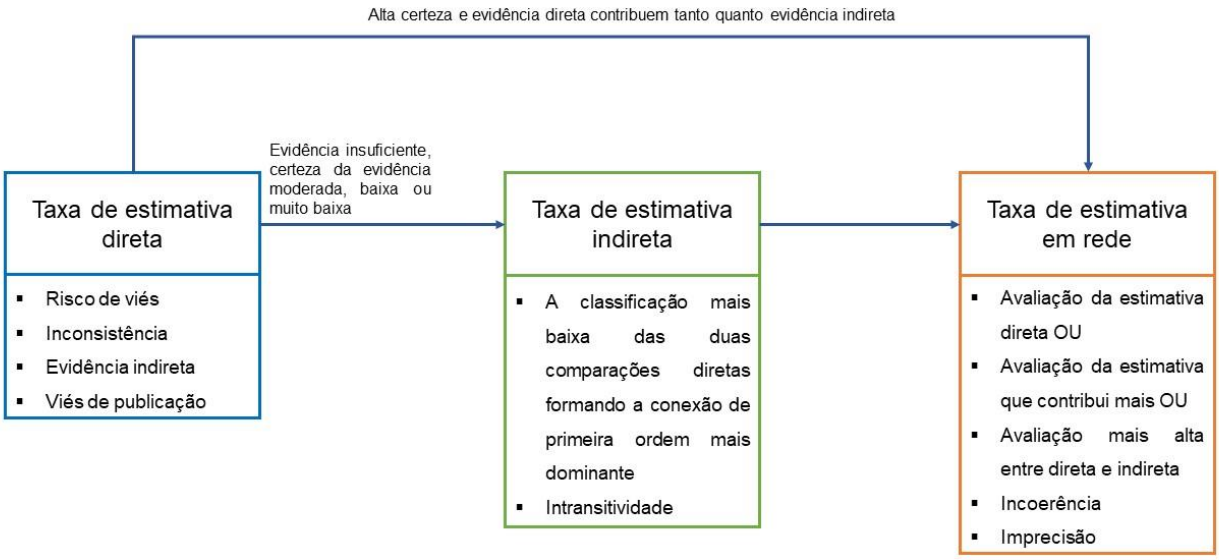
A avaliação das estimativas de efeitos indiretos depende do tamanho e da estrutura da rede de evidências. Nessa rede, pode haver um número variado de *loops* que podem contribuir com evidências indiretas para as comparações de interesse, com desde um *loop* até vários *loops* contribuindo. Para manter a avaliação da qualidade da evidência indireta gerenciável, sugere-se um foco em *loops* de primeira ordem, que geralmente contribuem com mais informações para a estimativa indireta (154). Para identificar os *loops* relevantes, é necessário um gráfico de rede similar ao apresentado na Figura 31 (a linha contínua vermelha representa o *loop* de primeira ordem para a comparação indireta de inibidores de DPP4 *versus* GLP1).

A avaliação da certeza da estimativa indireta é, então, baseada nas avaliações das duas estimativas pareadas (como GLP1 *versus* placebo e DPP4 *versus* placebo) que contribuem para a estimativa indireta da comparação de interesse (GLP1 *versus* DPP4). Essas avaliações podem seguir a orientação do sistema GRADE (Figura 32). A avaliação da certeza de evidência mais baixa das duas comparações diretas constitui a avaliação da certeza da evidência da comparação indireta. Nesse caso, para ambas as comparações, a avaliação da certeza da evidência é moderada e, portanto, a classificação inicial da evidência indireta é moderada.

Há, no entanto, um problema adicional que pode reduzir ainda mais a confiança das estimativas da comparação indireta: a intransitividade (Quadro 41). Se os estudos que formam a base da estimativa indireta (como o conjunto de evidências sobre os inibidores do cotransportador tipo 2 de sódio-glicose [SGLT2] *versus* placebo e de metformina *versus* placebo, Figura 33) diferem em aspectos importantes, a

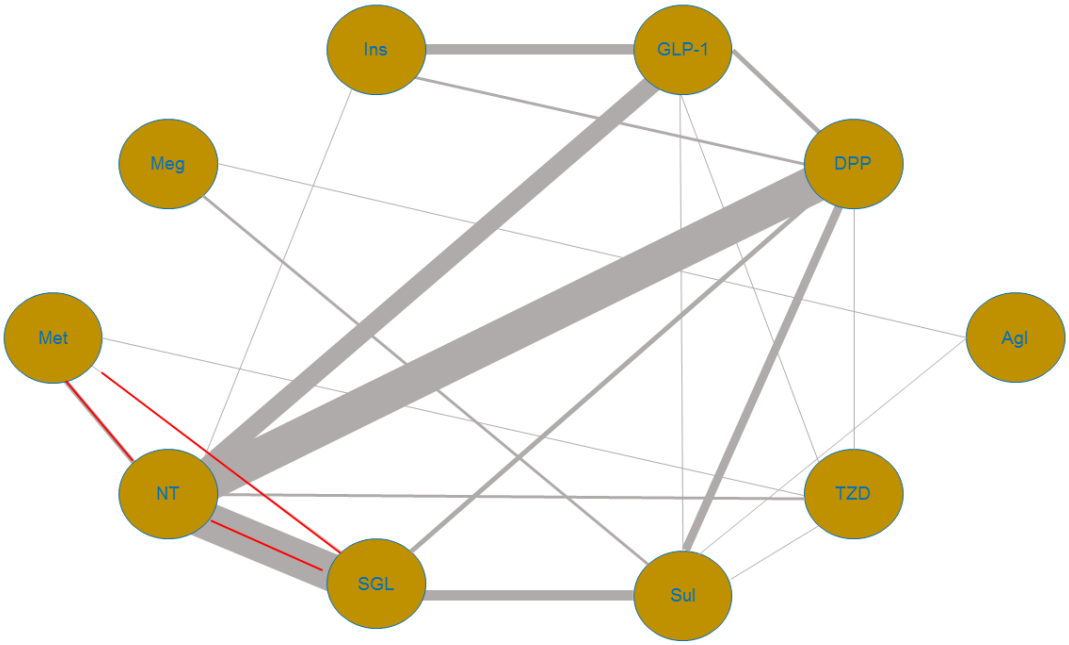
probabilidade de intransitividade pode ser alta. Como consequência, a estimativa indireta da comparação de interesse (SGLT2 *versus* metformina) pode ser tendenciosa. Na presença de intransitividade, a confiança das comparações diretas contribuintes seria rebaixada em um nível. Os estudos que utilizaram placebo como comparador comum fornecem a maioria das evidências indiretas. O SGLT2 foi testado em 27 estudos para redução do risco de morte em pacientes com DM2. Em metade desses estudos, os pacientes estavam recebendo uma cointervenção, como dieta e exercício. Isso contrasta com o estudo de metformina e placebo, no qual os pacientes só estavam recebendo como intervenção o medicamento. Como consequência dessas diferenças entre os estudos de SGLT2 *versus* placebo e metformina *versus* placebo, a evidência foi rebaixada em um nível devido à intransitividade de SGLT2 *versus* metformina.

Figura 32 – Processo para avaliar a certeza da estimativa de rede para cada comparação de pares em uma metanálise em rede



A Figura apresenta o processo para obter uma avaliação de estimativa de rede. As avaliações de evidências diretas e indiretas são ilustradas no contexto de influenciar a avaliação da estimativa da rede. Deve-se considerar a imprecisão de cada uma das estimativas para obter uma avaliação final da certeza da evidência direta e indireta.
Fonte: adaptado de Brignardello-Petersen et al. (136).

Figura 33 – Exemplo de rede formada pelos tratamentos medicamentosos cotransportador de sódio-glicose *versus* placebo e de metformina *versus* placebo para DM2



Fonte: Adaptado de Brasil (155).

Contudo, pode haver casos em que não há necessidade de avaliar a evidência indireta quando a certeza da evidência direta é alta e a contribuição da evidência direta para a estimativa da rede é pelo menos tão grande quanto a da evidência indireta. Isso pode ser verificado, por exemplo, por meio da comparação das larguras dos IC das estimativas diretas e indiretas, em que a estimativa que possui o IC mais estreito é a que mais contribui para a estimativa da rede (Quadro 43).

Etapas 3 e 4. Apresentação e avaliação da certeza da evidência de estimativas do efeito da NMA

Para iniciar a avaliação da certeza da evidência de estimativas do efeito da NMA, deve-se considerar a avaliação preliminar dos efeitos diretos e/ou indiretos. Não é necessário considerar o domínio imprecisão para informar a certeza inicial da estimativa da rede (Quadro 44). Se apenas evidências diretas **ou** indiretas estiverem disponíveis para uma determinada comparação, a classificação da qualidade da rede

será baseada nessa estimativa. Quando, para uma comparação específica, evidências diretas **e** indiretas estiverem disponíveis, sugere-se o uso da mais alta das duas avaliações para a classificação de qualidade da estimativa da NMA (por exemplo, se a certeza da estimativa direta for moderada e a certeza da estimativa indireta for baixa, a certeza da NMA será moderada).

Há duas razões pelas quais o grupo GRADE adota essa abordagem. Em primeiro lugar, se as estimativas diretas e indiretas forem semelhantes (coerentes), a estimativa de qualidade mais baixa só pode reforçar a mais alta (não faria sentido adicionar evidências que reduzissem a qualidade das estimativas). Segundo, em geral, espera-se que a estimativa da classificação mais alta seja o corpo de evidências mais preciso (e, portanto, dominante).

Outro domínio que deve ser considerado na avaliação da certeza da evidência de estimativas do efeito da NMA é a incoerência. A avaliação de coerência aborda a suposição de que as evidências diretas e indiretas são semelhantes o suficiente para serem agrupadas. Na avaliação de incoerência, os revisores devem considerar não apenas a significância estatística da diferença entre as estimativas diretas e indiretas (valor de p), mas também as diferenças nas estimativas pontuais e a sobreposição dos IC (41). Questões de geometria da rede (por exemplo, presença de estudos com multibraços) e contexto clínico também podem influenciar o julgamento. A avaliação desse domínio é descrita mais detalhadamente no Quadro 45.

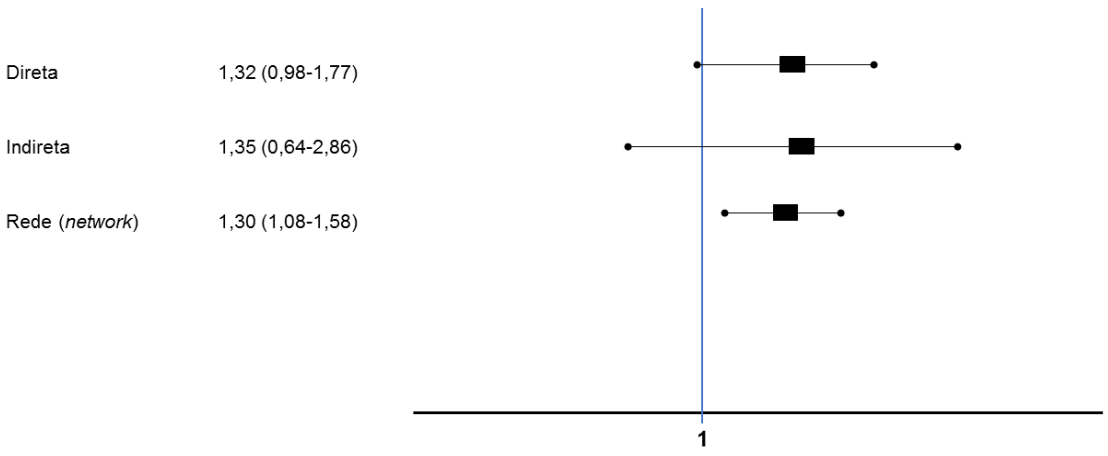
Por fim, deve-se avaliar a imprecisão. Ao avaliar a imprecisão das estimativas da rede, deve-se considerar o IC e sua relação com os limites de interesse e o tamanho ótimo da informação (Quadro 47). A imprecisão das estimativas da rede pode ser grave o suficiente para reduzir em três níveis a certeza da evidência. No entanto, é preciso prestar atenção a julgamentos espúrios de imprecisão em redes esparsas. Na realização de uma NMA de redes esparsas, a suposição em comum de heterogeneidade entre os estudos – necessária para realizar as análises em tais redes – pode resultar em estimativas de rede com IC implausivelmente amplos após a combinação de evidências diretas e indiretas. Para lidar com essa dificuldade, é sugerido o planejamento de análises de sensibilidade por meio de análises estatísticas (por exemplo, uso de modelos de efeitos fixos), obtendo-se, assim, estimativas de rede com maior probabilidade de serem úteis para a tomada de decisões (136).

Quadro 42 – Considerações sobre imprecisão

Não é necessário penalizar duplamente por imprecisão ao informar a certeza inicial (escolha entre a preliminar das evidências direta e indireta) da estimativa da rede. O avaliador deve classificar as evidências diretas e indiretas com base nos outros quatro domínios do sistema GRADE (risco de viés, inconsistência, evidência indireta e viés de publicação) e reservar a classificação de precisão para a certeza da estimativa da rede.

Por exemplo, em uma NMA que comparou agentes para prevenção de úlceras de estresse em pacientes críticos ventilados mecanicamente (158), os autores consideraram quatro tratamentos: antagonistas dos receptores de histamina-2 (H2RAs), inibidores da bomba de prótons, sucralfato e placebo. Na comparação H2RA *versus* sucralfato, para o desfecho de pneumonia, a evidência direta apresentou uma razão de chances (OR) de 1,32 (IC95% 0,98 a 1,77), exigindo penalização por imprecisão devido ao IC amplo (Figura 34). Ainda, os autores penalizaram essa evidência direta por risco de viés; portanto, a certeza da evidência direta foi classificada como baixa devido a esses dois domínios. Já a estimativa de rede apresentou um efeito relativo mais estreito (OR 1,30; IC95% 1,08 a 1,58), permitindo, agora, inferir com segurança que os H2RAs aumentam a incidência de pneumonia. Dessa maneira, não há necessidade de penalizar a estimativa de rede por imprecisão, resultando em uma classificação de certeza da evidência moderada.

Figura 34 – Estimativas diretas, indiretas e em rede de H2RA *versus* sucralfato para prevenção de úlceras de estresse em pacientes críticos ventilados mecanicamente



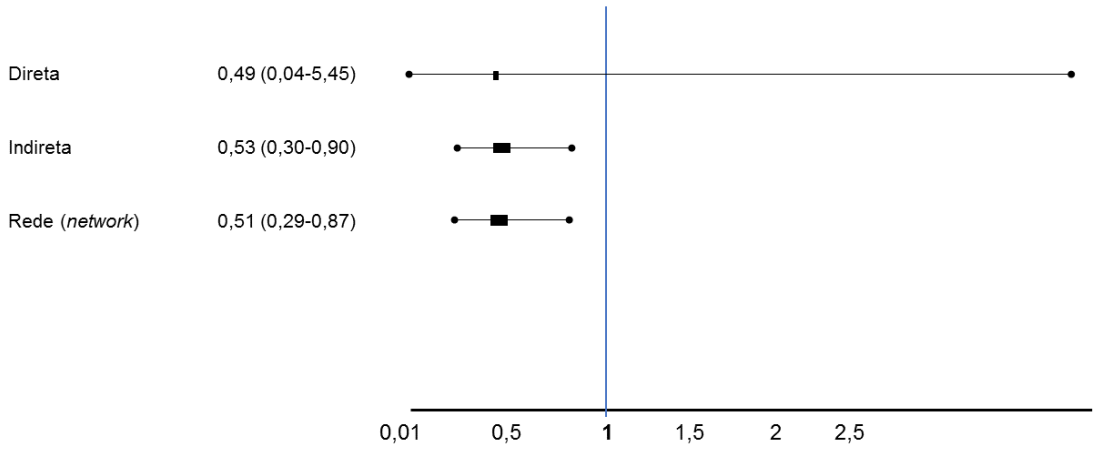
Fonte: adaptado de Alhazzani et al. (158).

Quadro 43 – Quando a certeza da evidência direta é alta e a contribuição da evidência direta para a estimativa da rede é pelo menos tão grande quanto a da evidência indireta

Caso escolham não avaliar a evidência indireta, os avaliadores devem estar cientes das possíveis limitações estatísticas associadas à comparação dos IC de estimativas diretas e indiretas. Por exemplo, a utilização do método *split node* para obtenção das estimativas diretas e indiretas e comparação das larguras dos seus IC levaria a estimativas direta e indireta cujos IC correspondentes foram calculados usando dois parâmetros de heterogeneidade diferentes. O IC da estimativa direta é calculado por meio de metanálise tradicional em pares, na qual a heterogeneidade que influencia a largura do IC agrupado é estimada usando apenas os estudos que compararam diretamente os dois tratamentos de interesse. Por outro lado, o IC da estimativa indireta é calculado por meio de NMA, na qual a heterogeneidade que influencia a largura do IC da comparação indireta é estimada usando todos os estudos incluídos na rede. Dependendo das diferenças na heterogeneidade do estudo, ao se considerar apenas um subconjunto de estudos (ou seja, aqueles que compararam diretamente os dois tratamentos de interesse) e não todos os estudos incluídos na rede, a contribuição da evidência direta à estimativa da rede pode ser incorretamente julgada como maior (ou menor) do que a da evidência indireta.

Para ilustrar, descreve-se o exemplo de uma RS que comparou o impacto de 11 agentes farmacológicos no risco de fraturas por fragilidade (159). Para a comparação direta entre alendronato e raloxifeno, não houve penalização em relação a risco de viés, inconsistência, evidência indireta ou viés de publicação. Portanto, a avaliação preliminar, sem a avaliação de imprecisão, foi classificada como certeza de evidência alta. Ao se analisar as estimativas dos efeitos diretos e indiretos, no entanto, fica claro que não se pode pular a avaliação da estimativa indireta. Ao se comparar as larguras dos IC, a estimativa indireta, por ter um IC muito mais estreito, domina a estimativa da rede (Figura 35) e, portanto, não seria apropriado desconsiderar a certeza da evidência associada à estimativa indireta.

Figura 35 – Estimativas diretas, indiretas e em rede para comparações de alendronato *versus* raloxifeno



Fonte: adaptado de Murad et al. (159).

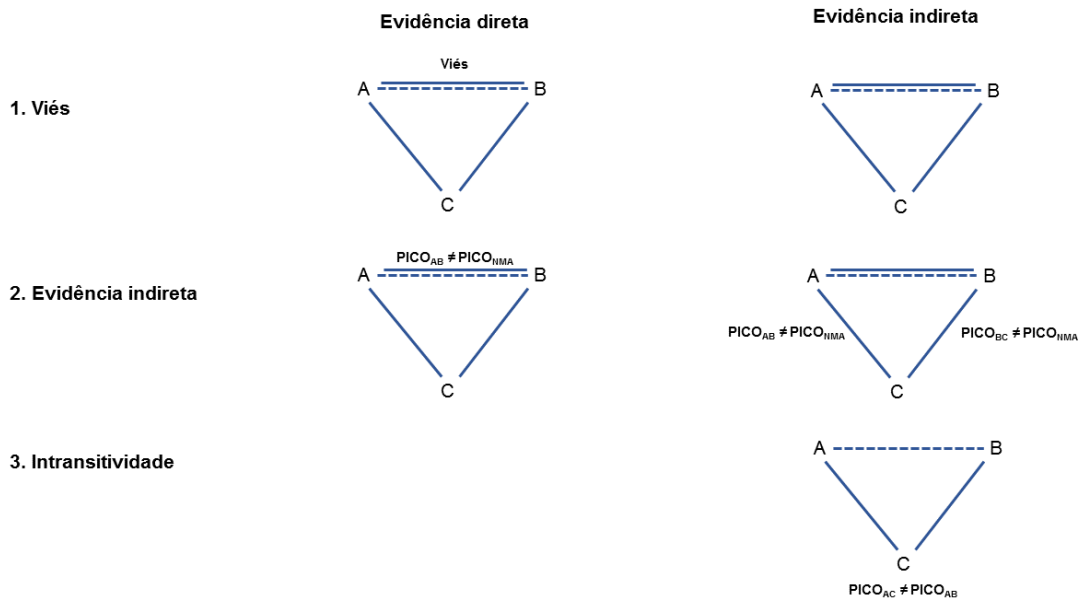
Quadro 44 – Avaliação da incoerência em relação à estimativa da rede

A coerência, um dos pressupostos centrais da NMA, refere-se à concordância entre as evidências diretas e indiretas. Ou seja, para cada comparação pareada de quaisquer duas intervenções, as estimativas de eficácia relativa das evidências diretas e indiretas devem ser semelhantes (160).

Para cada comparação pareada resultante de evidências diretas e indiretas, o sistema GRADE orienta os revisores a compararem a direção e a magnitude das estimativas pontuais das estimativas diretas e indiretas, avaliarem a extensão da sobreposição dos IC e considerarem os resultados de uma comparação estatística dessas duas estimativas. Se as estimativas diretas e indiretas forem suficientemente diferentes, os revisores devem penalizar a certeza da estimativa de rede por incoerência. A estimativa utilizada deve ser a mais confiável (aquela com a maior certeza de evidência, seja direta ou indireta) e apresentar a melhor estimativa de efeito relativo da comparação pareada em avaliação.

Existem várias razões para a presença de incoerência (Figura 36). Primeiro, as estimativas de efeitos diretos ou indiretos, ou ambos, podem ser tendenciosas devido às limitações no desenho dos estudos ou viés de publicação (no caso de evidência indireta, limitações no desenho do estudo ou viés de publicação em uma ou mais das comparações diretas que informam a comparação indireta) (Figura 37 [1]). Em segundo lugar, as estimativas diretas ou indiretas podem ser penalizadas por evidência indireta e, portanto, se aplicarem a pacientes, intervenções ou desfechos diferentes da questão clínica de interesse (Figura 36 [2]). Terceiro, a intransitividade pode resultar em uma estimativa indireta enviesada devido a diferenças, por exemplo, nas populações inscritas, que modificam o efeito das intervenções nas comparações diretas que informam as comparações indiretas (Figura 37 e Figura 36 [3]). Assim, a decisão de reduzir a certeza da evidência por incoerência representa um processo de três etapas (Figura 31) (161).

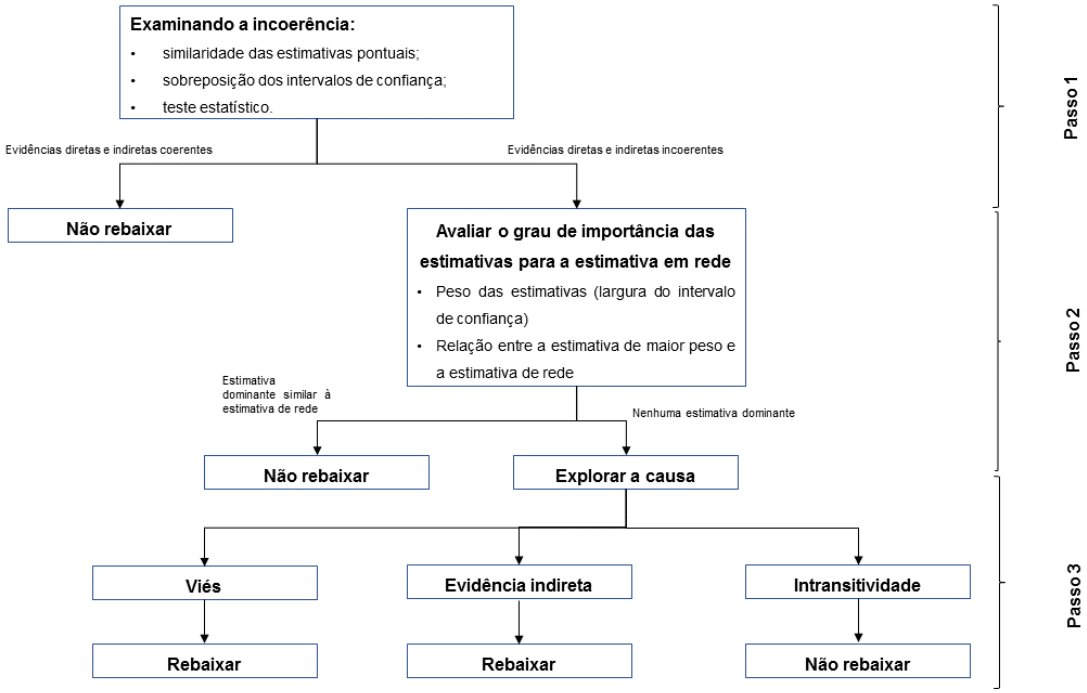
Figura 36 – Possíveis causas de incoerência



NMA = metanálise em rede; PICO = população, intervenção, comparador e desfecho.
Fonte: adaptado de Brignardello-Petersen et al. (161).

A comparação de interesse é A *versus* B. As linhas sólidas representam comparações diretas entre as intervenções, e as linhas pontilhadas representam a comparação indireta de A *versus* B através do comparador comum C.

Figura 37 – Rebaixar por incoerência



Fonte: adaptado de Brignardello-Petersen et al. (161).

Todas as referências podem ser rebaixadas em um nível. Muito raramente os revisores optam por rebaixar dois níveis por incoerência.

Um exemplo da aplicação do conceito em NMA pode ser encontrado no artigo Foote et al. (162).

Fonte: adaptado de Brignardello-Petersen et al. (161).

Quadro 45 – Avaliação da imprecisão em relação à estimativa da rede

Primeiramente, os limites de interesse devem ser definidos com base no grau de contextualização estabelecido (6). Usando uma abordagem minimamente contextualizada, o efeito nulo é utilizado para avaliar a certeza de que há um benefício ou um dano. De forma alternativa, pode-se escolher um efeito pequeno, mas importante, como seu limite. Usando uma abordagem parcialmente contextualizada, os limites para efeitos pequenos, moderados e grandes são utilizados para avaliar se há certeza de que o verdadeiro efeito tem uma magnitude particular (por exemplo, trivial ou nenhum, pequeno, moderado ou grande) (163). Os desenvolvedores de diretrizes clínicas que utilizam uma abordagem totalmente contextualizada devem definir um limite que represente um efeito grande o suficiente para exigir uma decisão entre as alternativas. Qualquer que seja o grau de contextualização, quando o IC ultrapassar um ou mais limites, a evidência deve ser reduzida por imprecisão (ou seja, sempre que um IC ultrapassar um limite, a certeza de que o verdadeiro efeito está acima ou abaixo do limite, ou dentro de dois limites, é menor) (6). Esse princípio se aplica tanto a metanálises pareadas quanto a estimativas de NMA.

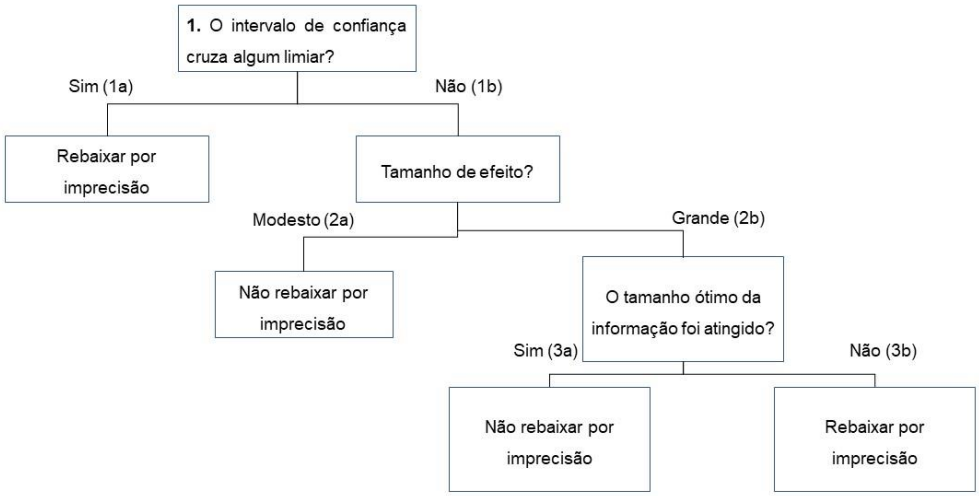
Porém, se o IC não ultrapassar um limite, é preciso observar o tamanho do efeito. Quando o tamanho do efeito é modesto e plausível, um IC estreito e que não ultrapassa um limite relevante indica um tamanho de amostra suficiente subjacente à estimativa da rede. Portanto, quando o IC não ultrapassar um limite e o tamanho do efeito for modesto, não se deve penalizar a evidência por imprecisão (Figura 38, cenário 2a).

Ainda, quando o tamanho do efeito é grande, deve-se avaliar se o tamanho ótimo da informação foi atendido. Um IC pode ser amplo e não incluir um limite apenas porque a estimativa pontual é grande e potencialmente implausível. Nesse caso, é importante avaliar o tamanho efetivo da amostra da estimativa da rede e determinar se o tamanho ótimo da informação foi atendido (

Figura 38, cenário 2b). O IC de uma estimativa de rede pode ajudar a determinar se o tamanho ótimo da informação foi atendido e se a imprecisão deve ser penalizada. Brignardello-Petersen et al. disponibilizaram um documento Excel com planilhas

que permitem calcular facilmente o tamanho ótimo da informação e o tamanho efetivo da amostra de uma estimativa de rede (planilha disponível em: https://ars.els-cdn.com/content/image/1-s2.0-S0895435621002195-mmc1.xlsx) (164).

Figura 38 – Processo para avaliar imprecisão em cada estimativa de rede



Fonte: adaptado de Brignardello-Petersen et al. (164).

Nota: Embora o tamanho do efeito (etapa 2) seja uma forma de estimar o quanto o tamanho ótimo da informação foi atingido, (etapa 3) é ilustrado aqui como uma etapa diferente, pois talvez possa ajudar a evitar os cálculos necessários para avaliar a etapa 2b. É necessário um julgamento específico para cada cenário sobre a consideração de efeito “modesto” (por exemplo, redução no risco relativo de 30%). O tamanho de efeito modesto pode ser definido como um resultado plausível que não cruza o limiar de efetividade; já o tamanho de efeito grande pode ser caracterizado como um resultado que não cruza o limiar de efetividade, porém com uma magnitude de efeito não plausível de ocorrência. A etapa 3 pode ser avaliada pelo cálculo do tamanho amostral subjacente a uma estimativa de rede, contrastando-o com o tamanho ótimo da informação; ou para avaliações de efeitos de razão de chances e risco relativo através das extremidades inferiores e

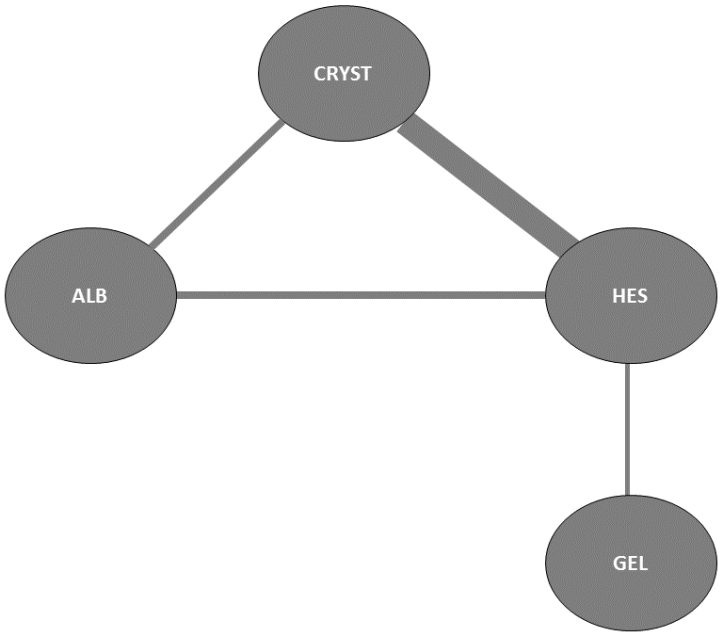
superiores do intervalo de confiança.

Fonte: adaptado de Brignardello-Petersen et al. (164).

8.4 Aplicação dos conceitos para avaliação da certeza da evidência

Como exemplo, será utilizada uma NMA que objetivou avaliar os efeitos da ressuscitação volêmica na mortalidade em pacientes com sepse (165). Os autores incluíram 14 estudos randomizados que compararam albumina, cristaloide, hidroxil-etil amido e soluções de gelatina entre si (Figura 39). A seguir, são descritas as etapas de avaliação da certeza da evidência para todas as combinações de tratamentos que apoiaram a interpretação dos achados, detalhes adicionais de estimativas e o ranqueamento (*ranking order*). É sugerida a apresentação desses resultados, pois complementam as informações das tabelas NMA-SoF (135).

Figura 39 – Rede de fluído de ressuscitação em sepse



Fonte: adaptado de Rochwerg et al. (165).

1. **Elaboração da tabela de combinações**

2. Utilize a ferramenta Excel ou outra alternativa semelhante para criar a tabela. Coloque os domínios a serem avaliados nas linhas e as combinações das tecnologias nas colunas, conforme indicado na Tabela 7. Recentemente, foi desenvolvida uma planilha com o intuito de automatizar determinadas etapas (156).

3. **Construção dos pares a serem avaliados**

Construa os pares a serem avaliados conforme a figura da rede (Figura 39). Sugere-se iniciar por uma tecnologia e fazer todas as combinações possíveis, disponibilizando-as lado a lado em colunas.

a) As combinações possíveis para essa rede são: albumina *versus* cristaloide; albumina *versus* hidroxil-etil amido; albumina *versus* soluções de gelatina; cristaloide *versus* hidroxil-etil amido; cristaloide *versus* soluções de gelatina; hidroxil-etil amido *versus* gelatina.

4. **Avaliação da evidência direta**

a) Inicie avaliando as evidências diretas, que, nesse exemplo, são os pares albumina *versus* cristaloide; albumina *versus* hidroxil-etil amido; cristaloide *versus* hidroxil-etil amido; e hidroxil-etil amido *versus* gelatina. Os pares albumina *versus* gelatina e cristaloide *versus* soluções de gelatina não possuem evidência direta, isto é, não foram encontrados estudos que comparassem essas intervenções diretamente.

b) Os domínios a serem avaliados seguindo a metodologia GRADE são: risco de viés, inconsistência, evidência indireta e viés de publicação.

c) No exemplo, após a avaliação desses domínios, a avaliação preliminar da evidência direta foi classificada como alta.

d) O domínio imprecisão deve ser avaliado somente para obtenção da avaliação final da certeza da evidência direta, conforme orientação da metodologia GRADE. No exemplo, a avaliação preliminar da evidência direta não foi penalizada nos domínios risco de viés, inconsistência, evidência direta e viés de publicação e foi classificada como alta. Porém, para a avaliação final da

evidência direta, a imprecisão foi penalizada em um nível por IC amplo (RR 0,81; IC95% 0,64 a 1,03) e foi classificada como moderada.

5. Avaliação da evidência indireta

a) Para a avaliação da evidência indireta, deve-se escolher o *loop* mais dominante; devem ser considerados *loops* de primeira ordem e o conjunto de evidências com o maior número de estudos e participantes.

b) Em seguida, observe a avaliação de cada uma das estimativas diretas. Por exemplo, na comparação entre albumina *versus* cristaloide, o comparador comum é a hidroxil-etil amido, e as comparações diretas são albumina *versus* hidroxil-etil amido e cristaloide *versus* hidroxil-etil amido. A certeza da evidência para ambas as comparações foi avaliada como alta, devendo ser escolhida a mais baixa das duas avaliações, que, nesse caso, permaneceu como certeza de evidência alta.

c) Em seguida, deve-se avaliar a intransitividade. Como explicado anteriormente, devem ser observadas diferenças nas características basais do conjunto de estudos que formam a base em uma estimativa indireta. No exemplo, o conjunto de evidências que deve ser observado são as comparações albumina *versus* hidroxil-etil amido e cristaloide *versus* hidroxil-etil amido. Devido às semelhanças entre os conjuntos de evidências, a intransitividade foi classificada como não grave.

d) Após a avaliação desses domínios, a avaliação preliminar da evidência indireta foi classificada como alta.

e) O domínio imprecisão deve ser avaliado somente para obtenção da avaliação final da evidência indireta, conforme orientação da metodologia GRADE. No exemplo, a imprecisão foi penalizada em dois níveis (RR 1,13; IC95% 0,18 a 7,32) por IC muito amplo. Assim, a avaliação final da evidência indireta foi classificada como baixa.

6. Para a avaliação da estimativa da rede

a) Para a avaliação da estimativa da rede, deve-se escolher a avaliação da estimativa que mais contribuiu em relação às evidências diretas e indiretas ou, caso ambas tenham contribuído de forma semelhante, a mais alta. No nosso exemplo, como a avaliação preliminar da evidência direta e indireta foi alta, trouxemos essa informação para a avaliação da estimativa da rede.

b) Em seguida, deve-se avaliar a incoerência. Para avaliar esse domínio, deve-se observar a concordância entre as estimativas diretas e indiretas, como similaridade entre as estimativas pontuais, sobreposição dos IC e valor de p para o teste de incoerência. No exemplo, a incoerência foi classificada como não grave.

c) Em seguida, deve-se avaliar a imprecisão, conforme orientação da metodologia GRADE. A estimativa da rede para a comparação albumina *versus* cristaloide foi classificada como grave (RR 0,83; IC 0,65-1,04) devido ao amplo IC. Dessa maneira, a avaliação final da estimativa da rede foi classificada como moderada.

7. Por fim, deve-se escolher a estimativa com a maior certeza de evidência entre a evidência direta, indireta ou da rede. No exemplo, a estimativa da rede foi a escolhida.

Tabela 7 – Detalhes da avaliação da certeza de estimativas de metanálises de rede de fluidoterapias sobre a mortalidade de pacientes com sepse

Comparações	*Starch* vs. cristaloide	Albumina vs. cristaloide	Gelatina vs. cristaloide	Albumina vs. *starch*	Gelatina vs. *starch*	Gelatina vs. albumina
Evidência direta						
Risco de viés	Não séria	Não séria		Não séria	Não séria	
Inconsistência	Não séria	Não séria		Não séria	Não séria	
Evidência indireta	Não séria	Não séria		Não séria	Não séria	
Viés de publicação	Não detectada	Não detectada		Não detectada	Não detectada	
Avaliação direta preliminar	Alta	Alta		Alta	Alta	

Comparações	***Starch* vs. cristaloide**	**Albumina vs. cristaloide**	**Gelatina vs. cristaloide**	**Albumina vs. *starch***	**Gelatina vs. *starch***	**Gelatina vs. albumina**
Contribui tanto quanto indireta	Sim	Sim		Não	Sim	
Necessidade de avaliação indireta	Não	Não		Sim	Não	
Imprecisão	Não séria	Séria		Muito séria	Muito séria	
Avaliação direta final	**Alta**	**Moderada**		**Baixa**	**Baixa**	
Evidência indireta						
Comparador comum	Albumina	*Starch*	*Starch*	Cristaloide		*Starch*

Comparações	***Starch* vs. cristaloide**	**Albumina vs. cristaloide**	**Gelatina vs. cristaloide**	**Albumina vs. *starch***	**Gelatina vs. *starch***	**Gelatina vs. albumina**
Tratamento 1 vs. avaliação do comparador comum	Alta	Alta	Alta	Alta		Alta
Tratamento 2 vs. avaliação do comparador comum	Alta	Alta	Alta	Alta		Alta
Mais baixo dos dois	Alta	Alta	Alta	Alta		Alta
Intransitividade	Não séria	Não séria	Não séria	Não séria		Não séria
Avaliação indireta preliminar	**Alta**	**Alta**	**Alta**	**Alta**		**Alta**
Imprecisão	Muito séria	Muito séria	Muito séria	Não séria		Muito séria

Comparações	***Starch* vs. cristaloide**	**Albumina vs. cristaloide**	**Gelatina vs. cristaloide**	**Albumina vs. *starch***	**Gelatina vs. *starch***	**Gelatina vs. albumina**
Avaliação indireta final	**Baixa**	**Baixa**	**Baixa**	**Alta**		**Baixa**
Evidência em rede						
Mais alto entre direta e indireta	Alta	Alta	Alta	Alta	Alta	Alta
Incoerência	Não séria	Não séria	NA	Não séria	NA	NA
Imprecisão	Não séria	Séria	Muito séria	Não séria	Muito séria	Muito séria
Avaliação em rede final	**Alta**	**Moderada**	**Baixa**	**Alta**	**Baixa**	**Baixa**
Estimativa mais plausível	**Rede**	**Rede**	**Rede**	**Rede**	**Rede**	**Rede**

Comparações	*Starch* vs. cristaloide	Albumina vs. cristaloide	Gelatina vs. cristaloide	Albumina vs. *starch*	Gelatina vs. *starch*	Gelatina vs. albumina
Itens em negrito representam a avaliação de certeza de estimativas diretas e indiretas, para informar a estimativa em rede (avaliação preliminar) e a avaliação final. Espaços em branco indicam que nenhum tipo de evidência contribui para a estimativa em rede. NA = não aplicável.						

Fonte: Adaptado de Rochwerg et al. (165).

8.5 Tabela de sumário dos resultados (SoF) para uma NMA

Com o intuito de fornecer informações relevantes em um formato simples e fácil de usar, foi desenvolvida uma tabela de sumário dos resultados (*summary of findings*, SoF), que exibe as informações críticas de uma NMA (139).

A tabela NMA-SoF desenvolvida inclui evidências para um comparador principal e outras intervenções para um único desfecho. Como a escolha do comparador é um desafio, os autores do estudo sugerem as seguintes opções para escolher o comparador de referência para a tabela NMA-SoF: (1) uma intervenção placebo, (2) um tratamento padrão-ouro para a condição em análise, (3) a intervenção mais custo-efetiva ou (4) a intervenção menos eficaz. Também é sugerido apresentar as intervenções por linha na tabela NMA-SoF com base na ordem de classificação.

A seguir, são apresentadas as seções que compõem a tabela NMA-SoF, a qual pode ser visualizada na Tabela 8:

- a seção superior exibe informações sobre os componentes PICO. Um desfecho precisa ser escolhido para cada tabela NMA-SoF. Além disso, uma intervenção precisa ser selecionada como um ‘‘comparador de referência’’, e as outras intervenções devem ser listadas sob o rótulo ‘‘intervenção’’. Um gráfico de rede está incluído nessa seção;
- a seção intermediária é composta por oito colunas, que relatam as seguintes informações: (1) intervenções para um desfecho específico. Abaixo de cada intervenção, há uma descrição do número de participantes incluídos na comparação direta e se a contribuição para o efeito relativo resultou da evidência direta, indireta ou ambas. (2) A estimativa de efeito relativo para cada intervenção, que é calculada em relação ao comparador de referência. (3) Três colunas, das quais duas relatam as informações de efeitos absolutos antecipados em relação ao comparador de referência, para cada uma das intervenções, e uma relata a diferença de risco. (4) A certeza da evidência com a justificativa para rebaixar o corpo da evidência. (5) O ranqueamento (*ranking order*), que pode ser expresso como mediana ou SUCRA (do inglês *surface under the cumulative ranking*) com o intervalo de credibilidade correspondente. (6) Por fim, a interpretação dos achados, que descreve o nível de superioridade ou inferioridade de cada intervenção em comparação com o comparador de

referência após a combinação da estimativa de efeito relativo, a certeza da evidência e os componentes de probabilidade do ranqueamento (*ranking order*).

- A seção inferior tem três células com (1) definições sobre terminologia e abreviaturas da NMA, (2) descrição de cada um dos julgamentos finais sobre a certeza da evidência no GRADE e (3) notas explicativas que descrevem em detalhes a justificativa para a certeza das avaliações de evidência.

Tabela 8 – Tabela sumária de resultados de metanálise em rede em formato final

Estimativas de efeito, intervalos de credibilidade e certeza de evidência
Tabela NMA-SoF bayesiana
Benefícios
Gráfico geométrico de metanálise em rede
Pacientes ou população:
Intervenções:
Comparador (referência):
Desfecho:
Contexto:

Total de estudos: Total de participantes:		Efeitos relativos**	Efeito absoluto antecipado*** (ICr95%)			Certeza de evidência	*Ranking* **** (ICr95%)	Interpretação dos achados
		(ICr95%)	Sem intervenção	Com intervenção	Diferença			

Definições de tabelas NMA-SoF

***** Linhas representam comparações diretas.

** Estimativas são apresentadas como *odds ratio.* ICr: intervalo de credibilidade. Resultados são expressos em intervalos de credibilidade como oposto ao intervalo de confiança, uma vez que a análise bayesiana foi conduzida.

*** Efeito absoluto antecipado. O efeito absoluto antecipado compara dois riscos por meio do cálculo da diferença entre os riscos do grupo intervenção e controle.

**** *Ranking* SUCRA e intervalos de credibilidade para eficácia são apresentados. As estatísticas de classificação são definidas como as probabilidades de um tratamento dentre *n* tratamentos em uma metanálise de rede ser o melhor, o segundo melhor, o terceiro melhor e assim em diante até o tratamento menos eficaz.

Graus de evidência propostos pelo Grade *Working Group* (ou certeza de evidência)

Qualidade alta: Nós estamos muito confiantes de que o efeito verdadeiro se encontra perto da estimativa de efeito.

Qualidade moderada: Nós estamos moderadamente confiantes de que o efeito verdadeiro está provavelmente próximo da estimativa de efeito, mas há uma possibilidade de que seja substancialmente diferente.

Qualidade baixa: A nossa confiança no efeito estimado é baixa: o efeito verdadeiro talvez seja substancialmente diferente do efeito estimado.

Qualidade muito baixa: Nós temos muito pouca confiança no efeito estimado: o efeito verdadeiro provavelmente é substancialmente diferente do efeito estimado.

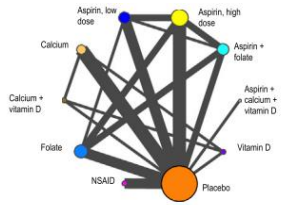
Explicações

-

Fonte: adaptado de Yepes-Nunez et al. (139).

Ainda, uma tabela alternativa foi desenvolvida para uso em decisões clínicas ou políticas. Esse formato é focado em um número menor de intervenções, idealmente as três principais opções em relação ao comparador principal escolhido. A tabela apresenta informações sobre os desfechos desejáveis e indesejáveis para essas opções, com o objetivo de facilitar a tomada de decisões (Tabela 9) (139).

Tabela 9 – Tabela sumária de resultados de metanálise em rede relatando informações sobre comparações de múltiplos tratamentos e múltiplos desfechos

Estimativas de efeito, intervalos de credibilidade e certeza de evidência para quimioprevenção de câncer colorretal em indivíduos com neoplasia colorretal prévia		
Benefícios		
Pacientes ou população: indivíduos com neoplasia colorretal prévia **Intervenções:** doses baixa e alta de ácido acetilsalicílico, anti-inflamatórios não esteroidais não salicílicos (NSAIDs), cálcio, vitamina D, ácido fólico **Comparador:** placebo **Contexto:** ambulatorial, tempo de seguimento de 3 a 5 anos.		
Desfecho	**Efeitos e confiança das estimativas de efeito**	**Comentários**

	Não salicilícos NSAIDs		Ácido acetilsalicílico, baixa dose		Ácido acetilsalicílico + cálcio + vitamina D		
Prevenção de neoplasia Seguimento: acima de 24 meses até 60 meses							
Placebo comparador	**OR 0,37** (0,24 a 0,53) Estimativa de rede	**47 menos por 1.000** (56 menos a 35 menos)	**OR 0,71** (0,41 a 1,23) Estimativa de rede	**21 menos por 1.000** (44 menos a 17 mais)	**OR 0,71** (0,18 a 2,49) Estimativa de rede	**21 menos POR 1.000** (61 menos a 110 mais)	Nenhum dos tratamentos classificados entre placebo contra NSAIDs, cálcio, vitamina D ou ácido fólico foram maiores do que os relatados aqui. Assim, não foram incluídas outras comparações na tabela.
74 por 1.000[1] (7,4%)	⨁⨁⨁⨁ **Alta** Confiança da estimativa		⨁⨁◯◯ Baixa Confiança da estimativa devido à imprecisão [2 3]		⨁⨁◯◯ Baixa Confiança da estimativa devido à imprecisão [2 3]		

Classificação 7 (4 a 9)	**Classificação**[4] 1 (1 a 2) Com base em 3.486 participantes (4 ECR)		**Classificação** 3 (2 a 9) Com base em 823 participantes (3 ECR)		**Classificação** 3 (1 a 10) Com base em 427 participantes (1 ECR)		
Eventos adversos graves Seguimento: acima de 24 meses até 60 meses							
Placebo comparador 74 por 1.000[1] (7,4%)	**OR 1,23** (0,95 a 1,64) Estimativa de rede	**34 menos por 1.000** (8 menos a 87 mais) ⨁⨁◯◯ Baixa	**OR 0,78** (0,43 a 1,38) Estimativa de rede	**35 menos por 1.000** (54 menos a 97 mais) ⨁⨁◯◯ Baixa	**OR 0,90** (0,54 a 1,51) Estimativa de rede	**15 mais por 1.000** (71 mais a 77 menos) ⨁⨁◯◯ Baixa Confiança da estimativa devido à imprecisão [2] [3]	Intervenções que relataram desfechos de dano foram escolhidas com base em intervenções que incluíram desfechos benéficos. Assim, não foram incluídas outras comparações na tabela.

	Confiança da estimativa devido à imprecisão [2] [3]	Confiança da estimativa devido à imprecisão [2] [3]		
Classificação	**Classificação** 2 (1 a 9)	**Classificação** 8 (3 a 10)	**Classificação** 4 (2 a 7)	
4 (2 a 7)	Com base em 3.964 participantes (3 ECR)	Com base em 12.098 participantes (1 ECR)	Com base em 714 participantes (1 ECR)	

Definições de tabelas NMA-SoF

Linhas representam comparações diretas.

Estimativas são apresentadas como *odds ratio.* ICr: intervalo de credibilidade. Resultados são expressos em intervalos de credibilidade como oposto ao intervalo de confiança, uma vez que a análise bayesiana foi conduzida.

A base para o **risco assumido** (por exemplo, a mediana do risco do grupo controle entre os estudos) está presente nas notas de rodapé. O **risco correspondente** (e o seu IC95%) é baseado no risco assumido no grupo de comparação e o **risco relativo** da intervenção (e o seu IC95%).

CI: intervalo de confiança; ECR = ensaio clínico randomizado; OR: *odds* ratio.

Graus de evidência propostos pelo Grade *Working Group* (ou certeza de evidência) **Qualidade alta:** É muito improvável que pesquisas futuras mudem a nossa confiança na estimativa de efeito. **Qualidade moderada:** É provável que pesquisas futuras tenham um importante impacto na nossa confiança na estimativa de efeito e talvez modifiquem a estimativa. **Qualidade baixa:** É muito provável que futuras pesquisas tenham um importante impacto na nossa confiança na estimativa de efeito e provavelmente mudem a estimativa. **Qualidade muito baixa:** Nós estamos muito incertos sobre a estimativa.
Explicações: 1 Risco basal (assumido do risco controle) obtido do projeto de associação do Instituto Nacional do Câncer. 2 Imprecisão muito séria porque o ICr95% cruza a unidade, e a amplitude do intervalo de credibilidade sugere alta possibilidade de dano. 3 Conceitualmente, não há intransitividade significante, com distribuição comparável de modificadores de efeito plausíveis em tentativas de diferentes agentes quimiopreventivos. 4 A classificação é mostrada como mediana (classificação 1-10) e ICr95%.

Fonte: adaptado de Yepes-Nunez et al. (139).

8.6 Classificações e conclusões para uma NMA

Em uma NMA, quanto maior for o número de intervenções e, portanto, o número de comparações, mais complexa e desafiadora será a interpretação dos resultados. Para auxiliar essa avaliação, o sistema GRADE descreveu orientações de como realizar uma interpretação ideal na NMA. Os investigadores podem conduzir esse processo usando duas abordagens: uma estrutura minimamente contextualizada ou uma estrutura parcialmente contextualizada.

Uma estrutura minimamente contextualizada minimiza julgamentos de valor sobre a magnitude dos efeitos da intervenção, baseando-se na posição do IC em relação a um limiar para categorizar as intervenções e, assim, enfatizando questões de precisão. As intervenções devem ser agrupadas em categorias da mais eficaz ou nociva para a menos eficaz ou nociva, e os julgamentos que colocam as intervenções em tais categorias devem considerar simultaneamente as estimativas de efeito, a certeza da evidência e o ranqueamento (138).

Uma abordagem parcialmente contextualizada deve estabelecer faixas de magnitudes de efeito que representem um efeito trivial ou nenhum efeito; efeito pequeno, mas importante; efeito moderado; e efeito grande. Os princípios que orientam essa estrutura incluem o agrupamento das intervenções em categorias com base na magnitude do efeito; e a consideração, nos julgamentos que colocam as intervenções em tais categorias, das estimativas de efeito, da certeza das evidências e do ranqueamento (137). O **Quadro 46** resume as semelhanças e diferenças entre as duas abordagens (138).

Quadro 46 – Similaridades e diferenças entre as estruturas GRADE minimamente e parcialmente contextualizadas tirando conclusões para metanálises em rede

Características	Estrutura minimamente contextualizada	Estrutura parcialmente contextualizada
Similaridades		
Objetivo geral	Para grupos de intervenções de diferentes categorias de acordo com os seus efeitos	
Informações consideradas	Estimativas de efeito, certeza de evidência e *rankings*	
Inferências gerais	Intervenções colocadas em categorias mais altas são mais prováveis de serem mais efetivas do que intervenções colocadas em categorias mais baixas	
Diferenças		
Estatística primária usada para categorizar intervenções	Intervalo de confiança	Estimativa pontual
Valor colocado na imprecisão	Mais alta do que outras limitações de evidência	Semelhante a outras limitações de evidência
Categorias resultantes	Do mais para o menos efetivo/prejudicial	Do benefício/dano grande ao trivial

Fonte: adaptado de Brignardello-Petersen, et al. (137)

Usando a estrutura minimamente contextualizada, por exemplo, em comparação com o padrão de referência e usando um limite de decisão sem efeito, a categorização seria diferente para uma redução de risco absoluto de 20% (IC95% 1 a 39) *versus* a mesma estimativa com um IC95% de -1 a 41. Usando a estrutura parcialmente contextualizada, a classificação inicial seria feita com base na estimativa pontual, sem avaliação do IC (se o IC cruzasse o nulo, seria irrelevante); com a mesma estimativa pontual relativa à referência, seria colocada na mesma categoria. Dessa maneira, a abordagem parcialmente contextualizada reconhece, por exemplo, que um efeito de intervenção com um IC95% de 1 a 39 rebaixado por risco de viés não deve ser mais confiável do que um efeito com um IC de -1 a 41% rebaixado por imprecisão (138).

A seguir, as duas abordagens são ilustradas utilizando uma NMA de intervenções farmacológicas e nutricionais para o tratamento de diarreia aguda e gastroenterite em crianças (Tabela 10, Tabela 11) (166).

Tabela 10 – Classificação final de 27 intervenções, com base em revisão sistemática com metanálise em rede de intervenções para diarreia aguda em criançasmetanálise

Certeza de evidência e classificação* da intervenção	**Intervenção****	**Intervenção vs. tratamento padrão ou placebo (diferença média [ICr95%])**	**Área sob a curva de classificação cumulativa**
Alta certeza (certeza de evidência moderada a alta)			
Categoria 2: entre as mais efetivas	*Saccharomyces boulardii* + zinco (M)	-39,45 (-52,5 a -26,7)	0,92
	Esmectita + zinco (M)	-35,63 (-57,6 a -13,2)	0,88
	Simbióticos (H)	-26,26 (-36,1 a -16,2)	0,77

Certeza de evidência e classificação* da intervenção	**Intervenção****	**Intervenção vs. tratamento padrão ou placebo (diferença média [ICr95%])**	**Área sob a curva de classificação cumulativa**
Categoria 1: da inferior até a mais efetiva, ou da superior até a menos efetiva	Zinco + fórmula livre de lactose (M)	-21,37 (-36,5 a -6,1)	0,61
	Zinco (M)	-18,38 (-23,4 a -13,5)	0,50
	Loperamida (M)	-17,79 (-30,4 a -5,7)	0,46
	Zinco + micronutrientes (M)	-17,76 (-31,8 a -4,1)	0,46
Categoria 0: entre as menos efetivas	Prebióticos (M)	-15,32 (-42,8 a 12)	0,38
Baixa certeza (certeza de evidência baixa a muito baixa)			
Categoria 2: pode estar entre os mais efetivos	*Lactobacillous rhamnosus GG + esmectita (VL)*	-51,08 (-64,3 a -37,9)	1,00
	Zinco + probióticos (L)	-29,39 (-40,3 a -18,6)	0,81
Categoria 1: pode ser inferior ao mais efetivo ou superior ao menos efetivo	Simbióticos + fórmula livre de lactose (VL)	-32,11 (-53 a -11,3)	0,85
	Esmectita (VL)	-23,90 (-30,8 a -17)	0,69
	L. rhamnosus GG (L)	-22,84 (-28,8 a -16,7)	0,65
	Probióticos (L)	-19,36 (-23,7 a -15,1)	0,54

Certeza de evidência e classificação* da intervenção	**Intervenção****	**Intervenção vs. tratamento padrão ou placebo (diferença média [ICr95%])**	**Área sob a curva de classificação cumulativa**
	Racecadotril (L)	-17,19 (-24,7 a -9,8)	0,46
	S. boulardii (L)	-16,48 (-23,3 a -9,7)	0,42
	Fórmula livre de lactose (VL)	-12,50 (-19 a -6)	0,31
Categoria 0: pode estar entre os menos efetivos	*S. boulardii* + zinco + fórmula livre de lactose (L)	-16,74 (-36,1 a 2,7)	0,42
	Iogurte + probióticos + zinco (VL)	-15,63 (-56,8 a 26,6)	0,38
	Fórmula livre de lactose + probióticos (VL)	-13,27 (-36 a 9,2)	0,31
	S. boulardii + fórmula livre de lactose (VL)	-12,32 (-30 a 6)	0,27
	Vitamina A (VL)	-5,95 (-21,4 a 9,3)	0,19
	Caolim-pectina (VL)	-5,32 (-33,8 a 22,8)	0,15
	Micronutrientes (L)	-0,68 (-33,3 a 32,8)	0,08
	Tratamento padrão ou placebo	-	0,08

Certeza de evidência e classificação* da intervenção	Intervenção**	Intervenção vs. tratamento padrão ou placebo (diferença média [ICr95%])	Área sob a curva de classificação cumulativa
	Iogurte (VL)	-16,43 (-30,5 a -2,1)	0,42
	Leite diluído (VL)	3,02 (-14,3 a 8,4)	0,04

* Categorias não informam juízos de valor sobre a importância dos efeitos. Um formato sugerido de apresentação pode incluir diferentes cores e tons.

** Letras em colchetes representam a certeza de evidência de cada intervenção em comparação com a referência. H = alta certeza da evidência; M = moderada; L = baixa; VL = muito baixa.

Fonte: adaptado de Florez, et al. (166).

Tabela 11 – Classificação das intervenções com base em metanálise em rede de intervenções para diarreia aguda e gastroenterite em crianças

Classificação da intervenção	Intervenção	Efeito sobre a duração da diarreia (horas; diferença média [IC95%])	Área sob a curva de classificação cumulativa (IC95%)	Certeza de evidência
Efeito grande de benefício	Fórmula livre de lactose + esmectita	-51,08 (-64,30 a -37,85)	1,00 (0,92 a 1,00)	Muito baixa
	Saccharomyces boulardii + zinco**	-39,45 (-52,45 a -26,73)**	0,92 (0,77 a 1,00)**	Moderada**
	Esmectita + zinco**	-35,63 (-57,57 a -13,16)**	0,88 (0,35 a 1,00)**	Moderada**
	Simbióticos + fórmula livre de lactose	-32,11 (-53,01 a -11,33)	0,85 (0,27 a 1,00)	Muito baixa
	Zinco + probióticos	-29,39 (-40,26 a -18,57)	0,81 (0,5 a 0,96)	Baixa
	Simbióticos**	-26,26 (-36,14 a -16,22)**	0,77 (0,38 a 0,92)**	Alta**
Efeito moderado de benefício	Esmectita	-23,90 (-30,80 a -16,96)	0,69 (0,42 a 0,88)	Muito baixa
	Lactobacillous rhamnosus GG	-22,74 (-28,81 a -16,68)	0,65 (0,38 a 0,85)	Baixa
	Zinco + fórmula livre de lactose**	-21,37 (-36,54 a -6,13)**	0,61 (0,19 a 0,92)**	Moderada**
	Todos os probióticos	-19,36 (-23,66 a -15,09)	0,54 (0,31 a 0,73)	Baixa

Classificação da intervenção	Intervenção	Efeito sobre a duração da diarreia (horas; diferença média [IC95%])	Área sob a curva de classificação cumulativa (IC95%)	Certeza de evidência
	Zinco**	-18,38 (-23,39 a -13,45)**	0,50 (0,27 a 0,69)**	Moderada**
	Loperamida**	-17,79 (-30,5 a -5,65)**	0,46 (0,15 a 0,85)**	Moderada**
	Zinco + micronutrientes**	-17,76 (-31,77 a -4,13)**	0,46 (0,15 a 0,85)**	Moderada**
	Racecadotril	-17,19 (-24,65 a -9,76)	0,46 (0,23 a 0,73)	Baixa
	S. boulardii + zinco + fórmula livre de lactose	-16,74 (-36,05 a 2,72)	0,42 (0,08 a 0,88)	Baixa
	S. boulardii	-16,48 (-23,33 a -9,69)	0,42 (0,19 a 0,69)	Baixa
	Iogurte	-16,43 (-30,49 a -2,05)	0,42 (0,11 a 0,85)	Muito baixa
	Iogurte + probióticos + zinco	-15,63 (-56,82 a 26,63)	0,38 (0,00 a 1,00)	Muito baixa
	Prebióticos	-15,62 (-42,42 a 11,28)	0,38 (0,00 a 0,96)	Muito baixa
	Fórmula livre de lactose + probióticos	-13,27 (-35,96 a 9,19)	0,31 (0,00 a 0,88)	Muito baixa
	Fórmula livre de lactose	-125,50 (-19,04 a -5,99)	0,31 (0,15 a 0,54)	Muito baixa

Classificação da intervenção	Intervenção	Efeito sobre a duração da diarreia (horas; diferença média [IC95%])	Área sob a curva de classificação cumulativa (IC95%)	Certeza de evidência
	S. boulardii + fórmula livre de lactose	-12,32 (-30,01 a 5,98)	0,27 (0,04 a 0,81)	Muito baixa
Efeito pequeno de benefício	Vitamina A	-5,95 (-21,43 a 9,32)	0,19 (0,00 a 0,61)	Muito baixa
	Caolim-pectina	-5,32 (-33,76 a 22,83)	0,15 (0,00 a 0,89)	Muito baixa
Trivial ou sem efeito (sem diferença em relação ao placebo)	Micronutrientes	-0,68 (-33,29 a 32,79)	0,08 (0,00 a 0,85)	Baixa
Pequeno efeito de dano	Leite diluído	3,02 (-14,32 a 8,41)	0,04 (0,00 a 0,23)	Muito baixa

* Um formato sugerido de apresentação pode incluir diferentes cores e tons; este formato de apresentação não foi testado pelo usuário e não é uma orientação do GRADE *Working Group*.

** Presença de alta ou moderada certeza de evidência.

Fonte: adaptado de Brignardello-Petersen et al. (138).

9. Sistema GRADE para modelagem

- Modelagem consiste em uma abordagem matemática que combina dados de diferentes fontes com o objetivo de fornecer estimativa para a questão na qual inexiste evidência direta;
- Um modelo pode possuir diferentes graus de complexidade, incluindo tanto estimativas de redução absoluta de risco para uma população quanto avaliações econômicas em saúde;
- Três abordagens são possíveis: desenvolver um modelo específico, adaptar um modelo existente ou então utilizar resultados de um ou mais modelos disponíveis;
- Estudos de modelagem apresentam estimativas e cada estimativa possui um grau de certeza subjacente que deve ser adequadamente avaliado e contextualizado;
- A certeza nas estimativas provenientes de um modelo depende tanto da adequabilidade de sua estrutura quanto da certeza da evidência de seus parâmetros. Os domínios de avaliação do GRADE são aplicáveis para esse tipo de evidência;
- O sistema GRADE para estudos de modelagem está em desenvolvimento, contudo estimulamos que seus princípios gerais sejam aplicados.

9.1 O que é um modelo?

Apesar das rigorosas revisões sistemáticas sobre a eficácia e segurança de intervenções de saúde, pacientes, profissionais de saúde e gestores de políticas podem permanecer em dúvida sobre qual conduta adotar devido à incerteza em relação aos benefícios e riscos, além das preferências conflitantes em relação ao uso e à disponibilidade de certas tecnologias. Um modelo tem como objetivo simular um determinado fenômeno, para o qual não há uma estimativa direta, em especial na impossibilidade ou na indisponibilidade de estudos clínicos. Esse objetivo é atingindo combinando dados de diferentes fontes, em um modelo que representa, de certa forma, o fenômeno de interesse.

Dessa forma, estudos de modelagem e simulação na área da saúde podem complementar a evidência procedente de revisões sistemáticas com informações úteis

para a tomada de decisão, em especial em situações na qual não há evidência direta para o desfecho de interesse (167).

Um modelo pode possuir, assim, diferentes graus de complexidade. A estimativa de redução absoluta de risco ao combinar o risco relativo de estudos clínicos com a incidência de um determinado desfecho observado em nossa população de interesse consiste em um modelo, frequentemente utilizado em diretrizes clínico-assistenciais. Da mesma forma, estudos de custo utilidade e avaliações de impacto orçamentário consistem também em modelos matemáticos frequentemente utilizados em avaliações de tecnologias em saúde.

Um exemplo dessa abordagem na área de diretrizes clínico-assistenciais são as recomendações do *U.S Preventive Services Task Force* (USPSTF), em especial em questões relacionadas a rastreamento, como uma forma de estimar o *net-benefit* (benefício líquido) das estratégias propostas (168). Questões de rastreamento são bastante propícias para o uso de modelos uma vez que englobam variáveis que dificilmente são abordadas em um estudo único. Uma vez que o propósito do rastreamento é detectar precocemente e intervir nos pacientes identificados, um modelo, por exemplo, necessita combinar a prevalência da condição na população, o prognóstico dessa condição, a acurácia do teste de rastreamento e a efetividade das intervenções disponíveis para o seu tratamento.

Na área de avaliação de tecnologias em saúde, é difundido o uso de avaliações econômicas baseadas em modelagem matemática. Os dados utilizados são provenientes de muitas fontes (como dados primários, revisões sistemáticas e estudos de custos) e são integrados em modelos (como árvores de decisão, coortes simuladas de Markov, microssimulação de pacientes individuais e simulação de eventos discretos), a fim de oferecer estimativas das reais relações de custo-efetividade existentes e do grau de incerteza dos valores obtidos. O uso de modelagem é comum porque estudos primários não costumam englobar todo o cenário de uma questão de pesquisa. Nesses casos, geralmente é apresentada como medida sumária uma razão de custo-efetividade incremental, sendo fundamental no processo de incorporação de tecnologias. Contudo, raramente é apresentada a certeza da evidência relacionada a essa estimativa. Assim, é necessário ter cautela quando os modelos são baseados em evidências insuficientes, de baixa qualidade ou extrapoladas, ou então, o modelo

matemático proposto não é adequado para representar a realidade da condição avaliada.

Neste capítulo, foi considerada a seguinte definição de modelo: “estrutura matemática que representa variáveis e suas inter-relações para descrever um fenômeno observado ou prever evento futuro”, usada em disciplinas relacionadas à saúde para tomada de decisão. Pode haver modelos mecanicistas, empíricos e híbridos. Não foram considerados aqui modelos estatísticos usados para estimar as associações entre variáveis medidas (por exemplo, modelos de riscos proporcionais ou modelos usados para metanálise).

9.2 Uso do sistema GRADE junto à modelagem

Na elaboração de diretrizes GRADE, o processo de desenvolvimento geralmente envolve a coleta e o uso de informações para cada questão de pesquisa para apoiar a construção das EtD, onde são apresentadas as devidas informações para cada desfecho crítico e importante definidos anteriormente a coleta de dados, pelo grupo elaborador de diretrizes (169). Em determinados cenários, em especial em relação ao uso de recursos e custo-efetividade (mas também em outros cenários como a própria efetividade) não há evidência direta, podendo ser utilizados modelos para esse propósito. Estudos de modelagem, portanto, podem ser usados para prever a dinâmica e a carga da doença, a probabilidade de uma exposição representar um risco à saúde, o impacto das intervenções nos benefícios e riscos à saúde e a eficiência econômica das intervenções de saúde, entre outros.

Há o entendimento que estudos de modelagem muitas vezes exige competências, financiamento e tempo significativos, devendo ser priorizado o seu uso em situações para as quais há necessidade de maior embasamento no processo de tomada de decisão.

Ao longo deste capítulo, é descrita em linhas gerais a abordagem sugerida para a incorporação de modelos no processo de tomada de decisão (170-172).

Uma vez definido que há benefício com o uso de estudos de modelagem, os pesquisadores devem iniciar o processo conceituando o problema e o modelo-alvo ideal que melhor represente o fenômeno real ou o problema de decisão que está sendo

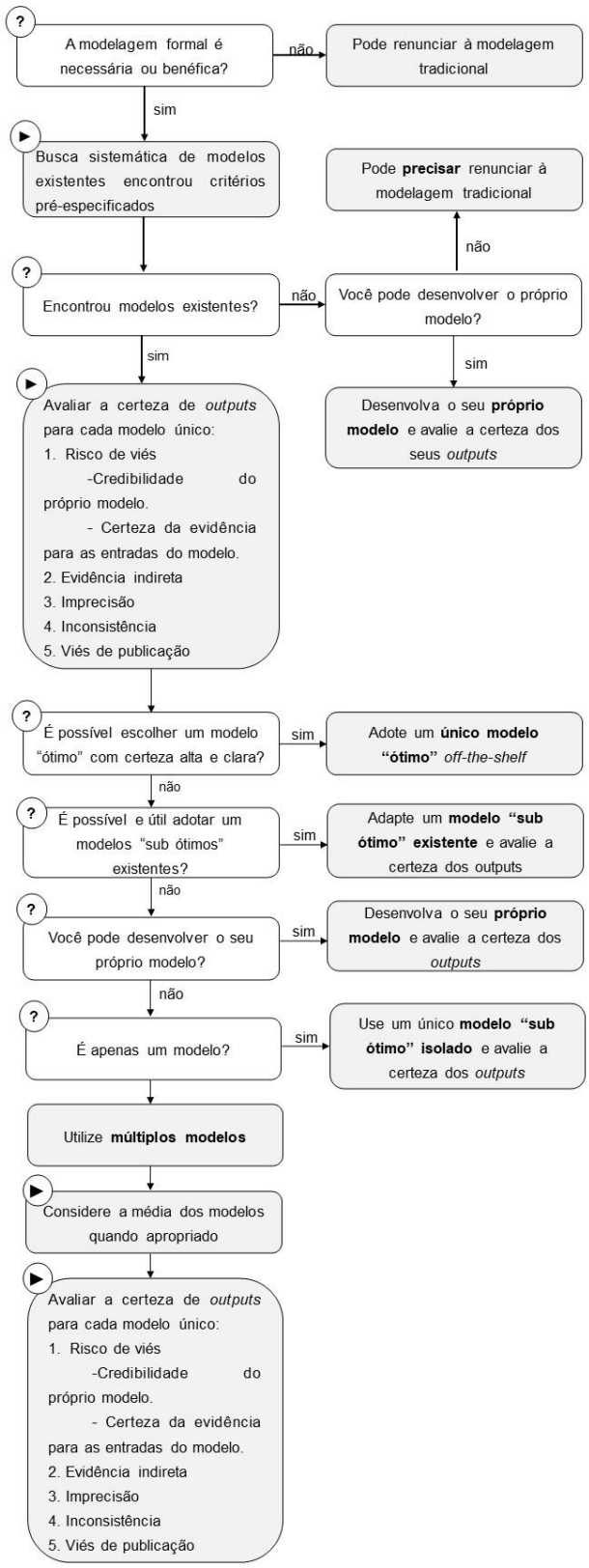
considerado. Essa conceituação pode orientar o desenvolvimento de um novo modelo ou servir de referência para comparação com os modelos existentes.

Existem basicamente três opções que os usuários podem utilizar para incorporação dos resultados obtidos do modelo na tomada de decisão em saúde (Figura 40):

1. Desenvolver um modelo específico (*de novo*) para responder a questões de interesse;
2. Procurar por um modelo já existente que responda à mesma questão ou a uma questão de pesquisa muito similar, adaptando-o;
3. Utilizar os resultados de diversos modelos encontrados na literatura.

Em relação ao uso de modelos encontrados na literatura, importante salientar que a transferibilidade de modelos econômicos é bastante limitada, em especial devido à evidência indireta, com as estimativas resultantes dificilmente sendo aplicáveis ao contexto nacional, contudo, em certos casos podem sinalizar direção e magnitude do impacto econômico (173-175).

Figura 40 – Abordagem para incorporar os resultados do modelo na tomada de decisões relacionadas à saúde



Fonte: adaptado de Brozek et al. (171).

Após escolher uma das três opções citadas acima, é necessário avaliar a certeza de evidência (ou seja, evidências geradas a partir desse modelo) utilizando os domínios do GRADE para elevar ou rebaixar a certeza de evidência. Quando os pesquisadores desenvolvem seu próprio modelo ou quando identificam um único modelo que é considerado suficientemente direto para o problema em questão, eles devem avaliar a certeza dos resultados. Se um modelo estimar vários *resultados* (por exemplo, custo por ano de vida ganho ajustado para a qualidade e custo por morte evitada), os pesquisadores precisam avaliar a certeza de cada um separadamente (37, 41, 55, 73, 87, 176).

Quando são incluídos resultados de vários modelos existentes, os pesquisadores devem avaliar a certeza dos resultados em todos os modelos incluídos, o que é mais complexo do que para modelos únicos. Os pesquisadores devem ter cuidado para evitar a "dupla contagem" do mesmo modelo como se fosse vários modelos. Por exemplo, o mesmo modelo (ou seja, mesma estrutura e premissas) pode ter sido usado em vários estudos de modelagem, nos quais os investigadores confiaram em diferentes parâmetros. Diante desse cenário, os pesquisadores podem precisar decidir quais dos modelos são os mais diretos para sua questão específica e incluir apenas esse modelo na revisão. É importante frisar que as informações de um modelo podem ser facilmente acessíveis, mas a obtenção das informações necessárias para avaliar o modelo desenvolvido por outros pesquisadores geralmente é mais difícil, muitas vezes inviabilizando a adequada avaliação.

Risco de viés

Na avaliação da certeza de evidência para um modelo único, o risco de viés dos resultados do modelo (ou seja, os resultados do modelo sendo sistematicamente superestimados ou subestimados) é determinado pela credibilidade do próprio modelo e pela certeza de evidência para cada um dos parâmetros do modelo. Quando os pesquisadores desenvolvem seu modelo *de novo*, para minimizar o risco de viés, eles precisam especificar previamente os parâmetros de entrada para os quais espera-se que os resultados sejam mais sensíveis. Por exemplo, em modelos econômicos, esses parâmetros-chave podem incluir medidas de efeito, uso de recursos e valores de

utilidade. Seguindo a lógica do sistema GRADE de que a certeza geral de evidência não pode ser maior do que a certeza mais baixa para qualquer corpo de evidências que seja crítico para uma decisão, a classificação geral da certeza de evidência do modelo deve ser limitada à classificação de certeza mais baixa para qualquer corpo de evidências (nesse caso, parâmetros) para qual os resultados do modelo se mostraram relevantes (parâmetros cuja variabilidade implica em modificação importante nos resultados do modelo).

Para a avaliação do risco de viés entre diversos modelos, devem ser realizados uma avaliação do domínio em cada modelo individual e, posteriormente, um julgamento sobre o risco geral de viés em todos os modelos incluídos.

Evidência indireta

Evidência direta ou relevância se refere a até que ponto os resultados do modelo representam diretamente o fenômeno que está sendo modelado. Há conceitualmente duas fontes separadas de evidência indireta: os parâmetros utilizados no modelo e a estrutura do modelo em si.

Para avaliar o modelo, é preciso compará-lo com o modelo conceitual ideal. Determinar a relevância dos resultados do modelo inclui avaliar em que medida a população modelada, as intervenções e os comparadores assumidos, o horizonte de tempo, a perspectiva analítica e os resultados que estão sendo modelados refletem aqueles que são de interesse atual. Avaliar a evidência indireta em um único modelo também requer avaliação dos parâmetros individuais utilizados.

Para avaliar a certeza em múltiplos modelos, os pesquisadores precisam novamente avaliar o caráter indireto para cada um dos modelos incluídos e, depois, integrar os julgamentos entre os modelos.

Inconsistência

A avaliação de inconsistência em modelos deve se concentrar em diferenças inexplicáveis entre os resultados do modelo para um determinado resultado. Um único modelo pode produzir resultados inconsistentes devido a uma variabilidade inexplicável nos resultados de estudos individuais que informam as estimativas agrupadas de

variáveis de entrada no modelo. Se não houver explicação plausível para a diferença nas estimativas de utilidade, os resultados de um modelo baseado nessas entradas também podem ser qualitativamente inconsistentes. A análise de sensibilidade pode ajudar a fazer um julgamento.

Se múltiplos modelos existentes abordando o mesmo problema produzirem resultados consideravelmente diferentes ou chegarem a conclusões contrastantes, a comparação cuidadosa dos modelos pode levar a uma compreensão mais profunda dos fatores que impulsionam os resultados e as conclusões. Os pesquisadores precisam avaliar se essas diferenças são importantes ou não, ou seja, se levariam a conclusões diferentes.

Imprecisão

A certeza geral dos resultados do modelo também pode ser menor quando os resultados do modelo não são estimados com precisão. Para resultados quantitativos, deve-se examinar não apenas a estimativa pontual (por exemplo, evento previsto médio), mas também a variabilidade dessa estimativa (por exemplo, resultados da análise de sensibilidade probabilística com base na distribuição dos parâmetros de utilizados). Orientações adicionais sobre como avaliar a imprecisão dos resultados do modelo precisam levar em consideração se as conclusões mudam de acordo com esse parâmetro específico (no caso de um modelo econômico, se os diferentes cenários resultam em resultados inferiores e superiores ao limiar de disposição a pagar). Quando os pesquisadores optam por realizar apenas um resumo qualitativo dos resultados entre modelos, é desejável que eles relatem alguma estimativa de variabilidade dos resultados de modelos individuais e uma avaliação do quão severa é a variabilidade (por exemplo, faixa de efeitos estimados).

Risco de viés de publicação

O risco de viés de publicação refere-se à probabilidade de modelos relevantes terem sido construídos, mas não terem sido disponibilizados ao público por meio de publicação ou outro meio. Quando se pretende utilizar um modelo existente, mas se sabe ou suspeita fortemente que outros modelos semelhantes foram desenvolvidos, mas não foram disponibilizados, pode-se inferir que os resultados dos outros modelos

diferiram sistematicamente do modelo disponível. No caso de modelos construídos *de novo*, esse risco de viés pode ser irrelevante. A avaliação do viés de publicação para modelos múltiplos é semelhante à avaliação no contexto de um único modelo.

Fatores que aumentam a certeza dos resultados

Os mesmos fatores considerados no sistema GRADE (grande magnitude do efeito estimado, presença de gradiente dose-resposta em um efeito estimado e confusão residual plausível oposta) podem ser considerados na avaliação dos modelos. A presença de um gradiente dose-resposta nos resultados do modelo pode ser aplicável em algumas situações, e uma grande magnitude de efeito nos resultados do modelo pode aumentar a certeza de evidência (por exemplo, razão de custo efetividade incremental muito distante do limiar de disposição a pagar indica maior certeza na ausência de custo-efetividade da intervenção). O efeito de um fator residual plausível também parece, em teoria, aplicável na avaliação da certeza dos resultados do modelo (ou seja, um modelo conservador que não incorpora parâmetros de dados de entrada em favor de uma intervenção, mas ainda encontra resultados favoráveis). As considerações para modelos múltiplos são semelhantes ao de um único modelo.

Considerações

Este capítulo traz conceitos gerais que devem ser considerados para a construção de estudos de modelagem e para avaliar a certeza dos seus resultados. É importante esclarecer que modelos proporcionam estimativas e essa estimativa possui um grau de certeza subjacente, muitas vezes negligenciado, que pode ser avaliado utilizando os mesmos domínios preconizados no GRADE.

A metodologia GRADE para estudos de modelagem atualmente está em desenvolvimento e há poucos exemplos de sua utilização na literatura (177, 178). Futuros estudos devem ser realizados com o intuito de prover informações mais detalhadas sobre a avaliação da certeza de estimativas oriundas de modelagem matemática. Apesar de seu uso não ser essencial em modelos, estimula-se que os grupos considerem seus princípios na avaliação, seja em pareceres técnico-científicos, diretrizes clínico-assistenciais ou avaliações de tecnologias em saúde.

10. Sistema GRADE para incorporação de tecnologias

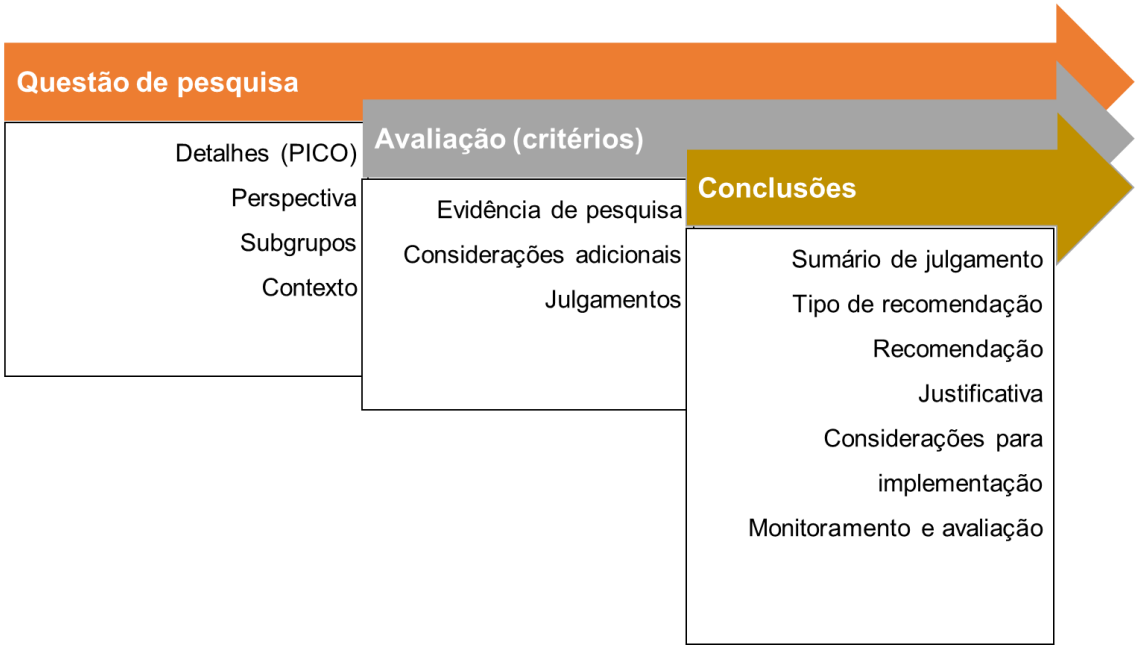
A ATS tem como principal objetivo auxiliar os gestores em saúde na tomada de decisões coerentes e racionais quanto à incorporação de novas tecnologias, evitando a introdução de tecnologias cujo valor é incerto para os sistemas de saúde e optando por uma abordagem política responsável pelas decisões para a população (179, 180). No Brasil, a ATS aplicada na CONITEC é usada para a tomada de decisão sobre a incorporação de tecnologias em saúde, como um medicamento ou teste diagnóstico, no SUS. Assim, as decisões sobre a incorporação de uma tecnologia dizem respeito às determinações feitas pelos pagadores de serviços de saúde (181, 182) sobre pagar, por exemplo, por exames, medicamentos ou procedimentos cirúrgicos, entre outros serviços. Normalmente, tais decisões são complexas e exigem a consideração de muitos fatores, entre eles a certeza das evidências, o balanço dos riscos e benefícios para os pacientes (*trade-off*), os custos e a equidade em saúde (183, 184). Devem, ainda, ser justas e relevantes para o público-alvo, sendo necessário o desenvolvimento de um processo transparente, possível de ser revisado e documentado. Esse processo de tomada de decisão deve ser baseado nas melhores evidências científicas disponíveis e, para isso, lança-se mão do uso de revisões sistemáticas, podendo incluir, ainda, análises econômicas e critérios relacionados à equidade e viabilidade de implementação da tecnologia considerada.

O sistema GRADE elaborou uma EtD específica para decisões sobre incorporação, que pode ajudar a garantir que os processos de tomada de decisão sigam os princípios citados acima. Na Figura 41, são apresentados os itens que compõem a estrutura de uma EtD para incorporação, semelhante à EtD citada no capítulo 5. Uso do GRADE para o desenvolvimento de recomendações, e inclui itens como formulação da questão de pesquisa, avaliação e conclusão (184).

A formulação da questão de pesquisa e os critérios que afetam uma decisão são semelhantes ao descrito no capítulo 2. Elaboração da questão de pesquisa e escolha dos desfechos. Porém, existem algumas diferenças importantes em relação aos julgamentos dos painelistas sobre o valor que as pessoas atribuem aos principais desfechos, equidade, aceitabilidade e viabilidade. Ainda, enquanto os painéis de diretrizes podem fazer recomendações clínicas da perspectiva de um paciente individual (111), as decisões de incorporação são sempre feitas na **perspectiva da população**.

Existem três pontos principais que devem ser considerados: formulação da questão de pesquisa, avaliação dos critérios a serem julgados e conclusão.

Figura 41 – Pontos principais para formulação da questão de pesquisa



Fonte: elaboração própria.

10.1 Formulação da questão de pesquisa

A formulação da questão de pesquisa para incorporação de uma tecnologia é muito semelhante à formulação de uma questão clínica, em que se deve levar em consideração o acrônimo PICO, assim como o cenário e a perspectiva em saúde.

10.2 Avaliação dos critérios para tomada de decisão (making an assessment)

A avaliação das tabelas EtD para incorporação de tecnologias pode envolver 12 fatores determinantes da recomendação que podem influenciar as decisões (Quadro 47). Para cada critério, os painelistas devem fazer um julgamento baseado na melhor evidência disponível. É importante destacar que os critérios não são obrigatórios e deve-se avaliar quais são importantes para o contexto da ATS em questão.

Quadro 47 – Fatores/critérios considerados na EtD de incorporação de uma nova tecnologia
1. O problema é uma prioridade?
2. Quão substanciais são os efeitos desejáveis?
3. Quão substanciais são os efeitos indesejáveis esperados?
4. 'Qual é a certeza do conjunto final de evidências?
5. Há incerteza importante sobre o valor que as pessoas atribuem aos desfechos principais?
6. O balanço entre os efeitos desejáveis e os efeitos indesejáveis favorece a intervenção ou o comparador?
7. Quão grandes são os recursos necessários (custos)?
8. Qual é a certeza da evidência de uso desses recursos?
9. A relação de custo-efetividade favorece a intervenção ou a comparação?
10. Qual seria o impacto na equidade em saúde?
11. 'Qual É a aceitabilidade da intervenção pelas principais partes interessadas?
12. 'A implementação da intervenção avaliada é viável?

Fonte: elaboração própria.

São apresentados a seguir alguns conceitos e exemplos que merecem destaque em relação às tabelas EtD para incorporação de novas tecnologias em saúde.

Priorização do problema

As decisões de incorporação de uma nova tecnologia são frequentemente influenciadas pela questão clínica que está sendo caracterizada prioridade. Por exemplo, tratamentos para doenças menos graves geralmente não são cobertos por

planos de saúde, enquanto tratamentos para doenças graves às vezes são priorizados para incorporação. A prevalência de uma condição também pode influenciar as decisões, mas não de forma independente, visto que alguns governos podem priorizar a disponibilização de novas terapias para doenças frequentes e que possuem grande impacto na saúde da população, enquanto outros podem priorizar a incorporação de tecnologias para doenças raras. Ambas as opções são razoáveis, mas algumas pessoas provavelmente considerarão uma delas injusta. Outro ponto que deve ser considerado em relação à priorização é quando uma determinada condição de saúde possui um custo atual alto e a incorporação de uma tecnologia alternativa poderia economizar recursos ou, ainda, substituir uma outra intervenção de alto custo com benefícios limítrofes. Nesse caso, é a economia potencial que torna o problema uma prioridade.

Avaliação da certeza geral da evidência

A certeza da evidência dos efeitos desejáveis e indesejáveis pode afetar decisões de incorporação em saúde, visto que as decisões consideram o balanço entre esses efeitos e o seu impacto na tomada de decisão por recomendar ou não determinada tecnologia. Por exemplo, se os efeitos desejáveis forem incertos, isso pode levar à decisão de não incorporar uma intervenção. Se uma tecnologia for promissora, mas há variabilidade no grau de incerteza e de outros fatores, pode-se tomar a decisão de incorporá-la com monitoramento de potenciais efeitos adversos, apenas no contexto de pesquisa (185, 186), ou de não incorporar a tecnologia até que haja mais evidências.

Na maioria dos casos, os desfechos mais relevantes para decisões de incorporação são aqueles que afetam os usuários da intervenção. Por exemplo, na avaliação sobre utilização de antibióticos, o desfecho da resistência bacteriana ao medicamento pode ser considerado como crítico para a tomada de decisão.

Incerteza sobre o valor que as pessoas atribuem aos desfechos principais

Em relação à incerteza sobre o valor que as pessoas atribuem aos desfechos principais, esse conceito refere-se ao valor que as pessoas acometidas pela intervenção atribuem, em média, aos desfechos que desempenham um papel

importante na decisão de incorporação, de forma semelhante às recomendações clínicas. Incerteza importante significa que não há certeza sobre o valor atribuído pelas pessoas aos desfechos críticos, inclusive sobre o quão importante os desfechos são em relação uns aos outros. Por exemplo, os eventos adversos da cirurgia de câncer de próstata, como impotência sexual e incontinência, podem ter maior peso na escolha por um tratamento.

Há variabilidade no valor que as pessoas atribuem aos desfechos críticos nas decisões de incorporação: enquanto uma proporção importante de pessoas pode valorizar os efeitos desejáveis mais do que os efeitos indesejáveis e achar justificável pagar pela intervenção, outras pessoas podem optar por não usar a intervenção.

Uso de recursos

Quanto ao uso de recursos, as avaliações econômicas desempenham um papel importante nas decisões de incorporação porque a informação sobre custos é central para essas decisões. Por exemplo, quanto mais altos forem os custos de uma intervenção e quanto menos custo-efetiva ela for, será menos provável que ela seja adicionada a um sistema de saúde (187). A certeza sobre o real valor incremental de determinada tecnologia quanto ao cuidado que já vem sendo disponibilizado à população pode ser motivo suficiente para adiar a incorporação de uma nova tecnologia ou incorporá-la sob a condição de monitoramento de seus efeitos e custos.

Algumas instituições, entre elas o Ministério da Saúde do Brasil, exigem uma avaliação econômica como parte dos requisitos de apresentação para decisões de incorporação, e a maioria também exige uma análise de impacto orçamentário (188, 189). Contudo, devido às limitações de recursos financeiros e/ou humanos para a elaboração desses relatórios, muitas vezes uma instituição não consegue desenvolver o próprio relatório de avaliação econômica, podendo utilizar evidências disponíveis na literatura. Como o uso de recursos é sempre relevante para as decisões de incorporação, os painéis de recomendação devem ser os mais sistemáticos e transparentes possível sobre como o uso de recursos foi considerado e sobre quaisquer suposições que foram feitas.

Para facilitar o julgamento sobre os custos de uma nova tecnologia, alguns países possuem limiares de disposição a pagar para decisões de incorporação. A

adoção de um limiar tem o potencial para contribuir com a gestão da oferta e da demanda por tecnologias, além de apoiar as análises que preconizam a utilização mais eficiente e efetiva dos recursos do orçamento (Quadro 48).

Equidade

O acesso e a cobertura universal equitativa (equidade) às intervenções em saúde são pontos importantes para as decisões de incorporação. A não disponibilização de uma intervenção para a qual os benefícios superam os riscos significa que as pessoas que não podem pagar pela intervenção não terão acesso a ela, aumentando, assim, a desigualdade. Por outro lado, cobrir uma intervenção pode aumentar as desigualdades, como, por exemplo, a disponibilização de tratamentos que requerem acesso a grandes centros urbanos.

Aceitabilidade

As considerações sobre aceitabilidade são semelhantes para recomendações clínicas e decisões de incorporação e incluem a distribuição de benefícios, danos e custos, bem como considerações éticas (111). Nelas, são avaliadas as preferências dos principais atores envolvidos, como paciente, profissional de saúde e gestor.

Decidir não incorporar uma intervenção devido aos custos ou ao custo-benefício pode ser inaceitável para os pacientes que se beneficiariam da intervenção e para os profissionais de saúde que cuidam deles. Por outro lado, decidir incorporar intervenções de alto custo pode não ser aceitável para aqueles que não se beneficiaram, mas compartilham os custos da intervenção. Embora essas considerações normalmente não devam alterar a decisão de incorporar uma tecnologia, elas devem ser consideradas ao se pensar na disseminação e implementação da decisão. Considerações semelhantes podem ser aplicadas a uma decisão sobre restringir a cobertura de determinada tecnologia caso algumas pessoas que não teriam direito à cobertura pudessem se beneficiar da intervenção.

Viabilidade

As considerações de viabilidade para decisões de incorporação são diferentes das considerações de viabilidade para recomendações clínicas. Para o cenário de incorporação, as considerações concentram-se principalmente na viabilidade de incorporação, na incorporação restrita a um cenário ou na não incorporação de uma intervenção.

Uma questão de viabilidade pode surgir a partir da dificuldade de implementação de um procedimento administrativo eficaz para garantir que a incorporação seja limitada a uma população específica de pacientes identificados por características clínicas. A capacidade de atender à demanda crescente por uma intervenção também pode ser uma consideração importante. Por exemplo, a capacidade de um sistema de saúde de fornecer um exame diagnóstico pode ser uma consideração importante para uma decisão sobre a incorporação desse exame. De forma semelhante às considerações sobre aceitabilidade, essas considerações podem não alterar a decisão sobre a incorporação de uma tecnologia, mas devem ser consideradas para a disseminação e implementação dessa decisão.

Quadro 48 – Limiares para decisões de incorporação de tecnologias em saúde

- Julgamentos sobre a prioridade de um problema, quão substanciais são os efeitos desejáveis e indesejáveis, quão grandes são os custos e a relação de custo-efetividade de uma intervenção requerem a existência de um comparador.
- As organizações que tomam decisões sobre a incorporação de novas intervenções em saúde podem usar limites ou padrões explícitos ou implícitos para fazer esses julgamentos. Entre os limiares identificados, destaca-se a referência comum de valores de 1 a 3 PIB per capita da OMS, os valores de £ 20.000 a £ 30.000 por QALY no Reino Unido e os valores de $ 50.000 e $ 100.000 por QALY no Canadá e nos Estados Unidos, respectivamente (190).
- Com o uso ou não de um limiar a pagar, é importante que sejam discutidas alternativas em situações que promovam a inovação e equidade em saúde para o sistema de saúde, priorizando um processo consistente e transparente.
- Apesar de os limiares poderem ser úteis para o estabelecimento de um limiar de custo-efetividade, apresentam desvantagens, podendo superestimar ou subestimar o valor do limiar.

PIB = produto interno bruto; QALY = *quality-adjusted life years* /anos de vida ajustados por qualidade de vida.
Fonte: elaboração própria.

A estrutura da tabela EtD para decisões de incorporação é diferente da estrutura da tabela para recomendações clínicas. As cinco opções são: não incorporar, incorporar com desenvolvimento de novas evidências (no contexto de pesquisa), incorporar com negociação de preços, incorporar de forma restrita e incorporar (**Quadro 49**).

Quadro 49 – Opções para a decisão de incorporação	
Decisão	**Considerações**
Não incorporar	Decisão de não incorporação de uma tecnologia.
Incorporar com desenvolvimento de novas evidências (no contexto de pesquisa)	A decisão de abranger uma tecnologia no contexto de pesquisa pode ser tomada quando houver uma incerteza importante sobre os efeitos de uma intervenção.
Incorporar com negociação de preços	A incorporação com negociação de preços é comum para medicamentos novos e eficazes que não atendem aos padrões de uso de recursos ou custo-benefício. A negociação de preços pode incluir acordos de compartilhamento de risco entre fabricantes e pagadores.
Incorporar de forma restrita	A incorporação restrita é comumente usada para intervenções que são apenas benéficas ou custo-efetivas para um subgrupo de pacientes. A cobertura restrita é semelhante a uma recomendação condicional que especifica um subgrupo de pacientes para os quais uma tecnologia é recomendada.
Incorporar	Decisão de incorporação de uma tecnologia.

Fonte: elaboração própria.

Ressalta-se que, enquanto os painéis de diretrizes podem fazer recomendações clínicas da perspectiva de um paciente individual, as decisões de cobertura são sempre feitas na **perspectiva da população** (111).

A ATS tem como objetivo auxiliar os gestores em saúde na tomada de decisões sobre a incorporação de novas tecnologias, garantindo que sejam decisões coerentes e baseadas em evidências. No contexto brasileiro, a ATS é aplicada pela Conitec para decidir sobre a incorporação de tecnologias no âmbito do SUS. Essas decisões envolvem considerações complexas, além dos benefícios e riscos para os pacientes, como a avaliação da certeza das evidências, os custos e a equidade em saúde. É essencial que o processo de tomada de decisão seja transparente, revisável e baseado nas melhores evidências científicas disponíveis. O sistema GRADE fornece uma estrutura para o desenvolvimento de recomendações, incluindo conceitos-chave a serem considerados nas decisões de incorporação de tecnologias em saúde.

11. Sistema GRADE em saúde pública

A saúde pública está direcionada para um objetivo central de prevenir doenças, prolongar a vida e promover a saúde por meio de esforços organizados da sociedade. Para isso, está organizada nos domínios de proteção da saúde, serviços de saúde e melhoria da saúde, dentro de uma abordagem que reconhece o impacto dos determinantes sociais nos resultados de saúde tanto na dimensão individual quanto populacional e a importância de reduzir as desigualdades de acesso nos sistemas (191). Algumas decisões de saúde pública estão relacionadas a marcos políticos decisivos, que podem incluir a reforma de um sistema de saúde, a regulamentação de produtos prejudiciais à saúde (como comercialização de drogas lícitas ou alimentos com alto teor de sódio), o desenvolvimento de infraestrutura e a previdência social.

Dessa forma, a perspectiva para a tomada de decisão em saúde pública é complexa. Tais decisões podem ser tomadas em âmbitos internacional (diretrizes da Organização Mundial da Saúde), nacional (diretrizes elaboradas pelo Ministério da Saúde) ou local (diretrizes para tratamento de uma condição de saúde dentro de um contexto local municipal). Um desafio encontrado ainda hoje é a indefinição dos critérios adotados para a tomada de decisão, que, por sua vez, pode negligenciar critérios importantes, dar importância inadequada a certos critérios ou, até mesmo, não adotar as melhores evidências disponíveis para fundamentar os julgamentos.

Assim, adotar um sistema estruturado e transparente para a tomada de decisão, como o sistema GRADE, pode auxiliar na garantia de que todos os critérios importantes sejam considerados e que a melhor evidência disponível seja usada por todos os atores na tomada de decisão (*stakeholders*) na implementação das recomendações ou na utilização das decisões do sistema de saúde e de saúde pública. A avaliação da certeza da evidência de uma tecnologia em saúde e a repercussão no sistema de saúde é uma estratégia fundamental no processo de escolha. Para esse propósito, o GRADE utiliza as tabelas EtD, que são uma forma de apresentação de dados estruturada e transparente para a obtenção de consenso entre o grupo elaborador e os especialistas em relação à interpretação das evidências disponíveis, sejam elas sobre efetividade, custos ou os demais domínios necessários para a tomada de decisão (para mais detalhes do uso de tabelas EtD, consultar o Capítulo 5. Uso do GRADE para o desenvolvimento de recomendações).

EtD para decisão em saúde pública utilizando o sistema GRADE

No contexto de tomada de decisão em saúde pública, o grupo de trabalho GRADE desenvolveu cinco tabelas EtD conforme os cinco tipos de cenários já elencados no capítulo 5. Observa-se que os critérios específicos para o sistema de saúde e as decisões em saúde pública diferem em alguns aspectos dos critérios para recomendações clínicas, decisões de incorporação e recomendações e decisões sobre testes diagnósticos (169, 182, 192, 193). Em relação à natureza das decisões, elas são tomadas por formuladores de políticas (*policy makers*) ou gestores, os quais definem as políticas de saúde de uma população, enquanto as decisões clínicas são normalmente tomadas por indivíduos (profissionais de saúde ou pacientes). Em relação à questão de pesquisa, as decisões no sistema de saúde e na saúde pública geralmente começam com um problema e incluem possíveis opções para abordar esse problema; assim, é indicado o uso do acrônimo POCO (problema, opção, comparação e desfecho).

A tabela é composta por 12 itens, e as principais diferenças estão relacionadas aos seguintes itens: importância do problema, recursos necessários, equidade, aceitabilidade e viabilidade. No item “importância do problema”, a consideração do número de pessoas acometidas é fundamental no julgamento de um problema no sistema de saúde pública, pois a prevalência pode influenciar uma recomendação em saúde (da perspectiva populacional) ou a decisão de incorporação devido ao impacto nas necessidades de recursos (quanto mais pessoas afetadas, maior o custo). De qualquer forma, um problema de saúde não é mais ou menos importante para as pessoas com o problema por causa do número de pessoas acometidas, e a maioria das pessoas não consideraria um problema grave mais ou menos importante para tratar dependendo do número de pessoas acometidas.

Devido às questões relacionadas ao custo, o item “recursos necessários” acaba se tornando um ponto importante, principalmente devido à limitação de recursos. Nesse âmbito, os formuladores de políticas e os gestores que tomam decisões sobre o sistema de saúde e de saúde pública devem considerar as implicações do uso de recursos na implementação das opções alternativas. Esse item, juntamente com os itens “equidade”, “aceitabilidade” e “viabilidade”, é mais característico e importante para decisões no contexto de saúde pública do que para outros tipos de decisões, visto que devem ser consideradas as várias partes interessadas na recomendação.

Por fim, os formuladores de políticas e gestores muitas vezes precisam tomar decisões sobre o sistema de saúde e definições de políticas públicas mesmo quando a certeza das evidências é baixa ou muito baixa, assim como também precisam considerar os efeitos diretos e indiretos da implementação de certa tecnologia em saúde. Dessa forma, estão elencados a seguir esses e outros desafios encontrados no processo de uso do GRADE em saúde pública.

Desafios da utilização do sistema GRADE em saúde pública

A aplicação do sistema GRADE em questões de saúde pública apresenta desafios reconhecidos (191). A partir de uma revisão de escopo, o Grupo GRADE de Saúde Pública elencou as prioridades a serem discutidas e solucionadas nesse tópico, descritas a seguir.

Muitas diretrizes de saúde destinam-se a públicos multidisciplinares, inclusive diferentes profissionais da saúde, gestores de saúde, formuladores de políticas de saúde, pacientes e cuidadores. A inclusão de todas os atores envolvidos é fundamental e, ao mesmo tempo, desafiadora devido à heterogeneidade de cada perspectiva (194). Além disso, para abordar perspectivas mais amplas em saúde, as diretrizes de saúde pública podem ser direcionadas a profissionais, formuladores de políticas e outras partes interessadas que não são relacionadas à saúde, cujas perspectivas sobre evidências em saúde podem variar. Como exemplo, pode-se citar a diretriz do *National Institute for Health and Care Excellence* sobre atividade física e meio ambiente, que destina-se tanto ao governo local e seus contratados quanto aos empregadores e a organizações comunitárias responsáveis por espaços públicos, autoridades habitacionais, planejadores e fornecedores de transporte público e organizações que apoiam pessoas com mobilidade limitada (195). A maneira como esses públicos enxergam as questões políticas e valorizam a proteção e a melhoria da saúde e as prioridades estabelecidas em diferentes abordagens políticas pode ser distinta (196), mas essas visões precisam ser compreendidas para que a diretriz atinja seus objetivos. Além disso, diferentes profissões e grupos de interessados podem representar diferentes “culturas de evidência”, nas quais a legislação, os regulamentos e outros fatores contextuais podem ter prioridade sobre a pesquisa científica (197).

Selecionar e priorizar os desfechos

As diferenças na perspectiva e na cultura de evidências também se traduzem em desafios na seleção de desfechos de relevância clínica e no acordo sobre quais desfechos são críticos para a tomada de decisão. Por exemplo, na elaboração de uma política de transporte, as recomendações sobre gerenciamento de tráfego devem considerar, idealmente, todos os desfechos relevantes em saúde, inclusive o impacto nas condições respiratórias, mudanças nos níveis de atividade física, admissões em um serviço de emergência e a segurança para usuários que sofreram alguma vulnerabilidade em seu percurso. No entanto, os formuladores de políticas de transporte público podem não ver esses desfechos como críticos quando comparados ao fluxo de tráfego, ao tempo de deslocamento, a colisões no trânsito, aos custos, ao fornecimento de insumos para operação ou à opinião pública sobre o tema.

Da mesma forma, desafios são encontrados quando a percepção dos benefícios de uma determinada tecnologia difere entre o indivíduo e a comunidade: a imunização individual que leva à imunidade de rebanho pode ser considerada mais relevante para a comunidade do que para aqueles que receberam o tratamento, por exemplo (198). Outro desafio está relacionado ao potencial de troca (*trade-off*) entre saúde da população e equidade em saúde, em que um benefício para a saúde da população pode ser alcançado à custa do aumento das desigualdades. Por exemplo, uma intervenção de saúde pública, como triagem ou exames de prevenção primária, pode ter sucesso na melhoria da morbidade e mortalidade geral em uma população; no entanto, se os melhores resultados forem alcançados predominantemente em um subgrupo com poder socioeconômico mais alto, a desigualdade em saúde pode aumentar, pois haverá uma ampliação na diferença entre os desfechos de saúde dos subgrupos com poder socioeconômico alto e baixo. Por fim, é importante ressaltar que o avanço da saúde implica na possível ponderação dos benefícios para a saúde em relação aos impactos ambientais, econômicos e sociais mais amplos e suas interações potenciais (199, 200).

Interpretar desfechos e identificar os limites para a tomada de decisão

Na tomada de decisão sobre uma determinada tecnologia, é importante definir o contexto no qual o impacto do efeito será interpretado, visto que um desfecho em

saúde que apresenta uma mudança considerada muito pequena na perspectiva individual pode ser percebido de forma diferente da perspectiva da população (201). Por exemplo, um programa de redução de sal pode gerar um pequeno efeito percebido pelo indivíduo, mas pode impactar de maneira significativa na incidência de doenças cardiovasculares observada na população (202).

Nessas circunstâncias, a tomada de decisão pode ser menos influenciada pelo tamanho do efeito de uma intervenção do que pela direção do impacto geral na saúde (por exemplo, resultados prováveis em ganhos ou danos à saúde). As considerações sobre impacto no âmbito populacional, determinantes sociais de saúde e desigualdades e falta de equilíbrio podem influenciar as opiniões sobre o que constitui uma diferença importante e o grau de precisão necessário para apoiar a tomada de decisão.

Avaliar a certeza das evidências de diversas fontes, inclusive de estudos não randomizados

Na saúde pública, alguns tipos de ensaios clínicos não randomizados podem fornecer maior certeza do que outros ao investigar os efeitos na saúde de políticas ou intervenções sociais, considerando que estudos randomizados são menos comuns e, muitas vezes, inviáveis. A orientação atual do grupo GRADE indica que os diferentes tipos de delineamentos de ensaio clínico não randomizado têm potencial para fornecer evidências de qualidade moderada, embora exemplos ainda não estejam bem definidos (203). Enfrentar esse desafio identificando exemplos relevantes de políticas de saúde pública e aplicando essas novas ferramentas parece, portanto, um direcionador promissor por meio do qual soluções para outros desafios também podem surgir.

Abordar as implicações para os tomadores de decisão, inclusive preocupações sobre recomendações condicionais

Uma preocupação dos desenvolvedores de diretrizes e formuladores de políticas em saúde é a predominância de recomendações “fracas” observadas nas diretrizes de saúde pública. O sistema GRADE abordou esse problema alterando a terminologia aplicada em saúde pública: as recomendações fracas são rotuladas predominantemente como condicionais e, em seguida, as condições são especificadas.

Sabe-se que a saúde pública produz diversas situações em que existe uma recomendação forte apesar de a certeza da evidência ser baixa ou muito baixa, como:

- situações de risco de vida;
- benefício incerto, mas dano certo;
- equivalência potencial de eficácia em que uma opção é claramente mais ou menos arriscada e/ou cara;
- potencial para danos catastróficos (204).

As primeiras recomendações fortes no contexto de evidência de certeza muito baixa afirmam que o GRADE havia reconhecido esse problema (205). No entanto, alguns exemplos ainda podem ser contestados, como na política fiscal, em que os economistas podem recomendar fortemente políticas de austeridade, enquanto a saúde pública pode produzir uma recomendação fraca em termos de desfechos de saúde. Classificar recomendações como “condicionais” em vez de “fracas” pode ser uma alternativa para amenizar esse problema, mas os desenvolvedores de diretrizes precisam contextualizar e fundamentar melhor tais recomendações para delinear novas perspectivas. Uma abordagem com base na teoria de decisão bayesiana foi proposta no contexto desse problema, em que a melhor estimativa do efeito na saúde de uma intervenção não precisa depender exclusivamente de evidências de ECR, mas também pode depender de crenças anteriores baseadas em teoria, observação e experiência (206). Um desafio-chave para o GRADE em saúde pública é, portanto, identificar como equilibrar a aplicação correta da metodologia para avaliação das evidências com as suas implicações na formulação de recomendações (ou seja, fortes e condicionais) na perspectiva dos tomadores de decisão em contextos políticos.

12. Sistema GRADE para recomendações em situação de urgência e emergência

- Em situações de urgência, recomendações são necessárias em um período de 1 a 2 semanas, muitas vezes, incompatível com o processo tradicional de desenvolvimento de recomendações.
- Grupos responsáveis por desenvolver essas recomendações necessitam equilibrar a necessidade de realizar uma recomendação em um curto período e a necessidade de seguir padrões metodológicos adequados.
- O desenvolvimento de revisões sistemáticas tradicionais geralmente não é factível, podendo ser utilizadas alternativas como a elaboração de revisões sistemáticas rápidas, uso de revisões sistemáticas existentes, adoção ou adaptação de recomendações de outras diretrizes, ou mesmo uso de evidências selecionadas pelos painelistas por meio de buscas não-sistemáticas.
- Apesar de idealmente ser aplicada para evidência sumarizadas a partir de revisões sistemáticas, o GRADE é um sistema flexível, podendo ser utilizado mesmo na ausência desta, como por exemplo em uma revisão narrativa ou em um conjunto de evidências compiladas para respostas de urgência ou emergência.

12.1 Níveis de urgência no desenvolvimento de recomendações

Em situações de urgência, como foi observado em 2020 com a Covid-19, o período usual de desenvolvimento de recomendações – que pode variar de alguns meses até mesmo alguns anos – não atende às necessidades das partes interessadas (pacientes, profissionais de saúde e gestores) (207). O desenvolvimento rápido de recomendações, seja no contexto de diretrizes clínico-assistenciais, saúde pública ou de incorporação de tecnologias, é desafiador. Nesse contexto, os grupos responsáveis por essas recomendações necessitam equilibrar as necessidades de realizar uma recomendação em um período curto de tempo e de seguir padrões metodológicos adequados (2, 208).

O nível de urgência para o desenvolvimento de uma recomendação pode ser classificado em quatro níveis (Quadro 39) (209). Em certos casos, como desastres naturais, incidentes químicos ou nucleares, recomendações necessitam ser realizadas

em questões de horas (210, 211). Nessas circunstâncias, entende-se que muitas vezes não é factível logisticamente desenvolver recomendações de forma estruturada do ponto de vista logístico; contudo, sempre que possível, a evidência disponível deve ser considerada e avaliada criticamente.

Em circunstâncias nas quais é necessária uma resposta rápida (um a três meses), usualmente recomenda-se seguir o padrão usual de desenvolvimento de recomendações, com maior alocação de recursos humanos e financeiros, com redução de escopo (priorização das questões mais relevantes), e, se possível, utilizando processos de adaptação, como o preconizado pelo GRADE-ADOLOPMENT(212-214). Em algumas situações, pode-se também considerar o desenvolvimento de revisões sistemáticas rápidas em vez do processo tradicional de RS (215).

Quadro 39 - Níveis de urgência e tempo no desenvolvimento de recomendações

- Emergência: algumas horas
- Urgência: 1 a 2 semanas
- Rápida: até 3 meses
- Rotina: Mais do que 3 meses

Fonte: Adaptado de Thayer e Schunemann (209)

12.2 Desenvolvimento de recomendações

Em situações de urgência, nas quais recomendações devem ser realizadas em 1 a 2 semanas, as premissas preconizadas no desenvolvimento de recomendações devem ser mantidas, como adequada declaração e manejo de conflitos de interesse. A necessidade de adequado manejo de conflitos de interesse em situações de urgência é ainda maior, uma vez que o método tende a ser menos robusto para o desenvolvimento de recomendações.

Sempre que possível, revisões sistemáticas rápidas devem ser realizadas, mas, uma vez que a busca, seleção e síntese de evidências é usualmente a etapa que consome mais tempo no desenvolvimento de recomendações, entende-se que nem sempre será factível seu desenvolvimento. Importante destacar que, apesar de

idealmente ser aplicada para evidência sumarizada a partir de revisões sistemáticas, o GRADE é um sistema flexível, podendo ser utilizado mesmo na ausência desta, como por exemplo em uma revisão narrativa ou em um conjunto de evidências compiladas para respostas de urgência ou emergência.

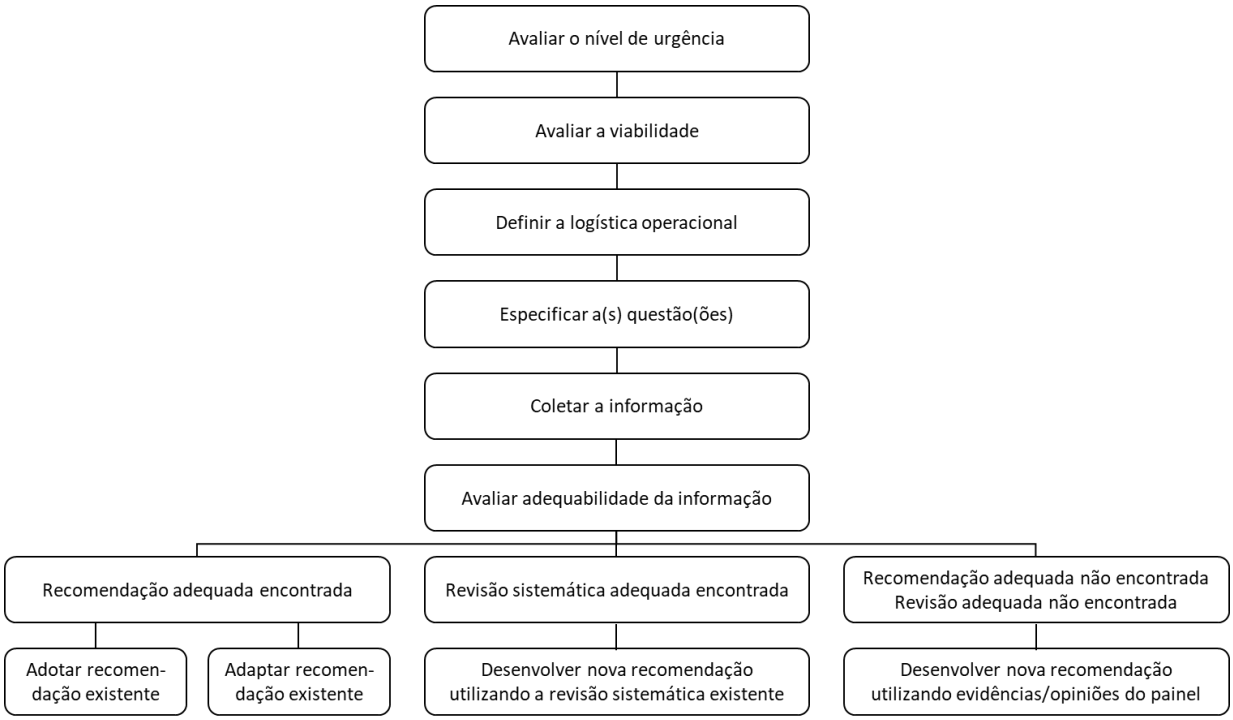
Apesar do processo geral (Figura 42) não diferir substancialmente do processo clássico na apresentação das recomendações, de forma sumarizada, destaca-se os seguintes passos para sistematizar o processo de desenvolvimento de recomendações em um contexto de urgência.

1) Avaliar o nível de urgência: consiste em certificar a real necessidade de elaborar as recomendações em um período de uma a duas semanas. No caso de uma diretriz com um escopo amplo, é importante verificar se todas as questões são sujeitas a serem respondidas por recomendações elaboradas em regime de urgência, ou se pode-se priorizar as questões para as quais uma recomendação urgente faz-se mais necessária.
2) Avaliar a viabilidade: frente a uma demanda que necessita de recomendações urgentes, é fundamental que os prazos sejam cumpridos. Frente a isso é importante que o grupo responsável pela recomendação avalie a viabilidade, considerando o número e a complexidade de questões, a carga de trabalho estimada e a disponibilidade de recursos humanos, incluindo painelistas.
3) Definir a logística operacional: importante constituir o grupo elaborador das recomendações, com função executiva, que possa ter reuniões frequentemente (por exemplo, ao menos 1x/dia), de forma a garantir o progresso do projeto dentro do cronograma proposto. Demais membros do grupo de trabalho devem ser selecionados rapidamente (ex. painelistas, revisores). Planos de comunicação e de documentação do processo devem ser definidos *a priori*. As atividades devem ser desenvolvidas conjuntamente ao demandante (ex. Ministério da Saúde), de forma a garantir a qualidade do processo e a adequabilidade do produto final.
4) Especificar a(s) questão(ões): o processo, o qual é semelhante ao desenvolvimento tradicional de recomendações, possui a necessidade de priorização de questões de forma a garantir um escopo factível. Além disso, situações de urgência são usualmente associadas a escassez de informações, em especial para desfechos clínicos; assim, muitas vezes será necessário

utilizar desfechos substitutos (ex. carga viral) e extrapolar evidências de outros contextos (ex. efetividade de intervenções em SARS e MERS foram inicialmente utilizadas como evidência indireta para Covid-19).

5) Coletar as informações: apesar de o desenvolvimento de novas revisões sistemáticas não ser geralmente factível, deve-se proceder com a busca por recomendações já disponíveis em outras diretrizes (e as suas evidências subjacentes) e revisões sistemáticas disponíveis. Além disso, é importante coletar outras informações necessárias para a tomada de decisão, como custos das intervenções propostas.
6) Avaliar a adequabilidade da informação coletada: deve-se avaliar a sua relevância (evidência direta) para a questão de interesse, credibilidade das revisões e/ou diretrizes identificadas, e atualidade da informação disponível.
7) Desenvolvimento das recomendações. Nesse caso, pode-se realizar uma das seguintes abordagens:
    a. Adotar recomendação existente
    b. Adaptar recomendação existente
    c. Desenvolver nova recomendação utilizando RS identificada (ou elaborada, se for o caso)
    d. Desenvolver nova recomendação, sem o uso de RS, utilizando evidências apresentadas pelos especialistas. Em relação a essa alternativa, os especialistas são solicitados a revisar a literatura de forma a destacar os domínios necessários da tabela de evidência para a decisão. Painelistas podem também descrever suas experiências, o que pode ser considerado como equivalente a séries de casos no processo de tomada de decisão (216). O grupo elaborador, prévio à reunião de recomendações, coleta e compila os dados obtidos dos especialistas em uma tabela de evidência para a decisão, a qual será utilizado para consenso. Apesar dessa prática ser controversa, entende-se que em situações de urgência e escassez de informações, especialistas no tema provavelmente selecionarão as evidências mais relevantes para o processo de tomada de decisão.
8) Planejamento da atualização das recomendações

Figura 42 - Etapas envolvidas no desenvolvimento de recomendações em contexto de urgência



Fonte: Akl et al (217)

12.3 Considerações adicionais

O processo de desenvolvimento de recomendações para urgência deve seguir um processo semelhante ao processo tradicional de elaboração de recomendações. Entendemos que esse processo é aplicável principalmente no desenvolvimento de diretrizes clínico-assistenciais e de recomendações a nível de saúde pública. Contudo, seu uso em recomendações relacionadas à incorporação de tecnologias em saúde não pode ser descartado.

O GRADE é um sistema flexível, podendo ser utilizado mesmo na ausência desta, como por exemplo em uma revisão narrativa ou em um conjunto de evidências compiladas para respostas de urgência ou emergência. Mesmo sem uma RS tradicional, expressar o nível de evidência com o GRADE proporciona informação válida no processo de tomada de decisão. As tabelas GRADE de evidência para decisão devem ser utilizadas, mesmo na ausência de revisões sistemáticas.

13. Considerações finais

O GRADE consiste em um sistema abrangente e transparente para a avaliação da evidência e para a formulação de recomendações, consistindo em uma ferramenta útil na elaboração de diferentes documentos em saúde, como diretrizes clínico-assistenciais, recomendações em saúde pública e avaliação de tecnologias em saúde. Desde 2014, quando foi publicada a primeira versão do manual da ferramenta GRADE, houve importante adoção dessa ferramenta no sistema de saúde brasileiro. A avaliação da certeza da evidência com o GRADE passou a ser utilizado como rotina em diferentes documentos do Ministério da Saúde, como relatórios de recomendação para incorporação de tecnologias e diretrizes clínicas baseadas em evidências.

A área de maior avanço provavelmente foi na elaboração de diretrizes clínicas, não sendo utilizado somente no processo de avaliação da certeza da evidência, mas na elaboração das recomendações, com adoção de tabelas de evidência para decisão e, em alguns casos, sendo utilizadas outras abordagens, como o GRADE-ADOLOPMENT (214, 218). Essas ações estão em linha com a visão estratégica do próprio Ministério da Saúde e Núcleos de Avaliação de Tecnologias em Saúde (NATS) para a elaboração de diretrizes clínico-assistenciais (189).

Desde a publicação das diretrizes em 2014, diversos avanços na metodologia foram observados. O sistema GRADE que, até então, estava direcionado principalmente a revisões sistemáticas de intervenções, testes diagnósticos, e recomendações a elas relacionadas, passou a ter diferentes desdobramentos, como seu uso para estimativas de incidência, prevalência, prognóstico, comparações indiretas, modelagem e incorporação de tecnologias (47, 127, 135, 171, 184).

Dentro do contexto do sistema de saúde brasileiro, entende-se haver benefício com sua adoção em áreas especialmente relacionadas à avaliação de tecnologias em saúde. Um aspecto importante no processo de incorporação de tecnologias são as incertezas presentes, tanto nas estimativas de custo-efetividade quanto de impacto orçamentário, que precisam de uma análise mais crítica sobre a confiança em suas estimativas, com o GRADE podendo ser um instrumento útil nesse processo. Espera-se que em nosso contexto o GRADE passe a ter adoção gradual nessas áreas.

Por fim, é importante entender que o GRADE é um sistema que vem apresentando constante evolução, estando já bem estabelecido em algumas áreas

como na avaliação de intervenções e de recomendações em diretrizes, com seu uso consistindo em boa prática nessas áreas. Sua adoção levou ao desenvolvimento de recomendações mais robustas e transparentes, refletindo em benefícios para o sistema de saúde e para seus usuários. Da mesma forma, entendemos que esses benefícios podem ser também transponíveis a outras áreas, como recomendações em saúde pública e avaliação de tecnologias em saúde, sendo sugerida sua adoção.

MATERIAL SUPLEMENTAR

Anexo I. Exemplo do processo de obtenção da estimativa do efeito absoluto da intervenção

Antagonistas da aldosterona comparado a placebo em pacientes com IC?

Avaliação da Certeza da evidência							Sumário de Resultados				
Participantes (estudos) Seguimento	Risco de viés	Inconsistência	Evidência indireta	Imprecisão	Outros	Certeza geral da evidência	Taxas de eventos do estudo (%)		Efeito relativo (IC 95%)	Efeitos absolutos potenciais	
							Com placebo	Com mineralocorticoides		Risco com placebo	Diferença de risco com mineralocorticoides
Mortalidade (seguimento: média 18 meses)											

IC: Intervalo de confiança; **RR:** Risco relativo
Explicações

-

Nota: área destacada em amarelo indica qual é a parte que está sendo modificada na etapa.

Fonte: elaboração própria.

Sumário de Resultados				
Taxas de eventos do estudo (%)		Efeito relativo (IC 95%)	Efeitos absolutos potenciais	
Com placebo	Com mineralocorticoides		Risco com placebo	Diferença de risco com mineralocorticoides
			População do estudo	
			Taxa anual Brasil (fonte: DATASUS, 2016)	

IC: Intervalo de confiança; **RR:** Risco relativo
Explicações

a. Dois estudos (AREA IN-CHF, 2000 e RALES, 1999) não reportam o modo de cegamento dos pacientes e como foi realizada a alocação sigilosa.
b. Em um estudo (RALES 1999), 29% da amostra apresentava classe IV (NYHA).

Nota: área destacada em amarelo indica qual a é parte que está sendo modificada na etapa.

Fonte: elaboração própria.

Sumário de Resultados				
Taxas de eventos do estudo (%)		**Efeito relativo (IC 95%)**	**Efeitos absolutos potenciais**	
Com placebo	**Com mineralocorticoides**		**Risco com placebo**	**Diferença de risco com mineralocorticoides**
1165/5769 (20.2%)	939/5730 (16.4%)	**RR 0.81** (0.74 a 0.88)	**População do estudo**	
			202 por 1.000	**38 menos por 1.000** (de 53 menos para 24 menos)
			Taxa anual Brasil (fonte: DATASUS, 2016)	

IC: Intervalo de confiança; **RR:** Risco relativo

Explicações

a. Dois estudos (AREA IN-CHF, 2000 e RALES, 1999) não reportam o modo de cegamento dos pacientes e como foi realizada a alocação sigilosa.
b. Em um estudo (RALES 1999), 29% da amostra apresentava classe IV (NYHA).

Risco com o placebo (taxa de eventos com controle)

20,2% ou 202/1000 pacientes com placebo

Risco com mineralocorticoides (taxa de eventos com a intervenção)

% placebo * RR = % intervenção

20,2% * 0,81 = 16,4% ou 164/1000 pacientes tratados

Diferença de risco com mineralocorticoides

% intervenção - % placebo

16,4% - 20,2% = -3,8% ou menos 38/1000 pacientes tratados

Quantos tratados para reduzir 1 evento?

38 casos menos – 1000 tratados

1 caso menos – x tratados

x = (1 * 1000) / 38

1 caso menos/26 pacientes tratados

Nota: área destacada em amarelo indica qual a é parte que está sendo modificada na etapa.

Fonte: elaboração própria.

Sumário de Resultados				
Taxas de eventos do estudo (%)		**Efeito relativo (IC 95%)**	**Efeitos absolutos potenciais**	
Com placebo	**Com mineralocorticoides**		**Risco com placebo**	**Diferença de risco com mineralocorticoides**
1165/5769 (20.2%)	939/5730 (16.4%)	**RR 0.81** (0.74 a 0.88)	**População do estudo**	
			202 por 1.000	**38 menos por 1.000** (de 53 menos para 24 menos)
			Taxa anual Brasil (fonte: DATASUS, 2016)	
			110 por 1.000	**21 menos por 1.000** (de 29 menos para 13 menos)

IC: Confidence interval; **RR:** Risk ratio

Explicações

a. Dois estudos (AREA IN-CHF, 2000 e RALES, 1999) não reportam o modo de cegamento dos pacientes e como foi realizada a alocação sigilosa.
b. Em um estudo (RALES 1999), 29% da amostra apresentava classe IV (NYHA).

Risco com o placebo (taxa de eventos com controle)

11,0% (110/1000 pacientes com placebo)

Risco com mineralocorticoides (taxa de eventos com a intervenção)

11,0% * 0,81 = 8,9% (89/1000 pacientes tratados)

Diferença de risco com mineralocorticoides

8,9% - 20,2% = -2,1% (menos 21/1000 pacientes tratados)

Quantos tratados para reduzir 1 evento?

21 casos menos – 1000 tratados

1 caso menos – x tratados

x = (1 * 1000) / 21

1 caso menos/48 pacientes tratados

Nota: área destacada em amarelo indica qual a é parte que está sendo modificada na etapa.

Fonte: elaboração própria.

Antagonistas da aldosterona comparado a placebo em pacientes com IC?

Avaliação da Certeza da evidência							Sumário de Resultados				
Participantes (estudos) Seguimento	Risco de viés	Inconsistência	Evidência indireta	Imprecisão	Outros	Certeza geral da evidência	Taxas de eventos do estudo (%)		Efeito relativo (IC 95%)	Efeitos absolutos potenciais	
							Com placebo	Com mineralocorticoides		Risco com placebo	Diferença de risco com mineralocorticoides
Mortalidade (seguimento: média 18 meses)											
11499 (4 ECRs)	não grave[a]	não grave	não grave[b]	não grave	nenhum	⊕⊕⊕⊕ Alta	1165/5769 (20.2%)	939/5730 (16.4%)	**RR 0.81** (0.74 a 0.88)	**População do estudo**	
										202 por 1.000	**38 menos por 1.000** (de 53 menos para 24 menos)
										Taxa anual Brasil (fonte: DATASUS, 2016)	
										110 por 1.000	**21 menos por 1.000** (de 29 menos para 13 menos)

IC: Intervalo de confiança; **RR:** Risco relativo
Explicações

a. Dois estudos (AREA IN-CHF, 2000 e RALES, 1999) não relataram o modo de cegamento dos pacientes e como foi realizada a alocação sigilosa.
b. Em um estudo (RALES 1999), 29% da amostra apresentava classe IV (NYHA).

Nota: área destacada em amarelo indica qual a é parte que está sendo modificada na etapa.

Fonte: elaboração própria.

Anexo II. Exemplo de avaliação de uma EtD para teste diagnóstico

Exemplo:

Questão: Deve ou não ser usada a triagem mamográfica organizada para a detecção precoce do câncer de mama em mulheres entre 45 e 49 anos de idade?

PIRO:

População: mulheres entre 45 e 49 anos de idade.

Teste índice (*index test*): triagem mamográfica organizada.

Teste referência (*reference test*): sem triagem mamográfica.

Desfecho (*outcome*): mortalidade por câncer de mama (diagnósticos durante a triagem); mortalidade por câncer de mama (diagnósticos de câncer de mama no maior período de seguimento disponível); mortalidade por outras causas; câncer de mama estágio IIA ou superior; câncer de mama estágio III+ ou tumor com tamanho ≥ 40 mm; taxa de mastectomias; realização de quimioterapia; diagnóstico (diagnósticos de câncer de mama ao longo do período de seguimento); qualidade de vida (inferida pelos efeitos psicológicos); efeitos adversos relacionados a falso-positivos (sofrimento psicológico); e efeitos adversos relacionados a falso-positivos (biópsias e cirurgias).

Recomendação 2. Para mulheres assintomáticas entre 45 e 49 anos de idade que apresentam risco médio de câncer de mama, o GDD da ECIBC sugere triagem mamográfica em vez de nenhuma triagem mamográfica, no contexto de um programa de triagem organizado (recomendação condicional, certeza da evidência moderada)

QUESTÃO

Deve ou não ser usada a triagem mamográfica organizada para a detecção precoce do câncer de mama em mulheres entre 45 e 49 anos de idade?	
POPULAÇÃO:	Mulheres entre 45 e 49 anos de idade
INTERVENÇÃO:	triagem mamográfica organizada
COMPARAÇÃO:	sem triagem mamográfica
DESFECHOS PRINCIPAIS:	Mortalidade por câncer de mama (contabilizando casos diagnosticados durante a triagem); mortalidade por câncer de mama (contabilizando todos os casos de câncer de mama no maior período de seguimento disponível); mortalidade por outras causas; câncer de mama estágio IIA ou superior; câncer de mama estágio III+ ou tumor com tamanho ≥40 mm; taxa de mastectomias; realização de quimioterapia; sobrediagnóstico (contabilizando os casos ocorridos ao longo do período de seguimento); qualidade de vida (inferida pelos efeitos psicológicos); efeitos adversos relacionados a falso-positivos (sofrimento psicológico); e efeitos adversos relacionados a falso-positivos (biópsias e cirurgias)
local:	União Europeia
PERSPECTIVA:	Populacional (Sistema Nacional de Saúde)
RETROSPECTIVA:	Embora a triagem mamográfica tenha tanto potenciais benefícios quanto riscos, muitos países organizaram programas para mulheres com idade superior a 50 anos. Entretanto, ainda existem debates sobre recomendações para triagem mamográfica, no geral (Jorgensen 2009, Arie 2014) e especialmente em se tratando de mulheres entre 40 e 49 anos de idade (Petitti 2010).
CONFLITO DE INTERESSEs:	Gestão dos conflitos de interesse (CIs): os CIs de todos os membros do Grupo de Desenvolvimento das Diretrizes (GDD) foram analisados e gerenciados pelo Centro Comum de Investigação (*Joint Research Center,* JRC), após um procedimento estabelecido segundo as normas da Comissão Europeia. A participação dos membros do GDD no desenvolvimento das recomendações esteve em conformidade com a declaração de CIs. Consequentemente, para esta questão em particular, os seguintes membros do GDD não puderam votar: Roberto d'Amico, Jan Danes, Axel Gräwingholt e Ruben van Engen.

Avaliação

Problema O problema é uma prioridade?		
DECISÃO	EVIDÊNCIAS DE PESQUISA	CONSIDERAÇÕES ADICIONAIS
○ Não ○ Provavelmente não ○ Provavelmente sim ● Sim ○ Varia ○ Não se sabe	O câncer de mama é o segundo câncer mais comum no mundo e, sem dúvida, o câncer mais frequente entre mulheres, com uma estimativa de 1,67 milhões de novos casos de câncer diagnosticados em 2012 – representando 25% do total de casos de câncer (GLOBOCAN 2012). O câncer de mama é a quinta maior causa de mortalidade por câncer no mundo e a segunda maior causa de mortalidade por câncer em regiões desenvolvidas (GLOBOCAN 2012). Na União Europeia, 367.090 mulheres foram diagnosticadas com câncer de mama, e 92.000 mulheres faleceram devido à doença em 2012 (Ferlay 2013). O câncer de mama ocupa o quarto lugar entre os tipos de câncer com maior carga de doença (Tsilidis 2016). A incidência anual de câncer de mama na UE em mulheres entre 45 e 49 anos de idade é de 1,7 por 1.000, e a mortalidade é de 0,2 por 1.000 por ano (GLOBOCAN 2012).	O GDD priorizou essa pergunta para a ECIBC.
Efeitos Desejáveis Quão substanciais são os efeitos desejáveis previstos?		
DECISÃO	EVIDÊNCIAS DE PESQUISA	CONSIDERAÇÕES ADICIONAIS

○ Triviais
○ Pequenos
● Moderados
○ Grandes
○ Variam
○ Não sabe

Triagem mamográfica organizada comparada a nenhuma triagem mamográfica para a detecção precoce do câncer de mama em mulheres entre 45 e 49 anos

paciente ou população: detecção precoce do câncer de mama em mulheres entre 45 e 49 anos
Contexto: União Europeia
Intervenção: triagem mamográfica organizada
Comparação: sem triagem mamográfica

Desfechos	Nº de participantes (estudos) Seguimento	Certeza da evidência (GRADE)	Efeito relativo (IC95%)	Efeitos potenciais absolutos	
				Risco sem triagem mamográfica	Diferença de risco com a triagem mamográfica organizada
Mortalidade por câncer mama (contabilizando casos diagnosticados durante a triagem) para mulheres abaixo de 50 anos seguimento: média 16,8 anos	348.478 (8 ECRs)[1,2,3,4,5,6,7,a]	⊕⊕⊕◯ Moderada[b,c,d]	**RR 0,88** (0,76 a 1,02)	Baixo	
				400 por 100.000[e]	**48 menos por 100.000** (96 menos a 8 mais)
				Alto	
				700 por 100.000	**84 menos por 100.000** (168 menos a 14 mais)
Qualidade de vida (inferida a partir dos efeitos psicológicos)[g]	(54 estudos observacionais)[8]	⊕⊕◯◯ Baixa[h]	Uma revisão sistemática incluindo 54 estudos – sem metanálise (Brett 2005). A triagem mamográfica parece não gerar ansiedade em mulheres que recebem um resultado claro após a mamografia e em seguida retornam para uma consulta de rotina. Houve resultados variáveis quanto às mulheres que voltaram para exames complementares: vários estudos relataram ansiedade transitória ou de longo prazo (de 6 meses a 1 ano após o retorno), enquanto outros estudos não relataram diferenças nos níveis de ansiedade. A natureza e a extensão dos exames complementares parecem determinar o grau de ansiedade.		

Esses estudos utilizaram uma análise por intenção de tratar; portanto, uma abordagem por protocolo levaria a maiores efeitos absolutos.

Os membros do GDD mencionaram que se devem considerar estudos de modelagem que descrevam a qualidade e a duração da "vida ganha".

A contabilização dos casos ocorridos durante o período de seguimento pode diluir o efeito da intervenção, pois, em alguns estudos, incluirá casos diagnosticados após a conclusão do estudo quando ambos os braços estão recebendo a mesma intervenção. Portanto, realizamos uma análise de sensibilidade incluindo apenas estudos que relataram estimativas dos casos ocorridos durante o seguimento, e observamos um efeito diluidor pequeno, embora não significativo (RR 0,92; IC95% 0,83 a 1,02).

Os membros do GDD concordam que os efeitos desejáveis para a saúde diferem em relação à idade na primeira triagem. Para as mulheres entre 45 e 49 anos de idade, os membros do GDD concordaram que elas teriam maiores efeitos benéficos previstos

	Efeitos adversos relacionados a falso-positivos (sofrimento psicológico)[g]	(24 estudos observacionais)[9,10]	⊕⊕◯◯ Baixa	Duas revisões sistemáticas. Uma revisão incluiu 17 estudos e identificou que mulheres que receberam um resultado de mamografia falso-positivo apresentaram maior estresse, medo, ansiedade e preocupação a respeito do câncer de mama (Saltz 2010). Em uma segunda revisão que incluiu 7 estudos, o RR de sofrimento psicológico, avaliado medidas específicas da doença, em mulheres (idade não especificada) com mamografia falso-positiva 35 meses após a última avaliação foi: para mulheres que necessitaram de mamografia complementar RR=1,28 (IC95% 0,82-2,00); para mulheres logo chamadas para uma consulta de retorno, RR=1,82 (IC95% 1,22-2,72); para mulheres de necessitaram de punção aspirativa por agulha fina RR=1,80 (IC95% 1,17-2,77); para mulheres que necessitaram de biópsia RR=2,07 (IC95% 1,22-3,52); não foram observadas diferenças em medidas genéricas de ansiedade geral e depressão 6 semanas após a avaliação e 3 meses após a triagem (Bond, 2013).	para a saúde (efeito moderado) em comparação às mulheres com idade entre 40 e 44 anos, devido à maior incidência absoluta e maior mortalidade por câncer de mama em mulheres de 45-49 anos em comparação às mulheres de 40-44 anos. A redução do percentual de mortalidade não diferiu significativamente daquela observada em mulheres de 50 a 69 anos; embora existam consideráveis evidências observacionais de benefício para mulheres de 45 a 49 anos (ver perfil das evidências). A precisão do exame é menor em mulheres mais jovens, principalmente devido à sua densidade mamária na mamografia. A mamografia digital, que não estava em uso na época da maioria dos estudos aqui revisados, pode aumentar a precisão do exame em mulheres de 45 a 49 anos. Na coorte da *Sweden Mammography Screening of Young Women* (SCRY), que comparou a mortalidade por câncer de mama entre mulheres convidadas e não convidadas para triagem, foram relatadas RRs de 0,82 (IC95%, 0,67-1,00) e 0,63 (IC95%, 0,54-0,75) para as faixas etárias de 40 a 44 e de 45 a 49 anos,
	Efeitos adversos relacionados a falso-positivos (biópsias e cirurgias)[g]	(4 estudos observacionais)[11]	⊕◯◯◯ Muito baixa[i]	Os resultados de uma revisão da literatura (4 estudos, 390.000 mulheres de 50 a 69 anos) demonstraram um percentual total de falso-positivos de 19,7% na triagem em mulheres que realizaram 10 exames bienais de triagem (estimativa agrupada de risco com base em 3 estudos; intervalo 8 - 21%). Isso esteve relacionado a um risco cumulativo agrupado de 2,9% de procedimento invasivo com desfecho benigno (intervalo 1,8% a 6,3%; com base em 2 estudos) e um risco de 0,9% de ser submetido a intervenção cirúrgica com desfecho benigno (com base em 1 estudo) (Hofvind 2012). Dados transversais do Projeto EUNICE (mulheres entre 50 e 69 anos): 17 países, 20 programas de triagem, 1,7 milhão de triagens iniciais, 5,9 milhões de triagens subsequentes; demonstrou que 2,2% e 1,1% de todos os exames de triagem resultaram em biópsia por agulha em mulheres sem câncer de mama (triagem inicial e triagens subsequentes, respectivamente). Além disso, 0,19% e 0,07% de todos os exames de triagem resultaram em intervenções cirúrgicas em mulheres sem câncer de mama (triagem inicial e triagens subsequentes, respectivamente).	

		respectivamente. A RR ponderada para o grupo de 40 a 49 anos não diferiu da estimativa não ponderada de 0,71 (IC95%, 0,62-0,80).
	* O risco no grupo de intervenção (e seu intervalo de confiança de 95%) é baseado no risco assumido do grupo comparador e o efeito relativo da intervenção (e seu IC95%). IC: intervalo de confiança; RR: razão de risco Graus de evidência do Grupo de Trabalho do GRADE *Working Group* Certeza alta: estamos muito confiantes de que o verdadeiro efeito está próximo do efeito estimado. Certeza moderada: estamos moderadamente confiantes no efeito estimado: o efeito verdadeiro provavelmente está próximo do estimado, mas existe alguma possibilidade de que seja substancialmente diferente. Certeza baixa: nossa confiança sobre o efeito estimado é limitada: o efeito verdadeiro pode ser substancialmente diferente do efeito estimado. Certeza muito baixa: temos muito pouca confiança no efeito estimado: é provável que o verdadeiro efeito seja substancialmente diferente do efeito estimado. Explicações a. As referências listadas no perfil de evidências correspondem às publicações específicas utilizadas para extrair os dados crus com a finalidade de estimar os tamanhos de efeito dos desfechos. As referências adicionais que descrevem as características dos estudos incluídos podem ser encontradas no texto principal do documento desta revisão sistemática. b. Alguns estudos utilizaram métodos que atualmente não seriam aceitáveis para a alocação aleatória. Um estudo realizou a avaliação não cega da "causa do óbito". O GDD considerou que o CNBSS-1 possivelmente apresentasse problemas em alcançar o equilíbrio prognóstico. O GDD considerou que a falta de sigilo da alocação nesse conjunto de estudos não acarretou um alto risco de viés. Devido à falta de ensaios clínicos únicos que motivassem os resultados gerais e à semelhança dos tamanhos de efeito (o teste para diferenças entre subgrupos - estudos com baixo vs alto risco de viés - não foi significativo) e os intervalos de confiança (ICs) sobrepostos, o risco de viés foi classificado como "não grave". c. O IC95% provavelmente cruza o limite de decisão clínica (como o IC é amplo, pode-se tomar uma decisão clínica diferente sobre a intervenção, dependendo se consideramos o limite inferior ou superior). d. Os ensaios clínicos foram realizados há mais de 20 anos. Atualmente, as mulheres têm maior adesão à triagem para câncer de mama e houve melhorias no controle de qualidade da triagem e no cuidado do câncer de mama. Um grande estudo não randomizado (Hellquist B 2011) demonstrou uma redução do risco de óbito por câncer de mama em mulheres entre 40 e 49 anos convidadas para a triagem, se comparadas com as mulheres que não foram convidadas (RR=0,74; IC95%, 0,66-0,83), o que é consistente com os resultados observados nos ECRs. O GDD não	

	diminuiu a classificação devido ao fato de os dados para mortalidade por câncer de mama serem indiretos, mas a considerou grave para outros desfechos. e. Mediana ou média do grupo controle dos estudos incluídos, salvo indicação em contrário. g. A importância do desfecho foi reduzida de "crítica" para "importante" porque os membros do GDD consideraram que esse desfecho não influenciou nem a direção nem a força da recomendação. h. Inconsistência inexplicável em relação à variabilidade no nível de ansiedade no grupo de mulheres chamadas para exames complementares. i. Os estudos incluíram mulheres entre 50 e 69 anos de idade. É provável que as estimativas para a faixa etária de 45-49 anos sejam maiores. Referências	
Efeitos Indesejáveis Quão substanciais são os efeitos indesejáveis previstos?		
Decisão	EVIDÊNCIAS DE PESQUISA	CONSIDERAÇÕES ADICIONAIS
○ Grandes ● Moderados ○ Pequenos ○ Triviais ○ Variam ○ Não se sabe	Triagem mamográfica organizada comparada a nenhuma triagem mamográfica para a detecção precoce de câncer de mama em mulheres de 45 a 49 anos paciente ou população: detecção precoce de câncer de mama em mulheres de 45 a 49 anos Contexto: União Europeia Intervenção: triagem mamográfica organizada Comparação: sem triagem mamográfica	O sobrediagnóstico e sua magnitude não são grandemente influenciados pela idade na primeira triagem. A estimativa de sobrediagnóstico tanto no CNBSS1 quanto no CNBSS2 pode ter sido superestimada pela triagem subsequente na população (tanto organizada quanto oportunista) após o término da CNBSS em 1988. Portanto, enquanto, no seguimento de 25 anos, observou-se um excesso estatisticamente não significativo de todos os casos de câncer de mama no braço de intervenção dos ensaios CNBSS (diferença 2,6; IC95% -0,8 a 5,9), a taxa de excesso de câncer de mama *in situ*/invasivo na verdade aumentou ao longo dos cinco anos após a triagem no CNBSS1, e diminuiu

Desfechos	Nº de participantes (estudos) Seguimento	Certeza da evidência (GRADE)	Efeito relativo (IC95%)	Efeitos absolutos potenciais	
				Risco sem triagem mamográfica	Diferença de risco com triagem mamográfica organizada
				Baixo	

	Mortalidade por câncer de mama (contabilizando casos diagnosticados durante a triagem) para mulheres abaixo de 50 anos seguimento: média 16,8 anos	348.478 (8 ECRs)[1,2,3,4,5,6,7,a]	⊕⊕⊕◯ Moderada[b,c,d]	**RR** (0.76 a 1.02)	**0.88**	400 por 100.000[e]	**48 menos por 100.000** (96 menos a 8 mais)	consideravelmente 10 anos após a triagem no CNBSS2. Devido à antecipação diagnóstica (tempo de diagnóstico antecipado com a realização da triagem), deve haver um grande número de casos de câncer a ser tratado no grupo triado, durante o período de observação, o que pode ocasionar um aumento da taxa de quimioterapia e de mastectomia no grupo triado. Observou-se que a taxa de falso-positivos é maior em mulheres abaixo de 50 anos do que em mulheres de 50 a 69 anos.
						Alto		
						700 por 100.000	**84 menos por 100.000** (168 menos a 14 mais)	
	Qualidade de vida (inferida pelos efeitos psicológicos)[g]	(54 estudos observacionais)[8]	⊕⊕◯◯ Baixa[h]	Uma revisão sistemática incluindo 54 estudos – sem metanálise (Brett 2005). A triagem mamográfica parece não gerar ansiedade em mulheres que recebem um resultado claro após a mamografia e em seguida retornam para uma consulta de rotina. Houve resultados variáveis quanto às mulheres que voltaram para exames complementares: vários estudos relataram ansiedade transitória ou de longo prazo(de 6 meses a 1 ano após o retorno), enquanto outros estudos não relataram diferenças nos níveis de ansiedade. A natureza e a extensão dos exames complementares parecem determinar o grau de ansiedade.				

	Efeitos adversos relacionados a falso-positivos (sofrimento psicológico)[g]	(24 estudos observacionais)[9,10]	⊕⊕◯◯ Baixa	Duas revisões sistemáticas. Uma revisão incluiu 17 estudos e identificou que mulheres que receberam um resultado de mamografia falso-positivo apresentaram maior sofrimento, medo, ansiedade e preocupação a respeito do câncer de mama (Saltz 2010). Em uma segunda revisão que incluiu 7 estudos, o RR de sofrimento psicológico, avaliado utilizando medidas específicas da doença, em mulheres (idade não especificada) com mamografia falso-positiva 35 meses após a última avaliação foi: para mulheres que necessitaram de mamografia complementar RR=1,28 (IC95% 0,82-2,00); para mulheres logo chamadas para uma consulta de retorno RR=1,82 (IC95% 1,22-2,72); para mulheres que necessitaram de punção aspirativa por agulha fina RR=1,80 (IC95% 1,17-2,77); para mulheres que necessitaram de biópsia RR=2,07 (IC95% 1,22-3,52); não foram observadas diferenças em médicas genéricas de ansiedade geral e depressão 6 semanas após a avaliação e 3 meses após a triagem (Bond, 2013).	
	Efeitos adversos relacionados a falso-positivos (biópsias e cirurgias)[g]	(4 estudos observacionais)[11]	⊕◯◯◯ Muito baixa[i]	Os resultados de uma revisão sistemática (4 estudos, 390.000 mulheres entre 50 e 69 anos) demonstraram um percentual total de 19,7% de falso-positivos em mulheres que realizaram 10 bienais de triagem (estimativa agrupada de risco com base em 3 estudos; intervalo 8 - 21%). Isso esteve relacionado a um risco cumulativo agrupado de 2,9% de procedimento invasivo com desfecho benigno (intervalo 1,8% a 6,3%; com base em 2 estudos) e um risco de 0,9% de ser submetido a intervenção cirúrgica com desfecho benigno (com base em 1 estudo) (Hofvind 2012). Dados transversais do Projeto EUNICE (mulheres de 50 a 69 anos): 17 países, 20 programas de triagem, 1,7 milhão de triagens iniciais, 5,9 milhões de triagens subsequentes; demonstrou que 2,2% e 1,1% de todos os exames de triagem resultaram em biópsia por agulha em mulheres sem câncer de mama (triagem inicial e triagens subsequentes, respectivamente). Além disso, 0,19% e 0,07% de todos os exames de triagem resultaram em intervenções cirúrgicas em mulheres sem câncer de mama (triagem inicial e triagens subsequentes, respectivamente).	

	* O risco no grupo de intervenção (e seu intervalo de confiança de 95%) é baseado no risco assumido do grupo comparador e o efeito relativo da intervenção (e seu IC95%). IC: Intervalo de confiança; RR: Razão de risco Graus de evidência do Grupo de Trabalho do *GRADE Working Group* Certeza alta: estamos muito confiantes de que o verdadeiro efeito está próximo do efeito estimado. Certeza moderada: estamos moderadamente confiantes no efeito estimado: o efeito verdadeiro provavelmente está próximo do estimado, mas existe alguma possibilidade de que seja substancialmente diferente. Certeza baixa: nossa confiança sobre o efeito estimado é limitada: o efeito verdadeiro pode ser substancialmente diferente do efeito estimado. Certeza muito baixa: temos muito pouca confiança no efeito estimado: é provável que o verdadeiro efeito seja substancialmente diferente do efeito estimado. Explicações a. As referências listadas no perfil de evidências correspondem às publicações específicas utilizadas para extrair os dados brutos com a finalidade de estimar os tamanhos de efeito dos desfechos. As referências adicionais que descrevem as características dos estudos incluídos podem ser encontradas no texto principal do documento desta revisão sistemática. b. Alguns estudos utilizaram métodos que atualmente não seriam aceitáveis para a alocação aleatória. Um estudo realizou a avaliação não cega da "causa do óbito". O GDD considerou que o CNBSS-1 possivelmente apresentasse problemas em alcançar o equilíbrio prognóstico. O GDD considerou que a falta de sigilo da alocação nesse conjunto de estudos não acarretou um alto risco de viés. Devido à falta de ensaios clínicos únicos que motivassem os resultados gerais e à semelhança dos tamanhos de efeito (o teste para diferenças entre subgrupos - estudos com baixo vs alto risco de viés - não foi significativo) e os intervalos de confiança (ICs) sobrepostos, o risco de viés foi classificado como 'não grave'. c. O IC95% provavelmente cruza o limite de decisão clínica (como o IC é amplo, pode-se tomar uma decisão clínica diferente sobre a intervenção, dependendo se consideramos o limite inferior ou superior). d. Os ensaios clínicos foram realizados há mais de 20 anos. Atualmente, as mulheres têm maior adesão à triagem para câncer de mama e houve melhorias no controle de qualidade da triagem e no cuidado do câncer de mama. Um grande estudo não randomizado (Hellquist B 2011) demonstrou uma redução do risco de óbito por câncer de mama em mulheres entre 40 e 49 anos convidadas para a triagem, se comparadas com as mulheres que não foram convidadas (RR=0,74; IC95%, 0,66-0,83), o que é consistente com os resultados observados nos ECRs. O GDD não diminuiu a classificação devido ao fato de os dados para mortalidade por câncer de mama serem indiretos, mas a considerou sério para outros desfechos. e. Mediana ou média do grupo controle dos estudos incluídos, salvo indicação em contrário.	

	g. A importância do desfecho foi reduzida de "crítica" para "importante" porque os membros do GDD consideraram que esse desfecho não influenciou nem a direção nem a força da recomendação. h. Inconsistência inexplicável em relação à variabilidade no nível de ansiedade no grupo de mulheres chamadas para exames complementares. i. Os estudos incluíram mulheres entre 50 e 69 anos de idade. É provável que as estimativas para a faixa etária de 45-49 anos sejam maiores. Referências 1.Miller AB, Baines CJ,To T,Wall C.. Canadian National Breast Screening Study: 1. Breast cancer detection and death rates among women aged 40 to 49 years. CMAJ; 1992. 2.Tabar L, Duffy SW,Yen MF,Warwick J,Vitak B,Chen HH,Smith RA.. All-cause mortality among breast cancer patients in a screening trial: support for breast cancer mortality as an end point. J Med Screen; 2002. 3.S, Shapiro. Periodic screening for breast cancer: the HIP Randomized Controlled Trial. Health Insurance Plan. J Natl Cancer Inst Monogr; 1997. 4.Bjurstam NG, Björneld LM,Duffy SW.. Updated results of the Gothenburg Trial of Mammographic Screening. Cancer; 2016. 5.Nyström L, Andersson I,Bjurstam N,Frisell J,Nordenskjöld B,Rutqvist LE. Long-term effects of mammography screening: updated overview of the Swedish randomised trials. Lancet; 2002. 6.Moss SM, Cuckle H,Evans A,Johns L,Waller M,Bobrow L,Group.,Trial,Management. Effect of mammographic screening from age 40 years on breast cancer mortality at 10 years´follow-up: a randomised controlled trial. Lancet Oncol; 2015. 7.Habbema JD, van Oortmarssen GJ,van Putten DJ,Lubbe JT,van der Maas PJ.. Age-specific reduction in breast cancer mortality by screening: an analysis of the results of the Health Insurance Plan of Greater New York study. J Natl Cancer Inst.; 1986. 8.Brett J, Bankhead C,Henderson B,Watson E,Austoker J.. The psychological impact of mammographic screening. A systematic review. Psychooncology; 2005. 9.Bond M, Pavey T,Welch K,Cooper C,Garside R,Dean S,et al.. Systematic review of the psychological consequences of false-positive screening mammograms. Health Technol Assess; 2013. 10.Salz T, Richman AR,Brewer NT.. Meta-analyses of the effect of false-positive mammograms on generic and specific psychosocial outcomes. Psychooncology; 2010. 11.Hofvind S1, Ponti A,Patnick J,Ascunce N,Njor S,Broeders M,et al.. False-positive results in mammographic screening for breast cancer in Europe: a literature review and survey of service screening programmes. J Med Screen; 2012.	
Certeza de evidência Qual é a certeza geral da evidência dos efeitos?		
DECISÃO	EVIDÊNCIAS DE PESQUISA	CONSIDERAÇÕES ADICIONAIS

○ Muito baixa ○ Baixa ● Moderada ○ Alta ○ Nenhum estudo incluído	A certeza geral (isto é, a qualidade) da evidência foi considerada moderada, visto que esta foi a qualidade mais baixa entre os desfechos críticos — a saber, mortalidade por câncer de mama e sobrediagnóstico.	
Valores Há uma incerteza ou variabilidade importante no quanto as pessoas valorizam os desfechos principais?		
DECISÃO	EVIDÊNCIAS DE PESQUISA	CONSIDERAÇÕES ADICIONAIS
○ Incerteza ou variabilidade importante ● Possivelmente há incerteza ou variabilidade importante ○ Possivelmente não há incerteza ou variabilidade importante ○ Nenhuma incerteza ou variabilidade importante	Uma revisão sistemática demonstra que os participantes dão pouco valor aos efeitos psicossociais e físicos dos resultados falso-positivos e do sobrediagnóstico (JRC *Technical Repo*rt PICO 10-11, contrato FWC443094012015; disponível mediante solicitação). As mulheres geralmente consideram esses efeitos indesejáveis aceitáveis (baixa confiança na evidência). Entretanto, esses achados têm valor limitado, principalmente devido a preocupações significativas a respeito da adequação das informações fornecidas pelas mulheres, a fim de tomar uma decisão informada sobre a participação. Além disso, a aceitabilidade dos resultados falso-positivos se baseia em estudos de participantes que já receberam um resultado falso-positivo. A preferência delas pode ser diferente daquela da população em geral. Também se observou que a triagem para câncer de mama representa um fardo significativo para alguns participantes, devido ao sofrimento psicológico e à inconveniência associados (confiança moderada na evidência). Quanto ao diagnóstico de câncer de mama, dados muito limitados estão disponíveis abordando as visões das pessoas. Um dos principais temas identificados na literatura é o fato de que as pessoas desvalorizam enormemente a ansiedade causada pela demora em receber os resultados diagnósticos, ou pela falta de entendimento dos exames devido à comunicação subótima com os médicos (confiança moderada na evidência). Além disso, as pessoas têm uma preferência geral maior por procedimentos diagnósticos mais confortáveis e breves (confiança moderada na evidência). (JRC *Technical Report* PICO 10-11, contrato FWC443094012015; disponível mediante solicitação).	
Saldo de efeitos O saldo de efeitos desejáveis e indesejáveis favorece a intervenção ou a comparação?		
DECISÃO	EVIDÊNCIAS DE PESQUISA	CONSIDERAÇÕES ADICIONAIS

○ Favorece a comparação ○ Provavelmente favorece a comparação ○ Não favorece a intervenção nem a comparação ● Provavelmente favorece a intervenção ○ Favorece a intervenção ○ Varia ○ Não se sabe	Não foram identificadas evidências de pesquisa.	Os membros do GDD concordam que a primeira triagem aos 45 anos de idade apresentou efeitos desejáveis moderados para a saúde e efeitos indesejáveis moderados para a saúde; entretanto, não foi alcançado consenso quanto ao saldo entre esses dois. Dezesseis membros votaram que o saldo provavelmente favorece a intervenção; cinco membros votaram que o saldo não favorece a intervenção nem a comparação; e um membro votante se absteve.
Recursos necessários Quais são os recursos necessários (custos)?		
DECISÃO	EVIDÊNCIAS DE PESQUISA	CONSIDERAÇÕES ADICIONAIS
○ Grandes custos ● Custos moderados ○ Custos e economia insignificantes ○ Economia moderadas ○ Grande economia ○ Variam ○ Não se sabe	Diferenças nos recursos necessários para triagem mamográfica versus nenhuma triagem em mulheres entre 40 e 49 anos nos estudos analisados podem estar relacionadas à inclusão ou não de custos relacionados ao processo de triagem, às técnicas diagnósticas, ao tratamento e ao seguimento das mulheres diagnosticadas (1) (2). Os custos da triagem para uma coorte de 10.000 mulheres entre 47 e 49 anos foram estimados em £ 420.000 no Reino Unido. O custo do diagnóstico para os resultados positivos seria de £ 70.000, e a triagem levaria a uma economia de £ 17.000 em custos de tratamento (£480 por câncer detectado pela triagem, calculada com base na diferença nos custos do tratamento entre os braços do estudo denominados controle e intervenção), levando a triagem a ter um custo líquido de £ 473.000 por 10.000 mulheres triadas (taxa de desconto de 3,5%) (1) Com base nos resultados do estudo (2), o custo total do diagnóstico de câncer de mama, tratamento e óbito sem triagem foi estimado em € 1.161.008 por 1.000 mulheres entre 50 e 74 anos, acompanhadas ao longo da vida (taxa de desconto de 3,5%). A triagem bienal custará € 1.298.065 por 1.000 mulheres (idade 50-74 anos) triadas, e o custo relatado para 1.000 entre 40 e 74 anos é de € 1.467.598. Portanto, o custo estimado da triagem de 1.000 mulheres entre 40 e 49 anos seria de € 169.533.	Varia de acordo com o intervalo de triagem, com o país, e com a presença de triagem oportunista. Os membros do GDD consideraram que o custo é pelo menos moderado. Entretanto, foi possível observar diferenças consideráveis em países europeus sem programas de triagem de base populacional ou naqueles programas com políticas de triagem diferentes.

		As estimativas se referem a programas de triagem organizados. Análises de recursos/custos no nível local/regional/nacional existem, ou são necessárias, para estimar o custo para cada cenário.
Certeza da evidência dos recursos necessários Qual é a certeza da evidência dos recursos necessários (custos)?		
DECISÃO	EVIDÊNCIAS DE PESQUISA	CONSIDERAÇÕES ADICIONAIS
○ Muito baixa ● Baixa ○ Moderada ○ Alta ○ Nenhum estudo incluído	A certeza da evidência dos recursos necessários é baixa devido ao delineamento de estudo dos estudos incluídos, os quais eram estudos de modelagem baseados em dados observacionais. Além disso, foram observadas as seguintes diferenças: no estudo (1), os parâmetros do modelo se basearam em dados de uma triagem trienal, enquanto os dados provenientes do estudo (2) corresponderam a uma triagem bienal. Os estudos relataram os custos da triagem, do diagnóstico e do tratamento. Com base nos dados desses estudos, os custos totais por extensão de uma rodada de triagem trienal seria de £47 por mulher no Reino Unido (valor de 2006), valor semelhante ao de €61,3 por uma rodada de triagem bienal nos Países Baixos (valor de 2014). A avaliação formal da certeza da evidência dos custos e recursos utilizados foi realizada utilizando os critérios GRADE e foi relatada no Perfil de Evidências (JRC *Technical Report* PICO 14-15, contrato FWC443094012015; disponível mediante solicitação).	Ambos os estudos avaliaram a extensão dos seus programas atuais de triagem de base populacional. Conforme mencionado anteriormente, podem-se observar diferenças consideráveis em países europeus sem programas de triagem de base populacional ou naqueles programas com políticas de triagem diferentes.
Custo-efetividade A relação custo-efetividade da intervenção favorece a intervenção ou a comparação?		
Decisão	EVIDÊNCIAS DE PESQUISA	CONSIDERAÇÕES ADICIONAIS

○ Favorece a comparação ○ Provavelmente favorece a comparação ○ Não favorece a intervenção nem a comparação ○ Provavelmente favorece a intervenção ○ Favorece a intervenção ● Varia ○ Nenhum estudo incluído	Com base nas evidências fornecidas pelo estudo (2), a ampliação da triagem mamográfica bienal com início aos 40 anos de idade parece ser custo-efetiva com um valor de "disposição a pagar" de €20 000 por ano de vida ganho (AVG) e uma razão de custo-efetividade incremental (RCEI) de €10,826 por AVG ao se iniciar a triagem aos 40 anos em vez de 45 anos. Por outro lado, com base nas evidências fornecidas pelo estudo (1), a ampliação da triagem mamográfica trienal em mulheres entre 47 e 49 anos parece não ser custo-efetiva com um valor de "disposição a pagar" de £20.000 por anos de vida ajustados por qualidade de vida (*quality-adjusted life years*, QALYs). A probabilidade de a triagem ser custo-efetiva nesse limite era pequena (29%). A RCEI por QALYs ganho com a triagem trienal foi de £27.400.	As diferenças na relação custo-efetividade poderiam ser explicadas pelas diferenças de cenário, política dos programas de triagem, medidas de desfecho e o tipo de tecnologia utilizada. Enquanto o estudo (2) relatou a RCEI por AVG nos Países Baixos, o estudo (1) relatou a RCEI por QALYs no Reino Unido. Os efeitos negativos dos resultados falso-positivos no Reino Unido reduziram significativamente o QALYs. O estudo (2) avaliou a mamografia digital, enquanto o estudo (1) avaliou a mamografia convencional. Os membros do GDD consideraram a relação custo-efetividade variável, com base nos resultados opostos obtidos nos estudos de modelagem.
Equidade Qual seria o impacto na equidade da saúde?		
DECISÃO	EVIDÊNCIAS DE PESQUISA	CONSIDERAÇÕES ADICIONAIS
○ Diminuiria ○ Provavelmente diminuiria ○ Provavelmente sem	Não foram identificadas evidências de pesquisa.	Não foram realizadas revisões sistemáticas a respeito desse tema. Entretanto, a utilização dos serviços de triagem de câncer de mama pode depender, em

impacto ○ Provavelmente aumentaria ○ Aumentaria ○ Varia ● Não se sabe		grande parte, da disponibilidade de programas públicos de triagem, embora achados europeus destaquem que as desigualdades são maiores em países sem programas de triagem de base populacional (Palència, 2010).
Aceitabilidade A intervenção é aceitável para os principais interessados?		
DECISÃO	EVIDÊNCIAS DE PESQUISA	CONSIDERAÇÕES ADICIONAIS
○ Não ○ Provavelmente não ○ Provavelmente sim ● Sim ○ Varia ○ Não se sabe	Uma revisão sistemática (JRC Technical Report PICO 16-17, contrato FWC443094032016; disponível mediante solicitação) encontrou as seguintes barreiras associadas à triagem para câncer de mama: (a) falta de conhecimento e percepções equivocadas sobre medicina preventiva e saúde da mama (alta confiança na evidência), (b) pouca habilidade de comunicação por parte dos profissionais de saúde (alta confiança na evidência), (c) dificuldades de acessibilidade à triagem mamária, especialmente entre mulheres com deficiência (alta confiança na evidência), (d) medo e estresse relacionados ao procedimento e à possibilidade do diagnóstico de câncer (alta confiança na evidência), (e) dor e desconforto durante o procedimento (confiança moderada na evidência), (f) constrangimento e timidez durante o procedimento (confiança moderada na evidência), (g) falta de apoio e encorajamento por parte dos familiares, cuidadores e rede social (confiança moderada na evidência), (h) falta de informação sobre os recursos disponíveis (baixa confiança na evidência) e (i) baixa priorização da triagem para câncer de mama (baixa confiança na evidência).	Alguns membros do GDD relataram que algumas categorias profissionais podem não achar um programa de triagem aceitável, devido a seus interesses financeiros.
Viabilidade A intervenção é viável de ser implementada?		
DECISÃO	EVIDÊNCIAS DE PESQUISA	CONSIDERAÇÕES ADICIONAIS
○ Não ○ Provavelmente não ● Provavelmente sim ○ Sim ○ Varia ○ Não se sabe	Não foram identificadas evidências de pesquisa.	Não foi realizada uma revisão sistemática a respeito desse assunto. Alguns países não têm programas de triagem principalmente devido à falta de recursos e também de infraestrutura. Visto que essa recomendação se somaria à triagem em faixas etárias mais elevadas (50

		a 69 anos), decidiu-se que a recomendação é provavelmente viável de ser implementada.

RESUMO DAS DECISÕES

	Decisão						
PROBLEMA	Não	Provavelmente não	Provavelmente sim	Sim		Varia	Não se sabe
EFEITOS DESEJÁVEIS	Triviais	Pequenos	Moderados	Grandes		Variam	Não se sabe
EFEITOS INDESEJÁVEIS	Grandes	Moderados	Pequenos	Triviais		Variam	Não se sabe
CERTEZA DA EVIDÊNCIA	Muito baixa	Baixa	Moderada	Alta			Nenhum estudo incluído
VALORES	Incerteza ou variabilidade importante	Possivelmente há incerteza ou variabilidade importante	Provavelmente não há incerteza ou variabilidade importante	Nenhuma incerteza ou variabilidade importante			
SALDO DE EFEITOS	Favorece a comparação	Provavelmente favorece a comparação	Não favorece nem a intervenção nem a comparação	Provavelmente favorece a intervenção	Favorece a intervenção	Varia	Não se sabe
RECURSOS NECESSÁRIOS	Grandes custos	Custos moderados	Custos e economia insignificantes	Economia moderada	Grande economia	Variam	Não se sabe
CERTEZA DA EVIDÊNCIA DOS RECURSOS NECESSÁRIOS	Muito baixa	Baixa	Moderada	Alta			Nenhum estudo incluído

	Decisão						
CUSTO-EFETIVIDADE	Favorece a comparação	Provavelmente favorece a comparação	Não favorece nem a intervenção nem a comparação	Provavelmente favorece a intervenção	Favorece a intervenção	Varia	Nenhum estudo incluído
EQUIDADE	Diminuiria	Provavelmente diminuiria	Provavelmente sem impacto	Provavelmente aumentaria	Aumentaria	Varia	Não se sabe
ACEITABILIDADE	Não	Provavelmente não	Provavelmente sim	Sim		Varia	Não se sabe
VIABILIDADE	Não	Provavelmente não	Provavelmente sim	Sim		Varia	Não se sabe

TIPO DE RECOMENDAÇÃO

Recomendação forte contra a intervenção	Recomendação condicional contra a intervenção	Recomendação condicional para intervenção ou comparação	Recomendação condicional para a intervenção	Forte recomendação para a intervenção
○	○	○	●	○

CONCLUSÕES

Recomendação

Para mulheres assintomáticas entre 45 e 49 anos de idade com risco médio de câncer de mama, o Grupo de Desenvolvimento de Diretrizes (GDD) da ECIBC sugere a triagem mamográfica em vez de nenhuma triagem mamográfica, no contexto de um programa de triagem organizado (recomendação condicional, certeza da evidência moderada).

Justificativa

Justificativa geral

A recomendação condicional favorável à triagem mamográfica, no contexto de um programa de triagem organizado, foi resultado do saldo dos efeitos para a saúde, que provavelmente favorece a triagem mamográfica, apesar de haver apenas uma certeza moderada sobre as evidências a respeito desses efeitos. Os membros do GDD concordaram que essas mulheres teriam maiores benefícios previstos para a saúde (efeitos moderados) se comparados às mulheres de 40 a 44 anos, devido a uma maior incidência absoluta e mortalidade por câncer de mama em mulheres entre 45 e 49 anos do que em mulheres entre 40 e 44 anos, além de evidências observacionais demonstrando maior benefício nessa faixa etária (Hellquist 2011).

Não foi possível chegar a um acordo sobre a direção dessa recomendação entre os membros do GDD, e o resultado da votação dos membros sem CIs foi o seguinte: 17 membros votaram pela recomendação condicional favorável à intervenção; 1 membro votou pela recomendação condicional contra a intervenção; 4 membros se abstiveram.

Justificativa detalhada

Efeitos desejáveis

A mamografia, se comparada a nenhuma triagem, não reduziu significativamente o risco de mortalidade por câncer de mama (77 menos óbitos por câncer de mama por 100.000, com um intervalo de 7 mais a 147 menos óbitos, ou menos 44 óbitos por câncer de mama por 100.000, com um intervalo de 4 mais a 84 menos óbitos por câncer de mama, utilizando um risco basal de 0,7% e 0,4%, respectivamente) em mulheres convidadas para uma triagem ao longo de um seguimento de 16,4 anos (evidência de qualidade moderada). Entretanto, existem consideráveis evidências observacionais sobre os benefícios em mulheres entre 45 e 49 anos. A mamografia, se comparada a nenhuma triagem, reduziu o risco de câncer de mama estágio IIA ou superior (46 menos casos de câncer de mama por 100.000 mulheres durante um seguimento médio de 13,6 anos) (evidência de qualidade muito baixa), mas não reduziu o risco de mortalidade por todas as causas (evidência de baixa qualidade), mortalidade por outras causas (evidência de qualidade muito baixa) e de câncer de mama estágio III+ ou com tumor de tamanho ≥ 40 mm (evidência de baixa qualidade).

Efeitos indesejáveis

Mulheres de 40 a 74 anos de idade randomizadas para triagem tinham maior probabilidade de realizar mastectomia (180 mais mastectomias por 100.000 mulheres) (evidência de baixa qualidade). O percentual de sobrediagnóstico estimado é de 12,4% (evidência de qualidade moderada) com base em uma perspectiva populacional e 22,7% com base na perspectiva das mulheres convidadas para a triagem (evidência de qualidade moderada). O número de falso-positivos dependerá da idade na primeira triagem. O risco cumulativo estimado de um resultado falso-positivo na triagem nas mulheres entre 50 e 69 anos de idade que realizaram 10 exames de triagem bienais foi de 19,7%. Entretanto, observou-se que as taxas de falso-positivo eram mais elevadas em mulheres abaixo de 50 anos do que em mulheres entre 50 e 69 anos. Além disso, 2,2% foram submetidas a biópsia por agulha após a triagem mamográfica inicial. Mamografias Falso-positivas também estiveram associadas a maior ansiedade e sofrimento a respeito do câncer de mama, assim como consequências psicológicas negativas que podem perdurar por até três anos (evidência de baixa qualidade). A triagem mamográfica, se comparada a nenhuma triagem, não aumentou o número de mulheres entre 43 e 74 anos tratadas com quimioterapia (evidência de muito baixa qualidade). As mulheres que fizeram exames complementares após a mamografia de rotina apresentaram ansiedade significativa no curto prazo.

Certeza da evidência

A certeza geral (isto é, a qualidade) da evidência foi considerada moderada, visto que esta foi a qualidade mais baixa (correspondendo à qualidade da evidência para mortalidade por outras causas) entre os desfechos considerados críticos (mortalidade por câncer de mama e sobrediagnóstico).

Considerações de subgrupo

Esta recomendação não se aplica a mulheres do grupo de risco (ver recomendações para mulheres com alta densidade mamária).

Considerações para a implementação

Os membros do GDD concordaram que há necessidade de técnicas de imagem complementares nessa faixa etária, além da necessidade de uma tomada de decisão compartilhada. A implementação nessa faixa etária deve ser feita de forma a possibilitar uma quantificação mais aprofundada dos benefícios e dos malefícios.

Monitoramento e avaliação

Os futuros monitoramento e avaliação dos serviços de triagem devem considerar os riscos e os benefícios no contexto dos protocolos de tratamento e manejo em evolução.
Os critérios de monitoramento e avaliação estão sendo elaborados no âmbito da ECIBC.

Prioridades de pesquisa

1. Realizar avaliações da eficácia da intervenção, dos intervalos de tempo, dos fatores de risco e da estratificação das mulheres, assim como do custo-efetividade específico para o contexto dessa faixa etária.
2. Realizar estudos sobre o papel de outros métodos de triagem (por exemplo, RM) nessa população.

RESUMO DAS REFERÊNCIAS

1. Madan J, Rawdin A,Stevenson M,Tappenden P.. A Rapid Response Economic Evaluation of the UK NHS Cancer Reform Strategy Breast Cancer Screening Program Extension via a Plausible Bounds Approach. Value Health; 2010.

2. Sankatsing VD, Heijnsdijk EA,van Luijt PA,van Ravesteyn NT,Fracheb

Anexo III. Contribuições dos autores desta edição

Capítulos	Autores
1. Saúde baseada em evidências: sistemas para avaliação da certeza da evidência e para a graduação da força da recomendação	• Cinara Stein • Gilson Pires Dorneles • Maicon Falavigna • Suena Medeiros Parahiba • Verônica Colpani
2. Elaboração da questão de pesquisa e escolha dos desfechos	• Cinara Stein • Karlyse Claudino Belli • Maicon Falavigna • Suena Medeiros Parahiba • Verônica Colpani
3. Avaliação da certeza da evidência	• Cinara Stein • Cintia Pereira de Araujo • Débora Dalmas Gräf • Gilson Pires Dorneles • Karlyse Claudino Belli • Maicon Falavigna • Suena Medeiros Parahiba • Verônica Colpani
4. Síntese de evidências	• Cinara Stein • Maicon Falavigna • Suena Medeiros Parahiba • Verônica Colpani
5. Uso do GRADE para o desenvolvimento de recomendações	• Cinara Stein • Karlyse Claudino Belli • Maicon Falavigna • Suena Medeiros Parahiba • Verônica Colpani
6. Sistema GRADE para testes e estratégias diagnósticos	• Cinara Stein • Gilson Pires Dorneles • Karlyse Claudino Belli

Capítulos	Autores
	• Maicon Falavigna • Suena Medeiros Parahiba • Verônica Colpani
7. Sistema GRADE para prognóstico, incidência e prevalência	• Cinara Stein • Gilson Pires Dorneles • Maicon Falavigna • Suena Medeiros Parahiba • Verônica Colpani
8. Sistema GRADE para metanálises em rede	• Cinara Stein • Gilson Pires Dorneles • Karlyse Claudino Belli • Maicon Falavigna • Suena Medeiros Parahiba • Verônica Colpani
9. Sistema GRADE para modelagem	• Cinara Stein • Maicon Falavigna • Suena Medeiros Parahiba • Verônica Colpani
10. Sistema GRADE para incorporação de tecnologias	• Cinara Stein • Karlyse Claudino Belli • Maicon Falavigna • Suena Medeiros Parahiba • Verônica Colpani
11. Sistema GRADE em saúde pública	• Cinara Stein • Karlyse Claudino Belli • Maicon Falavigna • Suena Medeiros Parahiba • Verônica Colpani
12. Sistema GRADE para recomendações em situação de urgência e emergência	• Cinara Stein • Gilson Pires Dorneles • Maicon Falavigna

Capítulos	Autores
	• Suena Medeiros Parahiba • Verônica Colpani
13. Considerações finais	• Cinara Stein • Gilson Pires Dorneles • Maicon Falavigna • Suena Medeiros Parahiba • Verônica Colpani
Coordenadores	• Cinara Stein • Cintia Pereira de Araujo • Gilson Pires Dorneles • Maicon Falavigna • Suena Medeiros Parahiba • Verônica Colpani
Revisão Técnica	• Airton Stein: Universidade Federal de Ciências da Saúde de Porto Alegre • Celina Borges Migliava: Instituto de Avaliações de Tecnologias em Saúde • Maicon Falavigna: Hospital Moinhos de Vento

Anexo IV. Lista de autoria da primeira versão do Sistema GRADE: manual de graduação da qualidade da evidência e força de recomendação para tomada de decisão em saúde

Brasil. Ministério da Saúde. Secretaria de Ciência, Tecnologia e Insumos Estratégicos. Departamento de Ciência e Tecnologia. Diretrizes metodológicas: Sistema GRADE – Manual de graduação da qualidade da evidência e força de recomendação para tomada de decisão em saúde / Ministério da Saúde, Secretaria de Ciência, Tecnologia e Insumos Estratégicos, Departamento de Ciência e Tecnologia. – Brasília: Ministério da Saúde, 2014. 72 p. ISBN 978-85-334-2186-8

Supervisão Geral:

Carlos Augusto Gabrois Gadelha (SECTICS/MS)

Antônio Carlos Campos de Carvalho (Decit/SCTIE/MS)

Jorge Otávio Maia Barreto (Decit/SCTIE/MS)

Elaboração:

Maicon Falavigna (IATS/UFRGS)

Airton Tetelbom Stein (GHC; UFCSPA, ULBRA)

Sérgio Sirena (GHC)

Marisa Santos (INC/MS)

Revisão de Especialista:

Anna Buehler (HAOC)

***Organização*:**

Kathiaja Miranda Souza (Decit/SCTIE/MS)

Maria Augusta Rodrigues de Oliveira (GHC)

Roberta Moreira Wichmann (Decit/SCTIE/MS)

Revisão Técnica:

Betânia Ferreira Leite (Decit /SCTIE/MS)

Carlos de Andrade (INI/Fiocruz)

Evelinda Trindade (InCOR)

Ivan Ricardo Zimmermann (DGITS/SCTIE/MS)

Júlia Souza Vidal (Anvisa)

Marcus Tolentino Silva (UFAM)

Sônia Venâncio (IS/SES-SP)

Taís Galvão (UFAM)

Editoração:

Eliana Carlan (Decit/SCTIE/MS)

Jessica Alves Rippel (Decit/SCTIE/MS)

Design Gráfico:

Gustavo Veiga e Lins (Decit/SCTIE/MS)

Normalização:

Amanda Soares Moreira (CGDI/ Editora MS)

REFERÊNCIAS

1. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. The Lancet. 2009;374(9683):86-9.
2. Schunemann HJ, Wiercioch W, Etxeandia I, Falavigna M, Santesso N, Mustafa R, et al. Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. CMAJ. 2014;186(3):E123-42.
3. Franco JVA, Arancibia M, Meza N, Madrid E, Kopitowski K. Clinical practice guidelines: Concepts, limitations and challenges. Medwave. 2020;20(3):e7887.
4. Armstrong JJ, Goldfarb AM, Instrum RS, MacDermid JC. Improvement evident but still necessary in clinical practice guideline quality: a systematic review. J Clin Epidemiol. 2017;81:13-21.
5. Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011;64(4):401-6.
6. Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, et al. The GRADE Working Group clarifies the construct of certainty of evidence. J Clin Epidemiol. 2017;87:4-13.
7. The periodic health examination. Canadian Task Force on the Periodic Health Examination. Can Med Assoc J. 1979;121(9):1193-254.
8. Centre of Evidence-Based of Evidence (CEBM). OCEBM Levels of Evidence [Available from: https://www.cebm.ox.ac.uk/resources/levels-of-evidence/ocebm-levels-of-evidence.
9. Schunemann HJ, Best D, Vist G, Oxman AD, Group GW. Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. CMAJ. 2003;169(7):677-80.
10. Otto CM, Nishimura RA, Bonow RO, Carabello BA, Erwin JP, 3rd, Gentile F, et al. 2020 ACC/AHA Guideline for the Management of Patients With Valvular Heart Disease: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Circulation. 2021;143(5):e72-e227.
11. Scotland HI. Antithrombotics: indications and management
A national clinical guideline 1993-2013 [Available from: https://www.sign.ac.uk/media/1067/sign129.pdf.
12. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. BMJ. 2004;328(7454):1490.
13. Kavanagh BP. The GRADE system for rating clinical guidelines. PLoS Med. 2009;6(9):e1000094.
14. GRADE Handbook 2013 [Available from: https://gdt.gradepro.org/app/handbook/handbook.html#h.47t67glagox6.
15. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. Journal of Clinical Epidemiology. 2011;64(12):1311-6.
16. GRADE Handbook2013.
17. Brasil. Ministério da Saúde. Diretrizes Metodológicas: Sistema GRADE - manual de graduação da qualidade da evidência e força de recomendação para tomada de decisão em saúde. 2014 [Available from: https://bvsms.saude.gov.br/bvs/publicacoes/diretrizes_metodologicas_sistema_grade.pdf.

18. Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol. 2011;64(4):395-400.
19. Freedman B. Equipoise and the ethics of clinical research. N Engl J Med. 1987;317(3):141-5.
20. Jönsson L, Sandin R, Ekman M, Ramsberg J, Charbonneau C, Huang X, et al. Analyzing Overall Survival in Randomized Controlled Trials with Crossover and Implications for Economic Evaluation. Value in Health. 2014;17(6):707-13.
21. Witherspoon JW, Vasavada RP, Waite MR, Shelton M, Chrismer IC, Wakim PG, et al. 6-minute walk test as a measure of disease progression and fatigability in a cohort of individuals with RYR1-related myopathies. Orphanet J Rare Dis. 2018;13(1):105.
22. Administration USFD. Table of Surrogate Endpoints That Were the Basis of Drug Approval or Licensure [Available from: https://www.fda.gov/drugs/development-resources/table-surrogate-endpoints-were-basis-drug-approval-or-licensure.
23. Cuello-Garcia CA, Santesso N, Morgan RL, Verbeek J, Thayer K, Ansari MT, et al. GRADE guidance 24 optimizing the integration of randomized and non-randomized studies of interventions in evidence syntheses and health guidelines. Journal of Clinical Epidemiology. 2022;142:200-8.
24. Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919.
25. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses [Available from: https://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.
26. Joanna Briggs Institute. Checklist for Cohort Studies 2020 [Available from: https://jbi.global/critical-appraisal-tools.
27. Joanna Briggs Institute. Checklist for Case Control Studies 2020 [Available from: https://jbi.global/critical-appraisal-tools.
28. Joanna Briggs Institute. Checklist for Analytical Cross Sectional Studies 2020 [Available from: https://jbi.global/critical-appraisal-tools.
29. Joanna Briggs Institute. Checklist for Case Series 2020 [Available from: https://jbi.global/critical-appraisal-tools.
30. Joanna Briggs Institute. Checklist for Case Reports 2020 [Available from: https://jbi.global/critical-appraisal-tools.
31. Joanna Briggs Institute. Checklist for Prevalence Studies 2020 [Available from: https://jbi.global/critical-appraisal-tools.
32. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36.
33. Joanna Briggs Institute. Checklist for Diagnostic Test Accuracy Studies 2020 [Available from: https://jbi.global/critical-appraisal-tools.
34. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017;358:j4008.
35. Whiting P, Savovic J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol. 2016;69:225-34.
36. Joanna Briggs Institute. Checklist for Systematic Reviews 2020 [Available from: https://jbi.global/critical-appraisal-tools.

37. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). J Clin Epidemiol. 2011;64(4):407-15.
38. Group RD. Revised Cochrane risk-of-bias tool for randomized trials (RoB 2). 2019.
39. Sterne JAC, Higgins JPT, Elbers RG, Reeves BC, ROBINS-I Tdgf. Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I): detailed guidance 2016 [Updated 12 October 2016:[Available from: https://www.riskofbias.info/.
40. GRADE Handbook: Inconsistency of results. 2022 [Available from: https://gdt.gradepro.org/app/handbook/handbook.html#h.g2dqzi9je57e.
41. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. J Clin Epidemiol. 2011;64(12):1294-302.
42. Guyatt G, Zhao Y, Mayer M, Briel M, Mustafa R, Izcovich A, et al. GRADE guidance 36: updates to GRADE's approach to addressing inconsistency. Journal of Clinical Epidemiology. 2023;158:70-83.
43. Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. BMC Med Res Methodol. 2008;8:79.
44. Zeng L, Brignardello-Petersen R, Hultcrantz M, Siemieniuk RAC, Santesso N, Traversy G, et al. GRADE guidelines 32: GRADE offers guidance on choosing targets of GRADE certainty of evidence ratings. Journal of Clinical Epidemiology. 2021;137:163-75.
45. Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I 2 in assessing heterogeneity may mislead. BMC Medical Research Methodology. 2008;8(1):79.
46. Migliavaca CB, Stein C, Colpani V, Barker TH, Ziegelmann PK, Munn Z, et al. Meta-analysis of prevalence: I(2) statistic and how to deal with heterogeneity. Res Synth Methods. 2022;13(3):363-7.
47. Foroutan F, Guyatt G, Zuk V, Vandvik PO, Alba AC, Mustafa R, et al. GRADE Guidelines 28: Use of GRADE for the assessment of evidence about prognostic factors: rating certainty in identification of groups of patients with different absolute risks. J Clin Epidemiol. 2020;121:62-70.
48. Iorio A, Spencer FA, Falavigna M, Alba C, Lang E, Burnand B, et al. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. BMJ. 2015;350(mar16 7):h870-h.
49. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. Bmj. 2010;340:c117.
50. Schandelmaier S, Briel M, Varadhan R, Schmid CH, Devasenapathy N, Hayward RA, et al. Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses. Canadian Medical Association Journal. 2020;192(32):E901-E6.
51. Zhang Y, Coello PA, Guyatt GH, Yepes-Nuñez JJ, Akl EA, Hazlewood G, et al. GRADE guidelines: 20. Assessing the certainty of evidence in the importance of outcomes or values and preferences-inconsistency, imprecision, and other domains. J Clin Epidemiol. 2019;111:83-93.
52. Torrance GW. Preferences for health states: a review of measurement methods. Mead Johnson Symp Perinat Dev Med. 1982(20):37-45.
53. Sepucha K, Ozanne EM. How to define and measure concordance between patients' preferences and medical treatments: A systematic review of approaches and recommendations for standardization. Patient Educ Couns. 2010;78(1):12-23.

54. Ryan M, Scott DA, Reeves C, Bate A, van Teijlingen ER, Russell EM, et al. Eliciting public preferences for healthcare: a systematic review of techniques. Health Technol Assess. 2001;5(5):1-186.
55. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. J Clin Epidemiol. 2011;64(12):1303-10.
56. Song F, Xiong T, Parekh-Bhurke S, Loke YK, Sutton AJ, Eastwood AJ, et al. Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study. BMJ. 2011;343:d4909.
57. GRADE Handbook: Imprecision 2022 [Available from: https://gdt.gradepro.org/app/handbook/handbook.html#h.ygojbnr1bi5y.
58. Zeng L, Brignardello-Petersen R, Hultcrantz M, Mustafa RA, Murad MH, Iorio A, et al. GRADE Guidance 34: update on rating imprecision using a minimally contextualized approach. Journal of Clinical Epidemiology. 2022;150:216-24.
59. Schünemann HJ, Neumann I, Hultcrantz M, Brignardello-Petersen R, Zeng L, Murad MH, et al. GRADE guidance 35: update on rating imprecision for assessing contextualized certainty of evidence and making decisions. J Clin Epidemiol. 2022;150:225-42.
60. Xiao Y, Guyatt G, Zeng L, RW Jayne D, A Merkel P, AC Siemieniuk R, et al. Comparative efficacy and safety of alternative glucocorticoids regimens in patients with ANCA-associated vasculitis: a systematic review. BMJ Open. 2022;12(2):e050507.
61. Rochwerg B, Oczkowski SJ, Siemieniuk RAC, Agoritsas T, Belley-Cote E, D'Aragon F, et al. Corticosteroids in Sepsis: An Updated Systematic Review and Meta-Analysis. Critical Care Medicine. 2018;46(9):1411-20.
62. Cohen J. A power primer. Psychol Bull. 1992;112(1):155-9.
63. GRADE Handbook: Publication bias 2022 [Available from: https://gdt.gradepro.org/app/handbook/handbook.html#h.xivvyiu1pr3v.
64. Egger M, Smith GD. Bias in location and selection of studies. Bmj. 1998;316(7124):61-6.
65. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database Syst Rev. 2009;2009(1):Mr000006.
66. Dickersin K, Min YI, Meinert CL. Factors influencing publication of research results. Follow-up of applications submitted to two institutional review boards. Jama. 1992;267(3):374-8.
67. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. Bmj. 1997;315(7109):640-5.
68. Bardy AH. Bias in reporting clinical trials. Br J Clin Pharmacol. 1998;46(2):147-50.
69. Hopewell S, Clarke M, Stewart L, Tierney J. Time to publication for results of clinical trials. Cochrane Database Syst Rev. 2007;2007(2):Mr000011.
70. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. Bmj. 1997;315(7109):629-34.
71. Jüni P, Holenstein F, Sterne J, Bartlett C, Egger M. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. Int J Epidemiol. 2002;31(1):115-23.
72. Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions. Cochrane Database Syst Rev. 2007;2007(2):Mr000010.

73. Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. J Clin Epidemiol. 2011;64(12):1277-82.
74. Altman DG. Systematic reviews of evaluations of prognostic variables. Bmj. 2001;323(7306):224-8.
75. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. Lancet. 1991;337(8746):867-72.
76. Cappelleri JC, Ioannidis JP, Schmid CH, de Ferranti SD, Aubert M, Chalmers TC, et al. Large trials vs meta-analysis of smaller trials: how do their results compare? Jama. 1996;276(16):1332-8.
77. Melander H, Ahlqvist-Rastad J, Meijer G, Beermann B. Evidence b(i)ased medicine--selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. Bmj. 2003;326(7400):1171-3.
78. Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. Bmj. 2003;326(7400):1167-70.
79. Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. Bmj. 2006;333(7568):597-600.
80. Cochrane Handbook for Systematic Reviews of Interventions 2022 [Available from: https://training.cochrane.org/handbook/current.
81. Sutton AJ, Duval SJ, Tweedie RL, Abrams KR, Jones DR. Empirical assessment of effect of publication bias on meta-analyses. Bmj. 2000;320(7249):1574-7.
82. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. J Clin Epidemiol. 2008;61(10):991-6.
83. Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. J Clin Epidemiol. 2005;58(9):894-901.
84. Ryan M, Hill S. How to GRADE the quality of the evidence.
: Cochrane Consumers and Communication Group; 2016.
85. Bross ID. Pertinency of an extraneous variable. J Chronic Dis. 1967;20(7):487-95.
86. Weiner SJ. Contextualizing medical decisions to individualize care: lessons from the qualitative sciences. J Gen Intern Med. 2004;19(3):281-5.
87. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol. 2011;64(12):1311-6.
88. Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. BMJ. 2007;334(7589):349-51.
89. Cannegieter SC, Rosendaal FR, Briet E. Thromboembolic and bleeding complications in patients with mechanical heart valve prostheses. Circulation. 1994;89(2):635-41.
90. Baudet EM, Puel V, McBride JT, Grimaud JP, Roques F, Clerc F, et al. Long-term results of valve replacement with the St. Jude Medical prosthesis. J Thorac Cardiovasc Surg. 1995;109(5):858-70.
91. Singer DE, Albers GW, Dalen JE, Fang MC, Go AS, Halperin JL, et al. Antithrombotic therapy in atrial fibrillation: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines (8th Edition). Chest. 2008;133(6 Suppl):546S-92S.
92. Brozek JL, Akl EA, Alonso-Coello P, Lang D, Jaeschke R, Williams JW, et al. Grading quality of evidence and strength of recommendations in clinical practice

guidelines. Part 1 of 3. An overview of the GRADE approach and grading quality of evidence about interventions. Allergy. 2009;64(5):669-77.
93. Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Crit Care Med. 2006;34(6):1589-96.
94. Ekelund U, Steene-Johannessen J, Brown WJ, Fagerland MW, Owen N, Powell KE, et al. Does physical activity attenuate, or even eliminate, the detrimental association of sitting time with mortality? A harmonised meta-analysis of data from more than 1 million men and women. Lancet. 2016;388(10051):1302-10.
95. Murad MH, Verbeek J, Schwingshackl L, Filippini T, Vinceti M, Akl E, et al. GRADE guidance 38: Updated guidance for rating up certainty of evidence due to a dose-response gradient. J Clin Epidemiol. 2023.
96. Devereaux PJ, Choi PT, Lacchetti C, Weaver B, Schunemann HJ, Haines T, et al. A systematic review and meta-analysis of studies comparing mortality rates of private for-profit and private not-for-profit hospitals. CMAJ. 2002;166(11):1399-406.
97. Salpeter SR, Greyber E, Pasternak GA, Salpeter EE. Risk of fatal and nonfatal lactic acidosis with metformin use in type 2 diabetes mellitus. Cochrane Database Syst Rev. 2010;2010(4):CD002967.
98. Carrasco-Labra A, Brignardello-Petersen R, Santesso N, Neumann I, Mustafa RA, Mbuagbaw L, et al. Improving GRADE evidence tables part 1: a randomized trial shows improved understanding of content in summary of findings tables with a new format. J Clin Epidemiol. 2016;74:7-18.
99. Guyatt GH, Oxman AD, Santesso N, Helfand M, Vist G, Kunz R, et al. GRADE guidelines: 12. Preparing summary of findings tables-binary outcomes. J Clin Epidemiol. 2013;66(2):158-72.
100. Brasil MdS. PORTARIA CONJUNTA Nº 17 2020 [Aprova as Diretrizes Brasileiras para Diagnóstico e Tratamento da Insuficiência Cardíaca com Fração de Ejeção Reduzida]. Available from: https://www.gov.br/conitec/pt-br/midias/relatorios/portaria/2020/20210825_portaria-conjunta-17_diretrizes-brasileiras-icfer.pdf.
101. Langendam M, Carrasco-Labra A, Santesso N, Mustafa RA, Brignardello-Petersen R, Ventresca M, et al. Improving GRADE evidence tables part 2: a systematic survey of explanatory notes shows more guidance is needed. J Clin Epidemiol. 2016;74:19-27.
102. Santesso N, Carrasco-Labra A, Langendam M, Brignardello-Petersen R, Mustafa RA, Heus P, et al. Improving GRADE evidence tables part 3: detailed guidance for explanatory footnotes supports creating and understanding GRADE certainty in the evidence judgments. J Clin Epidemiol. 2016;74:28-39.
103. Mustafa RA, Santesso N, Brozek J, Akl EA, Walter SD, Norman G, et al. The GRADE approach is reproducible in assessing the quality of evidence of quantitative evidence syntheses. J Clin Epidemiol. 2013;66(7):736-42; quiz 42.e1-5.
104. Santesso N, Glenton C, Dahm P, Garner P, Akl EA, Alper B, et al. GRADE guidelines 26: informative statements to communicate the findings of systematic reviews of interventions. J Clin Epidemiol. 2020;119:126-35.
105. Brasil. Ministério da Saúde. DIRETRIZES METODOLÓGICAS - ELABORAÇÃO DE DIRETRIZES CLÍNICAS 2020 [Available from: https://www.gov.br/conitec/pt-br/midias/artigos_publicacoes/diretrizes/diretrizes-metodologicas-elaboracao-de-diretrizes-clinicas-2020.pdf.
106. GRADE Handbook. Perspective. 2020.

107. Andrews J, Guyatt G, Oxman AD, Alderson P, Dahm P, Falck-Ytter Y, et al. GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. J Clin Epidemiol. 2013;66(7):719-25.
108. Brasil. Ministério da Saúde. Secretaria de Ciência TeIE, ,. Diretrizes metodológicas: diretriz de avaliação econômica. Brasília: Ministério da Saúde; 2014.
109. Dans AM, Dans L, Oxman AD, Robinson V, Acuin J, Tugwell P, et al. Assessing equity in clinical practice guidelines. J Clin Epidemiol. 2007;60(6):540-6.
110. Oxman AD, Schunemann HJ, Fretheim A. Improving the use of research evidence in guideline development: 12. Incorporating considerations of equity. Health Res Policy Syst. 2006;4:24.
111. Alonso-Coello P, Oxman AD, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 2: Clinical practice guidelines. BMJ. 2016;353:i2089.
112. GRADE Handbook. Going from evidence to recommendations 2020 [Available from: https://gdt.gradepro.org/app/handbook/handbook.html#h.33qgws879zw.
113. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet. 2005;365(9453):82-93.
114. Akl EA, Maroun N, Guyatt G, Oxman AD, Alonso-Coello P, Vist GE, et al. Symbols were superior to numbers for presenting strength of recommendations to health care consumers: a randomized trial. J Clin Epidemiol. 2007;60(12):1298-305.
115. Schünemann HJ, Mustafa RA, Brozek J, Steingart KR, Leeflang M, Murad MH, et al. GRADE guidelines: 21 part 1. Study design, risk of bias, and indirectness in rating the certainty across a body of evidence for test accuracy. J Clin Epidemiol. 2020;122:129-41.
116. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. Cmaj. 1988;138(8):697-703.
117. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. Bmj. 2006;332(7549):1089-92.
118. Mustafa RA, Wiercioch W, Arevalo-Rodriguez I, Cheung A, Prediger B, Ivanova L, et al. Decision making about healthcare-related tests and diagnostic test strategies. Paper 4: International guidelines show variability in their approaches. J Clin Epidemiol. 2017;92:38-46.
119. Schünemann HJ, Mustafa RA. Decision making about healthcare-related tests and diagnostic test strategies. Paper 1: a new series on testing to improve people's health. J Clin Epidemiol. 2017;92:16-7.
120. Mulrow CD, Linn WD, Gaul MK, Pugh JA. Assessing quality of a diagnostic test evaluation. J Gen Intern Med. 1989;4(4):288-95.
121. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. Cmaj. 2006;174(4):469-76.
122. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. Jama. 1999;282(11):1061-6.
123. Schünemann HJ, Mustafa RA, Brozek J, Steingart KR, Leeflang M, Murad MH, et al. GRADE guidelines: 21 part 2. Test accuracy: inconsistency, imprecision, publication bias, and other domains for rating the certainty of evidence and presenting it in evidence profiles and summary of findings tables. J Clin Epidemiol. 2020;122:142-52.
124. Schünemann HJ, Mustafa RA, Brozek J, Santesso N, Bossuyt PM, Steingart KR, et al. GRADE guidelines: 22. The GRADE approach for tests and strategies-from test

accuracy to patient-important outcomes and recommendations. J Clin Epidemiol. 2019;111:69-82.
125. Schünemann HJ, Mustafa R, Brozek J, Santesso N, Alonso-Coello P, Guyatt G, et al. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. J Clin Epidemiol. 2016;76:89-98.
126. Borges Migliavaca C, Stein C, Colpani V, Barker TH, Munn Z, Falavigna M, et al. How are systematic reviews of prevalence conducted? A methodological study. BMC Med Res Methodol. 2020;20(1):96.
127. Iorio A, Spencer FA, Falavigna M, Alba C, Lang E, Burnand B, et al. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. BMJ. 2015;350:h870.
128. Spencer FA, Iorio A, You J, Murad MH, Schünemann HJ, Vandvik PO, et al. Uncertainties in baseline risk estimates and confidence in treatment effects. Bmj. 2012;345:e7401.
129. Paulo Roberto S, Celina Borges M, Uwe S, Daniela S, Marjan A. Prevalence of Mycoplasma genitalium infection among HIV PrEP users: a systematic review and meta-analysis. Sexually Transmitted Infections. 2023;99(5):351.
130. Righy C, Rosa RG, da Silva RTA, Kochhann R, Migliavaca CB, Robinson CC, et al. Prevalence of post-traumatic stress disorder symptoms in adult critical care survivors: a systematic review and meta-analysis. Critical Care. 2019;23(1):213.
131. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383-94.
132. Hayden JA, van der Windt DA, Cartwright JL, Cote P, Bombardier C. Assessing bias in studies of prognostic factors. Ann Intern Med. 2013;158(4):280-6.
133. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. 2019;170(1):51-8.
134. Migliavaca CB, Stein C, Colpani V, Munn Z, Falavigna M, Prevalence Estimates Reviews - Systematic Review Methodology G. Quality assessment of prevalence studies: a systematic review. J Clin Epidemiol. 2020;127:59-68.
135. Brignardello-Petersen R, Bonner A, Alexander PE, Siemieniuk RA, Furukawa TA, Rochwerg B, et al. Advances in the GRADE approach to rate the certainty in estimates from a network meta-analysis. J Clin Epidemiol. 2018;93:36-44.
136. Brignardello-Petersen R, Murad MH, Walter SD, McLeod S, Carrasco-Labra A, Rochwerg B, et al. GRADE approach to rate the certainty from a network meta-analysis: avoiding spurious judgments of imprecision in sparse networks. J Clin Epidemiol. 2019;105:60-7.
137. Brignardello-Petersen R, Florez ID, Izcovich A, Santesso N, Hazlewood G, Alhazanni W, et al. GRADE approach to drawing conclusions from a network meta-analysis using a minimally contextualised framework. BMJ. 2020;371:m3900.
138. Brignardello-Petersen R, Izcovich A, Rochwerg B, Florez ID, Hazlewood G, Alhazanni W, et al. GRADE approach to drawing conclusions from a network meta-analysis using a partially contextualised framework. Bmj. 2020;371:m3907.
139. Yepes-Nunez JJ, Li SA, Guyatt G, Jack SM, Brozek JL, Beyene J, et al. Development of the summary of findings table for network meta-analysis. J Clin Epidemiol. 2019;115:1-13.
140. Rouse B, Chaimani A, Li T. Network meta-analysis: an introduction for clinicians. Internal and Emergency Medicine. 2017;12(1):103-11.

141. Cheung R, Sullens CM, Seal D, Dickins J, Nicholson PW, Deshmukh AA, et al. The paradox of using a 7 day antibacterial course to treat urinary tract infections in the community. Br J Clin Pharmacol. 1988;26(4):391-8.
142. Martinot JB, Carr WD, Cullen S, Heredia Budo JL, Bauer K, MacLeod C, et al. A comparative study of clarithromycin modified release and amoxicillin/clavulanic acid in the treatment of acute exacerbation of chronic bronchitis. Adv Ther. 2001;18(1):1-11.
143. Taggart AJ, Johnston GD, McDevitt DG. Does the frequency of daily dosage influence compliance with digoxin therapy? Br J Clin Pharmacol. 1981;11(1):31-4.
144. Matsumura K, Arima H, Tominaga M, Ohtsubo T, Sasaguri T, Fujii K, et al. Does a combination pill of antihypertensive drugs improve medication adherence in Japanese? A randomized controlled trial. Circ J. 2012;76(6):1415-22.
145. Mooney ME, Sayre SL, Hokanson PS, Stotts AL, Schmitz JM. Adding MEMS feedback to behavioral smoking cessation therapy increases compliance with bupropion: a replication and extension study. Addict Behav. 2007;32(4):875-80.
146. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol. 1997;50(6):683-91.
147. Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med. 2002;21(16):2313-24.
148. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med. 2004;23(20):3105-24.
149. Chaimani A, Higgins JP, Mavridis D, Spyridonos P, Salanti G. Graphical tools for network meta-analysis in STATA. PLoS One. 2013;8(10):e76654.
150. Chung H, Lumley T. Graphical exploration of network meta-analysis data: the use of multidimensional scaling. Clin Trials. 2008;5(4):301-7.
151. Rucker G, Schwarzer G. Automated drawing of network plots in network meta-analysis. Res Synth Methods. 2016;7(1):94-107.
152. Caldwell DM. An overview of conducting systematic reviews with network meta-analysis. Syst Rev. 2014;3:109.
153. Tonin FS, Rotta I, Mendes AM, Pontarolo R. Network meta-analysis: a technique to gather evidence from direct and indirect comparisons. Pharm Pract (Granada). 2017;15(1):943.
154. Puhan MA, Schunemann HJ, Murad MH, Li T, Brignardello-Petersen R, Singh JA, et al. A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. BMJ. 2014;349:g5630.
155. Brasil. Ministério da Saúde. Secretaria de Ciência T, Inovação e Insumos Estratégicos em Saúde., Saúde. DdGeIdTeIe. Protocolo Clínico e Diretrizes Terapêuticas do Diabete Melito Tipo 2. Brasília: Ministério da Saúde; 2020.
156. Izcovich A, Chu DK, Mustafa RA, Guyatt G, Brignardello-Petersen R. A guide and pragmatic considerations for applying GRADE to network meta-analysis. Bmj. 2023;381:e074495.
157. Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. Stat Med. 2010;29(7-8):932-44.
158. Alhazzani W, Alshamsi F, Belley-Cote E, Heels-Ansdell D, Brignardello-Petersen R, Alquraini M, et al. Efficacy and safety of stress ulcer prophylaxis in critically ill patients: a network meta-analysis of randomized trials. Intensive Care Med. 2018;44(1):1-11.
159. Murad MH, Drake MT, Mullan RJ, Mauck KF, Stuart LM, Lane MA, et al. Clinical review. Comparative effectiveness of drug treatments to prevent fragility fractures: a

systematic review and network meta-analysis. J Clin Endocrinol Metab. 2012;97(6):1871-80.
160. Lu G, Ades AE. Assessing Evidence Inconsistency in Mixed Treatment Comparisons. J Am Stat Assoc. 2006;201(474):447-59.
161. Brignardello-Petersen R, Mustafa RA, Siemieniuk RAC, Murad MH, Agoritsas T, Izcovich A, et al. GRADE approach to rate the certainty from a network meta-analysis: addressing incoherence. J Clin Epidemiol. 2019;108:77-85.
162. Foote CJ, Guyatt GH, Vignesh KN, Mundi R, Chaudhry H, Heels-Ansdell D, et al. Which Surgical Treatment for Open Tibial Shaft Fractures Results in the Fewest Reoperations? A Network Meta-analysis. Clin Orthop Relat Res. 2015;473(7):2179-92.
163. Zeng L, Brignardello-Petersen R, Hultcrantz M, Siemieniuk RAC, Santesso N, Traversy G, et al. GRADE guidelines 32: GRADE offers guidance on choosing targets of GRADE certainty of evidence ratings. J Clin Epidemiol. 2021;137:163-75.
164. Brignardello-Petersen R, Guyatt GH, Mustafa RA, Chu DK, Hultcrantz M, Schunemann HJ, et al. GRADE guidelines 33: Addressing imprecision in a network meta-analysis. J Clin Epidemiol. 2021;139:49-56.
165. Rochwerg B, Alhazzani W, Sindi A, Heels-Ansdell D, Thabane L, Fox-Robichaud A, et al. Fluid resuscitation in sepsis: a systematic review and network meta-analysis. Ann Intern Med. 2014;161(5):347-55.
166. Florez ID, Veroniki AA, Al Khalifah R, Yepes-Nunez JJ, Sierra JM, Vernooij RWM, et al. Comparative effectiveness and safety of interventions for acute diarrhea and gastroenteritis in children: A systematic review and network meta-analysis. PLoS One. 2018;13(12):e0207701.
167. Dahabreh IJ, Chan JA, Earley A, Moorthy D, Avendano EE, Trikalinos TA, et al. AHRQ Methods for Effective Health Care. Modeling and Simulation in the Context of Health Technology Assessment: Review of Existing Guidance, Future Research Needs, and Validity Assessment. Rockville (MD): Agency for Healthcare Research and Quality (US); 2017.
168. Owens DK, Whitlock EP, Henderson J, Pignone MP, Krist AH, Bibbins-Domingo K, et al. Use of Decision Models in the Development of Evidence-Based Clinical Preventive Services Recommendations: Methods of the U.S. Preventive Services Task Force. Ann Intern Med. 2016;165(7):501-8.
169. Moberg J, Oxman AD, Rosenbaum S, Schunemann HJ, Guyatt G, Flottorp S, et al. The GRADE Evidence to Decision (EtD) framework for health system and public health decisions. Health Res Policy Syst. 2018;16(1):45.
170. Riva JJ, Bhatt M, Martins CC, Brunarski DJ, Busse JW, Xie F, et al. Indirectness (transferability) is critical when considering existing economic evaluations for GRADE clinical practice guidelines: a systematic review. Journal of Clinical Epidemiology. 2022;148:81-92.
171. Brozek JL, Canelo-Aybar C, Akl EA, Bowen JM, Bucher J, Chiu WA, et al. GRADE Guidelines 30: the GRADE approach to assessing the certainty of modeled evidence-An overview in the context of health decision-making. J Clin Epidemiol. 2021;129:138-50.
172. Philips Z, Bojke L, Sculpher M, Claxton K, Golder S. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. Pharmacoeconomics. 2006;24(4):355-71.
173. Heupink LF, Peacocke EF, Sæterdal I, Chola L, Frønsdal K. Considerations for transferability of health technology assessments: a scoping review of tools, methods, and practices. International Journal of Technology Assessment in Health Care. 2022;38(1).

174. Drummond M, Barbieri M, Cook J, Glick HA, Lis J, Malik F, et al. Transferability of economic evaluations across jurisdictions: ISPOR Good Research Practices Task Force report. Value Health. 2009;12(4):409-18.
175. Riva JJ, Bhatt M, Martins CC, Brunarski DJ, Busse JW, Xie F, et al. Indirectness (transferability) is critical when considering existing economic evaluations for GRADE clinical practice guidelines: a systematic review. J Clin Epidemiol. 2022;148:81-92.
176. Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. J Clin Epidemiol. 2011;64(12):1283-93.
177. Morgano GP, Wiercioch W, Anderson DR, Brozek JL, Santesso N, Xie F, et al. A modeling approach to derive baseline risk estimates for GRADE recommendations: Concepts, development, and results of its application to the American Society of Hematology 2019 guidelines on prevention of venous thromboembolism in surgical hospitalized patients. J Clin Epidemiol. 2021;140:69-78.
178. Riva JJ, Bhatt M, Brunarski DJ, Busse JW, Martins CC, Xie F, et al. Guidelines that use the GRADE approach often fail to provide complete economic information for recommendations: A systematic survey. J Clin Epidemiol. 2021;136:203-15.
179. World Health Organization. Health technology assessment of medical devices. Geneva: World Health Organization; 2011.
180. Hailey D, Babidge W, Cameron A, Davignon LA. HTA Agencies and Decision Makers: an INAHTA guidance document2010.
181. Guindo LA, Wagner M, Baltussen R, Rindress D, van Til J, Kind P, et al. From efficacy to equity: Literature review of decision criteria for resource allocation and healthcare decisionmaking. Cost Eff Resour Alloc. 2012;10(1):9.
182. Lavis JN, Permanand G, Oxman AD, Lewin S, Fretheim A. SUPPORT Tools for evidence-informed health Policymaking (STP) 13: Preparing and using policy briefs to support evidence-informed policymaking. Health Res Policy Syst. 2009;7(Suppl 1):S13.
183. Dahm P, Oxman AD, Djulbegovic B, Guyatt GH, Murad MH, Amato L, et al. Stakeholders apply the GRADE evidence-to-decision framework to facilitate coverage decisions. J Clin Epidemiol. 2017;86:129-39.
184. Parmelli E, Amato L, Oxman AD, Alonso-Coello P, Brunetti M, Moberg J, et al. GRADE EVIDENCE TO DECISION (EtD) FRAMEWORK FOR COVERAGE DECISIONS. Int J Technol Assess Health Care. 2017;33(2):176-82.
185. Levin L, Goeree R, Levine M, Krahn M, Easty T, Brown A, et al. Coverage with evidence development: the Ontario experience. Int J Technol Assess Health Care. 2011;27(2):159-68.
186. Hutton J, Trueman P, Henshall C. Coverage with evidence development: an examination of conceptual and policy issues. Int J Technol Assess Health Care. 2007;23(4):425-32.
187. World Health Organization (WHO). Making fair choices on the path to universal health coverage: final report of the WHO consultative group on equity and universal health coverage. Geneva: World Health Organization (WHO); 2014.
188. Drummond M. Twenty years of using economic evaluations for drug reimbursement decisions: what has been achieved? J Health Polit Policy Law. 2013;38(6):1081-102.
189. Colpani V, Kowalski SC, Stein AT, Buehler AM, Zanetti D, Côrtes G, et al. Clinical practice guidelines in Brazil - developing a national programme. Health Res Policy Syst. 2020;18(1):69.

190. Brasil MdS. O USO DE LIMIARES DE CUSTO-EFETIVIDADE NAS DECISÕES EM SAÚDE: RECOMENDAÇÕES DA COMISSÃO NACIONAL DE INCORPORAÇÃO DE TECNOLOGIAS NO SUS. Brasília: Ministério da Saúde; 2022.
191. Boon MH, Thomson H, Shaw B, Akl EA, Lhachimi SK, Lopez-Alcalde J, et al. Challenges in applying the GRADE approach in public health guidelines and systematic reviews: a concept article from the GRADE Public Health Group. J Clin Epidemiol. 2021;135:42-53.
192. Treweek S, Oxman AD, Alderson P, Bossuyt PM, Brandt L, Brozek J, et al. Developing and Evaluating Communication Strategies to Support Informed Decisions and Practice Based on Evidence (DECIDE): protocol and preliminary results. Implement Sci. 2013;8:6.
193. The SURE Collaboration. SURE Guides for Preparing and Using Evidence-Based Policy Briefs: 1. Getting started: The SURE Collaboration; 2011.
194. Juarez SP, Honkaniemi H, Dunlavy AC, Aldridge RW, Barreto ML, Katikireddi SV, et al. Effects of non-health-targeted policies on migrant health: a systematic review and meta-analysis. Lancet Glob Health. 2019;7(4):e420-e35.
195. National Institute for Health and Care Excellence (NICE). Physical activity and the environment 2018 [Available from: https://www.nice.org.uk/guidance/ng90.
196. Stone D. Policy Paradox: The Art of Political Decision Making. New York1997.
197. Lorenc T, Tyner EF, Petticrew M, Duffy S, Martineau FP, Phillips G, et al. Cultures of evidence across policy sectors: systematic review of qualitative evidence. European Journal of Public Health. 2014;24(6):1041-7.
198. Welch VA, Akl EA, Guyatt G, Pottie K, Eslava-Schmalbach J, Ansari MT, et al. GRADE equity guidelines 1: considering health equity in GRADE guideline development: introduction and rationale. J Clin Epidemiol. 2017;90:59-67.
199. Griffiths SM. The Sustainable Development Goals: an agenda for us all. Perspect Public Health. 2019;139(5):224-5.
200. Rehfuess EA, Stratil JM, Scheel IB, Portela A, Norris SL, Baltussen R. The WHO-INTEGRATE evidence to decision framework version 1.0: integrating WHO norms and values and a complexity perspective. BMJ Glob Health. 2019;4(Suppl 1):e000844.
201. Rose G. Strategy of prevention: lessons from cardiovascular disease. Br Med J (Clin Res Ed). 1981;282(6279):1847-51.
202. Webster J, Waqanivalu T, Arcand J, Trieu K, Cappuccio FP, Appel LJ, et al. Understanding the science that supports population-wide salt reduction programs. J Clin Hypertens (Greenwich). 2017;19(6):569-76.
203. Schunemann HJ, Cuello C, Akl EA, Mustafa RA, Meerpohl JJ, Thayer K, et al. GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. J Clin Epidemiol. 2019;111:105-14.
204. Andrews JC, Schunemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, et al. GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. J Clin Epidemiol. 2013;66(7):726-35.
205. Schunemann HJ, Hill SR, Kakad M, Bellamy R, Uyeki TM, Hayden FG, et al. WHO Rapid Advice Guidelines for pharmacological management of sporadic human infection with avian influenza A (H5N1) virus. Lancet Infect Dis. 2007;7(1):21-31.
206. Fischer AJ, Threlfall A, Meah S, Cookson R, Rutter H, Kelly MP. The appraisal of public health interventions: an overview. J Public Health (Oxf). 2013;35(4):488-94.
207. Schunemann HJ, Santesso N, Vist GE, Cuello C, Lotfi T, Flottorp S, et al. Using GRADE in situations of emergencies and urgencies: certainty in evidence and

recommendations matters during the COVID-19 pandemic, now more than ever and no matter what. J Clin Epidemiol. 2020;127:202-7.
208. Qaseem A, Forland F, Macbeth F, Ollenschlager G, Phillips S, van der Wees P, et al. Guidelines International Network: toward international standards for clinical practice guidelines. Ann Intern Med. 2012;156(7):525-31.
209. Thayer KA, Schunemann HJ. Using GRADE to respond to health questions with different levels of urgency. Environ Int. 2016;92-93:585-9.
210. (NCEH) NCfEH. Nuclear Event Response 2018 [Available from: https://www.cdc.gov/nceh/radiation/emergencies/nuclearresponse.htm.
211. Emergency PH. Contaminated Water in Flint 2020 [Available from: https://www.phe.gov/emergency/events/Flint/Pages/default.aspx.
212. Kowalski SC, Morgan RL, Falavigna M, Florez ID, Etxeandia-Ikobaltzeta I, Wiercioch W, et al. Development of rapid guidelines: 1. Systematic survey of current practices and methods. Health Res Policy Syst. 2018;16(1):61.
213. Morgan RL, Florez I, Falavigna M, Kowalski S, Akl EA, Thayer KA, et al. Development of rapid guidelines: 3. GIN-McMaster Guideline Development Checklist extension for rapid recommendations. Health Res Policy Syst. 2018;16(1):63.
214. Schunemann HJ, Wiercioch W, Brozek J, Etxeandia-Ikobaltzeta I, Mustafa RA, Manja V, et al. GRADE Evidence to Decision (EtD) frameworks for adoption, adaptation, and de novo development of trustworthy recommendations: GRADE-ADOLOPMENT. J Clin Epidemiol. 2017;81:101-10.
215. King VJ, Stevens A, Nussbaumer-Streit B, Kamel C, Garritty C. Paper 2: Performing rapid reviews. Syst Rev. 2022;11(1):151.
216. Schunemann HJ, Zhang Y, Oxman AD, Expert Evidence in Guidelines G. Distinguishing opinion from evidence in guidelines. BMJ. 2019;366:l4606.
217. Akl EA, Morgan RL, Rooney AA, Beverly B, Katikireddi SV, Agarwal A, et al. Developing trustworthy recommendations as part of an urgent response (1-2 weeks): a GRADE concept paper. J Clin Epidemiol. 2021;129:1-11.
218. Alonso-Coello P, Schünemann HJ, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. Bmj. 2016;353:i2016.