METODOLOGIA

DE AVALIAÇÃO FARMACOTERAPÊUTICA DE TECNOLOGIAS DE SAÚDE

REPÚBLICA PORTUGUESA
SAÚDE

SNS SERVIÇO NACIONAL DE SAÚDE

Infarmed
Autoridade Nacional do Medicamento e Produtos de Saúde, I.P.

Metodologia de avaliação farmacoterapêutica de tecnologias de saúde

Versão	Data da publicação
3.0	05 de agosto de 2022

Sugestão de citação: Vinhas J, Dias S, Gouveia AM, Correia A, Dias CV, Sousa D, Oliveira J, Perelman J, Azevedo L, Marques N, Saramago P, Faria R, Couto S, Torres S, (2021) Metodologia de avaliação farmacoterapêutica, Versão 3.0. Comissão de Avaliação de Tecnologias de Saúde, INFARMED - Autoridade Nacional do Medicamento e Produtos de Saúde, I.P., Lisboa

Disponível *online* em www.infarmed.pt

Autores

José Vinhas
Centro Hospitalar de Setúbal e Comissão de Avaliação das Tecnologias de Saúde (INFARMED, I.P.) (coordenação)

Sofia Dias
Universidade de York e Comissão de Avaliação das Tecnologias de Saúde (INFARMED, I.P.) (coordenação)

Antonio Melo Gouveia
Instituto Português de Oncologia de Lisboa Francisco Gentil e Comissão de Avaliação das Tecnologias de Saúde (INFARMED, I.P.)

Catarina Viegas Dias
Universidade Nova Lisboa e Comissão de Avaliação das Tecnologias de Saúde (INFARMED, I.P.)

Diana Sousa
Centro Hospitalar Universitário Lisboa Norte e Comissão de Avaliação das Tecnologias de Saúde (INFARMED, I.P.)

João Oliveira
Instituto Português de Oncologia de Lisboa Francisco Gentil e Comissão de Avaliação das Tecnologias de Saúde (INFARMED, I.P.)

Julian Perelman (CATS / ENSP)
Escola Nacional de Saúde Pública, Universidade Nova de Lisboa e Comissão de Avaliação das Tecnologias de Saúde (INFARMED, I.P.)

Luís Azevedo
Faculdade de Medicina da Universidade do Porto e Comissão de Avaliação das Tecnologias de Saúde (INFARMED, I.P.)

Nuno Marques
Centro Hospitalar e Universitário do Algarve e Comissão de Avaliação das Tecnologias de Saúde (INFARMED, I.P.)

Pedro Saramago
Universidade York e Comissão de Avaliação das Tecnologias de Saúde (INFARMED, I.P.)

Rita Faria
Universidade York e Comissão de Avaliação das Tecnologias de Saúde (INFARMED, I.P.)

Sofia Torres
Hospital Universitário de Antuérpia e Comissão de Avaliação das Tecnologias de Saúde (INFARMED, I.P.)

Alex Correia
Direção de Avaliação de Tecnologias de Saúde, INFARMED, I.P.

Sara Couto
Direção de Avaliação de Tecnologias de Saúde, INFARMED, I.P.

Revisão Institucional

Claudia Furtado, INFARMED, I.P.

Rui Santos Ivo, INFARMED, I.P.

António Faria Vaz, INFARMED, I.P.

Cláudia Belo Ferreira, INFARMED, I.P.

ÍNDICE GERAL

ÍNDICE DE FIGURAS E TABELAS

Índice de Figuras

Tabelas

LISTA DE ABREVIATURAS

AIC	Critério de Informação de Akaike/ Akaike's Information Criterion
AIM	Autorização de Introdução no Mercado
ATS	Avaliação de Tecnologias de Saúde
CATS	Comissão de Avaliação de Tecnologias de Saúde
CE-CATS	Comissão Executiva da Comissão de Avaliação de Tecnologias de Saúde
CHMP	Comité para Medicamentos de Uso Humano/ Committee for Medicinal Products for Human Use
CiNeMA	Confidence in Network Meta-Analysis
CSR	Relatório do Estudo Clínico/ Clinical Study Report
DATS	Direção de Avaliação das Tecnologias de Saúde
DCI	Denominação Comum Internacional
DIC	Critério de Informação de desviância/ Deviance Information Criterion
EMA	Agência Europeia do Medicamento/ European Medicines Agency
EUDRACT	European Union Drug Regulating Authorities Clinical Trials Database
EUnetHTA	Rede Europeia de Avaliação de Tecnologias de Saúde/ European Network for Health Technology Assessment
GAE	Grupo de Avaliação da Evidência
GRADE	Grading of Recommendations, Assessment, Development and Evaluation
HRQoL	Qualidade de vida relacionada com a saúde/ Health-Related Quality of Life,
IC	Intervalo de Confiança
INFARMED, I.P.	Autoridade Nacional do Medicamento e Produtos de Saúde, I.P.
MAIC	Comparações Indiretas Ajustadas por Correspondência/ Matching Adjusted Indirect Comparisons
ML-NMR	Método de Meta-Regressão em Rede Multiníveis/ Multi-Level Network Meta-Regression
NICE	National Institute for Health and Care Excellence
MOOSE	Meta-analysis of Observational Studies
NNT	Número Necessário a Tratar
PICO	Populção, Intervenção, Comparação, controle ou comparador, Medidas de resultado/ Patient, Intervention, Comparator and Outcome
PRISMA	Preferred Reporting Items for Systematic Reviews
RAFT	Relatório de Avaliação Farmacoterapêutico
RCM	Resumo das Características do Medicamento
RD	Diferença de riscos/ Risk Difference
SiNATS	Sistema Nacional de Avaliação de Tecnologias de Saúde
SNS	Serviço Nacional de Saúde
STC	Comparações de Tratamentos Simulados/ Simulated Treatment Comparisons
STE	Limiar do Efeito no Sub-Rogado/ Surrogate Threshold Effect
TAIM	Titular de Autorização de Introdução no Mercado
TOI	Tamanho Ótimo de Informação

LISTA DE DEFINIÇÕES

Biomarcador	Uma característica que é medida objetivamente e que é avaliada como um indicador de processos biológicos normais, de processos patogénicos, ou de respostas farmacológicas a uma intervenção terapêutica.
Efetividade	Refere-se ao (à medição do) efeito desejado/benéfico da intervenção em condições de prática clínica.
Eficácia	Refere-se ao (à medição do) efeito desejado/benéfico da intervenção em condições ideais (em contexto de ensaio clínico).
Medida de resultado subrogada	Um biomarcador que pretende substituir uma medida de resultado clínico.
Medida de resultado clínico	Uma característica ou variável que reflete os sintomas, a capacidade funcional ou a expectativa de vida do doente.
Medidas de efeito absolutas	Medidas que expressam o resultado (medida de resultado) através da diferença (subtração) entre os riscos observados nos dois grupos. São exemplos de medidas de efeito a diferença de riscos (risk difference [RD]) e o número necessário a tratar (NNT). As medidas de efeito absolutas traduzem o risco basal de uma população, ao contrário das medidas relativas, e são clinicamente mais úteis em decisões terapêuticas.
Medidas de efeito relativas	Medidas que expressam o resultado (medida de resultado) num grupo relativamente a outro, geralmente sob a forma de uma divisão/rácio. São exemplos de medidas de efeito relativas o risco relativo (relative risk ou risk ratio [RR]), a razão de chances (odds ratio [OR]) ou a razão de riscos (hazard ratio [HR]).
Medidas de segurança	São os resultados relevantes para a segurança do utente. Podem ser globais (ex: eventos adversos graves, eventos adversos globais) ou específicos, no caso dos eventos adversos de especial interesse, que são eventos que podem estar potencialmente relacionados com a doença em estudo (ex: incidência de neoplasias, infeções genito-urinárias).
Prática clínica	Padrões de prática dos profissionais de saúde em Portugal, aferidos através das fontes disponíveis.
Rede ligada (conexa)	Rede em que é possível estabelecer um caminho (usando as arestas) de qualquer intervenção (vértice) para qualquer outra.
Rede em estrela	Rede de intervenções só ligadas por um comparador comum.
Variáveis binárias	Variáveis que têm apenas dois valores possíveis (por exemplo: morte).
Variáveis categóricas	Variáveis que contêm um número finito e geralmente fixo de valores (superior a 2), que correspondem a categorias ou grupos distintos. Os dados categóricos podem não ter uma ordem lógica. (por exemplo: categorias de índice de massa corporal [IMC]).
Variáveis contínuas	Variáveis numéricas que podem ter um número infinito de valores entre dois valores quaisquer (por exemplo: estatura)

1 INTRODUÇÃO

1.1 Sobre este documento

A metodologia apresentada neste documento destina-se a todos os interessados no processo de avaliação de tecnologias de saúde em Portugal e, nomeadamente, aos avaliadores da Comissão de Avaliação das Tecnologias de Saúde (CATS), requerentes (titulares de AIM), decisores, associações de doentes, profissionais de saúde, investigadores e demais partes interessadas.

1.2 Objetivo do documento

O objetivo deste documento é orientar a avaliação e reavaliação farmacoterapêutica de medicamentos e outras tecnologias de saúde realizado pela CATS, de modo a clarificar os desafios metodológicos encontrados, descrevendo também o processo de avaliação. Introduz-se, assim, maior consistência e transparência no processo e metodologia utilizada.

1.3 Grupo de Trabalho

De acordo com o estatuto do INFARMED, I.P., a Comissão de Avaliação de Tecnologias de Saúde (CATS), que apoia a Direção de Avaliação de Tecnologias da Saúde (DATS), criada pelo Decreto-Lei n.º 97/2015, de 1 de junho, na sua atual redação, em exercício de funções desde junho de 2016, tem competências para genericamente emitir pareceres em matérias relacionadas com a avaliação e reavaliação de tecnologias de saúde, no âmbito do seu financiamento e propor medidas adequadas aos interesses da saúde pública e do SNS relativamente a tecnologias de saúde, no âmbito do SiNATS.

Decorrente da experiência adquirida, o Doutor José Vinhas, na sua qualidade de presidente da Comissão Executiva da CATS (CE-CATS), veio propor um grupo de trabalho com o propósito de rever a metodologia para a avaliação farmacoterapêutica de medicamentos, o qual veio a coordenador. A proposta do grupo de trabalho incluiu 12 membros que integravam a CATS aquando do início deste processo de revisão.

A opção pelos peritos da CATS é explicada essencialmente por duas razões: 1) experiência na avaliação farmacoterapêutica dos diversos processos de medicamentos submetidos para efeitos de financiamento, tendo assim enfrentado na prática as dificuldades que exigiam uma orientação adequada; 2) enquanto intervenientes na avaliação, estão cientes da necessidade de harmonização dos processos, de modo a assegurar a consistência do processo de avaliação.

O processo de revisão contou também com a participação e coordenação da Professora Sofia Dias, membro da CATS, professora na Universidade de York, e parte da equipa que elabora relatórios de avaliação de tecnologias de saúde para o National Institute for Health and Care Excellence (NICE), académica internacionalmente reconhecida com larga experiência na síntese de evidência de tecnologias de saúde.

1.4 Metodologia do processo de revisão

O processo de revisão iniciou-se com a identificação dos temas a incluir na nova versão da metodologia farmacoterapêutica com vista à atualização da versão publicada em novembro de 2016, de acordo com os avanços mais recentes nesta área. A lista de temas circulou e foi discutida entre os autores, até à obtenção de uma lista consensual. Em seguida, cada tema foi atribuído a um grupo de pelo menos

duas pessoas e dois temas foram revistos individualmente por dois peritos, em função dos seus interesses e especializações individuais.

Foi pedido a cada grupo para elaborar uma breve revisão da literatura e uma lista de opções para eventuais alterações dentro de cada tema.

Após esta fase preparatória, em janeiro de 2020 teve lugar, em Lisboa, uma reunião de dois dias com todos os autores. Cada grupo apresentou resumidamente a sua revisão bibliográfica e os seus argumentos sobre o modo como a metodologia deveria ser ou não revista, seguindo-se um debate até ser atingido um consenso sobre o conteúdo de cada ponto. Nesta reunião estiveram também presentes os vice-presidentes da CE-CATS.

Após esta a reunião, os autores elaboraram uma versão preliminar da revisão da metodologia. A versão preliminar foi discutida pela equipa coordenadora e remetida aos autores, que reviram as suas versões à luz das sugestões e comentários entretanto recebidos.

Em agosto de 2020, o documento apresentando a proposta de revisão da metodologia de avaliação farmacoterapêutica foi objeto de consulta alargada às entidades interessadas. Para o efeito, foi contactado um grupo de entidades e individualidades, para se pronunciarem por escrito sobre a nova proposta das orientações metodológicas. Foram recebidos comentários escritos das seguintes entidades e individualidades: Associação Nacional de Farmácias (ANF), Associação Portuguesa da Indústria Farmacêutica (APIFARMA), Associação dos Profissionais de Registos e Regulamentação Farmacêutica (APREFAR), ARS Norte, Associação Portuguesa de Bioindústria (P-BIO), Defesa do Consumidor (DECO), Ordem dos Enfermeiros, Ordem dos Médicos, Registo Oncológico Nacional (RON).

Após receção destes comentários, foi organizada em abril de 2021 uma reunião de discussão entre o grupo de trabalho da CATS para discussão dos comentários recebidos. A nova versão da metodologia farmacoterapêutica foi revista, quando tal foi identificado como necessário, após discussão entre a equipa de autores, tendo a mesma sido finalizada pelo grupo de trabalho da CATS em maio de 2021.

A equipa de autores elaborou resposta aos comentários recebidos, tendo a mesma sido remetida pelo INFARMED, I.P. a cada uma das entidades em momento anterior ao da publicação da nova versão da metodologia farmacoterapêutica.

1.5 Enquadramento

Autorização de Introdução no Mercado

A comercialização de medicamentos no território nacional está sujeita a uma Autorização de Introdução no Mercado (AIM). De acordo com o n.º 2 do art.º 14.º do Decreto-Lei n.º 176/2006, de 30 de agosto, na sua redação atual, a decisão de AIM para um medicamento deve assentar exclusivamente em critérios científicos objetivos, de qualidade, segurança e eficácia terapêuticas do medicamento em questão, independentemente de quaisquer considerações de caráter económico. Para este efeito, além da avaliação da qualidade do medicamento, é efetuada uma avaliação da relação benefício-risco, ou seja, a avaliação dos efeitos terapêuticos positivos de um medicamento face aos seus próprios riscos no que toca à saúde dos doentes ou à saúde pública. Esta avaliação é efetuada por uma Autoridade Nacional Competente (por exemplo: INFARMED, I.P.) ou pela Agência Europeia do Medicamento (EMA), consoante o procedimento de avaliação aplicável, o qual depende do tipo de medicamento/ área terapêutica. Esta autorização é o único requisito para a comercialização do medicamento na jurisdição em que é válida.

Financiamento

No seguimento desta AIM, as decisões sobre o financiamento e o preço de um medicamento podem ser realizadas a nível nacional, regional ou local em cada Estado-Membro da União Europeia. Portugal dispõe de um Serviço Nacional de Saúde (SNS) que financia tecnologias de saúde em parte ou na sua totalidade. No caso dos medicamentos, apenas poderão ser financiados aqueles que obtiverem a respetiva AIM. Para apoiar a decisão de financiamento é efetuada uma Avaliação de Tecnologias de Saúde (ATS). Em Portugal esta avaliação também é realizada pelo INFARMED, I.P. enquanto Agência de ATS (organicamente através da DATS e da CATS), independentemente do organismo que avaliou a AIM.

Diferença entre Autorização de Introdução no Mercado e Financiamento

Ainda que sumariamente, e porque (ainda) é motivo de frequente confusão, parece importante fazer notar neste documento as diferenças na avaliação de medicamentos e outras tecnologias de saúde (daqui em diante designados por tecnologias de saúde) entre as Agências reguladoras e as Agências de ATS as quais resultam da existência de objetivos, perspetivas e metodologias de avaliação diferentes. Enquanto os reguladores avaliam a qualidade, a eficácia e a segurança das tecnologias de saúde, na perspetiva de existir uma relação positiva entre os efeitos terapêuticos dessa tecnologia de saúde e os respetivos riscos, na indicação terapêutica em avaliação, as Agências de ATS efetuam uma recomendação quanto ao financiamento do medicamento ou outra tecnologia de saúde, tendo em conta, nomeadamente a existência de outras alternativas terapêuticas já financiadas e em utilização na prática clínica, através de uma análise comparativa da eficácia e segurança do novo medicamento ou outra tecnologia de saúde face às alternativas habitualmente utilizadas na prática clínica nacional.

Assim, é essencial por parte dos requerentes um planeamento e desenvolvimento antecipado da evidência necessária para fornecer a informação adequada. Esta informação deverá permitir responder quer às questões das Agências Reguladoras quer às questões das Agências de ATS, que prosseguem objetivos distintos e avaliam diferentes perspetivas das tecnologias de saúde através de metodologias de avaliação também diferentes. A existência de lacunas na quantidade e qualidade da evidência clínica disponível trazem desafios adicionais às avaliações destas Agências e levam à tomada de decisões com maior incerteza, podendo constituir um obstáculo ao acesso das tecnologias de saúde ou outras tecnologias de saúde por parte das pessoas que deles necessitam.

A Avaliação de Tecnologias de Saúde (ATS)

Conforme acima referido, a ATS tem como objetivo apoiar a decisão de utilização e financiamento (comparticipação e/ou avaliação prévia) das tecnologias de saúde no SNS. Esta decisão baseia-se não só nos critérios de qualidade, segurança e eficácia exigidos a todos os medicamentos, mas também em critérios de eficácia e segurança comparativas, de forma a otimizar a utilização dos recursos disponíveis.

Genericamente, em Portugal, o processo de financiamento público pode ser dividido nas seguintes fases: instrução do pedido, avaliação farmacoterapêutica, avaliação farmacoeconómica, negociação e decisão. Na fase farmacoterapêutica da ATS, à qual este documento se refere, pretende-se assegurar que não é recomendado o financiamento pelo SNS de tecnologias de saúde que não sejam úteis e/ou necessárias.

A ATS tem uma longa tradição em Portugal e em 2015 foi alvo de uma atualização com a criação do Sistema Nacional de Avaliação de Tecnologias de Saúde (SiNATS), através do Decreto-Lei n.º 97/2015, de 1 de junho. O SiNATS é constituído pelo conjunto de entidades e meios que procedem à ATS, incumbindo a sua gestão ao INFARMED, I. P. Este enquadramento legal estatuiu a criação da CATS, uma comissão consultiva do INFARMED, I. P. de apoio ao SiNATS, nos termos e condições previstas

no artigo 8.º do Decreto-Lei n.º 46/2012, de 24 de fevereiro, alterado pelo Decreto-Lei n.º 97/2015, de 1 de junho, na sua redação atual.

1.6 Âmbito de aplicação da metodologia

Conforme previsto nos n.º 2 e 4 do art.º 7º da Portaria n.º 195-A/2015, de 30 de junho, na sua redação atual, a avaliação farmacoterapêutica de tecnologias de saúde é objeto de parecer/ deliberação da CATS no caso de se tratar de um medicamento cuja denominação comum internacional (DCI) ou indicação terapêutica ainda não esteja comparticipada ou com autorização de utilização nas instituições e serviços tutelados pelo membro do Governo responsável pela área da saúde e/ou sempre que solicitado pelos serviços competentes do INFARMED, I.P.. Este documento apresenta a metodologia de avaliação genericamente utilizada pela CATS para emissão destas recomendações farmacoterapêuticas.

Importa referir que a metodologia aqui prevista pode sofrer adaptações para permitir a avaliação de tecnologias de saúde/ áreas terapêuticas específicas, podendo também não ser justificada a sua aplicação a medicamentos já com uso clínico bem estabelecido a nível nacional ou produtos-fronteira (por exemplo: dispositivos médicos medicamentados). Algumas destas situações encontram-se previstas na secção 5. Adicionalmente, encontram-se, por princípio, excluídos desta metodologia, as associações de dose fixa para substituição de componentes isolados já financiados nas mesmas doses e respetivas indicações terapêuticas, as vacinas profiláticas, e os medicamentos derivados do plasma humano (ou as suas versões recombinantes ou modificadas).

Este documento deve ser lido em conjunto com o enquadramento legal da Avaliação de Tecnologias de Saúde em Portugal e outros documentos normativos sobre esta temática.

1.7 Principais alterações da nova versão

A atual versão da Metodologia de Avaliação Farmacoterapêutica (versão 3.0), sofreu uma profunda revisão em relação à versão 2.0 de 23 novembro de 2016, tendo sido alterada a estrutura do documento, e introduzidas novas secções, enquanto outras secções sofreram profundas alterações que implicaram, em geral, um novo conteúdo e maior desenvolvimento e detalhe.

No capítulo 2 (Operacionalização da avaliação), foi criada uma nova secção (2.2. Definição da matriz de avaliação), que inclui os critérios para seleção dos comparadores, onde é desenvolvido em detalhe os motivos que levaram à revisão desses critérios.

No capítulo 3 (Metodologia geral), foi criada uma nova secção (3.3. Medidas de resultado) onde se detalham os diferentes tipos de medidas de resultado, e se descrevem os requisitos para validação de uma medida de resultado sub-rogada (ponto 3.3.3.3.). Foi adicionada uma nova secção (secção 3.4. Revisões sistemáticas) que faz recomendações sobre como conduzir o processo de revisão sistemática da literatura e se salienta a sua importância para o processo de avaliação.

Foi criado um novo capítulo (4. Métodos de comparação), onde se descreve em detalhe os métodos de comparação, incluindo a meta-análise convencional e meta-análise em rede, e métodos de comparação indireta ajustados (MAIC, STC) para utilização no contexto de doenças raras e ultra-raras (Capítulo 5).

O capítulo 6 (Análise de subgrupos) foi expandido, incluindo agora um maior detalhe, nomeadamente quanto aos princípios e critérios a verificar para a especificação de subgrupos no âmbito da matriz inicial de avaliação.

O capítulo 10 (Valor terapêutico acrescentado) foi totalmente reescrito, sendo a nova versão mais detalhada em relação ao processo de reconhecimento de valor terapêutico acrescentado com o objetivo de clarificar a recomendação da CATS relativa à síntese de resultados.

2 OPERACIONALIZAÇÃO DA AVALIAÇÃO

2.1 Introdução

A avaliação de tecnologias de saúde inicia-se pela definição da matriz de avaliação, ou seja, pela definição do PICO. O PICO é um acrónimo usado na prática baseada em evidência (e especificamente na Medicina Baseada em Evidência) para estruturar e responder a uma pergunta clínica ou de cuidados médicos. A estrutura do PICO também é usada para desenvolver estratégias de busca na literatura, por exemplo, em revisões sistemáticas. O acrónimo PICO significa: P – População; I – Intervenção; C - Comparação, controle ou comparador; O – Medidas de resultado. É esta matriz que vai definir os termos em que a avaliação se vai realizar.

2.2 Definição da matriz de avaliação

Definição da(s) população(ões)

A população deve ser definida tendo em conta as características clínicas da(s) população(ões) incluídas na indicação terapêutica em avaliação. No caso da indicação em avaliação incluir diferentes populações que se distinguem pela presença de modificadores de efeito, ou por habitualmente receberem tratamentos diferenciados, deve ser ponderada a possibilidade de dividir a população incluída na indicação aprovada em duas ou mais populações, e de avaliar o efeito do tratamento separadamente para cada uma dessas populações (ver também secção 6 Análise de subgrupos).

Intervenção

A intervenção deve incluir apenas a intervenção em avaliação, não devendo ser incluídos outros fármacos que não façam parte da indicação de interesse. No entanto, se a tecnologia em avaliação deve ser utilizada em combinação com outras tecnologias, estas devem ser parte da definição da intervenção.

Seleção de comparadores

Os comparadores são todas as alternativas terapêuticas habitualmente utilizadas na prática clínica em Portugal para tratar a indicação para a qual o medicamento em avaliação tem autorização de introdução no mercado (AIM). Os comparadores são opções contra as quais o medicamento novo é comparado com o objetivo de aferir se o medicamento novo apresenta benefício adicional e é custo-efetivo.

A seleção de uma dada intervenção para comparador não se traduz num julgamento sobre a sua eficácia, sendo o único critério de inclusão o seu uso habitual na prática clínica em Portugal. Deste modo, os comparadores relevantes não deverão ser constrangidos pelos comparadores utilizados para o grupo de controlo nos ensaios clínicos sobre o medicamento em avaliação.

Os comparadores podem incluir:

- medicamentos com AIM para a indicação;
- opções terapêuticas inativas (p.e.: melhores cuidados de suporte, monitorização), se habitualmente utilizados na prática clínica portuguesa para a indicação;
- opções ativas não farmacoterapêuticas (p.e.: cirurgia), se utilizados na prática clínica portuguesa para a indicação;

- sequências terapêuticas em que o medicamento sob avaliação é utilizado em segunda linha ou outra linha subsequente, se aplicável à indicação e se permitido na sua AIM e, se relevante, permitido na AIM dos comparadores.

Em casos excecionais, podem ser incluídos como comparadores medicamentos sem AIM para a indicação, desde que tenham o seu uso bem estabelecido na prática clínica portuguesa para a indicação.

Na fase de definição da matriz de avaliação, deverão ser identificados todos os comparadores relevantes para a indicação.

Justificação para a identificação de todos os comparadores relevantes

A avaliação farmacoterapêutica tem como objetivo determinar a eficácia, segurança e o valor terapêutico acrescentado do medicamento sob avaliação em relação a cada um dos seus comparadores. Sendo a avaliação farmacoterapêutica comparativa, os resultados e as conclusões necessariamente dependem da eficácia e segurança de todos os comparadores relevantes porque os benefícios adicionais do medicamento sob avaliação dependem da evidência sobre várias medidas de eficácia e segurança, nomeadamente da magnitude das diferenças em relação aos comparadores e da incerteza sobre estas diferenças.

Justificação para a inclusão de comparadores que sejam medicamentos sem AIM para a indicação, mas que são comummente utilizados na prática clínica portuguesa para a indicação

Há situações clínicas específicas em que a prática clínica inclui a utilização de um medicamento sem AIM para uma indicação. De modo a que a avaliação terapêutica reflita os comparadores em prática clínica portuguesa, é necessário incluir estes medicamentos como comparadores. Se estes medicamentos não forem incluídos, as avaliações terapêutica e de custo-efetividade poderão concluir que o valor acrescentado do medicamento é superior do que é na realidade.

Justificação para a inclusão de opções terapêuticas inativas

Opções terapêuticas inativas, tais como melhores cuidados de suporte ou monitorização, são relevantes para indicações em que a prática clínica portuguesa as inclui como opções terapêuticas. Mesmo que existam medicamentos com AIM para uma indicação em que melhores cuidados de suporte ou monitorização sejam opções, é necessário incluir estas na avaliação do medicamento. A exclusão destas opções pode levar a que os benefícios adicionais do medicamento sejam sobrestimados ou os seus custos subestimados porque estas opções poderão oferecer vantagens em termos de efeitos adversos ou nos custos. Como tal, a sua exclusão poderá ter consequências nefastas para a conclusão sobre o seu valor acrescentado e para o preço a que é comparticipado pelo SNS.

Justificação para a inclusão de opções ativas não farmacoterapêuticas

Opções ativas não farmacoterapêuticas são tratamentos que não envolvem medicamentos, tais como cirurgia, fisioterapia, aconselhamento psicológico, etc. Tal como discutido acima, a exclusão destas opções, quando estas fazem parte do arsenal terapêutico na prática clínica portuguesa, poderá levar a que os benefícios adicionais do medicamento sejam sobrestimados ou os seus custos subestimados, com consequências nefastas para a decisão de comparticipação.

Justificação para a consideração de sequências terapêuticas

As sequências terapêuticas são relevantes quando os doentes podem ser tratados com uma das opções terapêuticas pré-existentes em primeira linha, estando o medicamento novo reservado se o tratamento de primeira linha não for efetivo. Nos casos em que a eficácia do medicamento novo em segunda linha é elevada, e as opções terapêuticas pré-existentes têm alguma eficácia, elevada segurança e a duração do tratamento é curta, a sequência terapêutica poderá ter eficácia semelhante ao medicamento novo. Sequências terapêuticas são relevantes quando a AIM não restringe o medicamento a uma determinada linha terapêutica.

Definição das medidas de eficácia e segurança e classificação da sua importância

Deve ser proposto um conjunto de medidas de resultado, relacionadas com a eficácia e segurança da intervenção. Estas medidas deverão permitir uma perspetiva compreensiva do efeito do tratamento e deverão incluir medidas relevantes para o doente. Com este objetivo, considera-se como relevantes para o doente as medidas que avaliam os sintomas, a capacidade funcional ou a expectativa de vida do doente, ou seja, mortalidade, morbilidade (sintomas e complicações), duração da doença, e qualidade de vida relacionada com a saúde (*health-related quality of life*, HRQoL).

As medidas de eficácia terapêutica e de segurança devem ser classificadas, segundo o grau de importância que lhe é atribuído, em críticas e importantes, mas não críticas, de acordo com a metodologia de avaliação da CATS. As medidas de eficácia terapêutica e de segurança devem ser consideradas críticas quando, na perspetiva do avaliador, podem influenciar o sentido da avaliação. Como regra geral, deverão ser consideradas críticas as medidas que avaliam os sintomas, a capacidade funcional ou a expectativa de vida do doente, ou seja, mortalidade, morbilidade (sintomas e complicações), duração da doença, e qualidade de vida relacionada com a saúde.

A classificação é feita numa escala de um a nove: as medidas classificadas como importantes deverão ser quantificadas com uma pontuação entre quatro e seis e as medidas classificadas como críticas entre sete e nove. A pontuação final atribuída às medidas de resultado deve ser a média das pontuações atribuídas por cada um dos elementos do Grupo de Avaliação durante a reunião de discussão da matriz de avaliação. A pontuação é arredondada para a unidade, desprezando-se as casas decimais e, no caso de o número a seguir à vírgula ser cinco ou superior, aumenta-se uma unidade a esse número (1).

2.3 Conclusões

A avaliação de tecnologias de saúde inicia-se pela definição da estrutura do PICO.

Devem ser selecionados para comparadores todas as alternativas terapêuticas habitualmente utilizadas na prática clínica em Portugal para tratar a indicação para a qual o medicamento em avaliação tem autorização de introdução no mercado e o requerente solicitou o seu financiamento. A seleção de uma dada intervenção para comparador não se traduz num julgamento sobre a sua eficácia, sendo o único critério de inclusão o seu uso habitual na prática clínica em Portugal.

Deve ser proposto um conjunto de medidas de resultado, relacionadas com a eficácia e segurança da intervenção. Estas medidas deverão permitir uma perspetiva compreensiva do efeito do tratamento e deverão incluir medidas relevantes para o doente. Com este objetivo, considera-se como relevantes para o doente as medidas que avaliam os sintomas, a capacidade funcional ou a expectativa de vida do doente, ou seja, mortalidade, morbilidade (sintomas e complicações), duração da doença, e qualidade de vida relacionada com a saúde.

3 METODOLOGIA GERAL

3.1 A relevância da certeza de resultados

O objetivo da ATS é dar informação aos decisores, com a maior confiança possível, sobre se existe evidência disponível que prove os benefícios ou danos de uma intervenção específica face às alternativas já utilizadas na prática clínica.

Para a avaliação do valor terapêutico acrescentado, é utilizada a metodologia da Medicina Baseada na Evidência. De salientar, que em medicina o benefício de intervenções é avaliado em termos de probabilidade: o benefício é demonstrado quando a intervenção aumenta a probabilidade de um determinado resultado benéfico ou reduz a probabilidade de um resultado não-benéfico.

A Medicina Baseada na Evidência permite avaliar até que ponto a evidência disponível é confiável. Com este objetivo, utiliza um conjunto de regras e de instrumentos internacionalmente aceites e que constituem a base das avaliações de benefício. Das avaliações faz parte analisar um conjunto de detalhes sobre a forma como os estudos foram planeados, conduzidos, analisados e publicados.

Os ensaios clínicos comparativos e aleatorizados são considerados o método mais apropriado para estimar medidas do efeito relativo do tratamento. Estes devem ser integrados numa revisão sistemática e sintetizados através de meta-análise, convencional ou em rede (ver secção 4 Métodos de comparação). A incerteza na evidência deve ser identificada e explorada em análises de sensibilidade. Evidência não aleatorizada só poderá ser usada em situações específicas, que devem ser adequadamente justificadas (ver secção 5 Métodos de comparação em situações excecionais).

Recomenda-se que a ATS se baseie apenas em estudos com suficiente certeza de resultados. É responsabilidade da empresa titular da Autorização de Introdução no Mercado, a submissão dos processos para avaliação pelo INFARMED, I.P. incluindo toda a evidência que considerar relevante. Se da análise dessa evidência resultar que os estudos incluídos no processo de submissão não respondem às perguntas de investigação, poderá ser concluído que, com a documentação submetida, não existe evidência disponível que prove o benefício adicional de uma intervenção específica.

3.2 A ligação entre a certeza dos resultados e a proximidade às condições do dia a dia

É frequentemente referido que os estudos com uma elevada certeza de resultados (para fins deste documento, “certeza de resultados” significa elevada confiança nas estimativas de efeito) apresentam uma elevada validade interna, mas nem sempre representam a população na prática corrente, que é normalmente mais heterogénea. Ou seja, os resultados têm uma baixa validade externa e, por conseguinte, não são generalizáveis.

Contudo, esta crítica não resulta da metodologia utilizada no estudo mas do facto de os critérios de elegibilidade para esses estudos serem em geral muito restritivos, excluindo frequentemente doentes idosos ou com múltiplas comorbilidades; ou o protocolo do ensaio aleatorizado não refletir a prática clínica (por exemplo: diferenças na posologia, diferenças na regras de parar-recomeçar o tratamento, diferenças nas terapêutica prévia dos doentes; diferenças nas terapêuticas subsequentes, diferenças na intensidade de monitorização). Assim, aumentar a validade externa não implica reduzir o grau de certeza dos resultados, mas antes incluir os grupos de doentes considerados relevantes.

Desta forma, elevada certeza de resultados e proximidade às condições do dia-a-dia não se excluem mutuamente. Estudos comparativos, aleatorizados, com elevada validade interna e externa (por exemplo, estudos pragmáticos) são preferíveis.

3.3 Medidas de resultado

Deve existir uma definição prévia de quais as medidas de eficácia terapêutica e segurança que vão ser utilizadas na avaliação. As medidas de eficácia terapêutica e segurança utilizadas devem ser relevantes para o doente. Para este fim recomenda-se o uso das seguintes medidas: mortalidade, morbilidade (sintomas e complicações), duração da doença, e qualidade de vida relacionada com a saúde.

3.3.1. Medidas de efeito clínico

As medidas de efeito clínico são características ou variáveis que refletem os sintomas, a capacidade funcional ou a expectativa de vida do doente. No contexto da avaliação de uma intervenção (por exemplo, um fármaco), o efeito do tratamento nestas medidas materializa-se numa alteração que é detetável pelo doente, como melhoria de sintomas, melhoria da capacidade funcional, diminuição da probabilidade de desenvolver uma doença ou complicação dessa doença, ou um aumento da sobrevivência.

Na avaliação de tecnologias de saúde, e na definição da matriz de avaliação, devem ser valorizadas preferencialmente as medidas de efeito clínico. Assim, e como regra geral, apenas deve ser atribuída a máxima importância (medidas cuja importância deve ser classificada como crítica) às medidas de efeito clínico.

3.3.2. Medidas de efeito sub-rogadas

Para fins deste documento, define-se biomarcador, como uma característica que é medida e avaliada objetivamente como um indicador de resposta farmacológica a uma intervenção terapêutica.

Uma medida de efeito sub-rogada é um biomarcador que pretende substituir uma medida de resultado clínico, ou seja, um biomarcador que se espera ser capaz de prever o benefício clínico, o dano, ou a falta de benefício ou de dano. Esta expectativa deve ser suportada por evidência robusta (‘validação’).

Na avaliação de tecnologias de saúde, e na definição da matriz de avaliação, as medidas de efeito sub-rogadas devem ser menos valorizadas do que as medidas de efeito clínico. Assim, e como regra geral, não deve ser atribuída a máxima importância (ou seja, importância crítica) às medidas de efeito sub-rogadas. Estas medidas devem em geral ser classificadas como ‘importantes, mas não críticas’.

3.3.3. Validação de medidas de efeito sub-rogadas

3.3.3.1. Introdução

A substituição de uma medida de resultado clínico por uma medida de resultado sub-rogada num estudo aleatorizado, tem por objetivo permitir uma inferência estatística válida em relação à eficácia de uma intervenção sobre uma medida de resultado clínico, sem que o efeito sobre essa medida de resultado clínico tenha sido observado. Consequentemente, o uso de uma medida de resultado sub-rogada requer uma extrapolação que ultrapassa os dados observados, que permita fazer uma estimativa dos verdadeiros benefícios que são expectáveis para os doentes.

Na investigação médica é frequente o uso de medida de resultados sub-rogados como substitutos de medidas de eficácia terapêutica relevantes para o doente, com o objetivo de obter conclusões sobre a

eficácia de intervenções sobre medida de resultados clínicos mais precocemente e com menores custos. O uso de medida de resultados sub-rogados em estudos clínicos permite uma redução no número de participantes e na duração dos estudos, em comparação com o uso de medida de resultados clínicos.

Na perspetiva da Comissão, o uso de medida de resultados sub-rogados tem a vantagem potencial de acelerar o acesso a tecnologias inovadoras que oferecem valor acrescentado para os doentes. Contudo, frequentemente, as medidas de resultados sub-rogados não são capazes de prever de forma confiável o efeito global nas medidas de resultado clínico. O objetivo desta secção é recomendar o uso de uma metodologia que assegure que as medidas de resultado sub-rogadas utilizadas são capazes de prever de forma confiável o efeito global da intervenção nas medidas de resultado clínico.

Caso a evidência submetida pelo Titular de Autorização de Introdução no Mercado (TAIM) utilize medidas de resultado sub-rogadas, deve também conter informação sobre qual é a medida de resultado clínico que a medida sub-rogada substitui, e incluir demonstração da validação das medidas sub-rogadas utilizadas, utilizando a metodologia aqui recomendada.

Estudos que utilizam uma medida de resultados sub-rogada como base para a decisão de comparticipação/ financiamento de medicamentos

Os estudos que utilizam medida de resultados sub-rogados frequentemente sobrestimam o efeito do tratamento (2). Nos últimos anos, diferentes países aprovaram um número substancial de fármacos baseado em medida de resultados sub-rogados (3)(4).

A necessidade de avaliar estudos que utilizam medidas de resultados sub-rogados pode ter particular relevância no contexto da avaliação precoce de benefício de fármacos.

3.3.3.2. Requisitos de uma medida de resultado sub-rogada

Para que uma medida de resultado sub-rogada seja um substituto efetivo de uma medida de resultado clínico, os efeitos da intervenção na medida de resultado sub-rogada devem ser capazes de prever de forma confiável o efeito global na medida de resultado clínico, mas, na prática, esta condição frequentemente não é observada. Entre outras explicações para este facto, existe a possibilidade de o processo patológico afetar a medida de resultado clínico através de vários mecanismos causais não mediados pelo sub-rogado, sendo o efeito da intervenção nestes mecanismos causais diferente do seu efeito no sub-rogado (5). Ainda mais provável, a intervenção pode afetar a medida de resultado clínico por mecanismos de ação não reconhecidos, não previstos, e não intencionais, que operam independentemente do processo patológico (5).

É importante notar que, em alguns casos, o biomarcador é fortemente preditivo de sobrevivência, mas não prevê o efeito do tratamento na sobrevivência. As contagens de CD4+, usadas em estudos de HIV, são um exemplo desses marcadores.

Assim, na avaliação de benefício adicional de uma intervenção, medidas sub-rogadas de eficácia terapêutica podem ser consideradas como substitutas de medidas de eficácia clínica desde que tenham sido previamente validadas.

3.3.3.3. Validação de uma medida de resultado sub-rogada

Não existem procedimentos padronizados que permitam validar uma medida de resultado sub-rogada. A literatura metodológica defende frequentemente a utilização de métodos de correlação para validação

de sub-rogados, recomendando que as correlações sejam estimadas a nível do indivíduo e a nível dos estudos (3). Assim, nas suas avaliações de benefício, deve ser dada preferência a validações que utilizem estes procedimentos. Estes procedimentos requerem geralmente a realização de meta-análises de estudos aleatorizados, reportando os resultados sub-rogados e finais, em que é avaliado o efeito da intervenção na medida de resultado sub-rogada e na medida de resultado clínico. Apenas em casos excecionais são considerados métodos alternativos.

Assim, a validação passa por três etapas. Primeiro, avaliar a plausibilidade biológica da relação entre a medida de resultado sub-rogada e a medida de resultado clínico (nível três). Segundo, avaliar se existe uma correlação forte entre a medida de resultado sub-rogada e a medida de resultado clínico em diferentes coortes ou a nível do doente individual (esta correlação não valida a medida sub-rogada mas pode identificar bons marcadores de prognóstico) [nível dois]. Terceiro, avaliar se existe demonstração de uma relação entre o efeito do tratamento no sub-rogado e o efeito na medida de resultado clínico, preferencialmente, em vários estudos aleatorizados (nível um). No caso de novas tecnologias de saúde, que utilizam habitualmente medida de resultados sub-rogados, deve ser procurada evidência de outros estudos que avaliem a mesma tecnologia de saúde, ou tecnologias similares (incluindo fármacos da mesma classe ou, se esta evidência não estiver disponível, incluindo fármacos de classes diferentes) (6). Embora o segundo critério seja facilmente cumprido, o terceiro não é. Não existe um consenso sobre os valores de correlação (limiares) necessários à validação de um sub-rogado, mas frequentemente são apresentados valores de coeficiente de correlação (R_{estudo} ou $R_{indivíduo}$) entre 0,85 e 0,955. Se não existe uma correlação elevada, pode ainda ser usado o limiar do efeito no sub-rogado (*surrogate threshold effect* - STE). Este parâmetro também é baseado na análise de vários estudos aleatorizados, e define qual é o valor absoluto mínimo do efeito no sub-rogado que tem de ser observado para deduzir um efeito na medida de resultado clínico (3). Assim, pode ser calculado o STE em que um certo nível de variação no biomarcador se transforma em benefício clínico. Em ambos os casos, a certeza nas conclusões depende dos níveis pré-especificados de significância.

Para validar um sub-rogado, deve ser utilizada primariamente a correlação entre o efeito do tratamento no sub-rogado e o efeito do tratamento na medida de resultado clínico, avaliada ao nível dos estudos, utilizando os limiares definidos anteriormente.

De notar que os estimadores de correlação são sensíveis a pequenas alterações nos dados e o cálculo do seu intervalo de confiança é problemático quando as amostras dos estudos incluídos são pequenas ou moderadas(7). Por outro lado, as medidas de correlação só refletem relações (aproximadamente) lineares entre os efeitos do tratamento no sub-rogado e na medida de resultado clínico e não podem ser usadas para demonstrar relações com outras formas. É por isso importante considerar as 3 etapas de validação de uma medida sub-rogada, já que a utilização exclusiva da correlação pode exagerar ou diluir a importância da relação entre as medidas em consideração(8).

Pode ser considerado aceitável que, em situações excecionais, possam ser aceites medidas de resultado sub-rogadas não validadas, nos casos em que exista uma razoável probabilidade de o marcador ser capaz de prever o benefício clínico, e desde que seja demonstrada a impossibilidade prática de validar a medida de resultado sub-rogada, por exemplo, por o tempo necessário para observar o evento (medida de resultado clínico) ser excessivamente longo. Um exemplo prático desta situação é o uso da resposta virológica sustentada como uma medida de resultado sub-rogada de mortalidade ou carcinoma hepatocelular na hepatite C crónica. Para fins desta “razoabilidade” é necessário que exista, pelo menos, plausibilidade biológica (nível três de validação), e que se observe uma correlação entre o sub-rogado e a medida de resultado clínico (nível dois de validação).

De salientar, que a existência de uma correlação entre o efeito do tratamento no sub-rogado e o efeito do tratamento na medida de resultado clínico numa intervenção com um modo específico de ação, não significa necessariamente que essa correlação se observe com outras intervenções usadas para tratar

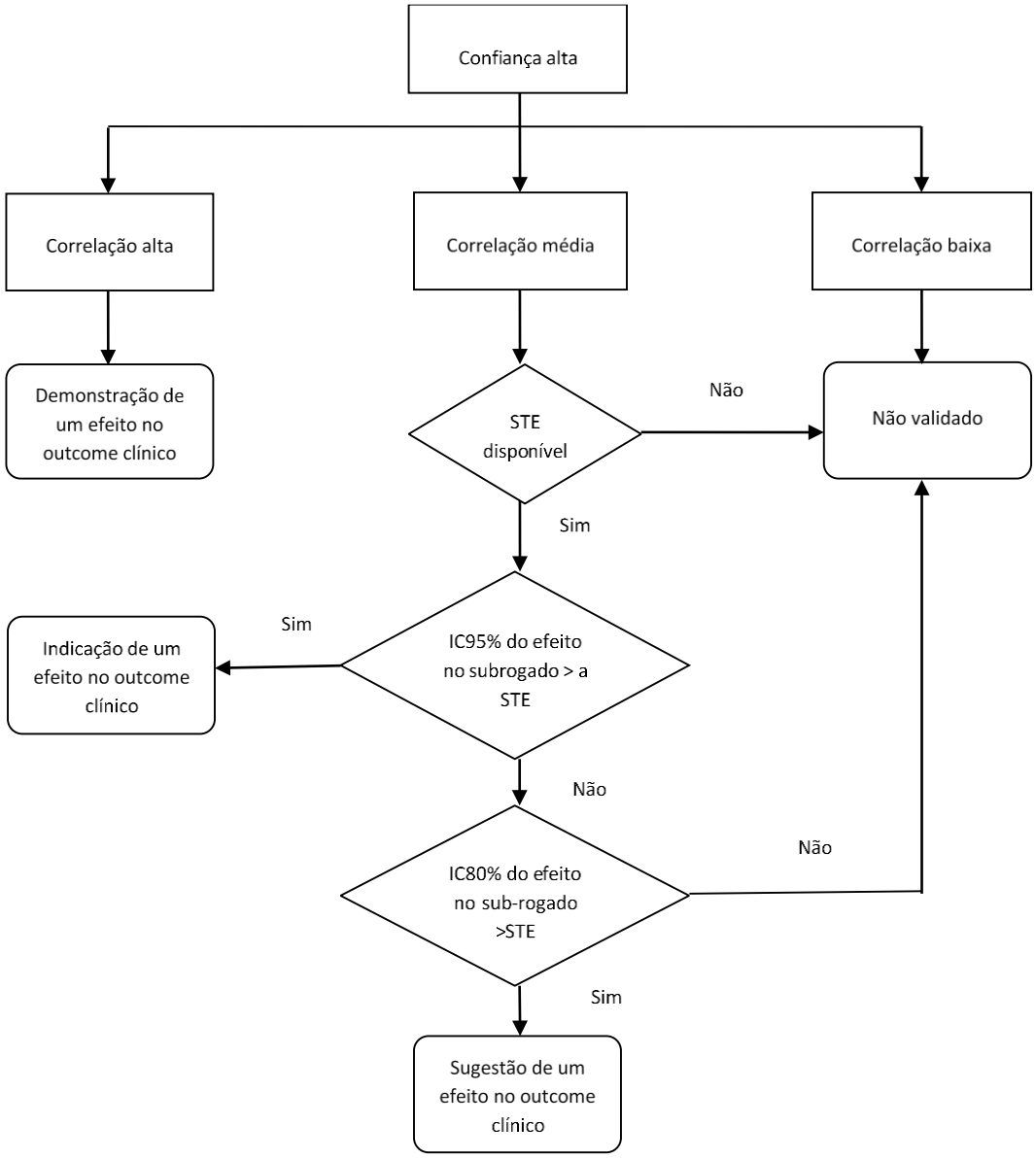
a mesma doença, que tenham um mecanismo de ação diferente. Assim, a validação de uma medida de resultado sub-rogada é normalmente feita numa população específica, e para uma intervenção específica, ou seja, a validação é específica para uma doença, para uma população, e para uma área terapêutica.

Uma vez que as novas tecnologias de saúde (utilizando novos/diferentes mecanismos de ação) são inicialmente avaliadas em estudos utilizando medidas de resultados sub-rogados, pode não existir evidência baseada em medida de resultados clínicos. Assim, a validação do sub-rogado só pode ter origem em estudos com fármacos com diferentes mecanismos de ação/de diferentes classes farmacêuticas. Recomenda-se que o uso de medida de resultados sub-rogados que apenas foram validados para fármacos utilizados na mesma indicação, mas com diferentes mecanismos de ação, apenas tenha lugar quando não existam alternativas de tratamento para a indicação em avaliação ou quando existe indicação, com razoável probabilidade, de que o novo fármaco pode apresentar benefício adicional em relação às alternativas existentes e a doença seja grave ou potencialmente fatal.

A conclusão sobre a validação de um sub-rogado depende de dois fatores: a qualidade da evidência que suporta a validação e a força da correlação entre o efeito da intervenção no sub-rogado e o efeito da intervenção na medida de resultado clínico.

No caso de o estudo de validação ser classificado como de alta qualidade, a conclusão sobre a validação do sub-rogado depende da força da correlação entre o efeito do tratamento no sub-rogado e na medida de resultado clínico ou do valor de STE. O diagrama de fluxo na Figura 1 descreve em detalhe o processo de classificação (9).

Figura 1: ***Conclusões sobre a validação do sub-rogado no caso de evidência de alta qualidade***



Modificado de Ref. 9

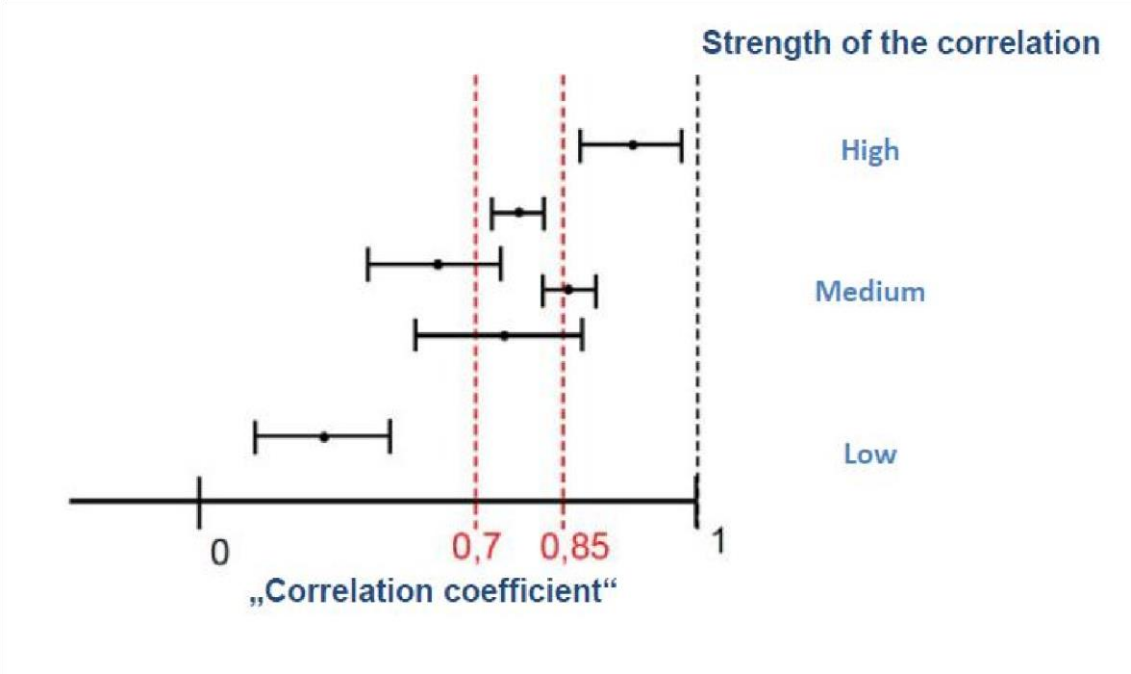
IC: intervalo de confiança; STE: Limiar do efeito no sub-rogado

Em relação à força da correlação, uma correlação é classificada como forte se o limite inferior do intervalo de confiança do coeficiente de correlação R for ≥0,85, é classificada como fraca se o limite superior do intervalo de confiança do R for ≤0,70, e é classificada como média se intervalo de confiança do R se sobrepõe, ainda que parcialmente, ao intervalo entre <0,85 e >0,70 (Figura 2) (9).

No caso de o estudo de validação ser classificado como de alta qualidade, considera-se que existe demonstração de validação do sub-rogado se existe uma correlação forte entre o efeito da intervenção no sub-rogado e o efeito na medida de resultado clínico; que não existe demonstração de validação do sub-rogado se se observou uma correlação fraca entre o efeito da intervenção no sub-rogado e o efeito na medida de resultado clínico; e considera-se que não é claro se o sub-rogado está validado, se existe uma correlação média (Figura 1) (9).

Neste caso, utiliza-se o limiar do efeito no sub-rogado (STE) e o intervalo de confiança do efeito da intervenção no sub-rogado para chegar a uma conclusão sobre a validação (Figura 2).

Figura 2*: Classificação da força da correlação em função da correlação entre o efeito do tratamento no sub-rogado e na medida de resultado clínico*



No caso de o estudo de validação ser classificado como de moderada ou baixa qualidade, considera-se que não é claro se o sub-rogado está validado (Tabela 1).

Tabela 1: *Validade da medida de resultado sub-rogada em função da qualidade da evidência (estudo de validação) e da correlação entre o efeito do tratamento no sub-rogado e na medida de resultado clínico*

Qualidade da evidência	**Correlação**	**Validade**
Alta	Forte	Sim
	Média	Não claro - usar STE
	Fraca	Não
Moderada		Não claro - usar STE
Baixa		Não claro
Muito Baixa		

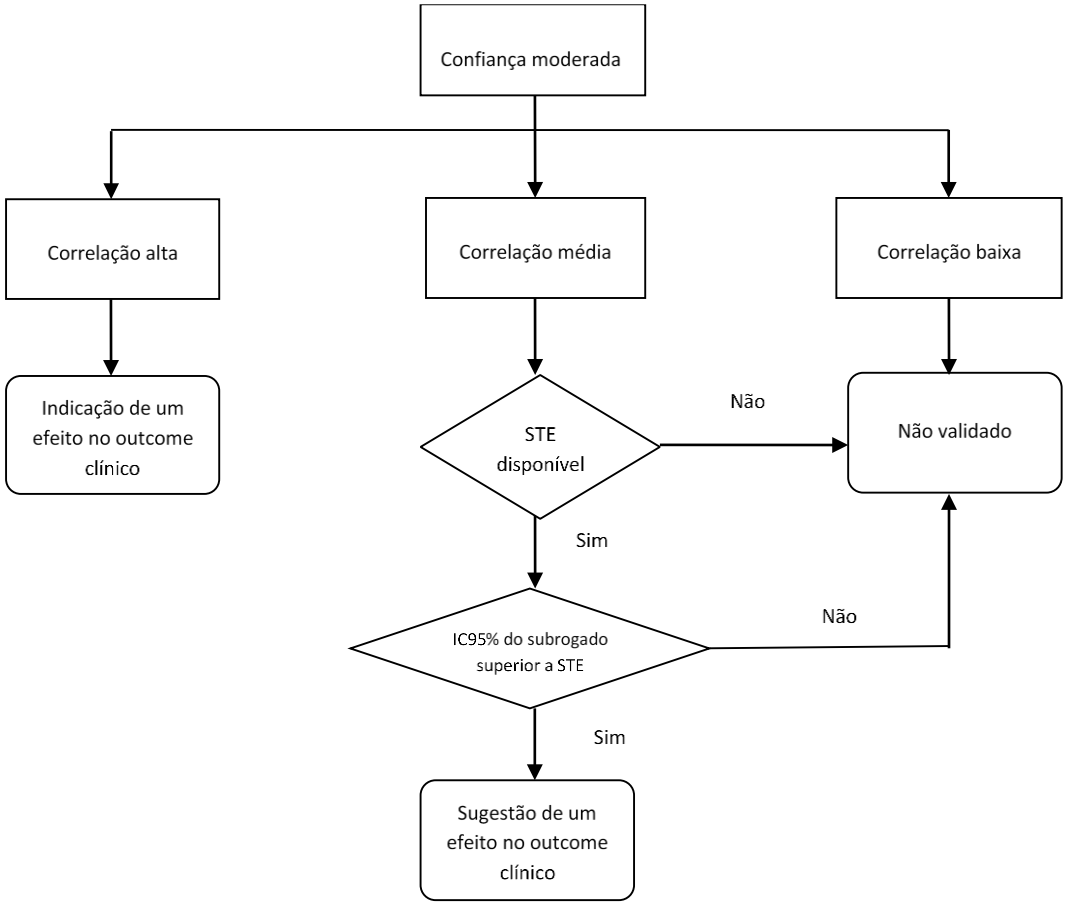
Fonte: adaptado de IQWiG Reports – Commission No. A10-05. Validity of surrogate endpoints in oncology. Version 1.1. 21.11.2011. (9)

No caso de o estudo de validação ser classificado como de moderada qualidade, a conclusão sobre a validação do sub-rogado depende da força da correlação entre o efeito do tratamento no sub-rogado e

na medida de resultado clínico ou do valor de STE. O diagrama de fluxo da Figura 3 descreve em detalhe o processo de classificação (9).

No caso de a aplicação desta metodologia conduzir a diferentes resultados em diferentes estudos de validação do sub-rogado, considera-se que o resultado da validação é o que resulta da maioria dos estudos de alta qualidade.

Figura 3**:** *Conclusões sobre a validação do sub-rogado no caso de evidência de moderada qualidade*



Fonte: Modificado de Ref. (IQWiG Reports – Commission No. A10-05. Validity of surrogate endpoints in oncology. Version 1.1. 21.11.2011. (9)

IC: intervalo de confiança; STE: Limiar do efeito no sub-rogado

3.3.3.4. Conclusões

A validação de uma medida de resultado sub-rogada passa por três etapas. Primeiro, avaliar a plausibilidade biológica da relação entre a medida de resultado sub-rogada e a medida de resultado clínico (nível três de validação). Segundo, avaliar se existe uma correlação forte entre a medida de resultado sub-rogada e a medida de resultado clínico em diferentes coortes ou a nível do doente individual (nível dois de validação). Terceiro, avaliar se existe demonstração de uma relação entre o efeito do tratamento no sub-rogado e o efeito na medida de resultado clínico, preferencialmente, em vários estudos aleatorizados (nível um de validação).

Não existe um consenso sobre os valores de correlação (limiares) necessários à validação de um sub-rogado, mas frequentemente são apresentados valores de coeficiente de correlação (R_{estudo} ou $R_{indivíduo}$) entre 0,85 e 0,955. Se não existe uma correlação elevada, pode ainda ser usado o limiar do efeito no sub-rogado (STE). Este parâmetro define qual é o valor absoluto mínimo do efeito no sub-rogado que tem de ser observado para deduzir um efeito na medida de resultado clínico.

Uma vez que as novas tecnologias de saúde (utilizando novos/diferentes mecanismos de ação) são inicialmente avaliadas considerando estudos que utilizam medidas de resultados sub-rogados, pode não existir evidência baseada em medida de resultados clínicos. Assim, a validação do sub-rogado só pode ter origem em estudos com fármacos com diferentes mecanismos de ação/de diferentes classes

farmacêuticas. Neste caso, e no sentido de avaliar a transferibilidade, os estudos de validação que incluam várias intervenções na mesma indicação devem, pelo menos, incluir dados sobre heterogeneidade. Contudo, o uso de medidas de resultados sub-rogados que apenas foram validados para fármacos utilizados na mesma indicação, mas com diferentes mecanismos de ação, apenas se justifica quando não existem alternativas de tratamento para a indicação em avaliação ou quando existe indicação, com razoável probabilidade, de que o novo fármaco pode apresentar benefício adicional em relação às alternativas existentes e a doença seja grave ou potencialmente fatal.

Podem ser aceites medida de resultados sub-rogados não validados, nos casos em que exista uma razoável probabilidade de o marcador ser capaz de prever o benefício clínico, desde que seja demonstrada a impossibilidade prática de validar a medida de resultado sub-rogada, por exemplo, por o tempo necessário para observar o evento (medida de resultado clínico) ser excessivamente longo. Para fins desta "razoabilidade" é necessário que exista, pelo menos, plausibilidade biológica (nível 3 de validação), e que se observe uma correlação entre o sub-rogado e a medida de resultado clínico (nível 2 de validação).

Caso a evidência submetida pelo TAIM utilize medidas de resultado sub-rogadas, deve também conter informação sobre qual é a medida de resultado clínico que a medida sub-rogada substitui, e incluir demonstração da validação das medidas sub-rogadas utilizadas, utilizando a metodologia aqui recomendada.

3.4 Revisões sistemáticas

3.4.1. Introdução

A revisão sistemática é um processo metodológico especificamente devolvido para identificar, selecionar e avaliar criticamente os estudos disponíveis sobre uma questão claramente formulada. É fundamental para o processo de avaliação que a base de evidência considerada seja abrangente e completa. A revisão sistemática deve ser transparente e objetiva, por forma a reduzir o viés e garantir que é obtida a resposta mais válida.

Assim, a evidência científica a considerar para informar a avaliação e determinar a eficácia comparativa do tratamento, inclui estudos secundários, nomeadamente revisão sistemática dos estudos clínicos sobre a intervenção em análise, e estudos primários. Dentro dos estudos primários, os ensaios aleatorizados fornecem o padrão mais elevado de evidência em relação à eficácia comparada de um tratamento e devem ser preferidos sempre que possível. No entanto, dados de estudos não aleatorizados podem ser necessários para suplementar os dados disponíveis ou para que seja obtida informação sobre outros parâmetros da avaliação, como efeitos adversos e o custo. Os dados dos estudos incluídos podem ser sintetizados através de meta-análise (ver secção 4 Métodos de comparação).

3.4.2. Protocolo de pesquisa

Cabe ao titular da AIM a revisão sistemática da evidência relevante para informar o processo de avaliação de tecnologias de saúde. Este processo deve ser conduzido conforme os melhores padrões internacionalmente estabelecidos para a realização de revisões sistemáticas, nomeadamente as definidas pelos consensos PRISMA (10) (*Preferred Reporting Items for Systematic Reviews*) e MOOSE (11) (*Meta-analysis of Observational Studies).* Mencionam-se igualmente as recomendações da EUnetHTA (*European Network for Health Technology Assessment*) para o processo de recolha de informação para revisões sistemáticas e avaliações de eficácia clínica de tecnologias de saúde (12).

Impõe-se assim a necessidade de formular um protocolo para a realização da revisão sistemática, em que são declarados antecipadamente os critérios de inclusão e exclusão, medida de efeitos, estratégia de pesquisa e análises planeadas. Estes aspetos deverão ser orientados pelo que foi previamente definido como população, intervenção, comparadores e medidas de resultado de interesse para a avaliação, e devem ser documentados com detalhe suficiente para garantir a transparência e reprodutibilidade da revisão sistemática realizada.

3.4.3. Bases de dados

No que diz respeito às fontes de informação a pesquisar deve ser incluído um conjunto alargado de bases de dados bibliográficas que permita garantir a identificação de todos os estudos relevantes para o tópico em avaliação. Deve ainda ser sistematicamente pesquisada e reportada a existência de ensaios clínicos concluídos ou em curso com relevância para a intervenção em avaliação, incluindo as interfaces clinicaltrials.gov e EUDRACT *(European Union Drug Regulating Authorities Clinical Trials Database*). Para diferentes aspetos da avaliação poderão ainda ser consideradas outras fontes de informação, para além da literatura publicada, como registos, conforme apropriado.

3.4.4. Estratégia de pesquisa e seleção de estudos

A estratégia de pesquisa deve ser reprodutível, estabelecida em linha com o enquadramento e o objetivo final da avaliação. A seleção da literatura deve ser baseada em critérios de inclusão e exclusão explícitos, usando metodologias padronizadas e reconhecidas. Os estudos não elegíveis devem ser listados, junto com a justificação pela qual foram excluídos e um diagrama de fluxo que resuma estas informações. Devem ser feitos os esforços possíveis para incluir toda a evidência relevante, independentemente da linguagem.

3.4.5. Avaliação da qualidade da evidência

A revisão deve revelar a melhor e mais atualizada evidência sobre a eficácia clínica da intervenção em relação aos seus comparadores. É assim fundamental a avaliação crítica da evidência científica usada para realizar a avaliação relativamente à sua validade, qualidade e aplicabilidade. Qualquer viés potencial que resulte do desenho dos estudos usados na avaliação deve ser explorado e documentado. Deve ainda ser considerada a validade externa dos resultados dos estudos incluídos na revisão, assim como a sua aplicabilidade para a prática clínica Portuguesa.

As estimativas de eficácia clínica dos tratamentos em comparação devem ser baseadas nos dados provenientes dos estudos disponíveis com melhor qualidade e aplicar-se, dentro daquilo que é a indicação em avaliação, ao doente típico, em circunstâncias clínicas normais, avaliando medida de efeitos clínicos relevantes e estabelecendo comparação com os comparadores apropriados, utilizando medidas relativas e absolutas de eficácia e medidas de incerteza adequadas.

Muitos fatores podem afetar a estimativa global de efeito relativo do tratamento que é obtida a partir da revisão sistemática. As diferenças entre os estudos incluídos podem resultar de diferenças nas características dos doentes (por exemplo, idade, sexo, gravidade da doença) ou noutros fatores, como diferenças na medição das medidas de efeitos ou no contexto de prestação de cuidados, por exemplo. Os modificadores potenciais do efeito do tratamento devem ser identificados antes da análise dos dados, através da extensa revisão do tema e discussão com peritos da disciplina clínica em questão.

Quando existam dados válidos e relevantes suficientes, expressos em medidas de medida de efeito comparáveis, é possível e apropriada a realização de uma síntese quantitativa através de meta-análise. Da mesma forma, quando os tratamentos em comparação não foram avaliados num mesmo ensaio

clínico aleatorizado, deve ser considerada a realização de meta-análise em rede, conforme apropriado. Estas metodologias e a sua aplicação são detalhadas na secção 4.4 Meta-análise em rede.

4 MÉTODOS DE COMPARAÇÃO

4.1 Introdução

A avaliação de tecnologias de saúde avalia o benefício adicional e a custo-efetividade de uma intervenção em relação aos comparadores de interesse. Com esse objetivo utiliza diferentes métodos de comparação.

Os ensaios clínicos aleatorizados, sintetizados numa meta-analise (convencional ou em rede) são preferíveis para estimar efeitos comparativos da intervenção em estudo e seus comparadores. Evidência não aleatorizada poderá ser aceite em situações especificas, que devem ser adequadamente justificadas (ver secção 5 Métodos de comparação em situações excecionais).

4.2 Comparações diretas e indiretas: definições

Entende-se por comparação direta entre dois tratamentos específicos, a comparação num estudo desses dois tratamentos, ou a combinação de múltiplos estudos desses mesmos tratamentos, para gerar uma estimativa combinada (meta-análise) da eficácia relativa dos dois tratamentos.

Comparação indireta é a estimativa da eficácia relativa entre dois ou mais tratamentos na ausência de estudos que os comparem diretamente. Comparação mista de tratamentos define-se como a estimativa da eficácia relativa de 3 ou mais tratamentos usando simultaneamente comparações diretas e indiretas. O termo meta-análise em rede engloba as comparações diretas, indiretas e mistas.

Quando a evidência disponível inclui vários estudos que comparam os tratamentos diretamente, por vezes combinam-se os resultados desses estudos utilizando técnicas meta-analíticas, para gerar uma estimativa combinada (*pooled estimate*) da eficácia relativa dos dois tratamentos.

Contudo, por vezes não existem dados suficientes para estimar de forma confiável a eficácia relativa de dois tratamentos ou pode haver necessidade de comparar mais de dois tratamentos simultaneamente, situação em que é necessário utilizar métodos de comparação de múltiplos tratamentos, por exemplo meta-análise em rede.

Assim, os métodos de comparação de múltiplos tratamentos podem ser usados para inferir a eficácia relativa de dois ou mais tratamentos na ausência de estudos que os comparem diretamente ou através da combinação de comparações diretas e indiretas.

É importante salientar que, o método de meta-análise (convencional ou em rede) usado deve manter intacta a aleatorização original dos estudos primários incluídos. As comparações que não mantenham a aleatorização, têm o mesmo valor das comparações utilizando estudos observacionais e não são recomendadas.

Só devem ser realizadas meta-análises convencionais ou em rede (incluindo comparações indiretas) se os estudos disponíveis forem comparáveis, homogéneos e consistentes, de modo que os resultados obtidos possam ser confiáveis. Estes assuntos são abordados com mais detalhe nas secções subsequentes.

4.3 Meta-análise convencional

4.3.1. Introdução

Meta-análise consiste na aplicação de um conjunto de metodologias estatísticas que permitem agregar os resultados provenientes de um conjunto de estudos primários, de forma a gerar uma ou mais medidas meta-analíticas de sumário. Também é possível analisar a presença, magnitude e potenciais moderadores da heterogeneidade na base de evidência identificada (13)(14)(15). Esta metodologia de síntese e análise quantitativa da base de evidência surge na sequência natural de uma revisão sistemática, sendo habitualmente a sua última fase.

Quando a evidência disponível incluiu dois ou mais estudos efetuando uma comparação direta das intervenções de interesse, os resultados desses estudos podem ser combinados utilizando técnicas meta-analíticas, para gerar uma estimativa combinada da eficácia relativa dos dois tratamentos. Contudo, a meta-análise convencional apenas permite comparar dois tratamentos entre si. A agregação de intervenções distintas para formar um único "tratamento" para efeitos de meta-análise não é aconselhável e tem que ser clinicamente justificada. Quando há múltiplos tratamentos em consideração, devem ser usados métodos de meta-análise em rede (secção 4.4 Meta-análise em rede).

Uma vez identificadas as medidas de efeito e extraídos os dados de cada estudo primário que permitem o seu cálculo e associadas medidas de precisão (erro padrão ou intervalos de confiança), será possível calcular medidas de sumário meta-analíticas que representam de forma agregada os resultados quantitativos dos estudos primários incluídos. Essas medidas resultam da agregação das medidas de efeito dos vários estudos primários incluídos, considerando ponderações específicas e distintas para cada estudo. Estas distintas ponderações têm em conta a precisão de cada estudo incluído (amostras de tamanho diferente, variabilidade diferente) e a heterogeneidade entre os estudos. Naturalmente, por exemplo, será expectável que, se um estudo tem um maior número de indivíduos analisados, o seu resultado tenha maior "peso", na altura do cálculo da medida de sumário, do que os resultados de outros estudos mais pequenos.

As vantagens de uma meta-análise incluem o aumento do poder estatístico e da precisão, a possibilidade de responder a questões não colocadas diretamente nos estudos individuais e a resolução de controvérsias, quando os estudos individuais chegam a conclusões contraditórias.

No entanto, é necessário ter em conta que os resultados de uma meta-análise podem ser afetados por diferenças no desenho e características dos estudos incluídos e por diferentes tipos de viés.

Quando se realiza uma meta-análise é necessário especificar a medida de efeito usada para descrever a eficácia da intervenção (por exemplo, risco relativo), o método estatístico de ponderação (por exemplo, ponderação pelo inverso da variância), modelo utilizado (efeito-fixo, efeitos-aleatórios), e o método de inferência estatística (frequentista ou Bayesiano). Deve também ser avaliada a variabilidade, magnitude e relevância das diferenças entre estudos (heterogeneidade). A consistência dos resultados dos estudos individuais pode influenciar a decisão de os combinar através de uma meta-análise e a decisão sobre o modelo analítico a ser usado.

4.3.2. Fatores que afetam a precisão

Por vezes, os estudos individuais são demasiado pequenos para estimar a eficácia relativa de dois tratamentos com suficiente precisão. O uso de técnicas meta-analíticas para combinar os resultados de vários estudos gera uma estimativa combinada da eficácia relativa dos dois tratamentos, podendo resultar num aumento da precisão da estimativa do efeito do tratamento.

O método de ponderação pelo inverso da variância é um método simples e comum de realização de meta-análises, usado tanto para variáveis dicotómicas como contínuas. É assim chamado porque o peso dado a cada estudo é o inverso da variação da estimativa do efeito (ou seja, um sobre o quadrado do seu erro padrão). É atribuído mais peso a estudos maiores, com erros padrão mais pequenos, que a estudos mais pequenos, que têm erros padrão maiores. É assim aproveitada a evidência proveniente de todos os estudos primários incluídos e é minimizada a imprecisão (incerteza) da estimativa combinada da eficácia dos tratamentos.

4.3.3. Modelos de efeito-fixo e efeitos-aleatórios

Dois modelos estatísticos são frequentemente usados em meta-análises:

- o modelo de efeito-fixo, pressupõe que todas estimativas do efeito da intervenção provenientes de diferentes estudos primários estimam o mesmo (verdadeiro) efeito na população de interesse, e que as diferenças observadas entre os diferentes estudos incluídos na meta-análise refletem unicamente uma variação aleatória (devido a serem provenientes de uma amostra);

- o modelo de efeitos-aleatórios, assume que existe uma variação nas estimativas do efeito da intervenção entre os estudos incluídos, para além da variação aleatória de natureza amostral. Este modelo assume que cada estudo estima o verdadeiro valor do efeito na população estudada e que esses verdadeiros valores do efeito seguem uma distribuição particular. Em geral, essa distribuição de verdadeiros efeitos de cada estudo assume-se como sendo normal (Gaussiana) e procura-se a estimação estatística da média e da variância dessa distribuição no contexto de um modelo hierárquico com dois níveis distintos. Neste modelo, considera-se que os estudos incluídos representam uma amostra aleatória de uma população teórica de estudos que respondem à questão de interesse.

O modelo de efeitos-aleatórios permite lidar com a heterogeneidade entre estudos que não pode ser explicada por outros fatores, incorporando-a no cálculo do sumário meta-analítico. Num conjunto de estudos heterogéneos, o modelo de efeitos-aleatórios atribui mais peso aos resultados de estudos mais pequenos que o modelo de efeito-fixo. A existência de uma grande heterogeneidade entre os estudos incluídos, pode causar problemas na interpretação dos resultados da meta-análise, em particular, ao atribuir mais peso aos resultados de estudos mais pequenos, o modelo de efeitos-aleatórios pode causar problemas na interpretação dos resultados da meta-análise (viés devido a estudos pequenos, *small studies bias*).

O modelo de efeito-fixo considera apenas a variabilidade dentro de cada estudo, enquanto o modelo de efeitos-aleatórios também considera a variabilidade entre estudos. Consequentemente, o modelo de efeito-fixo origina intervalos de confiança mais estreitos (isto é, melhor precisão). Na ausência de heterogeneidade entre estudos, os resultados obtidos com ambos os modelos coincidem.

De uma forma geral, as medidas meta-analíticas podem ser vistas como as melhores respostas disponíveis para a questão de investigação em apreço, desde que resultem da adequada síntese da melhor evidência disponível, cuja seleção tenha sido feita de maneira abrangente e não enviesada.

Nos casos em que a hipótese de homogeneidade entre os estudos não seja plausível, o modelo de efeitos-aleatórios deve ser usado. No entanto, quando estamos perante heterogeneidade grave, há uma sugestão de que os estudos incluídos estimam medidas de efeito aparentemente provenientes de realidades muito distintas. Nesse caso, as medidas meta-analíticas devem ser interpretadas com particular cuidado. Adicionalmente, importa procurar identificar as causas da heterogeneidade – ou seja,

é fundamental identificar as diferenças clínicas e/ou metodológicas entre os estudos que poderão explicar a heterogeneidade observada.

Se houver razões clínicas ou estatísticas para assumir homogeneidade nos efeitos relativos estimados pelos diferentes estudos, o método de efeitos-fixos pode ser utilizado. Se possível, uma análise de sensibilidade usando o modelo de efeitos-aleatórios deve ser também apresentada. No entanto, em geral não é possível estimar a heterogeneidade entre estudos com suficiente precisão em meta-análises com poucos estudos, pelo que um mínimo de 3 estudos deve ser considerado para uma meta-análise com modelo de efeitos-aleatórios.

4.3.4. Heterogeneidade

Heterogeneidade pode ser definida como qualquer tipo de variabilidade entre estudos:

- variabilidade nos participantes, intervenções e medidas de resultado (heterogeneidade clínica);
- variabilidade no desenho do estudo, na medição da estimativa do efeito e no risco de viés (heterogeneidade metodológica);
- variabilidade no efeito da intervenção a ser avaliada entre diferentes estudos ou no risco basal das diferentes populações (heterogeneidade estatística). Esta pode ser uma consequência tanto da heterogeneidade clínica como metodológica.

Uma meta-análise deve ser apenas realizada quando um grupo de estudos é suficientemente homogéneo em termos de participantes, intervenções e medidas de resultado.

No caso de haver estudos classificados como tendo elevado risco de viés, estes devem ser excluídos da meta-análise e só estudos com baixo risco de viés devem ser incluídos na análise principal.

A heterogeneidade estatística será referida a partir deste ponto como apenas heterogeneidade. É esperada alguma variação (inconsistência) nos resultados de diferentes estudos devido apenas ao acaso. A variabilidade que não pode ser atribuída ao acaso, reflete verdadeiras diferenças nos resultados dos estudos, ou seja, heterogeneidade. Se os intervalos de confiança dos resultados dos diferentes estudos apresentam pouca sobreposição, é uma indicação da presença de heterogeneidade. Esta pode ser avaliada mais formalmente através de um teste estatístico.

A heterogeneidade pode ser avaliada através do teste X^2 (Qui-quadrado), baseado na estatística Q de Cochran. Este teste avalia se as diferenças entre os resultados se devem apenas ao acaso. Um valor-p baixo (ou um teste X^2 elevado relativamente aos graus de liberdade) é evidência da existência de heterogeneidade nas estimativas dos efeitos da intervenção. No entanto, é necessária prudência na interpretação dos resultados desde teste, pois tem baixo poder quando existem poucos estudos incluídos na meta-análise ou quando o tamanho da amostra é pequeno e poder em excesso quando existem muitos estudos incluídos na meta-análise.

Em contraste, a estatística I^2 quantifica a percentagem da variação total dos efeitos estimados dos diferentes estudos incluídos na meta-análise que se deve à heterogeneidade e não à variabilidade aleatória de natureza amostral e, portanto, deverá ser sempre considerada em complemento ao teste de hipótese baseado na estatística Q de Cochran. Alguns autores consideram um valor I^2 inferior a 25% como baixo. Não existem pontos de corte universalmente aceites, no entanto, considera-se em geral que um valor de I^2 superior a 40-50% configura uma situação de heterogeneidade moderada a grave, que deverá merecer particular atenção e exploração. No entanto, a estatística I^2 também sofre de grande incerteza quando apenas alguns estudos estão disponíveis e é sensível à precisão dos estudos

incluídos. Reportar o grau de incerteza do I^2 (com um intervalo de confiança de 95%) é recomendável. A estimativa de heterogeneidade, τ (tau) ou τ^2 e o seu intervalo de confiança devem também ser tidos em consideração (16). Quando existem poucos estudos, as inferências acerca da heterogeneidade devem ser cautelosas. Como regra prática, quando existe heterogeneidade grave a interpretação das medidas meta-analíticas deve sempre ser feita com extremo cuidado, pois elas podem, nessa altura, não corresponder exatamente à melhor estimativa do efeito do tratamento que se pretende avaliar.

Quando se observa heterogeneidade considerável ou grave, é importante considerar as razões que possam explicá-la. Em particular, a heterogeneidade pode ser devida a diferenças entre subgrupos dos estudos. Também, erros na execução da revisão sistemática e na extração dos dados são uma causa comum de heterogeneidade nos resultados.

4.3.5. Análise de subgrupos e meta-regressão

Quando a estimativa do efeito da intervenção varia com diferentes populações ou com características da intervenção como a dose ou a duração do tratamento, essa variação é conhecida como interação ou modificação de efeito. Análise de subgrupos e meta-regressão são métodos usados para determinar se existe interação ou se os resultados são robustos. As regras de definição e avaliação da credibilidade de análise de subgrupos são detalhadas na secção 6 Análise de subgrupos.

Ajustamentos por meta-regressão são considerados resultados observacionais e devem ser usados para análises exploratórias de identificação de modificadores de efeito, ou de sensibilidade, e não como resultados principais.

A análise de subgrupos é realizada para investigar resultados heterogéneos e para responder a perguntas específicas acerca de um determinado grupo de doentes, tipo de intervenção ou tipo de estudo. Os resultados das análises de subgrupo podem induzir em erro, pois não são baseadas em comparações randomizadas.

A meta-regressão é um método alternativo para testar diferenças entre subgrupos. Neste caso um modelo de efeitos-aleatórios é preferível, devido ao risco de resultados falso-positivos quando um modelo de efeito-fixo é usado para comparar subgrupos.

A meta-regressão é uma extensão da análise de subgrupos que permite que o efeito tanto de características contínuas como categóricas seja investigado, e que o efeito de múltiplos comparadores seja investigado simultaneamente (se existir um número adequado de estudos). A meta-regressão não deve ser considerada se existirem menos de 10 estudos na meta-análise.

Numa meta-regressão, a medida de resultado é a estimativa do efeito da intervenção e as variáveis explanatórias são características dos estudos que podem influenciar o tamanho do efeito da intervenção. Para evitar risco de viés a meta-regressão com dados de doentes individuais é preferível, mas raramente possível no contexto da avaliação farmacoterapêutica. Os riscos de potencial viés de agregação devem ser tidos em conta ao interpretar resultados de meta-regressão com base em dados agregados.

4.3.6. Meta-análise com dados individuais

A meta-análise de dados individuais de participantes ou doentes (*individual patient or participant data – IPD*) é um tipo de meta-análise que envolve a obtenção e síntese de dados individuais dos participantes de vários estudos clínicos relacionados (17). É considerada a metodologia de referência das meta-análises e é particularmente relevante para determinar a eficácia das intervenções atendendo às características específicas dos participantes. No entanto, esta abordagem IPD não é frequentemente encontrada na prática, pois a obtenção dos dados dos participantes individuais de cada estudo é uma tarefa complicada do ponto de vista operacional e muitas vezes impossível de realizar. Quando não é possível obter dados

individualizados para todos os estudos da meta-análise, podem ser usados métodos que combinam dados individuais de participantes com dados agregados (18).

Este tipo de meta-análise deve ser realizado quando a meta-análise convencional não é adequada para responder à questão clínica pré-definida. Neste caso, o uso de dados individuais dos participantes nos diferentes estudos aleatorizados permite aumentar o poder estatístico para detetar efeitos de tratamento distintos. A disponibilidade de dados individuais de cada participante facilita a estandardização da análise estatística entre estudos e a obtenção direta da informação pretendida, independentemente da significância estatística ou de como foi reportada nos estudos individuais. É possível analisar os dados com mais detalhe, obter resultados com mais tempo de seguimento, incluir mais participantes, e investigar hipóteses diferentes das dos estudos originais. Também diminui o risco de viés associado ao uso de dados agregados na meta-regressão. A meta-análise de dados individuais é considerada mais fiável que a meta-análise convencional e pode originar diferentes conclusões. No entanto, é uma técnica organizacionalmente mais complexa, mais dispendiosa e mais demorada.

Os métodos estatísticos usados na meta-análise de dados individuais devem preservar o agrupamento (*clustering*) dos participantes de cada estudo (19). O agrupamento dos participantes é mantido durante a análise, podendo ser usadas duas possíveis abordagens, uma com um único passo e outra usando dois passos. Na abordagem de dois passos, em primeiro lugar, os dados individuais dos participantes são analisados de forma independente em cada estudo individual utilizando o método estatístico apropriado para o tipo de dados a ser analisado, o que produz resultados agregados para cada estudo. Depois, num segundo passo, estes dados são sintetizados usando um modelo adequado para análise de dados agregados, de forma semelhante à meta-análise convencional. Assim, modelos de efeito-fixo e de efeitos-aleatórios podem ser usados para estimar o efeito da intervenção. Como alternativa, pode ser usada uma técnica com um único passo, isto é, num único modelo em que os dados individuais dos participantes nos vários estudos podem ser analisados em simultâneo utilizando técnicas específicas para manter o agrupamento dos doentes em cada estudo (tipicamente uma regressão com um termo separado para cada estudo ou um que varia entre estudos via efeitos aleatórios). Mais uma vez é necessário usar um modelo específico para o tipo de dados a ser analisado e respeitar os pressupostos da meta-análise. Estas duas técnicas de meta-análise em um ou dois passos, originam habitualmente resultados semelhantes. No entanto, quando os estudos incluídos são pequenos e/ou o efeito é grande ou os eventos são raros, existe risco de viés com a técnica em dois passos, porque algumas pré-especificações do segundo passo podem não ser apropriadas (20).

É também importante reconhecer que a qualidade dos dados individuais para a meta-análise é dependente da qualidade dos estudos originais, assim como de uma revisão sistemática corretamente efetuada. Assim, uma meta-análise de dados individuais deve seguir um protocolo pré-definido, e incluir uma avaliação da qualidade dos estudos originais (20). Se apropriado, deve ficar claro como a inclusão de dados de menor qualidade afeta as conclusões.

4.3.7. Medidas de efeito e sua interpretação

Para medidas de resultado binárias, as medidas mais comuns para estimar o efeito da intervenção incluem a razão de riscos (*hazard ratio*) a razão de chances (*odds ratio*) e a diferença de risco (*risk difference*).

Para medidas de resultado contínuas, a medida usada para estimar o efeito da intervenção é a diferença média *(mean difference*). O seu uso é apropriado quando a estimativa do efeito da intervenção é feita na mesma escala para os diferentes estudos.

Quando os estudos a serem combinados usam escalas diferentes o efeito da intervenção em cada estudo pode ser dividido pelo desvio padrão para formar uma diferença média padronizada (*standardised mean difference*) que reflete a magnitude do efeito da intervenção em cada estudo relativamente a

variância da escala. A escolha do desvio padrão a ser utilizado não é consensual e pode causar viés e heterogeneidade, em especial no caso em que os estudos incluídos são pequenos – o desvio padrão observado em cada estudo é normalmente usado. No entanto, isto assume que este desvio padrão é idêntico em todos os estudos incluídos o que raramente é realista (13)(21). A utilização de desvios padrão estimados externamente para cada escala pode ser uma solução e facilita a conversão do efeito relativo estimado para uma das escalas originais o que facilita a interpretação (13).

Para medidas de resultado tempo até ao evento, a razão de riscos (*hazard ratio*) é a medida mais comum para estimar o efeito da intervenção. É necessário incluir na meta-análise o logaritmo da razão de riscos (*hazard ratio*) e o erro padrão para cada estudo. Risco relativo (*risk ratio*) e razão de chances (*odds ratio*) (relacionados com eventos que ocorrem num determinado tempo) não são equivalentes a razão de riscos (*hazard ratio*), e os tempos de sobrevivência mediana não devem ser usados numa meta-análise.

A seleção da medida usada para descrever a eficácia combinada da intervenção depende de 3 fatores:

- consistência: a medida escolhida deve originar estimativas semelhantes em todos os estudos da meta-análise e subpopulações em que a intervenção vai ser aplicada. Quanto mais consistente a medida escolhida, mais robusta é a justificação para descrever a estimativa combinada do efeito da intervenção com um único valor. As medidas relativas são habitualmente mais consistentes que as medidas absolutas, pelo que se deve evitar realizar meta-análises utilizando a diferença de risco. Razão de chances e razão de risco, são habitualmente equivalentes em termos de consistência; a meta-análise de razões de chances tem melhores propriedades estatísticas, mas razões de risco são mais fáceis de interpretar (22)(23);
- propriedades matemáticas: a medida escolhida deve ter as propriedades matemáticas necessárias para realizar uma meta-análise válida sendo a propriedade mais relevante a existência de um estimador adequado e de fácil aplicação da variância da medida (22)(23);
- a medida usada para descrever a eficácia da intervenção deve ser facilmente interpretável e aplicável ao objetivo da meta-análise. Pode ser indicado realizar uma análise de sensibilidade para determinar se a medida escolhida para estimar a eficácia combinada da intervenção influencia as conclusões da meta-análise.

4.3.8. Formas de reportar os resultados de uma meta-análise

Os resultados de uma meta-análise são habitualmente ilustrados usando um gráfico em floresta (*forest plot*):

- um gráfico em floresta inclui a estimativa do efeito da intervenção e os intervalos de confiança para estudos individuais e meta-análise;
- cada estudo é representado por um bloco com a estimativa do efeito da intervenção e uma linha horizontal que se estende para cada lado dos blocos;
- a área do bloco indica a estimativa pontual de cada estudo incluído na meta-análise, enquanto a linha horizontal representa o intervalo de confiança (habitualmente, um intervalo de confiança de 95%);

- o intervalo de confiança representa o intervalo do efeito da intervenção compatível com o resultado do estudo;
- o tamanho do bloco pode ser também indicativo do peso do estudo (quando o tamanho do bloco é maior, os intervalos de confiança são habitualmente mais estreitos) para a estimativa combinada do efeito da intervenção;
- a estimativa combinada do efeito da intervenção é normalmente representada por um diamante no final do gráfico em floresta;
- para cada meta-análise, deve também ser apresentada uma medida de consistência dos resultados dos estudos incluídos, como por exemplo I^2 e τ (medidas de heterogeneidade), com os respetivos intervalos de confiança.

4.4 Meta-análise em rede

4.4.1. Introdução

Quando não existem dados suficientes para estimar de forma confiável a eficácia relativa de duas intervenções ou quando há necessidade de comparar mais de duas intervenções simultaneamente, é necessário utilizar métodos para meta-análise em rede. Por exemplo, uma comparação indireta pode ser necessária quando duas intervenções não foram comparadas diretamente em ensaios clínicos, mas têm um comparador comum (p. ex. placebo). As comparações indiretas são casos especiais de meta-análise em rede em que não há estudos que comparem diretamente nenhuma das intervenções em causa e a meta-análise convencional é um caso especial que só inclui duas intervenções.

Deverá ser tido em conta que os métodos de meta-análise em rede são particularmente relevantes nos casos em que na prática clínica corrente se utilizam várias (mais do que uma) intervenções para a mesma indicação, e em que, por conseguinte, podem existir vários comparadores selecionados para a avaliação. Quando há múltiplos comparadores com evidência relevante, estes devem ser comparados simultaneamente numa análise de meta-analise em rede, de forma a ter em consideração a evidencia direta e indireta e assegurar a coerência das conclusões para todos os comparadores.

O método de escolha das intervenções a incluir na rede deve ser especificado previamente, ser reproduzível e garantir uma rede única de intervenções a comparar (24) (25).

Todos os estudos que comparem duas ou mais intervenções mencionadas no PICO (I ou C) devem ser incluídos desde que respeitem os outros critérios pré-especificados (ver secção 2.2). Só devem ser realizadas meta-análises em rede (incluindo comparações indiretas) quando os estudos disponíveis forem comparáveis (isto é suficientemente homogéneos) e de suficiente qualidade (ver secção 9.3), de modo que os resultados obtidos possam ser confiáveis.

A rede de comparações deve ser desenhada de forma a que as intervenções incluídas definam vértices e a existência de estudos comparando duas intervenções diretamente definam arestas (de notar que estudos com mais de 2 braços definem múltiplas comparações e todas devem ser incluídas na representação da rede). A meta-análise em rede produz efeitos relativos para todas as comparações de pares de intervenções incluídas na rede desde que formem uma rede ligada (conexa), ou seja, uma rede em que seja possível estabelecer um caminho (usando as arestas) de uma intervenção (vértice) para qualquer outra. Intervenções que não estejam ligadas por arestas não poderão ser comparadas. Numa rede com ciclos (isto é, em que é possível definir um caminho que comece e acabe no mesmo vértice) as comparações incluídas em cada ciclo são compostas por evidência direta e indireta o que aumenta a precisão mas requer a avaliação de consistência da evidência (ver secção 4.4.2 e 4.4.3).

No caso de a rede estar desligada (não conexa), o critério de inclusão pode ser estendido para incluir estudos aleatorizados que comparem intervenções adicionais com uma (ou mais) das intervenções em avaliação, que possam ligar a rede. Estes estudos adicionais devem respeitar todos os outros critérios de inclusão, nomeadamente terem uma população comparável à especificada no PICO. Quando há mais do que uma intervenção adicional que possa ligar as intervenções em consideração, todas as intervenções de ligação devem ser incluídas, de forma a que toda a evidência relevante seja considerada (24).

A inclusão de intervenções adicionais numa rede ligada pode ser justificada para aumentar a precisão dos resultados. Neste caso, os mesmos princípios de inclusão de todos as intervenções e estudos adicionais que possam aumentar a precisão da rede devem ser seguidos. Deve ser apresentada uma análise de sensibilidade com os resultados da meta-análise à rede original (incluindo somente as intervenções mencionados no PICO).

Quando não é possível ligar todos as intervenções da rede em consideração usando estudos aleatorizados realizados na mesma população, a base de evidência pode ser alargada a outras populações onde se possa pressupor que os efeitos relativos são comparáveis ou podem ser ajustados, por exemplo por meta-regressão. O uso de estudos não-aleatorizados ou observacionais para formar arestas que liguem redes desligadas é desaconselhado devido ao seu elevado potencial de viés, que poderá contaminar qualquer comparação que utilize essa aresta para ligar as intervenções. Dado o elevado potencial de viés, a utilização de métodos de ajustamento de populações para ligar redes desligadas e para incluir dados de estudos não comparativos (só com um braço), não é recomendada, exceto em situações excecionais. Estas situações excecionais encontram-se descritas na secção 5 e devem ser adequadamente justificadas.

Existem vários métodos de comparações múltiplas, genericamente denominadas de meta-análise em rede (*network meta-analysis*), que incluem (entre outros) o método de Bucher (26)para comparações indiretas, e métodos frequentistas e Bayesianos para comparações mistas de intervenções (27). A escolha do método deve ser individualizada para cada caso tendo em conta o tipo de evidência e a estrutura da rede. A descrição detalhada destas metodologias está para além do âmbito deste documento.

Métodos de comparação indireta podem ser usados para inferir a eficácia relativa de duas intervenções, na ausência de estudos que as comparem diretamente, se houver uma outra intervenção (por exemplo placebo) que permita ligar as duas intervenções em consideração. O método de Bucher (28)deve ser aplicado exclusivamente a situações de comparações indiretas entre duas intervenções com uma única intervenção de ligação, onde só um estudo está disponível para cada comparação. No caso de haver múltiplos estudos disponíveis para cada comparação, o método só pode ser usado quando estes são combinados usando modelos de meta-análise com efeito-fixo. Caso múltiplos estudos sejam agrupados usando meta-análise com efeito-aleatório, o método de Bucher não é apropriado. Métodos de meta-análise em rede (Bayesianos ou frequentistas) devem ser utilizados para efetuar comparações indiretas com efeitos-aleatórios.

Métodos de comparação indireta são também aplicáveis à comparação de múltiplas intervenções ligadas a um único comparador comum formando uma rede em estrela (rede de intervenções só ligadas por um comparador comum). A utilização de métodos para comparações mistas (de meta-análise em rede) é recomendada já que são mais eficientes no caso de modelos de efeito-fixo, e permitem melhor estimação da heterogeneidade em modelos de efeito-aleatório. Numa rede em estrela, as estimativas de efeito para a comparação de quaisquer duas intervenções por via do comparador comum só são afetadas pelos estudos que compõem essa comparação indireta. Quando um modelo de efeito-fixo é usado, estudos componentes da rede em estrela que não envolvam estas duas intervenções não afetam a estimativa do efeito relativo. No entanto, se um modelo de efeitos-aleatórios for usado, todos os

estudos contribuem para a estimação da heterogeneidade comum à rede, o que poderá afetar os intervalos de confiança para todas as comparações.

A escolha do método de meta-análise em rede deve ser individualizada para cada caso. A descrição detalhada destas metodologias está para além do âmbito deste documento, no entanto, notamos os seguintes pontos (ver também secção 4.4.3):

- só devem ser utilizados métodos que garantam a consistência de resultados e que usem princípios meta-analíticos, ou seja, que combinem efeitos relativos (e não efeitos absolutos) das intervenções. Os modelos de meta-análise em rede propostos por Lu & Ades (29) e descritos nos documentos do NICE DSU (30) são recomendados e podem ser estimados usando métodos Bayesianos (31) (32) ou frequentistas (33);
- o modelo de meta-análise em rede proposto por Rücker et al (34) também é adequado, mas a estimação é efetuada de forma diferente, pelo que os modelos descritos no ponto anterior são preferíveis para modelos com efeitos-aleatórios;
- os modelos propostos por Lumley (35) e Hong (36) incluem pressupostos diferentes e não devem ser usados.

4.4.2. Pressupostos de uma meta-análise em rede

Os métodos de meta-análise em rede são extensões dos métodos de meta-análise direta para comparações de mais do que duas intervenções. Consequentemente, todos os pressupostos subjacentes à validade das meta-análises diretas convencionais (ver secção 4.3) também se aplicam a meta-análises em rede e comparações indiretas.

Os participantes devem ser comparáveis entre os estudos incluídos, e devem ser relevantes para a avaliação em curso. Os estudos incluídos devem ser suficientemente homogéneos e não diferir substancialmente em características que possam alterar os efeitos relativos das intervenções. Os estudos devem incluir as intervenções especificados no PICO. A extensão de intervenções incluídas para ligar a rede ou aumentar a precisão dos resultados deve ser adequadamente justificada (ver secção 4.4.1).

A extensão dos pressupostos referidos na secção 4.3 à meta-análise em rede implicam os seguintes pontos adicionais:

- não deve haver diferenças entre os participantes incluídos em estudos que comparem intervenções diferentes, ou seja, em princípio qualquer participante poderia ter sido aleatorizado para qualquer intervenção e incluído em qualquer dos estudos;
- comparações diretas e indiretas estimam o mesmo efeito relativo na população incluída. Ou seja, para qualquer par de intervenções, o efeito da intervenção X comparada com a intervenção A e o efeito da intervenção Y comparada com A é o mesmo que seria observado num estudo que incluísse as intervenções A, X e Y.

Ao assegurar que os estudos incluídos na meta-análise em rede são suficientemente homogéneos no que respeita a aspetos clínicos relevantes (e que todos os pressupostos mencionados se verificam) está, em teoria, assegurada a consistência dos resultados da meta-análise em rede ou comparação indireta a efetuar. No entanto, esta consistência deve ser verificada estatisticamente sempre que possível, ou seja, em redes com ciclos.

4.4.3. Aspetos técnicos na meta-análise em rede

Modelos para meta-análise em rede com efeito-fixo ou efeitos-aleatórios podem ser usados, dependo dos pressupostos clínicos e da variabilidade dos estudos incluídos. No caso de modelos de efeitos-aleatórios, os modelos mais utilizados assumem o mesmo nível de heterogeneidade para todas as comparações, ou seja, estimam uma heterogeneidade comum a todas as comparações. Modelos que estimam níveis de heterogeneidade diferentes para diferentes comparações são mais complexos e na maioria dos casos não existe informação suficiente para os estimar (37). Modelos que estimam um parâmetro de heterogeneidade comum são, por isso, aceitáveis.

Para assegurar resultados estáveis na computação de resultados de meta-análise em rede, devem ser tidos em consideração os seguintes pontos:

- a rede de intervenções deve estar ligada (ver secção 4.4.1);
- em redes com estudos pequenos ou que investiguem eventos binários raros, é comum observar zero eventos em um ou mais braços do estudo. Estudos com zero eventos em todos os braços devem ser removidos já que não contribuem para a estimação de efeitos relativos. De notar que a rede pode ficar desligada quando estes estudos são removidos (ver secção 4.4.1);
- o tratamento de referência deve ser escolhido de forma a que seja um dos vértices com mais ligações aos outros tratamentos incluídos na rede, ou seja, o tratamento com o maior número de arestas e que esteja *no centro* da rede. Normalmente intervenções de controle ou placebos satisfazem esta condição e devem ser usadas como referência. Teoricamente a escolha da intervenção de referência não afeta os resultados da meta-análise em rede, já que todas as comparações são estimadas, no entanto, a escolha de um tratamento de referência com poucas comparações ou mais distante dos outros na rede (isto é, com mais arestas a percorrer para efetuar comparações) pode levar a problemas computacionais causando dificuldades na estimação, por exemplo uma reduzida velocidade de convergência dos algoritmos usados e elevada autocorrelação entre estimativas.

A escolha da escala dos efeitos relativos para a meta-análise em rede deve ter em conta o tipo de medida de resultado e as propriedades estatísticas dos efeitos relativos apropriados para esse tipo de medida de resultado (ver secção 4.3.7).

Na maioria das situações, a escolha do método de estimação dos efeitos (Bayesiano ou frequentista) não afeta o resultado da meta-análise em rede, desde que:

- a rede esteja ligada;
- seja usado *software* apropriado para a modelo a estimar, por exemplo WinBUGS, OpenBUGS, JAGS ou Stan com código apropriado para métodos Bayesianos e Stata ou R usando funções apropriadas para métodos frequentistas;
- haja um número suficiente de estudos para estimar o nível de heterogeneidade no caso de usar um modelo com efeitos-aleatórios;
- não haja estudos com medida de resultados discretos que não tenham observado eventos em um ou mais braços (ou seja com a presença de resultados com zero eventos);
- não seja usada evidência externa numa análise Bayesiana com distribuições *a priori* informativas.

Métodos de estimação Bayesiana devem ser preferidos quando:

- há estudos com zero eventos – os modelos de estimação Bayesiana aceitam resultados com zeros e não estão sujeitos a introduzir viés com a necessidade de adicionar 0.5 às células dos braços com zeros (28). Quando a única ligação de uma intervenção à rede é feita por um ou mais estudos com zero eventos observados num dos braços, a rede pode estar desligada. Nestes casos a conexão da rede deve ser reavaliada com estes estudos excluídos. Caso a rede fique desligada, métodos que adicionam 0.5 às células dos estudos com zeros necessários para ligar a rede podem ser considerados. No entanto, note-se que os resultados ficarão ligeiramente enviesados (28). Estudos em que não haja eventos em nenhum dos braços devem ser excluídos;

- há um número insuficiente de estudos para estimar a heterogeneidade mas é necessário considerar um modelo de efeitos-aleatórios devido às características dos estudos incluídos – neste caso o uso de distribuições *a priori* informativas para o parâmetro de heterogeneidade é recomendado (38);

- há informação externa relevante que deve ser utilizada como distribuição *a priori* na estimação dos efeitos relativos, por exemplo para ligação de redes desligadas. No contexto da avaliação farmacoterapêutica é rara a existência de informação externa validada, pelo que este cenário só deve ser aplicado em situações especiais e com adequada justificação.

No caso de utilização de métodos de estimação Bayesianos, a escolha das distribuições *a priori* deve ser justificada para todos os parâmetros a estimar e sujeita a análise de sensibilidade. Normalmente, distribuições não informativas devem ser escolhidas para os parâmetros que estimam efeitos relativos (e absolutos) das intervenções.

Em modelos com efeitos-aleatórios não é possível definir distribuições verdadeiramente não informativas para o parâmetro da heterogeneidade, pelo que distribuições consideradas pouco informativas devem ser usadas. Análises de sensibilidade usando diferentes distribuições para este parâmetro devem ser efetuadas. Distribuições *a priori* informativas para o parâmetro da heterogeneidade (39) podem ser utilizadas, mas o seu impacto deve ser explorado em análises de sensibilidade.

A qualidade do ajustamento do modelo aos dados deve ser avaliada. Isto é particularmente importante quando uma análise de efeito-fixo é usada, já que é essencial validar o pressuposto de efeito comum dos estudos incluídos, mas deve ser efetuado também para modelos de efeitos-aleatórios. Um modelo com falta de ajustamento aos dados incluídos, indica a possível falha de um, ou mais, dos pressupostos inerentes à síntese (por exemplo, excesso de heterogeneidade ou inconsistência na rede de evidência). O método de avaliação de ajustamento a usar depende do tipo de modelo utilizado e deve ser justificado tendo em conta o método de síntese utilizado. Em geral métodos de análise de resíduos ou análise de desviância são recomendados. O critério de informação de Akaike (*Akaike's information criterion, AIC*) ou o critério de informação de desviância (*deviance information criterion, DIC*) podem ser usados para comparação de modelos, respetivamente para métodos de síntese frequentista ou Bayesiana.

Os modelos de meta-análise em rede pressupõem a homogeneidade e consistência das estimativas provenientes de evidência direta e indireta. Quando a rede contém ciclos isto significa que há evidência direta e indireta para as comparações envolvidas nesse ciclo. Nestes casos o pressuposto de consistência pode, e deve ser avaliado estatisticamente.

Há métodos para avaliar a consistência entre evidência direta e indireta num ciclo de cada vez (localmente) ou na totalidade da rede. O método mais apropriado para avaliar a consistência depende da estrutura da rede e do número de ciclos, devendo ser determinado caso a caso:

- para avaliação de consistência localmente, o método Bucher (40) é adequado para redes com ciclos independentes que sejam estimadas com modelo de efeito-fixo, e o método de *node-splitting* é adequado para redes mais complexas (41) (42);
- para avaliação de inconsistência global podem ser usados modelos de inconsistência (*inconsistency models*) (43) (44);
- a avaliação da qualidade do ajustamento do modelo usado para avaliar consistência à evidência deve ser inspecionada e comparada com o modelo de meta-análise em rede original;
- em modelos de efeitos aleatórios o parâmetro da heterogeneidade deve ser avaliado. Uma redução na sua estimativa no modelo de inconsistência em relação à meta-análise em rede original, é informativo e sugere a existência de inconsistência entre a evidência direta e indireta.

Em caso de ser detetada inconsistência entre a evidência direta e indireta, a inclusão de todos os estudos deve ser revista para assegurar que cumprem os requisitos da revisão sistemática e são relevantes; os dados extraídos e incluídos no modelo devem ser verificados para excluir a possibilidade de erro; e a existência de risco de viés ou a presença de variáveis modificadoras de efeito deve ser explorada. Métodos usados para explicar a heterogeneidade entre estudos podem ser também usados para explicar a inconsistência entre evidência direta e indireta, por exemplo métodos de meta-regressão e consideração de subgrupos (ver secções 4.3.5 e 4.4.4).

Resultados estimados em redes com inconsistência têm um reduzido nível de confiança e estão sujeitos a viés. Em alguns casos é possível isolar partes da rede que não são afetadas pela inconsistência pelo que algumas comparações podem ser de maior qualidade – isto depende da estrutura da rede e dos resultados da análise de inconsistência pelo que deve ser investigado e justificado em cada caso.

Por exemplo, se uma comparação na rede é formada unicamente por um (ou mais) estudo(s) de pequena dimensão com zero (ou 100%) eventos num dos braços, efeitos relativos extremos e implausíveis podem ser estimados pela evidência direta nessa comparação. No entanto, a evidência indireta pode estimar efeitos relativos mais realistas se for baseada em estudos maiores e com mais (ou menos) eventos. Neste caso pode ser detetada inconsistência quando a evidência direta é comparada com a evidência indireta sendo esta somente causada pela dimensão extrema dos resultados diretos (ou seja, ambos os tipos de evidência demonstram efeitos na mesma direção, mas a dimensão do efeito direto é implausível). Nestes casos a aceitação do efeito relativo estimado pela na meta-análise em rede pode ser considerada credível, se for adequadamente justificada.

Se a estrutura da rede incluir uma sub-rede onde a inconsistência está localizada, mas esta sub-rede não inclui as comparações de interesse para a avaliação, essa inconsistência pode ser ignorada se a restante rede não apresentar evidência de inconsistência. Neste caso a análise principal deve incluir unicamente a sub-rede consistente com a rede completa incluída apenas como análise de sensibilidade (num modelo de efeitos-aleatórios a rede completa pode aumentar a precisão dos resultados permitir mais precisão na estimação da heterogeneidade).

Os resultados completos da meta-análise em rede devem ser reportados (31) (45) (46) , incluindo:

- diagrama com a estrutura da rede e tabela incluindo os dados usados na síntese;
- tabela com todos os efeitos relativos calculados e seus intervalos de confiança, acompanhado de um gráfico em floresta com estes resultados (se possível);
- medida de heterogeneidade e seu intervalo de confiança;

- medidas de ordenação das intervenções e suas medidas incerteza (de notar que a probabilidade de uma intervenção ser “a melhor” ou a medida de SUCRA (47) não são suficientes para caracterizar a incerteza na ordenação das intervenções, por isso o posto (*rank*) de cada intervenção e seu intervalo de confiança devem ser também apresentados);

- detalhes completos do modelo estatístico utilizado, incluindo *software* usado, qual o tratamento de referência, qualidade do ajustamento do modelo e distribuições *a priori* usadas em modelos Bayesianos.

4.4.4. Meta-regressão e ajustamento de viés

Elevada heterogeneidade (clínica ou estatística) entre os estudos incluídos, indica a presença de variáveis modificadoras de efeito que interagem com o efeito do tratamento. Estas variáveis podem refletir dois tipos de variação: variação clínica entre os efeitos dos tratamentos devido à variabilidade de populações, protocolos ou contexto nos estudos incluídos; ou variação devido a diferente qualidade dos estudos e seu risco de viés. Estudos classificados como tendo elevado risco de viés devem ser excluídos da rede e só estudos com baixo risco de viés devem ser incluídos na análise principal (48). A inclusão de estudos adicionais pode ser apresentada numa análise de sensibilidade.

Métodos de meta-regressão podem ser usados para obter resultados ajustados para variáveis modificadoras de efeito observáveis. O seu exemplo mais simples é a análise de subgrupos, mas variáveis contínuas, por exemplo o nível de risco basal, também podem ser consideradas (ver secção 4.3.5).

Numa meta-análise em rede, o excesso de variabilidade devido a variáveis modificadoras de efeito pode causar tanto heterogeneidade como inconsistência. Métodos de meta-regressão podem ser usados para explicar (e eliminar) heterogeneidade e inconsistência nos resultados (49,50)

Embora as razões para heterogeneidade sejam equivalentes na meta-análise em rede e convencional, devido à inclusão de um maior número de estudos e intervenções que podem abranger um horizonte temporal maior, a meta-análise em rede pode potencialmente incluir estudos mais heterogéneos, por exemplo em termos do risco basal absoluto dos doentes incluídos, o que pode ser um importante modificador de efeito. A investigação de potencial variação de riscos absolutos nos estudos incluídos, apesar de não ser conclusiva, é uma indicação de potencial heterogeneidade ou inconsistência entre evidência direta e indireta. Nesse caso, deve ser avaliado se o nível de risco basal é um potencial modificador de efeito (51).

Vários modelos de meta-regressão são possíveis numa meta-análise em rede (51)(52). Os modelos que assumem uma interação comum a todas as comparações são os mais relevantes para a avaliação farmacoterapêutica, desde que tenham validade clínica. No entanto, os seus resultados são considerados observacionais e só devem ser usados para análises exploratórias ou de sensibilidade.

4.5 Conclusões

Quando a evidência disponível incluiu dois ou mais estudos os resultados desses estudos podem ser combinados utilizando técnicas meta-analíticas, para gerar uma estimativa combinada da eficácia relativa das intervenções com maior poder estatístico e precisão.

A meta-análise convencional permite comparar dois tratamentos investigados em múltiplos estudos entre si. A meta-análise em rede permite a comparação de múltiplas intervenções comparadas em múltiplos estudos, desde que formem uma rede ligada. O método de escolha dos tratamentos a incluir

na rede deve ser especificado previamente, ser reproduzível e garantir uma rede única de tratamentos a comparar.

Os métodos de meta-análise em rede são particularmente relevantes nos casos em que na prática clínica corrente se utilizam vários tratamentos para a mesma indicação, e em que, por conseguinte, podem existir vários comparadores selecionados para a avaliação. Estes métodos devem ser preferidos, desde que seja possível formar uma rede de tratamentos ligada, baseada em estudos aleatorizados relevantes e sem elevado risco de viés. No caso de a rede estar desligada, o critério de inclusão pode ser estendido para incluir estudos aleatorizados que comparem tratamentos adicionais com um (ou mais) dos tratamentos em avaliação, que possam ligar a rede.

Só devem ser realizadas meta-análises convencionais ou em rede quando os estudos disponíveis forem suficientemente homogéneos, ou seja, comparáveis (não diferindo substancialmente em características que possam alterar os efeitos relativos dos tratamentos) e de qualidade, de modo que os resultados obtidos possam ser confiáveis.

5 MÉTODOS DE COMPARAÇÃO EM SITUAÇÕES EXCECIONAIS

5.1 Comparação indireta ajustada ancorada (MAIC, STC)

Qualquer meta-análise, simples ou em rede, tem como pressuposto principal que não há diferenças na distribuição de variáveis modificadoras de efeito nos estudos incluídos. No entanto, este pressuposto nem sempre é verificado, nomeadamente quando há um elevado nível de heterogeneidade clínica ou estatística entre os estudos. Comparações indiretas, e meta-análises que incluam poucos estudos, são particularmente vulneráveis à existência destas diferenças.

A utilização de modelos que ajustem resultados com base nas variáveis modificadoras de efeito, pode produzir efeitos relativos mais relevantes e credíveis. A meta-análise (simples ou em rede) com meta-regressão usando dados individuais dos participantes de todos os estudos é o método preferível (ver secção 4.3.6).

No contexto da avaliação farmacoterapêutica, o TAIM que submete o dossier, só tem normalmente acesso aos dados individuais dos seus estudos. Os métodos de comparações indiretas ajustadas por correspondência (*matching adjusted indirect comparisons*, MAIC) (53) e comparações de tratamentos simulados (*simulated treatment comparisons*, STC) (54) foram desenvolvidos para lidar com situações em que:

- uma comparação indireta entre dois tratamentos é necessária;
- há diferenças em uma ou mais características modificadoras de efeito entre a população dos estudos que irão formar a comparação indireta;
- a empresa tem acesso aos dados individuais do seu estudo, mas não dos outros estudos.

O método MAIC usa ponderação pelo inverso do *score* de propensão (*inverse propensity score*) para ponderar o efeito dos tratamentos usados na população para a qual os dados individuais estão disponíveis, para o efeito que seria observado na população do estudo para o qual os dados individuais não estão disponíveis. Métodos tipicamente usados para esta ponderação dão resultados equivalentes (55). O método STC usa regressão para ajustar o efeito do tratamento na população para a qual os dados individuais estão disponíveis, para o efeito que seria observado na população do estudo para o qual os dados individuais não estão disponíveis. Amostras aleatórias da distribuição conjunta das covariáveis no estudo com dados agregados são usadas para calcular o efeito previsto nessa população, usando um modelo de regressão. No entanto, estes métodos:

- em geral não efetuam comparações na escala de efeitos relativos que seria preferida numa meta-análise convencional (simples ou em rede). Recomenda-se que as comparações sejam efetuadas na escala escolhida para os efeitos relativos (56,57).
- produzem efeitos relativos aplicáveis à população de um dos estudos (o estudo sem dados individuais disponíveis), mas não garantem que os resultados sejam aplicáveis à população mais relevante para a avaliação (a população definida no PICO).
- assumem uma distribuição para os modificadores de efeito no estudo comparador baseado somente em sumários estatísticos descritos em publicações.
- só é possível proceder ao ajustamento de variáveis com sumários descritos nas publicações.

De notar que, como o ajustamento é efetuado com base em comparações aleatorizadas, não é necessário (e é até desaconselhado) proceder ao ajustamento de variáveis meramente prognósticas.

Além disso, o método MAIC só pode ser aplicado quando existe suficiente sobreposição nas distribuições das variáveis na população do estudo com dados individuais e o estudo comparador. Quando há pouca sobreposição o método não produz resultados credíveis (56,57). No entanto, quando há larga sobreposição nas distribuições das variáveis na população do estudo com dados individuais e o estudo comparador, não é necessário usar métodos de ajustamento já que se espera que as populações sejam comparáveis.

A utilização de simulação no método STC introduz variação adicional. O efeito estimado não é o efeito médio na população com dados agregados, mas o efeito previsto num individuo selecionado aleatoriamente dessa população (isto é, da distribuição preditiva), o que leva à sobrestimação da incerteza na comparação indireta final.

A necessidade da utilização de MAIC ou STC deve ser justificada com referência às características dos estudos incluídos e à evidência existente para suportar a modificação de efeito. Este tipo de análise utiliza métodos e pressupostos que representam um desvio em relação aos métodos tipicamente usados na avaliação farmacoterapêutica, pelo que devem ser consideradas de menor credibilidade do que a meta-análise (simples ou em rede) baseada em estudos aleatorizados sem evidência de modificadores de efeito. De notar que as estimativas obtidas são referentes à população do estudo comparador. Será necessário comparar essa população com a população em estudo.

Os métodos MAIC e STC não se estendem à situação em que há múltiplas comparações indiretas possíveis usando comparadores diferentes e os seus pressupostos são difíceis de validar (56,57). Há por isso a possibilidade de heterogeneidade de resultados em avaliações de produtos diferentes para a mesma área terapêutica, que derivam da escolha dos estudos usados para o ajustamento e que tinham dados individuais disponíveis em cada circunstância.

O método de meta-regressão em rede multiníveis (*multi-level network meta-regression*, ML-NMR) (58) é uma alternativa que permite ajustar os efeitos de uma rede de tratamentos à população em estudo, evita o risco de viés por agregação e produz resultados diretamente interpretáveis apesar de ter uma implementação mais complexa. A escolha da população para a qual os efeitos são ajustados deve ser justificada adequadamente.

5.2 Uso de estudos não aleatorizados

Quando há evidencia de qualidade proveniente de estudos comparativos aleatorizados, evidência não aleatorizada pode ser utilizada para complementar a evidência dos estudos aleatorizados, por exemplo para validar a sua aplicação ao contexto Português, mas não a substitui.

No entanto, em situações excecionais pode ser necessário, por falta de estudos aleatorizados, considerar evidência não aleatorizada, ou em “contexto real”. De notar que este tipo de evidência sofre de elevado risco de viés. No entanto, estudos não aleatorizados poderão ser aceites para informar parâmetros específicos (por exemplo, segurança a longo prazo), desde que baseada no tipo de estudo mais adequado para o objetivo e que minimize o risco de viés do resultado (por exemplo: historial de estudos controlados, devidamente ajustados). Este tipo de evidência pode ser usado para definir a história clínica do doente na ausência de tratamento, para comparar dados de segurança e para informar parâmetros de eficácia em casos em que estudos aleatorizados sejam manifestamente impossíveis de realizar, por exemplo no caso de doenças raras ou ultra-raras.

Os métodos MAIC e STC também podem ser usados para efetuar comparações entre estudos de um só braço ou para ligar redes desligadas. Dado o elevado potencial de viés, a utilização de métodos de ajustamento de populações para ligar redes desligadas e para incluir dados de estudos não comparativos (só com um braço), não é recomendada exceto em situações excecionais que devem ser detalhadamente justificadas (56,57). Em particular, métodos de comparação indireta não ancorada não devem ser usados quando uma comparação ancorada é possível já que têm um maior potencial para viés e menos precisão do que as comparações ancoradas.

Destas situações excecionais destacam-se:

- doenças raras, definidas por uma prevalência inferior a cinco em 10 000 pessoas, em que não existam alternativas terapêuticas, ou em que o efeito dessas alternativas é não provado ou incerto, ou em que o tratamento inclui medicamentos de uso bem estabelecido.

- doenças ultra-raras, definidas como uma doença com uma prevalência ≤ a um doente por 100.000 pessoas.

Nestes casos, quando não exista evidência de estudos aleatorizados, considera-se aceitável como demonstração de prova de benefício adicional, a utilização de comparações indiretas ajustadas (MAIC e STC) não ancoradas, de estudos só com um braço.

Deverão ser utilizados simultaneamente a MAIC e a STC. A demonstração de prova de benefício adicional implicará que os dois métodos (MAIC e STC) deem resultados concordantes. A utilização de métodos alternativos deverá ser adequadamente justificada.

Uma comparação não ancorada assume que o efeito absoluto da intervenção pode ser previsto com base nas características da população, ou seja pressupõe que todas as variáveis prognósticas e modificadoras de efeito são incluídas no modelo de previsão. Este pressuposto é mais forte do que o pressuposto usado nas comparações ancoradas nas quais não é necessário considerar as variáveis prognósticas, e praticamente impossível de se verificar. A falha deste pressuposto implica um nível de viés nas comparações efetuadas de difícil quantificação. Quando medidas de efeito baseadas em comparações não ancoradas são usadas, é necessário demonstrar a dimensão de erro plausível devido à falta de inclusão de variáveis no ajustamento, no efeito relativo estimado (56).

6 ANÁLISE DE SUBGRUPOS

6.1 Introdução

Os doentes numa indicação podem variar em características que afetam a magnitude dos benefícios da nova tecnologia de saúde, bem como os custos associados com o seu tratamento. Esta variação nas características é conhecida na literatura como heterogeneidade. A heterogeneidade pode influenciar a escolha do tratamento, dado que é possível selecionar o tratamento que mais beneficia o doente (ou que seja mais custo-efetivo) dado as suas características.

A existência de heterogeneidade pode ter razões variadas, sendo as mais frequentes a heterogeneidade no efeito relativo do tratamento (ou seja, modificação do efeito terapêutico) e a heterogeneidade no risco basal, tal como no risco de progressão ou no risco de eventos. Podem ainda existir situações em que o risco basal está correlacionado com o efeito relativo de tratamento (59).

Na maioria dos ensaios clínicos e revisões sistemáticas, os efeitos de tratamento não são homogéneos em toda a população incluída. As análises de subgrupo permitem avaliar essas diferenças na resposta ao tratamento, bem como outras fontes de heterogeneidade, de modo a possibilitar uma maior personalização nas decisões em saúde.

Estas recomendações têm como objetivo informar a avaliação farmacoterapêutica, e não impedem a avaliação de outros subgrupos na avaliação farmacoeconómica, em linha com as recomendações das orientações farmacoeconómicas.

6.2 Definição/ especificação de subgrupos

A especificação de subpopulações deve obedecer aos critérios estabelecidos pela matriz de avaliação (PICO) (ver secção 2.2).

Idealmente, estas subpopulações identificadas na matriz inicial seriam avaliadas em estudos separados, ou estudos desenhados para ter poder estatístico adequado para estudar as subpopulações dentro de um mesmo estudo. Contudo, por vezes estas subpopulações são incluídas num mesmo estudo, sendo realizado pelo titular de AIM e/ou equipa de investigação uma análise de subgrupos no sentido de detetar diferentes efeitos de tratamento. Quando isto acontece, é importante realizar uma avaliação da credibilidade da análise de subgrupos na evidência submetida (ver secção 6.4).

Deve propor-se a separação em subpopulações se existem características que sejam potenciais modificadores de efeito de tratamento, de preferência documentada em estudos anteriores na mesma patologia (termo de interação). Qualquer definição de uma subpopulação que não esteja prevista duma forma explícita na indicação aprovada deverá ser justificada pelo proponente desta subpopulação – seja a CATS ou o requerente. Se houver dúvidas sobre uma potencial modificação de efeito que não tenha sido estudada previamente, pode optar-se por anotar em nota de rodapé na tabela da matriz inicial de avaliação o efeito de subgrupo que se pretende verificar.

Não se deve propor a separação da população em subgrupos apenas por existirem características clínicas ou fatores de prognóstico diferente, se isso não influenciar previsivelmente o efeito do tratamento. A divisão em subgrupos apenas por heterogeneidade clínica pode até levantar questões de equidade e/ou éticas. Por exemplo, a utilização de idade ou classe etária pode ser apropriada se o efeito da idade é modificador do efeito do tratamento (por exemplo, tratamento menos eficaz em doen-

tes mais idosos) ou progressão da doença. Por outro lado, se a idade não reflete um efeito no tratamento (ou doença) o seu uso pode não ser equitativo e tais considerações devem ser explicitamente feitas aquando da definição dos subgrupos.

As análises de subgrupos possuem limitações metodológicas importantes e frequentemente não cumprem os critérios metodológicos necessários, levando a resultados erróneos (60). Ensaios clínicos aleatorizados são o tipo de evidência preferencial para a identificação de modificadores do efeito do tratamento. Evidência não aleatorizada (por exemplo, estudos observacionais longitudinais de grande dimensão), pode, no entanto, ser apropriada para a identificação de outros tipos de subgrupos acima referidos, como é o caso de informação de heterogeneidade de risco basal e prognóstico. Independentemente do desenho do estudo, o tamanho da amostra nos subgrupos é frequentemente reduzido, e sem poder estatístico para detetar diferenças entre grupos da mesma população, mesmo que estas diferenças existam.

6.3 Recomendações para a análise de subgrupos - perspetiva do titular de AIM

É recomendado o uso de um conjunto de regras na análise de subgrupos, na perspetiva do titular de AIM:

- as análises de subgrupos devem ser definidas antes do estudo iniciar e devem ser limitadas a um pequeno número de questões clinicamente relevantes;
- o protocolo de estudo deve incluir informação sobre a forma como os subgrupos foram selecionados, e sobre qual o motivo por que foram selecionados;
- as definições e as categorias exatas das variáveis de subgrupo devem ser definidas explicitamente à partida. Para variáveis contínuas ou categóricas, os limiares para análise devem estar pré-definidos;
- a direção e a magnitude do efeito esperado no subgrupo devem ser definidas *a priori*;
- no desenho do estudo, deve ser considerada a possibilidade de estratificação da aleatorização por variáveis de subgrupos importantes;
- no caso de se preverem importantes interações subgrupo - efeito do tratamento, o estudo deve ter poder estatístico suficiente para detetar de forma confiável essas interações;
- as regras de interrupção do estudo devem ter em conta as interações esperadas subgrupo - efeito do tratamento e não apenas no efeito global do tratamento;
- se for provável que o efeito relativo do tratamento esteja relacionado com o risco basal, o plano de análise deve incluir uma estratificação dos resultados em função do risco previsível. O modelo ou *risk score* deve ser pré-selecionado, de forma que os dados basais relevantes sejam registados;
- o significado do efeito do tratamento em subgrupos individuais não deve ser relatado, uma vez que as percentagens de falsos positivos e falsos negativos são extremamente elevadas. A única abordagem estatística confiável é testar a interação subgrupos - efeito do tratamento, ou seja, a análise correta não é a significância estatística do efeito do tratamento num ou outro subgrupo particular, mas se o efeito diferiu significativamente entre subgrupos (teste de interação subgrupo - efeito do tratamento). Em termos epidemiológicos, interação significa modificação de efeito.

Importante notar que o teste de interação deve utilizar as medidas relativas de efeito (risco relativo ou razão de riscos ou razão de chances) e não redução absoluta de risco. Isto porque a propriedade intrínseca do tratamento é representada pelas medidas relativas, que tendem a ser constantes nos diferentes estratos de risco absoluto;

- a significância estatística das interações efeito do tratamento - subgrupos deve ser adequadamente ajustada quando são feitas múltiplas análises de subgrupos;
- as análises de subgrupos devem ser relatadas como reduções de risco relativo e reduções de risco absoluto.
- idealmente, apenas deve ser estudado uma medida de resultado, de preferência a medida de resultado primário do estudo;
- a comparabilidade dos fatores prognósticos entre grupos de tratamento deve ser confirmada nos subgrupos;
- no caso de serem identificadas múltiplas interações subgrupos - efeito do tratamento, são necessárias análises adicionais que verifiquem se os seus efeitos são independentes;
- descrições da significância estatística do efeito do tratamento em subgrupos individuais devem ser ignoradas, sobretudo relatos de ausência de benefício num subgrupo particular num estudo em que se observou benefício global, a não ser que exista interação significativa subgrupo - efeito do tratamento;
- interações subgrupo - efeito do tratamento genuínas não esperadas, são raras. Por isso, as interações aparentes que são descobertas *post hoc* devem ser interpretadas com cuidado. Neste caso, nenhum teste de significância é confiável;
- análises de subgrupos *pre-hoc* não são intrinsecamente válidas e devem ser interpretadas com cuidado. A probabilidade de falsos positivos aumenta com o número de testes e pode ser avaliada pela fórmula 1-(1-p)^c, onde p é o nível de significância e c o número de testes;
- o melhor teste de validade de interações subgrupos - efeito do tratamento é a sua reprodutibilidade em outros estudos;
- poucos estudos têm poder estatístico para detetar efeitos nos subgrupos, pelo que a percentagem de testes de interação falsos negativos é elevada. Se existir uma interação genuína subgrupo – efeito do tratamento, a probabilidade de um resultado falso negativo com um teste formal de interação será muito superior aos 5% de falsos positivos observados num estudo em que não existe uma verdadeira interação;
- a incerteza nos subgrupos identificados deve ser devidamente quantificada e expressa de forma adequada (por exemplo: intervalo de confiança, desvio padrão). Uma quantificação adequada da incerteza numa análise de subgrupos é usualmente conseguida via análise de dados ao nível dos doentes. Modelação formal (por exemplo: via modelos de regressão, meta-regressão) facilita estabelecimento de interações subgrupo – efeito do tratamento, com estimação da incerteza de parâmetros e entre parâmetros.

6.4 Avaliação e classificação da credibilidade das análises de subgrupos

Face às inúmeras limitações inerentes à análise de subgrupos, torna-se fundamental avaliar criticamente as análises de subgrupos de forma a inferir o seu grau de credibilidade. Desta forma, recomenda-se avaliar a credibilidade das análises de subgrupos em duas etapas:

- análise do cumprimento dos critérios que permitem avaliar a credibilidade de uma análise de subgrupos;
- classificação do grau de credibilidade da análise desde o 'extremamente improvável' até 'extremamente plausível'.

Tabela 2: *Critérios para avaliar a credibilidade da análise de subgrupos*

Critério a avaliar	**S/N/NC**
Desenho	
1. A variável do subgrupo é uma característica medida após aleatorização ou no início do estudo?	
2. O efeito é sugerido por comparações dentro do estudo mais do que entre estudos?	
3. A hipótese foi especificada *a priori*?	
4. Foi testado um pequeno número de hipóteses?	
5. A direção do efeito no subgrupo foi especificada *a priori*?	
Análise	
6. O teste para interação sugere uma baixa probabilidade de o efeito aparente do subgrupo ser explicado por acaso?	
7. O efeito do subgrupo é independente?	
Contexto	
8. A magnitude do efeito do subgrupo é grande?	
9. A interação é consistente entre os estudos?	
10. A interação é consistente nas medidas de resultados estreitamente relacionados do estudo?	
11. Existe evidência indireta que suporte a hipotética interação (racional biológico)?	

Legenda: S=Sim; N=Não; NC=Não conhecido

Sugere-se a verificação dos 11 critérios constantes na Tabela 2 para avaliar a credibilidade de uma análise de subgrupos. De notar que a avaliação da credibilidade da análise de subgrupos não é uma questão dicotómica, mas uma avaliação que resulta num espectro de credibilidade desde o 'extremamente improvável' até 'extremamente plausível'. Sugere-se a adesão à valorização da credibilidade considerada pelo "User's Guide to the Medical Literature" (61) (62) considerando que os critérios 1, 3 e 9 são os mais relevantes para uma correta avaliação da credibilidade da análise de subgrupos.

Se a análise de subgrupos não for credível, sugere-se analisar a população total do ensaio, se houver confiança que a característica do subgrupo estudado não parecer ser um modificador do efeito de tratamento. Se esta análise não for viável ou se houver dúvidas sobre uma potencial modificação do efeito, não será possível avaliar o benefício adicional do fármaco na subpopulação em questão. Esta conclusão pode ou não levar a uma restrição da indicação em avaliação.

6.5 Conclusões

- Os doentes numa indicação podem variar em características que afetam a magnitude dos benefícios do medicamento novo, bem como os custos associados com o seu tratamento;
- Na perspetiva do titular de AIM, devem ser seguidas regras para definir análises de subgrupos de forma a apresentar análises mais credíveis;

- Na perspetiva do grupo de avaliação de evidência, a análise de subgrupos submetida deve ser analisada e classificada quanto à sua credibilidade, de acordo com os 11 critérios definidos na Tabela 2, num *continuum* entre extremamente improvável' e 'altamente plausível'.
- As análises de subgrupos são análises exploratórias de identificação de modificadores de efeito, ou de sensibilidade dos resultados, condicionadas pelo seu grau de credibilidade.

7 ASPECTOS PARTICULARES NA AVALIAÇÃO DE BENEFÍCIO

7.1 Impacto dos resultados de estudos não publicados nas conclusões

Um pré-requisito essencial para a validade de uma avaliação de benefício é a disponibilidade completa dos resultados dos estudos realizados sobre um tópico. Uma avaliação baseada em dados incompletos ou possivelmente até dados compilados seletivamente pode produzir resultados enviesados.

Adicionalmente, os vieses resultantes do viés de publicação e viés de relatório de resultados foram descritos de forma abrangente na literatura. Para minimizar as consequências deste problema, recomenda-se que a pesquisa de informação, para além de incluir uma pesquisa em bancos de dados bibliográficos, deverá também incluir, por exemplo, pesquisa de plataformas internacionais de registros de ensaios (ver secção 3.4).

7.2 Efeito dramático

Se o curso de uma doença é certamente ou quase certamente previsível, e nenhuma opção de tratamento está disponível para influenciar este curso, a prova de um benefício de uma intervenção médica pode também ser proporcionado pela observação de uma reversão do curso (mais ou menos) determinístico da doença em séries de casos bem documentadas de doentes. Se, por exemplo, se souber que é altamente provável que uma doença leve à morte dentro de um curto período de tempo após o diagnóstico, e é descrito em uma série de casos que, após a aplicação de uma intervenção específica, a maioria dos afetados sobrevive por um longo período de tempo, esse “efeito dramático” pode ser suficiente para fornecer prova de um benefício. Um exemplo desse efeito é a substituição de hormonas vitais em doenças com falta na produção hormonal (por exemplo, terapia com insulina em doentes com diabetes *mellitus* Tipo 1). Um pré-requisito essencial para a classificação como um "efeito dramático" é a documentação suficientemente confiável do curso fatídico da doença na literatura e do seu diagnóstico nos doentes incluídos no estudo a ser avaliado. Nesse contexto, possíveis danos da intervenção também devem ser levados em consideração. Dados empíricos sugerem que um risco relativo observado de cinco a dez não pode ser explicado apenas por fatores de confusão. Se, no período que antecede a avaliação, houver informações suficientes disponíveis indicando que um efeito dramático causado pela intervenção a ser avaliada pode ser esperado (por exemplo, devido a uma pesquisa preliminar da literatura), a avaliação deverá incluir os estudos que demonstrem maior certeza nos resultados devido ao seu desenho.

7.3 Duração do estudo

A duração do estudo é um critério essencial na seleção de estudos relevantes para a avaliação de benefício. Na avaliação de uma intervenção terapêutica para doenças agudas, onde o objetivo principal é, por exemplo, reduzir a duração da doença e aliviar os sintomas agudos, não tem sentido exigir estudos de longo prazo, a menos que sejam esperadas complicações tardias. Por outro lado, na avaliação de intervenções terapêuticas para doenças crónicas, os estudos de curto prazo geralmente não são adequados para obter uma avaliação completa dos benefícios da intervenção. Isso aplica-se sobretudo se o tratamento for necessário por vários anos, ou mesmo vitalício. Nesses casos, estudos que abrangem um período de tratamento de vários anos são particularmente relevantes e desejáveis. Como benefícios e danos podem ser distribuídos de maneira diferente ao longo do tempo em intervenções de longo prazo, a comparação dos benefícios e malefícios de uma intervenção só é possível com certeza suficiente se forem realizados estudos de duração suficiente.

8 ESTUDOS DE SUPERIORIDADE, NÃO INFERIORIDADE, E EQUIVALÊNCIA: DEFINIÇÕES E CRITÉRIOS PARA MUDANÇA DE OBJETIVOS

8.1 Introdução

A evidência de eficácia pode ser obtida a partir de diferentes tipos de ensaios controlados. Os ensaios de superioridade procuram mostrar que uma intervenção é superior ao controlo (placebo, ausência de tratamento, menor dose da intervenção). Outro tipo de ensaios são os que comparam a intervenção com um tratamento ativo (controlo ativo). Apesar deste tipo de ensaios também poder ter como objetivo demonstrar superioridade, é frequente que o seu objetivo seja mostrar que a diferença entre o novo tratamento e o controlo ativo é pequena e que, com base no seu desempenho em estudos prévios e na eficácia assumida do controlo ativo no estudo atual, é possível concluir que a nova intervenção é também efetiva. No entanto, o desenho e a interpretação dos resultados destes estudos coloca desafios específicos, pelo que se torna necessário estabelecer algumas considerações sobre esta matéria.

8.2 Demonstração de equivalência

Um dos erros graves mais frequentes na interpretação de dados médicos é classificar um resultado não significativo de um teste de significância como evidência de que a hipótese nula é verdade.

Para demonstrar “equivalência” é necessário utilizar métodos que permitam testar a hipótese de equivalência, ou seja, o estudo deve ter um desenho de equivalência que permita confirmar a ausência de uma diferença significativa entre tratamentos (por exemplo, que o valor médio da diferença entre 2 grupos é exatamente zero).

Este objetivo obtém-se através do cálculo e observação dos intervalos de confiança, uma vez que não é possível o uso de testes estatísticos. Na altura da elaboração do protocolo é necessário definir uma margem (Δ) de equivalência clínica, através da definição da maior diferença que é clinicamente aceitável, de tal modo que uma diferença maior seria relevante na prática clínica. Os dois tratamentos são considerados equivalentes se o intervalo de confiança a 95% (bilateral), que define o intervalo das diferenças plausíveis entre os dois tratamentos, está dentro do intervalo – Δ a + Δ. Na prática, o que se demonstra não é a existência de uma equivalência exata (que a diferença entre os valores médios dos 2 grupos é exatamente zero), mas que a diferença entre os dois grupos é irrelevante. Como nos estudos de superioridade, é necessário estimar o tamanho da amostra nos estudos de equivalência na altura da elaboração do protocolo.

No caso de estudos de bioequivalência tem sido aceite como o padrão a utilização de intervalos de confiança a 90% da diferença entre tratamentos, na avaliação dos valores médios de parâmetros farmacocinéticos. No caso de um genérico inalado ou de um produto aplicado topicamente, em que estudos de bioequivalência são impossíveis, aceita-se a realização de estudos de bioequivalência clínica, utilizando intervalos de confiança a 95%.

Quando o intervalo de confiança a 95% que define o intervalo das diferenças plausíveis entre os dois tratamentos é unilateral, o estudo é referido como de não inferioridade.

8.3 Demonstração de não inferioridade

Nos estudos de não inferioridade pretende-se demonstrar que o efeito do novo tratamento não é inferior (significando que pode ter a mesma eficácia ou ter mais eficácia) ao efeito do tratamento já existente, por um valor especificado, chamado de margem (Δ) de não inferioridade. Como referido, também aqui

se aplica a abordagem dos intervalos de confiança, mas agora apenas estamos interessados em avaliar a possível diferença numa única direção. Assim, o intervalo de confiança a 95% (bilateral) da diferença entre tratamentos deve estar completamente para a direita do valor – Δ. O teste estatístico é dado pela comparação do limite superior do intervalo de confiança (bilateral) para a comparação dos dois tratamentos com a margem previamente especificada. Se o limite superior do intervalo de confiança é inferior à margem, é estabelecida a não-inferioridade.

Como já referido, os estudos de não inferioridade são por vezes designados, erroneamente, como estudos de equivalência, representando uma fonte de confusão.

8.4 Intervalos de confiança unilaterais e bilaterais

De acordo com o *ICH E9 Note for Guidance* (63,64), deverão sempre ser usados intervalos de confiança a 95% bilaterais em todos os estudos clínicos independentemente do seu objetivo. No caso de serem utilizados intervalos de confiança unilaterais devem ser usados para cobrir uma probabilidade de 97,5%. No caso especial de estudos de bioequivalência, tem sido recomendada a utilização de intervalos de confiança a 90% bilaterais.

Uma possibilidade para a definição da margem (Δ) é estabelecer um valor igual ao efeito conhecido do tratamento existente em relação ao placebo, baseado em ensaios aleatorizados prévios. Com esta escolha de margem, e assumindo que o fármaco em avaliação atinge este nível de eficácia no estudo de não-inferioridade, a não inferioridade significa que o fármaco em teste tem um efeito superior a 0. No entanto, uma escolha usual é estabelecer que a margem (Δ) é equivalente a uma porção clinicamente relevante do efeito conhecido do tratamento existente em relação ao placebo, nomeadamente a porção do efeito do tratamento de controlo que é importante preservar no fármaco em avaliação, com base no julgamento clínico.

8.5 Estudos de superioridade

Os estudos de superioridade são desenhados para detetar uma diferença entre tratamentos, e a demonstração de superioridade faz-se através da utilização de um teste de significância estatística. O teste de significância estatística testa o efeito nulo, ou seja, coloca a hipótese de que não existem diferenças nos efeitos clínicos entre dois tratamentos. O grau de significado estatístico (valor de p) indica a probabilidade de que a diferença observada tenha ocorrido por acaso na hipótese de que, na realidade, não existe diferença entre os tratamentos. No entanto, os resultados não devem ser simplesmente reportados como tendo ou não “significado estatístico”, mas sim ser interpretados no contexto do tipo de estudo e do risco de viés que lhes está associado.

Uma vez que se considere que a hipótese de não-diferença entre tratamentos é insustentável por pouco provável (um $p < 0,05$ indica que essa probabilidade é inferior a 5%), é importante estimar a magnitude da diferença para avaliar se essa magnitude é clinicamente relevante. Com esse objetivo, é necessário calcular a melhor estimativa da magnitude da diferença entre tratamentos, habitualmente designada por estimativa pontual (*point estimate*), que, em dados com distribuição normal, corresponde à diferença nos valores médios da medida de eficácia utilizada. Deve-se também calcular o intervalo de confiança, que corresponde ao intervalo de valores plausíveis da verdadeira diferença. Este intervalo não deve incluir o efeito nulo, que é o valor zero no caso da dimensão do efeito (*effect size*) ser uma variável contínua (por exemplo, a diferença média padronizada (*standardized mean difference*), ou que é o valor um no caso dos rácios (*risk ratio, odds ratio ou hazard ratio*), uma vez que a hipótese nula (diferença zero entre tratamentos) já foi rejeitada. Assim, as seguintes afirmações são consideradas equivalentes: o intervalo de confiança a 95% da diferença entre tratamentos exclui o valor nulo e os

dois valores são estatisticamente diferentes no nível de significância de 5% bilateral *(two-sided)* (p <0,05).

Nos estudos de superioridade, avaliar se uma diferença entre tratamentos com significado estatístico é clinicamente relevante requer um juízo de valor a posteriori. Pelo contrário, nos estudos de equivalência e de não inferioridade a relevância clínica é definida previamente, na altura da elaboração do protocolo de estudo, através da definição do Δ (margem de não inferioridade [-Δ] ou margem de equivalência [±Δ]).

8.6 Relevância da pré-definição do Δ nos estudos de não inferioridade e de equivalência

A demonstração de "equivalência" ou de "não inferioridade" depende do valor Δ selecionado que deve representar a diferença máxima aceitável para o objetivo de interesse. É importante salientar que, após inspeção dos dados, é sempre possível selecionar um valor de Δ que conduz à conclusão de "equivalência" ou de "não inferioridade".

Uma vez que a escolha do Δ requer sempre julgamento clínico, existe nesta escolha sempre um risco de viés, mas esse risco aumenta exponencialmente se o Δ for selecionado após inspeção dos dados. Assim, a seleção do Δ deve ser sempre feita na altura da elaboração do protocolo de estudo, e justificada baseada em argumentos plausíveis.

8.7 Relevância da pré-definição do estudo como de superioridade, não-inferioridade ou equivalência

De acordo com as orientações estabelecidas pela Agência Europeia do medicamento (EMA) no documento *"Points to consider on switching between superiority and non-inferiority"* (65) *do "Committee for Proprietary Medicinal Products (CPMP)"*, é fundamental a pré-definição do estudo como de superioridade, de não inferioridade, ou de equivalência pelos seguintes motivos:

- assegura que os comparadores, doses dos medicamentos, populações a incluir, e medida de resultados são apropriados;
- permite que a estimativa do tamanho da amostra seja baseada em cálculos apropriados do poder estatístico;
- assegura que os critérios de equivalência ou não inferioridade são pré-definidos;
- permite que o protocolo descreva em detalhe a análise estatística apropriada;
- assegura que o estudo tem sensibilidade suficiente para atingir os seus objetivos.

8.8 É possível mudar o objetivo de uma comparação?

A mudança entre superioridade e não inferioridade é a única mudança com relevância prática. Os estudos de equivalência são tão específicos que não existe a possibilidade de mudança, seja entre equivalência e superioridade, seja entre equivalência e não inferioridade.

8.9 Interpretação um estudo de não-inferioridade como um estudo de superioridade

Nos casos em que todo o intervalo de confiança a 95% da diferença entre tratamentos está não apenas para a direita de – Δ mas também para a direita (acima de) do valor nulo (zero no caso das variáveis contínuas, e um no caso dos rácios), existe evidência de superioridade em termos de significado estatístico ao nível de 5% (p <0,05). Neste caso é aceitável calcular o valor de p associado a um teste de superioridade (teste estatístico de significância) e avaliar se tem significado estatístico. A interpretação deste teste não é afetada pelo problema de multiplicidade (erro de tipo I), uma vez que corresponde a um único teste estatístico de significância.

Assim, de acordo com o CPMP da EMA é possível mudar o objetivo de um estudo de não inferioridade para um estudo de superioridade desde que:

- o estudo tenha sido desenhado e conduzido de acordo com os requisitos de um estudo de não inferioridade;
- sejam apresentados os valores de p para a superioridade;
- o estudo tenha sido analisado de acordo com o princípio intenção de tratar.

8.10 Interpretação um estudo de superioridade como um estudo de não-inferioridade

Se um estudo de superioridade não mostra uma diferença estatisticamente significativa entre tratamentos, poderia haver interesse num objetivo menor de estabelecer não inferioridade. Se a diferença entre tratamentos num estudo de superioridade é apresentada sob a forma de intervalo de confiança, o limite inferior do intervalo de confiança fornece uma estimativa do efeito mínimo do novo tratamento em relação ao comparador. Se o protocolo de estudo define uma margem de não inferioridade considerada como aceitável, mudar o objetivo do estudo de superioridade para não inferioridade é possível e aceitável.

Nos estudos de superioridade em que a margem de não inferioridade não foi definida na altura da elaboração do protocolo, não é aceitável mudar o objetivo do estudo, uma vez que, após inspeção dos dados, é sempre possível selecionar um valor de Δ que conduz à conclusão de equivalência ou de não inferioridade. Assim, uma definição *post hoc* do Δ está associado a um risco de viés elevado pelo que não é aceitável.

Assim, um estudo de superioridade em que o teste de significância para a diferença entre tratamentos não mostrou significado estatístico, caso não tenha definido a margem de não inferioridade na altura de elaboração do protocolo, deve ser considerado apenas como um estudo de superioridade negativo, não sendo aceitável mudar o objetivo de superioridade para não inferioridade.

No entanto, a mudança do objetivo de um estudo de superioridade para não-inferioridade pode ser exequível desde que sejam cumpridos os seguintes requisitos, conforme recomendado pela EMA:

- a margem de não inferioridade em relação ao tratamento controlo foi pré-definida ou pode ser justificada;
- as análises de acordo com o princípio intenção de tratar e conforme protocolo, com intervalos de confiança e valores-p para a hipótese nula de inferioridade, mostram resultados semelhantes;

- o ensaio foi desenhado adequadamente e conduzido de acordo com os requisitos para ensaios de não inferioridade;
- s sensibilidade do ensaio é suficientemente elevada para assegurar capacidade para detetar diferenças relevantes, caso estas existam;
- existe evidência direta e indireta que o tratamento controlo demonstra o seu habitual nível de eficácia.

9 AVALIAÇÃO DA QUALIDADE DA EVIDÊNCIA

9.1 Avaliação do risco de viés por estudo

A avaliação da qualidade da evidência deve iniciar-se pela avaliação do risco de viés de cada um dos estudos incluídos na avaliação da intervenção. Para a avaliação do risco de viés de cada estudo são utilizados seis domínios: geração de sequência, ocultação da alocação, ocultação de participantes, investigadores e adjudicadores de medidas de resultado, dados incompletos de medidas de resultado, reporte seletivo de medidas de resultado e outras fontes de viés.

A classificação do risco de viés de cada estudo deve ser explicada de forma detalhada para cada um dos seis domínios referidos anteriormente e, para o conjunto de estudos avaliados, recomenda-se que seja sintetizada utilizando uma tabela e/ou uma figura com um gráfico de barras (66).

9.2 Avaliação da qualidade da evidência (certeza da evidência) na meta-análise convencional

A metodologia aqui descrita aplica-se essencialmente no contexto de uma meta-análise entre pares, não sendo em geral aplicável no contexto de meta-análise em rede. Contudo, em situações de redes simples, como é o caso do método de Bucher (40), esta metodologia pode igualmente ser aplicada.

A qualidade da evidência de cada comparação deve ser avaliada, para cada medida de resultado, mas a classificação final deve referir-se ao conjunto de estudos avaliados. A “qualidade” reflete a nossa confiança de que as estimativas de efeito estão corretas.

A qualidade da evidência deve ser classificada em 4 níveis: alta, moderada, baixa ou muito baixa. Esta classificação aplica-se não a estudos individuais, mas a cada comparação (certeza da evidência) e para cada medida de resultado. Na avaliação inicial, a evidência proveniente de estudos aleatorizados incluídos em cada comparação começa como evidência de alta qualidade, mas esta classificação inicial pode ser reduzida por cinco fatores (risco de viés, imprecisão, heterogeneidade, evidência não diretamente relevante e viés de publicação). Detalhes sobre a metodologia a utilizar para avaliação da qualidade da evidência serão descritos nos pontos 9.2.2 a 9.2.6. A classificação da qualidade da evidência permite hierarquizar a certeza de resultados:

- qualidade alta significa elevada certeza de resultados (significado: estamos muito confiantes de que o verdadeiro efeito está muito próximo das estimativas de efeito);
- qualidade moderada significa moderada certeza de resultados (significado: estamos moderadamente confiantes na estimativa de efeito. O verdadeiro efeito é provável que esteja próximo da estimativa de efeito, mas existe a possibilidade que possa ser substancialmente diferente);
- qualidade baixa significa baixa certeza de resultados (significado: a nossa confiança nas estimativas de efeito é limitada. O verdadeiro efeito pode ser substancialmente diferente da estimativa de efeito);
- qualidade muito baixa significa muito baixa certeza de resultados (significado: a nossa confiança nas estimativas de efeito é muito limitada. O verdadeiro efeito pode ser muito diferente da estimativa de efeito).

9.2.1. Classificação global da qualidade da evidência

As conclusões anteriores obtidas separadamente para cada medida de resultado, são depois sumarizadas numa avaliação final única sobre os benefícios e os danos da intervenção.

Nesta avaliação final, em geral, a classificação atribuída à qualidade da evidência global é a mesma que foi atribuída à medida de resultado "crítico" que fornece a confiança mais baixa.

9.2.2. Classificação da qualidade da evidência: risco de viés

A qualidade da evidência baseada em estudos aleatorizados é inicialmente classificada como alta, mas pode ser reduzida pelos cinco fatores referidos anteriormente. De salientar que a classificação da qualidade da evidência não se refere a estudos individuais, mas a cada medida de resultado utilizada em cada comparação (avaliação para cada medida de resultado), sendo esta a base para uma posterior avaliação global da qualidade (para todas as medidas de resultado).

O primeiro fator (risco de viés) resulta de problemas metodológicos no desenho ou condução do estudo e inclui um conjunto de cinco problemas metodológicos:

- ausência de alocação oculta: os investigadores têm conhecimento prévio do grupo a que vai ser alocado o próximo doente incluído;
- ausência de ocultação do tratamento: os doentes, os investigadores, os que registam as medidas de resultados, os que adjudicam as medidas de resultados, e/ou os que analisam os dados têm conhecimento do braço a que os doentes estão alocados [ou da medicação que estão atualmente a receber no caso de um estudo com desenho cruzado (*crossover*)];
- inclusão incompleta de doentes: perda de doentes para seguimento e não adesão ao princípio intenção-de-tratar nos estudos de superioridade. Historicamente, os metodologistas têm sugerido limiares arbitrários para uma perda de seguimento aceitável (por exemplo, menos de 20%). Contudo, o significado de uma perda para seguimento depende da relação entre perda para seguimento e número de eventos. Como regra geral, quanto maior for a diferença entre a percentagem de perda para seguimento e a percentagem de eventos nos grupos de intervenção e controlo, maior o risco de viés. Por exemplo, se os eventos forem 2% e 4% nos grupos de intervenção e controlo, uma perda para seguimento de 5% é preocupante;
- reporte seletivo de medidas de resultado: reporte incompleto ou ausente de algumas medidas de resultado influenciado pelos resultados. Para avaliar este domínio devem ser analisados os protocolos registados em bases de dados de ensaios clínicos (por exemplo, https://clinicaltrials.gov) e avaliar se todas as medidas de resultado registadas foram analisadas;
- outras limitações: interrupção precoce do estudo por benefício, uso de medidas de resultado sub-rogadas não validadas, etc.). A evidência empírica sugere que os estudos interrompidos precocemente por benefício sobrestimam o efeito do tratamento (67).

Na classificação inicial da qualidade da evidência, deve ser tido em conta que se trata de uma avaliação específica para cada medida de resultado, pelo que o impacto destes problemas metodológicos sobre cada uma dessas medidas pode variar substancialmente. O risco de viés por ausência de ocultação do tratamento ou por ausência de alocação oculta é maior em estudos com medidas de resultado subjetivas. Por exemplo, a ausência de ocultação do tratamento não é um problema grave se a medida de

resultado a avaliar for a mortalidade global, pelo que neste caso a classificação da qualidade não deverá ser reduzida.

Para cada comparação, e para cada medida de resultado, a qualidade da evidência deve ser reduzida em um nível (para moderada), ou em dois níveis (para baixa), se se considerar que existem limitações graves ou muito graves, respetivamente.

Contudo, se as limitações graves não forem ao nível da medida de resultado, mas ao nível de um estudo individual, deverá considerar-se a possibilidade de excluir esse estudo do processo de avaliação. De salientar que no caso de comparações por meta-análise em rede esta exclusão pode desconectar a rede.

9.2.3. Classificação da qualidade da evidência: imprecisão

Para cada medida de resultado, o principal critério para avaliar a precisão é o intervalo de confiança a 95% à volta da estimativa do efeito relativo do tratamento. Conceptualmente, o intervalo de confiança a 95% pode ser interpretado como o intervalo dentro do qual, em 95% dos casos, o verdadeiro valor se encontra. Na avaliação da qualidade da evidência, a questão é saber se o intervalo de confiança à volta da estimativa do efeito relativo da intervenção é suficientemente estreito.

A aleatorização permite equilibrar as variáveis prognósticas nos grupos intervenção e controlo. No entanto, este equilíbrio só é atingido se o tamanho da amostra for suficientemente grande. Grandes efeitos de tratamento na presença de um pequeno tamanho da amostra podem simplesmente ser o resultado de um desequilíbrio das variáveis de prognóstico entre grupos de tratamento mesmo em estudos aleatorizados com resultados com intervalos de confiança estreitos.

Assim, para avaliar se os resultados são suficientemente precisos deverão ser utilizados, cumulativamente, os dois critérios que se seguem:

- o intervalo de confiança a 95% é suficientemente estreito e exclui o efeito nulo;
- o número de participantes incluídos nos estudos em análise é igual ou superior ao “tamanho ótimo de informação (TOI)”. O TOI é obtido calculando o número de doentes necessários a incluir num estudo com poder estatístico suficiente. Trata-se no fundo de fazer uma estimativa do tamanho da amostra de um estudo com suficiente poder estatístico.
- Se os critérios definidos nos dois parágrafos anteriores não forem, cumulativamente cumpridos, deverá ser reduzida a classificação da qualidade da evidência por imprecisão.

9.2.4. Classificação da qualidade da evidência: heterogeneidade

Os critérios aqui utilizados para estabelecer a existência de “heterogeneidade” referem-se a medidas relativas (risco relativo, razão de riscos ou razão de chances), não a medidas absolutas, e aplicam-se no contexto de uma meta-análise em pares.

A classificação da qualidade da evidência pode ser reduzida por heterogeneidade, mas não é aumentada pela sua ausência.

Deve utilizar-se um conjunto de quatro critérios para avaliar a heterogeneidade nos resultados no contexto de uma meta-análise em pares:

- as estimativas de efeito variam substancialmente entre estudos;

- os intervalos de confiança não mostram qualquer sobreposição ou mostram apenas uma sobreposição mínima;
- os testes estatísticos de heterogeneidade (geralmente o teste Q) – que testam a hipótese nula de que todos os estudos incluídos numa meta-análise apresentam a mesma magnitude de efeito – mostram um valor de P significativo;
- o I^2 apresenta um valor elevado. O valor do I^2 pode variar entre 0 e 100%. I^2 informa-nos de qual é a proporção (%) da variância dos efeitos observados que reflete a variância dos verdadeiros efeitos [e, por conseguinte, que não resulta apenas de erro na amostragem (*sampling error*) situação em que o I^2 seria 0%].

Se utilizando os critérios anteriores se se chegar à conclusão de que os resultados apresentam problemas de heterogeneidade, deverá ser reduzida a classificação da qualidade da evidência em um ou dois níveis.

9.2.5. Classificação da qualidade da evidência: evidência não diretamente relevante (*indirectness*)

A evidência pode ser não diretamente relevante de três modos:

- a população de estudo é diferente da população de interesse;
- a intervenção testada é diferente da intervenção de interesse;
- as medidas de resultados são diferentes das medidas de resultados de interesse, por exemplo, o uso de medida de resultados sub-rogados.

Deve ser avaliado o impacto possível da evidência não diretamente relevante (diferentes populações ou intervenções e uso de sub-rogados) nos resultados e decidir se deve reduzir a classificação da qualidade da evidência.

9.2.6. Classificação da qualidade da evidência: reporte seletivo de medidas de resultado

O reporte seletivo de medidas de resultado é definido como a seleção de apenas uma parte das variáveis inicialmente definidas, com base nos resultados, para inclusão no relatório publicado do estudo. O reporte seletivo de medidas de resultado pode surgir de várias maneiras, algumas afetando o estudo como um todo e outras relacionadas com medidas de resultado específicas:

- omissão seletiva de algumas medidas de resultados no relatório: neste caso, apenas algumas das medidas de resultado analisadas são incluídas no relatório. Se a escolha for baseada nos resultados e, em particular, no resultado estatístico, é provável que as estimativas (meta-analíticas) correspondentes sejam igualmente enviesadas;
- escolha seletiva de dados para uma medida de resultado: para cada medida de resultado específica, podem existir diferentes momentos em que essa medida de resultado foi observada, ou podem ter sido usados diferentes instrumentos para medir a medida de resultado num determinado momento (por exemplo, diferentes escalas);
- reporte seletivo de análises usando os mesmos dados: existem múltiplas formas diferentes de analisar o efeito do tratamento numa medida de resultado. Mudar a análise da forma inicialmente prevista para outras formas de análise pode enviesar os resultados;

- reporte incompleto de dados: algumas vezes os dados são reportados de forma incompleta não permitindo, por exemplo, que esses dados possam ser incluídos numa meta-análise.

As medidas de efeito reportadas devem ser comparadas com as medidas de efeito previstas no protocolo de estudo. O não reporte, no relatório do estudo, do efeito de tratamento das medidas de resultado pré-especificadas no protocolo de estudo, indicia a existência de reporte seletivo de dados. A ausência de dados do efeito do tratamento em medidas de resultado consideradas fundamentais no contexto em avaliação deve também ser considerada um indício de reporte seletivo de dados.

9.3. Avaliação da qualidade da evidência na meta-análise em rede

A avaliação da qualidade da evidência baseada numa meta-análise em rede, pela sua complexidade, requer métodos específicos de avaliação, que têm em conta o facto de as estimativas para cada par de intervenções poderem ser baseadas em evidência direta e indireta e a complexidade da estrutura da rede.

Os métodos CiNeMA (*Confidence in Network Meta-Analysis*) (68,69) e de análise de limiares (70) (71) (*Threshold analysis*) têm em consideração a natureza mista (direta e indireta) da evidência e incorporam a influência de cada estudo na estimativa final. A qualidade de cada estudo não está diretamente relacionada com a sua contribuição para o resultado final. Por exemplo, um estudo de alta qualidade pode ter pouca influência nas estimativas finais da meta-análise em rede ou vice-versa (72).

Os métodos CiNeMA ou de análise de limiares devem ser usados para descrever a confiança nos resultados das meta-análises em rede.

O método CiNeMA considera seis critérios: risco de viés interno em cada estudo, risco de viés de relato seletivo de medida de resultados, evidência indireta, imprecisão, heterogeneidade e incoerência/inconsistência.

Assim, importa mencionar o seguinte:

- a matriz de contribuições, que descreve a percentagem de informação que cada estudo contribui para os resultados da meta-análise em rede, é usada para produzir classificações (semiautomáticas) do risco de viés e evidência indireta (69);
- ao contrário da metodologia GRADE, não são usadas classificações para julgar os critérios de imprecisão, heterogeneidade e inconsistência, mas o impacto destes campos na decisão clínica é avaliado;
- o método CiNeMA é implementado usando o *software* R (*package netmeta*) (73) pelo que só é aplicável a meta-análises em rede efetuadas usando o método frequentista de Rücker (2012) (74).

A análise de limiares quantifica até que ponto a evidência poderia ser alterada (por exemplo, devido a ajustamentos de viés ou variação amostral) sem alterar a recomendação, e identifica qual a nova recomendação caso a evidência saia fora dos limiares calculados.

Destaca-se, em seguida, o impacto da análise de limiares:

- a análise de limiares deve ser efetuada para cada estudo incluído na meta-análise, e para cada efeito relativo calculado pela meta-análise;

- a análise de limiares é implementada no *software* R (nmathresh) (75) e pode ser usada para avaliar análises frequentistas ou Bayesianas;
- o resultado da meta-análise em rede é considerado robusto se for considerado improvável que a evidência possa sair dos limiares calculados; caso contrário, o resultado é sensível a prováveis alterações na evidência;
- no caso de haver estudos identificados como suscetíveis de alterar as recomendações da meta-análise em rede, estes devem ser inspecionados em detalhe para determinar a plausibilidade de alterações ao seu efeito estimado para alem dos limiares calculados, tendo em conta o risco de viés e relevância do estudo para a população em avaliação;
- o grupo de avaliação está normalmente apenas interessado nas comparações da tecnologia em avaliação com os comparadores em uso. Os limiares calculados para estas comparações devem ser inspecionados em detalhe para determinar a plausibilidade de alterações destes efeitos para além dos limiares calculados, tendo em conta a qualidade dos estudos que compõem a rede.

10 VALOR TERAPÊUTICO ACRESCENTADO

10.1. Introdução

De acordo com o Decreto-Lei nº 97/2015, na sua atual redação, no seu artigo 14º, nº1, e artigo 25º, nº3, a comparticipação/avaliação prévia de medicamentos está condicionada, cumulativamente, à demonstração técnico-científica da inovação terapêutica ou da sua equivalência terapêutica, para as indicações terapêuticas reclamadas, e à demonstração da sua vantagem económica.

De acordo o artigo 14º, nº6, e artigo 25º, nº7, do mesmo Diploma, cabe ao titular da AIM do medicamento o ónus da prova quanto à eficácia, ao valor terapêutico acrescentado ou à sua equivalência terapêutica e à sua vantagem económica.

A avaliação de uma tecnologia de saúde utilizada para tratar uma determinada indicação, inclui a avaliação do benefício adicional dessa tecnologia de saúde, em comparação com as alternativas terapêuticas habitualmente utilizadas na prática clínica para tratar essa mesma indicação. Nos casos em que existe demonstração de benefício adicional, considera-se que existe demonstração técnico-científica de valor terapêutico acrescentado.

10.2. Critérios de demonstração do valor terapêutico acrescentado

O processo de avaliação compara o efeito do tratamento da tecnologia de saúde em avaliação, com o efeito do tratamento dos comparadores, num conjunto de medidas de eficácia terapêutica e de segurança que foram definidas na fase de definição da matriz de avaliação (ver secção 2.2).

As medidas de eficácia terapêutica e de segurança utilizadas, devem ser relevantes para o doente. Para este fim, consideram-se medidas relevantes para o doente, a mortalidade, morbilidade (sintomas e complicações), duração da doença, qualidade de vida, e segurança.

Podem ser utilizadas medidas de efeito de tratamento não clínicas, como substitutos de medidas clínicas (medidas sub-rogadas), desde que tenham sido previamente validadas. No caso de utilização de medidas sub-rogadas, a empresa deve submeter evidência que demonstre essa validação.

As medidas selecionadas na fase de pré-avaliação, para avaliar o efeito dos tratamentos, são pontuadas entre um e nove, consoante o grau de importância que lhes é atribuída pelos avaliadores, sendo pontuadas de um a três no caso de medidas não importantes, pontuadas de quatro a seis no caso de medidas importantes, mas não críticas, e pontuadas de sete a nove no caso de medidas críticas. Esta pontuação permite hierarquizar as medidas de eficácia e as medidas de segurança, em função da importância que lhes é atribuída. Recomenda-se que as medidas de efeito clínico [mortalidade, morbilidade (sintomas e complicações), duração da doença, qualidade de vida] sejam classificadas como críticas, e que as medidas de efeito sub-rogadas sejam classificadas como não críticas (pontuação inferior a sete).

Para cada comparação (para este fim “comparação” significa comparação de efeito de tratamento entre dois fármacos ou regimes terapêuticos), deve ser avaliado o efeito do tratamento nas medidas de resultado, utilizando as medidas selecionadas na fase de pré-avaliação. Assim, desta avaliação resulta que, para cada comparação, existirá uma estimativa de efeito relativo de tratamento sobre cada uma das medidas de resultado selecionadas. Para cada medida de resultado, é assim possível determinar se o efeito do tratamento do fármaco em avaliação apresenta ou não superioridade, em relação a cada comparador.

É necessário depois fazer uma síntese destes resultados, de forma a expressar o efeito global.

Em termos de determinação de existência ou não de superioridade e para cada comparação, o efeito relativo global de tratamento é avaliado pela estimativa de efeito relativo de tratamento observada sobre a medida de resultado à qual foi atribuída maior importância. No caso de existirem várias medidas de resultado com a mesma pontuação, deverá ser utilizada a estimativa de efeito de tratamento cujo resultado seja mais confiável, de entre as medidas com a maior pontuação. No caso de existirem várias medidas que cumpram estes critérios (em igualdade de pontuação e de credibilidade), a determinação de existência ou não de superioridade em termos globais, é feita utilizando as estimativas de efeito relativo de tratamento observada sobre essas medidas de resultado, sendo a existência ou não de superioridade apresentada de forma descritiva (efeito do tratamento nas medidas de resultado com a maior pontuação, e que apresentam igualdade de pontuação e de credibilidade) em caso de divergência.

As estimativas de efeito de tratamento sobre as outras medidas de resultado classificadas como críticas, são valorizadas no sentido de robustecer ou fragilizar as conclusões sobre superioridade, determinada pelo efeito de tratamento na medida de resultado mais pontuada e mais credível, mas não são usadas, por si só, para determinar a existência ou não de superioridade.

Pode acontecer que uma medida de efeito sub-rogada, inicialmente classificada como importante, mas não crítica, seja considerada determinante para o sentido da avaliação, por ausência de dados do efeito do tratamento em medidas de efeito clínicas. Neste caso, essa medida sub-rogada poderá ser reclassificada, sendo-lhe atribuída pontuação e grau de importância adequados, desde que essa medida sub-rogada tenha sido previamente validada.

10.3. Redação das conclusões sobre valor terapêutico acrescentado

Com o objetivo de determinar o valor terapêutico acrescentado, e com base na análise científica dos dados disponíveis, recomenda-se que as conclusões sejam expressas tendo por base o grau de certeza dos resultados: “prova” (elevada certeza de resultados quando a qualidade da evidência é alta), “indicativo” (moderada certeza de resultados quando a qualidade da evidência é moderada), “sugestivo” (baixa certeza de resultados quando a qualidade da evidência é baixa), ou nenhum dos anteriores quando não existem dados disponíveis ou a qualidade da evidência é muito baixa. O resultado da avaliação da existência de valor terapêutico acrescentado deve ser expresso numa das seguintes formas: existe prova, indicação ou sugestão de valor terapêutico acrescentado de uma intervenção.

10.4. Critérios para determinação de “equivalência terapêutica”

Da avaliação descrita anteriormente pode resultar o seguinte cenário: para cada comparação, o efeito global de tratamento mostra que o tratamento em avaliação não é superior ao comparador, mas a Comissão ficou convencida do efeito benéfico do fármaco, pelo que, utilizando apenas este critério, recomendaria o seu financiamento. Nestes casos é considerado que existe “equivalência terapêutica” para termos de fixação de preços.

10.5. Critérios para recomendação de não comparticipação / financiamento

Da avaliação descrita anteriormente pode resultar o seguinte cenário: para cada comparação, o efeito global de tratamento mostra que o tratamento em avaliação não é superior ao comparador, e a Comissão não ficou convencida do efeito benéfico do fármaco. Nestes casos, a Comissão recomenda o não financiamento da tecnologia de saúde.

Se da avaliação descrita não forem apresentados dados que permitam concluir a existência de superioridade do tratamento face ao comparador e a Comissão não ficar convencida do efeito benéfico do fármaco, a Comissão irá, também nestes casos, recomendar o não financiamento da tecnologia de saúde.

10.6. Classificação da magnitude do valor terapêutico acrescentado

Para classificar a magnitude do valor terapêutico acrescentado recomenda-se que seja tido em conta a estimativa do efeito global do tratamento e respetivo intervalo de confiança, usando o limite superior do intervalo de confiança a 95% e os limiares definidos na Tabela abaixo, e ser classificada numa das seguintes formas:

- valor terapêutico acrescentado substancial (maior);
- valor terapêutico acrescentado moderado;
- valor terapêutico acrescentado marginal (menor);
- valor terapêutico acrescentado não quantificável.

Medida de resultados binários: A determinação da extensão do valor terapêutico acrescentado deve ter em conta a qualidade da evidência em relação ao efeito do tratamento na medida de resultado, e ser baseada no risco relativo tendo em consideração a Tabela 3 e a Tabela 4.

Tempo até ao evento: O intervalo de confiança a 95% da razão de riscos é necessário para determinar a amplitude do efeito do tratamento no caso de medidas de resultado avaliadas por “tempo até ao evento”. Se existe uma meta-análise de vários estudos em que a medida de resultado é tempo até ao evento, deverá ser usado a razão de riscos. Para determinar a amplitude do valor terapêutico acrescentado utilizam-se os mesmos limiares da Tabela 3 e da Tabela 4. Se não existe uma razão de riscos ou este não for calculável, deve considerar-se a possibilidade de se calcular um risco relativo. Se apropriado, deve calcular-se o risco relativo numa determinada data.

Magnitude do Valor Terapêutico Acrescentado: nem sempre é possível quantificar a magnitude do valor terapêutico acrescentado ao nível da medida de resultado ou a nível global. Por exemplo, se se observar um efeito estatisticamente significativo num sub-rogado considerado como validado, mas não existirem dados de confiança do efeito do tratamento sobre uma medida de resultado relevante para o doente, não será possível quantificar a magnitude do efeito de tratamento. Neste caso, o valor terapêutico acrescentado será considerado como ‘não quantificável’. Esta mesma classificação se aplica aos casos de análises imaturas, em que a estimativa de efeito do tratamento pode estar sobrestimada.

Tabela 3: *Classificação da magnitude do valor terapêutico acrescentado (qualitativo)*

		Categoria de *outcome*			
		Mortalidade global	**Sintomas (ou complicações tardias) e eventos adversos graves (ou severos)**	**Qualidade de vida-relacionada com a saúde**	**Sintomas (ou complicações tardias) e eventos adversos não graves (ou não severos)**
Categorias de extensão	**Substancial (maior)** **Grande melhoria de forma sustentada** na medida de avaliação de benefício, que não foi anteriormente atingida em relação ao comparador apropriado	Aumento importante (maior) no tempo de sobrevida	Supressão ou evicção prolongada	Melhoria maior	Não aplicável
	Moderado **Melhoria marcada** na medida de avaliação de benefício, que não foi anteriormente atingida em relação ao comparador apropriado	Aumento moderado no tempo de sobrevida	Supressão ou evicção relevante	Melhoria importante	Evicção importante
	Marginal (menor) **Melhoria moderada e não apenas marginal** na medida de avaliação de benefício, que não foi anteriormente atingida em relação ao comparador apropriado	Qualquer aumento no tempo sobrevida	Qualquer redução	Melhoria relevante	Evicção relevante

Tabela 4: *Classificação da magnitude do valor terapêutico acrescentado (quantitativo)*

		Categoria de *outcome*		
		Mortalidade global	**Sintomas (ou complicações tardias) e eventos adversos graves (ou severos)**	**Sintomas (ou complicações tardias) e eventos adversos não graves (ou não severos)**
Categorias de extensão	**Substancial (maior)** **Grande melhoria de forma sustentada** na medida de avaliação de benefício, que não foi anteriormente atingida em relação ao comparador apropriado	0,85	0,75 e risco ≥5%	Não aplicável
	Moderado **Melhoria marcada** na medida de avaliação de benefício, que não foi anteriormente atingida em relação ao comparador apropriado	0,95	0,90	0,80
	Marginal (menor) **Melhoria moderada e não apenas marginal** na medida de avaliação de benefício, que não foi anteriormente atingida em relação ao comparador apropriado	1,00	1,00	0,90

Os valores na Tabela referem-se a um limiar superior abaixo do qual deve estar o intervalo de confiança a 95% do risco relativo. O intervalo de confiança a 95% estimado do risco relativo deve ser inferior (mais afastado de 1) ao limiar definido, ou seja, o limite superior do intervalo de confiança deve ser inferior ao limiar referido. Por exemplo, em relação à mortalidade global, para que o valor terapêutico acrescentado seja considerado disruptivo, o limite superior do intervalo de confiança a 95% do risco relativo deve ser pelo menos de 0,84, ou seja, a redução do risco relativo de morte em relação ao comparador apropriado deve estar incluído num intervalo de confiança a 95% cujo valor superior é igual ou inferior a 16% (0,84).

Fonte: Modificado de Ref. (IQWiG General Methods – Benefit assessment. Disponível em https://www.iqwig.de/en/about-us/methods/methods-paper/

11 REFERÊNCIAS

1. 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. Journal of Clinical Epidemiology. 2011;64(4):395–400.

2. Ciani O, Buyse M, Garside R, Pavey T, Stein K, Jonathan AC, et al. Comparison of treatment effect sizes associated with surrogate and final patient relevant outcomes in randomised controlled trials: Meta-epidemiological study. BMJ (Online). 2013;346(7898):1–12.

3. Ciani O, Buyse M, Drummond M, Rasi G, Saad ED, Taylor RS. Time to Review the Role of Surrogate End Points in Health Policy: State of the Art and the Way Forward. Value in Health. 2017;20(3):487–95.

4. Ciani O, Buyse M, Drummond M, Rasi G, Saad ED, Taylor RS. Use of surrogate end points in healthcare policy: A proposal for adoption of a validation framework. Nature Reviews Drug Discovery. 2016;15(7):516.

5. Stevens LA, Greene T, Levey AS. Surrogate end points for clinical trials of kidney disease progression. Clinical journal of the American Society of Nephrology : CJASN. 2006;1(4):874–84.

6. Weir CJ, Walley RJ. Statistical evaluation of biomarkers as surrogate endpoints: A literature review. Statistics in Medicine. 2006;25(2):183–203.

7. Wilcox R. Correlation and Tests of Independence. In 2012. p. 441–69.

8. Hung M, Bounsanga J, Voss MW. Interpretation of correlations in clinical research. Postgraduate Medicine. 2017 Nov 17;129(8).

9. IQWiG Reports – Commission No. A10-05. Validity of surrogate endpoints in oncology. Version 1.1. [Internet]. 2011. Available from: https://www.iqwig.de/en/projects-results/projects/drug-assessment/a10-05-validity-of-surrogate-endpoints-in-oncology-rapid-report.1325.html

10. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. BMJ (Online). 2009;339(7716):332–6.

11. Greenwood DC. Meta-analysis of observational studies. Modern Methods for Epidemiology. 2012;173–89.

12. EUnetHTA. European Network for Health Technology Assessment. Process of information retrieval for systematic reviews and health technology assessments on clinical effectiveness. 2014;(December 2019).

13. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ WV (editors). Chapter 6: Choosing effect measures and computing estimates of effect. Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019) [Internet]. Cochrane, 2019; Available from: www.training.cochrane.org/handbook.

14. Tanniou J, Van Der Tweel I, Teerenstra S, Roes KCB. Subgroup analyses in confirmatory clinical trials: Time to be specific about their purposes. BMC Medical Research Methodology. 2016;16(1).

15. Wang R, Lagakos SW, D P, Ware JH, Hunter DJ, Drazen JM. spe ci a l r ep or t Statistics in Medicine — Reporting of Subgroup Analyses in Clinical Trials. Health (San Francisco). 2007;2189–94.

16. Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I2 in assessing heterogeneity may mislead. BMC Medical Research Methodology. 2008;8:1–9.

17. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: Rationale, conduct, and reporting. BMJ (Online). 2010;340(7745):521–5.

18. Riley RD, Lambert PC, Staessen JA. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. 2008;(April):4267–78.

19. Simmonds MC, Higgins JPT, Stewart LA, Tierney JF, Clarke MJ, Thompson SG. Meta-analysis of individual patient data from randomized trials: A review of methods used in practice. Clinical Trials. 2005;2(3):209–17.

20. Tierney JF, Vale C, Riley R, Smith CT, Stewart L, Clarke M, et al. Individual participant data (IPD) metaanalyses of randomised controlled trials: Uidance on their use. PLoS Medicine. 2015;12(7):1–16.

21. Greenland S, Maclure M, Schlesselman JJ, Poole C, Morgenstern H. Standardized Regression Coefficients: A Further Critique and Review of Some Alternatives. Epidemiology [Internet]. 1991;2(5):387–92. Available from: http://www.jstor.org/stable/20065707

22. Deeks JJ HJAD (editors)., In: Higgins JPT TJCJCMLTPMWV (editors). Chapter 10: Analysing data and undertaking meta-analyses . In: Cochrane Handbook for Systematic Reviews of Interventions version 62 (updated February 2021) Cochrane, 2021 Available from www.training.cochrane.org/handbook.

23. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Statistics in Medicine. 2002 Jun 15;21(11).

24. Dias S, Welton NJ, Sutton AJ, Ades A. NICE DSU Technical Support Document 1: Introduction to evidence synthesis for decision making. 2011;(April 2011):1–24.

25. Dias S, Ades A, Welton NJ, Jansen JP, Sutton AJ. Network Meta-Analysis for Decision-Making. Wiley, editor. 2018.

26. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. Journal of Clinical Epidemiology. 1997 Jun;50(6).

27. Hoaglin DC, Hawkins N, Jansen JP, Scott DA, Itzler R, Cappelleri JC, et al. Conducting indirect-treatment-comparison and network-meta-analysis studies: Report of the ISPOR task force on indirect treatment comparisons good research practices: Part 2. Value in Health. 2011;14(4):429–37.

28. J. Sweeting M, J. Sutton A, C. Lambert P. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Statistics in Medicine. 2004 May 15;23(9).

29. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Statistics in Medicine. 2004;23(20):3105–24.

30. NICE. Evidence Synthesis TSD series [Internet]. [cited 2020 Jul 1]. Available from: http://nicedsu.org.uk/technical-support-documents/evidence-synthesis-tsd-series/

31. Dias S, Ades A, Welton NJ, Jansen JP, Sutton AJ. Network Meta-Analysis for Decision-Making. Wiley, editor. 2018.

32. van Valkenhoef G, Lu G, de Brock B, Hillege H, Ades AE, Welton NJ. Automating network meta-analysis. Research Synthesis Methods. 2012;3(4):285–99.

33. White IR. Network meta-analysis. Stata Journal [Internet]. 2015;15(4):951–85. Available from: //<? echo(www) ?>.stata-journal.com/article.html?article=st0410

34. Rücker G, Schwarzer G, Krahn U KJ. netmeta: Network Meta-Analysis using Frequentist Methods. 2015.

35. Lumley T. Network meta-analysis for indirect treatment comparisons. Statistics in Medicine. 2002;21(16):2313–24.

36. Hong H, Chu H, Zhang J, Carlin BP. A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. Research Synthesis Methods. 2016;7(1):6–22.

37. Turner RM, Domínguez-Islas CP, Jackson D, Rhodes KM, White IR. Incorporating external evidence on between-trial heterogeneity in network meta-analysis. Statistics in Medicine. 2019;38(8):1321–35.

38. Rhodes KM, Turner RM, Higgins JPT. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. Journal of Clinical Epidemiology. 2015;68(1):52–60.

39. Turner RM, Jackson D, Wei Y, Thompson SG, Higgins JPT. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. Statistics in Medicine. 2015;34(6):984–98.

40. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. Journal of Clinical Epidemiology. 1997;

41. Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. Statistics in Medicine. 2010;29(7–8):932–44.

42. White IR, Barrett JK, Jackson D, Higgins JPT. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. Research Synthesis Methods. 2012;3(2):111–25.

43. White IR, Barrett JK, Jackson D, Higgins JPT. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. Research Synthesis Methods. 2012;3(2):111–25.

44. van Valkenhoef G, Dias S, Ades AE, Welton NJ. Automated generation of node-splitting models for assessment of inconsistency in network meta-analysis. Research Synthesis Methods. 2016;7(1):80–93.

45. Ades AE, Caldwell DM, Reken S, Welton NJ, Sutton AJ DS. NICE DSU Technical Support Document 7: Evidence synthesis of treatment efficacy in decision making: a reviewer's checklist. 2012.

46. Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: Checklist and explanations. Annals of Internal Medicine. 2015;162(11):777–84.

47. Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: An overview and tutorial. Journal of Clinical Epidemiology. 2011;64(2):163–71.

48. Boutron I PMHJADLAHA, In: Higgins JPT TJCJCMLTPMWV (editors). Chapter 7: Considering bias and conflicts of interest among the included studies. . In: Cochrane Handbook for Systematic Reviews of Interventions version 62 (updated February 2021) Cochrane, 2021 Available from www.training.cochrane.org/handbook.

49. Donegan S, Welton NJ, Tudur Smith C, D'Alessandro U, Dias S. Network meta-analysis including treatment by covariate interactions: Consistency can vary across covariate values. Research Synthesis Methods. 2017;8(4):485–95.

50. Donegan S, Williamson P, D'Alessandro U, Smith CT. Assessing the consistency assumption by exploring treatment by covariate interactions in mixed treatment comparison meta-analysis: Individual patient-level covariates versus aggregate trial-level covariates. Statistics in Medicine. 2012;31(29):3840–57.

51. Cooper NJ, Sutton AJ, Morris D, Ades AE WN. Addressing between-study heterogeneity and inconsistency in mixed treatment com-parisons: Application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. Stat Med. 2009;(April):1861–81.

52. Dias S, Sutton AJ, Welton NJ, Hall C, Road W, Unit DS, et al. NICE DSU Technical Support Document 3 : Heterogeneity : Subgroups , Meta-Regression , Bias and Bias-Adjustment. Tropical Medicine. 2011;(September):1–75.

53. Signorovitch JE, Sikirica V, Erder MH, Xie J, Lu M, Hodgkins PS, et al. Matching-adjusted indirect comparisons: A new tool for timely comparative effectiveness research. Value in Health. 2012;15(6):940–7.

54. Caro JJ, Ishak KJ. No head-to-head trial? Simulate the missing arms. PharmacoEconomics. 2010;28(10):957–67.

55. Caro JJ, Ishak KJ. No head-to-head trial? Simulate the missing arms. PharmacoEconomics. 2010;28(10):957–67.

56. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. NICE DSU Technical Support Document 18: Methods for Population-Adjusted Indirect Comparisons in Submissions To NICE. Nice Dsu Technical Support Document 18. 2016;(December):1–82.

57. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for Population-Adjusted Indirect Comparisons in Health Technology Appraisal. Medical Decision Making. 2018;38(2):200–11.

58. Phillippo DM, Dias S, Ades AE, Belger M, Brnabic A, Schacht A, et al. Multilevel network meta-regression for population-adjusted treatment comparisons. Journal of the Royal Statistical Society Series A: Statistics in Society. 2020;1189–210.

59. Dias S, Sutton AJ, Welton NJ, Hall C, Road W, Unit DS, et al. NICE DSU Technical Support Document 3 : Heterogeneity : Subgroups , Meta-Regression , Bias and Bias-Adjustment. Tropical Medicine. 2011;(September):1–75.

60. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ (Online). 2010;340(7751):850–4.

61. Graves RS. Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice. Journal of the Medical Library Association [Internet]. 2002 Oct;90(4):483. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC128970/

62. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ. 2010 Mar 30;340(mar30 3).

63. European Medicines Agency. ICH E9: Note for Guidance on Statistical Principles for Clinical Trials. 1998;

64. European Medicines Agency. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials - Step 2b. 2017;44(August):1–23.

65. The European Agency for the Evaluation of Medicinal Products C for PMP (CPMP). Points to consider on switching between superiority and non-inferiority. London; 2000.

66. Higgins J, Green S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.1. The Cochrane Collaboration and John Wiley & Sons Ltd; 2012.

67. Montori VM, Devereaux PJ, Adhikari NKJ, Burns KEA, Eggert CH, Briel M, et al. Randomized Trials Stopped Early for Benefit. Jama. 2005;294(17):2203.

68. Salanti G, Giovane C Del, Chaimani A, Caldwell DM, Higgins JPT. Evaluating the quality of evidence from a network meta-analysis. PLoS ONE. 2014;9(7).

69. Institute of Social and Preventive Medicine U of B. CINeMA: Confidence in Network Meta-Analysis. 2017.

70. Phillippo DM, Dias S, Welton NJ, Caldwell DM, Taske N, Ades AE. Threshold analysis as an alternative to grade for assessing confidence in guideline recommendations based on network meta-analyses. Annals of Internal Medicine. 2019;170(8):538–46.

71. Phillippo DM, Dias S, Ades AE, Didelez V, Welton NJ. Sensitivity of treatment recommendations to bias in network meta-analysis. Journal of the Royal Statistical Society Series A: Statistics in Society. 2018;181(3):843–67.

72. Berlin JA. How confident should we be about recommendations based on network meta-analyses? Annals of Internal Medicine. 2019;170(8):571–2.

73. G. Rücker, G. Schwarzer, U. Krahn and JK. netmeta: Network Meta-Analysis using Frequentist Methods. 2017.

74. Rücker G. Network meta-analysis, electrical networks and graph theory. Research Synthesis Methods. 2012;3(4):312–24.

75. Phillippo DM, Dias S, Ades AE, Didelez V, Welton NJ. Sensitivity of treatment recommendations to bias in network meta-analysis. Journal of the Royal Statistical Society Series A: Statistics in Society. 2018;181(3):843–67.

Parque de Saúde de Lisboa, Av. do Brasil 53
1749-004 Lisboa, Portugal

www.infarmed.pt
infarmed@infarmed.pt