METHODOLOGY

FOR PHARMACOTHERAPEUTIC ASSESSMENT OF HEALTH TECHNOLOGIES

REPÚBLICA PORTUGUESA
SAÚDE

SNS SERVIÇO NACIONAL DE SAÚDE

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

Pharmacotherapeutic assessment methodology

Methodology for pharmacotherapeutic assessment of health technologies

Version	Date of publication
3.0	05 August 2022 (PT); 10 February 2023 (EN)

Suggested citation: 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) Pharmacotherapeutic assessment methodology, Version 3.0. Committee for Health Technology Assessment, INFARMED - Autoridade Nacional do Medicamento e Produtos de Saúde, I.P., Lisbon

Available online at www.infarmed.pt

Authors

José Vinhas
Setúbal Hospital Centre and Committee for Health Technology Assessment (INFARMED, I.P.) (coordination)

Sofia Dias
University of York and Committee for Health Technology Assessment (INFARMED, I.P.) (coordination)

Antonio Melo Gouveia
Francisco Gentil Portuguese Institute of Oncology of Lisbon and Committee for Health Technology Assessment (INFARMED, I.P.)

Catarina Viegas Dias
NOVA University Lisbon and the Committee for Health Technology Assessment (INFARMED, I.P.)

Diana Sousa
North Lisbon University Hospital Centre and Committee for Health Technology Assessment (INFARMED, I.P.)

João Oliveira
Francisco Gentil Portuguese Institute of Oncology of Lisbon and Committee for Health Technology Assessment (INFARMED, I.P.)

Julian Perelman (CATS / ENSP)
National School of Public Health, NOVA University Lisbon and Committee for Health Technology Assessment (INFARMED, I.P.)

Luís Azevedo
Faculty of Medicine of the University of Porto and Committee for Health Technology Assessment (INFARMED, I.P.)

Nuno Marques
Algarve Hospital and University Centre and Committee for Health Technology Assessment (INFARMED, I.P.)

Pedro Saramago
University of York and Committee for Health Technology Assessment (INFARMED, I.P.)

Rita Faria
University of York and Committee for Health Technology Assessment (INFARMED, I.P.)

Sofia Torres
Antwerp University Hospital and Committee for Health Technology Assessment (INFARMED, I.P.)

Alex Correia
Directorate of Health Technology Assessment, INFARMED, I.P.

Sara Couto
Directorate of Health Technology Assessment, INFARMED, I.P.

Institutional Review

Claudia Furtado, INFARMED,I.P.

Rui Santos Ivo, INFARMED,I.P.

António Faria Vaz, INFARMED,I.P.

Cláudia Belo Ferreira, INFARMED, I.P.

TABLE OF CONTENTS

TABLE OF CONTENTS 4

TABLE OF FIGURES AND TABLES 6

LIST OF ABBREVIATIONS 7

LIST OF DEFINITIONS 8

1 INTRODUCTION 9

1.1 About this document 9

1.2 Objective of the document 9

1.3 Working Group 9

1.4 Methodology of the review process 9

1.5 Framework 10

1.6 Scope of application of the methodology 12

1.7 Main changes of the new version 12

2 OPERATIONALISATION OF THE ASSESSMENT 13

2.1 Introduction 13

2.2 Definition of the assessment matrix 13

2.3 Conclusions 15

3 GENERAL METHODOLOGY 16

3.1 The relevance of certainty of outcome 16

3.2 The link between certainty of outcome and proximity to everyday conditions 16

3.3 Outcome measures 17

3.3.1. Measures of clinical effect 17

3.3.2. Surrogate effect measures 17

3.3.3. Validation of surrogate effect measures 17

3.3.3.1. Introduction 17

3.3.3.2. Requirements of a surrogate outcome measure 18

3.3.3.3. Validation of a surrogate outcome measure 18

3.3.3.4. Conclusions 24

3.4 Systematic reviews 25

3.4.1. Introduction 25

3.4.2. Research protocol 25

3.4.3. Databases 26

3.4.4. Research strategy and study selection 26

3.4.5. Assessment of the quality of evidence 26

4 METHODS OF COMPARISON 27

4.1 Introduction 27

4.2 Direct and indirect comparisons: definitions 27

4.3 Conventional meta-analysis 27

4.3.1. Introduction 27

4.3.2. Factors affecting accuracy 28

4.3.3. Fixed-effect and random-effects models 28

4.3.4. Heterogeneity 29

4.3.5. Subgroup analysis and meta-regression 30

4.3.6. Meta-analysis with individual data 31

4.3.7. Measures of effect and their interpretation 32

4.3.8. Ways to report the results of a meta-analysis 33

4.4 Network meta-analysis 33

4.4.1. Introduction 33

4.4.2. Assumptions of a network meta-analysis 35

4.4.3. Technical aspects in network meta-analysis 36

4.4.4. Meta-regression and bias adjustment 39

4.5 Conclusions 39

5 METHODS OF COMPARISON IN EXCEPTIONAL SITUATIONS 41
5.1 Anchored Adjusted Indirect Comparison (MAIC, STC) 41
5.2 Use of non-randomised studies 42
6 SUBGROUP ANALYSIS 44
6.1 Introduction 44
6.2 Definition/ specification of subgroups 44
6.3 Recommendations for subgroup analysis - MAH perspective 45
6.4 Assessment and rating the credibility of subgroup analyses 46
6.5 Conclusions 47
7 PARTICULAR ASPECTS IN BENEFIT ASSESSMENT 49
7.1 Impact of study outcomes not published in the conclusions 49
7.2 Dramatic effect 49
7.3 Duration of study 49
8 SUPERIORITY, NON-INFERIORITY, AND EQUIVALENCE STUDIES: DEFINITIONS AND CRITERIA FOR CHANGING OBJECTIVES 50
8.1 Introduction 50
8.2 Demonstration of equivalence 50
8.3 Demonstration of non-inferiority 50
8.4 Unilateral and bilateral confidence intervals 51
8.5 Superiority studies 51
8.6 Relevance of Δ pre-definition in non-inferiority and equivalence studies 52
8.7 Relevance of pre-definition of the study as superiority, non-inferiority or equivalence 52
8.8 Is it possible to change the purpose of a comparison? 52
8.9 Interpreting a non-inferiority study as a superiority study 52
8.10 Interpretation of a superiority study as a non-inferiority study 53
9 ASSESSMENT OF THE QUALITY OF EVIDENCE 54
9.1 Assessment of risk of bias by study 54
9.2 Assessment of quality of evidence (certainty of evidence) in conventional meta-analysis 54
9.2.1. Overall rating the quality of evidence 54
9.2.2. Rating the quality of evidence: risk of bias 55
9.2.3. Rating the quality of evidence: inaccuracy 56
9.2.4. Rating quality of evidence: heterogeneity 56
9.2.5. Rating the quality of evidence: not directly relevant evidence (indirectness) 57
9.2.6. Rating quality of evidence: selective reporting of outcome measures 57
9.3. Quality of evidence assessment in network meta-analysis 57
10 ADDED THERAPEUTIC VALUE 59
10.1. Introduction 59
10.2. Criteria for demonstrating added therapeutic value 59
10.3. Drafting of conclusions on added therapeutic value 60
10.4. Criteria for determining 'therapeutic equivalence' 60
10.5. Criteria for recommending no co-payment / funding 60
10.6. Rating the magnitude of added therapeutic value 61
11 REFERENCES 64

TABLE OF FIGURES AND TABLES

Table of Figures

Figure 1: *Conclusions on the validation of the surrogate in the case of high-quality evidence* 21
Figure 2*: Classification of correlation strength as a function of the correlation between treatment effect on the surrogate and the clinical outcome measure* 22
Figure 3**:** *Conclusions on the validation of the surrogate in the case of high-quality evidence* 24

Tables

Table 1: *Validity of the surrogate outcome measure as a function of the quality of evidence (validation study) and the correlation between the treatment effect on the surrogate and on the clinical outcome measure*..... 22
Table 2: *Criteria for assessing the credibility of subgroup analysis* 47
Table 3: *Rating the magnitude of added therapeutic value (qualitative)*..... 62
Table 4: *Rating the magnitude of added therapeutic value (quantitative)*..... 63

LIST OF ABBREVIATIONS

AIC	Akaike's Information Criterion
MA	Marketing Authorisation
HTA	Health Technology Assessment
CATS	Committeefor Health Technology Assessment / Comissão de Avaliação de Tecnologias de Saúde
CE-CATS	Executive Committee of the Committeefor Health Technology Assessment / Comissão Executiva da Comissão de Avaliação de Tecnologias de Saúde
CHMP	Committee for Medicinal Products for Human Use / Comité para Medicamentos de Uso Humano
CiNeMA	Confidence in Network Meta-Analysis
CSR	Clinical Study Report
DATS	Health Technology Assessment Department / Direção de Avaliação das Tecnologias de Saúde
INN	International Nonproprietary Name
DIC	Deviance Information Criterion
EMA	European Medicines Agency
EUDRACT	European Union Drug Regulating Authorities Clinical Trials Database
EUnetHTA	European Network for Health Technology Assessment
GAE	Evidence Assessment Group / Grupo de Avaliação da Evidência
GRADE	Grading of Recommendations, Assessment, Development and Evaluation
HRQoL	Health-Related Quality of Life
CI	Confidence Interval
INFARMED, I.P.	National Authority for Medicines and Health Products / Autoridade Nacional do Medicamento e Produtos de Saúde, I.P.
MAIC	Matching Adjusted Indirect Comparisons
ML-NMR	Multi-Level Network Meta-Regression Method
NICE	National Institute for Health and Care Excellence
MOOSE	Meta-analysis of Observational Studies
NNT	Numbers Needed to Treat
PICO	Patient, Intervention, Comparator and Outcome
PRISMA	Preferred Reporting Items for Systematic Reviews
PAR	Pharmacotherapeutic Assessment Report
SPC	Summary of Product Characteristics
RD	Risk Difference
SiNATS	National Health Technology Assessment System / Sistema Nacional de Avaliação de Tecnologias de Saúde
SNS	National Health Service / Serviço Nacional de Saúde
STC	Simulated Treatment Comparisons
STE	Surrogate Threshold Effect
MAH	Marketing Authorisation Holder
OIS	Optimal Information Size

LIST OF DEFINITIONS

Biomarker	A characteristic that is objectively measured and that is assessed as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention.
Effectiveness	Refers to (the measurement of) the desired/beneficial effect of the intervention under clinical practice conditions.
Efficacy	Refers to (the measurement of) the desired/beneficial effect of the intervention under optimal conditions (in the context of a clinical trial).
Surrogate outcome measure	A biomarker that is intended to replace a clinical outcome measure.
Clinical outcome measure	A characteristic or variable that reflects the patient's symptoms, functional capacity, or life expectancy.
Absolute effect measures	Measures that express the result (outcome measure) through the difference (subtraction) between the risks observed in the two groups. Examples of effect measures are risk difference (RD) and the number needed to be treated (NNT). Absolute effect measures reflect the baseline risk of a population, unlike relative measures, and are clinically more useful in therapeutic decisions.
Relative effect measures	Measures expressing the result (outcome measure) in one group over another, usually in the form of a split/ratio. Examples of relative effect measures are relative risk or risk ratio (RR), odds ratio (OR) or hazard ratio (HR).
Safety measures	These are the results relevant to the safety of the patient. They may be overall (e.g. serious adverse events, overall adverse events) or specific, in the case of adverse events of special interest, which are events that may be potentially related to the disease being studied (e.g. incidence of neoplasms, genitourinary infections).
Clinical practice	Standards of practice of health professionals in Portugal, measured through available sources.
Connected network	Network in which it is possible to establish a path (using the edges) from any intervention (vertex/node) to any other.
Star network	Intervention network only linked by a common comparator.
Binary variables	Variables that have only two possible values (for example: death).
Categorical variables	Variables that contain a finite and generally fixed number of values (greater than 2), which correspond to distinct categories or groups. Categorical data may not have a logical order. (e.g., body mass index [BMI] categories).
Continuous variables	Numeric variables that can have an infinite number of values between any two values (for example: height)

1 INTRODUCTION

1.1 About this document

The methodology presented in this document is intended for all those interested in the process of assessing health technologies in Portugal and, in particular, for the assessors of the Committeefor Health Technology Assessment (CATS), applicants (MA holders), decision-makers, patient associations, health professionals, researchers and other interested parties.

1.2 Objective of the document

The purpose of this document is to guide the pharmacotherapeutic assessment and reassessment of medicines and other health technologies carried out by CATS, in order to clarify the methodological challenges encountered, also describing the assessment process. Thus, greater consistency and transparency is introduced in the process and methodology used.

1.3 Working Group

According to INFARMED, I.P. statute, the Committeefor Health Technology Assessment (CATS), which supports the Directorate of Health Technology Assessment (DATS), created by Decree-Law No. 97/2015, of 1 June 2015, in its current wording, in office since June 2016, is responsible for issuing opinions on matters related to the assessment and re-assessment of health technologies, in the context of their reimbursement and proposing measures appropriate to the interests of public health and the SNS regarding health technologies, within the scope of SiNATS.

As a result of this experience, Dr. José Vinhas, in his capacity as Chairman of the Executive Committee of CATS (CE-CATS), proposed a working group with the purpose of reviewing the methodology for the pharmacotherapeutic assessment of medicines, which he coordinated. The proposed working group included 12 members who were part of CATS at the beginning of this review process.

The decision to use CATS experts is essentially for two reasons: 1) they have experience in the pharmacotherapeutic assessment of a number of medicine's dossiers submitted for reimbursement, and have thus encountered in practice the difficulties that required adequate guidance; 2) as actors in the assessment, they are aware of the need to harmonise the dossiers in order to ensure consistency in the assessment process.

The review process also included the participation and coordination of Professor Sofia Dias, CATS member, professor at the University of York, and part of the team that produces health technology assessment reports for the National Institute for Health and Care Excellence (NICE), an internationally recognised academic with extensive experience in evidence synthesis of health technologies.

1.4 Methodology of the review process

The review process began with the identification of the topics to be included in the new version of the pharmacotherapeutic methodology with a view to updating the version published in November 2016, according to the most recent advances in this area. The list of topics was circulated and discussed among the authors until a consensus list was obtained. Each topic was then assigned to a group of at least two people and two topics were reviewed individually by two experts, according to their individual interests and expertise.

Each group was asked to prepare a brief literature review and a list of options for possible changes within each topic.

After this preparatory phase, a two-day meeting with all the authors took place in Lisbon in January 2020. Each group briefly presented their literature review and their arguments on how the methodology should or should not be reviewed, followed by a discussion until consensus was reached on the content of each point. The deputy chairmen of CE-CATS were also present at the meeting.

After this meeting, the authors prepared a preliminary version of the methodology review. The draft version was discussed by the coordinating team and sent to the authors, who revised their versions in light of suggestions and comments received in the meantime.

In August 2020, the document presenting the proposed review of the pharmacotherapeutic assessment methodology was the subject of a broad consultation with stakeholders. To this end, a group of entities and individuals were contacted to give their views in writing on the new proposal for the methodological guidelines. Written comments were received from the following entities and individuals: Associação Nacional de Farmácias (ANF), Associação Portuguesa da Indústria Farmacêutica (APIFARMA), Associação dos Profissionais de Registros e Regulamentação Farmacêutica (APREFAR), ARS Norte, Associação Portuguesa de Bioindústrias (P-BIO), Defesa do Consumidor (DECO), Ordem dos Enfermeiros, Ordem dos Médicos, Registro Oncológico Nacional (RON).

After receiving these comments, a discussion meeting was organised in April 2021 between CATS working group to discuss the comments received. The new version of the pharmacotherapeutic methodology was reviewed, when this was identified as necessary, after discussion among the authoring-team, and it was finalised by the CATS working group in May 2021.

The team of authors prepared a response to the comments received, which was sent by INFARMED, I.P. to each of the entities prior to the publication of the new version of the pharmacotherapeutic methodology.

1.5 Framework

Marketing Authorisation

The marketing of medicines within the national territory is subject to a Marketing Authorisation (MA). In accordance with paragraph 2 of Article 14 of Decree-Law 176/2006, of 30 August 2006, in its current wording, the decision to grant a marketing authorisation to a medicine must be based exclusively on objective scientific criteria of quality, safety and therapeutic efficacy of the medicine in question, regardless of any economic considerations. To this end, in addition to the evaluation of the quality of the medicinal product, a benefit-risk assessment is carried out, i.e. the assessment of the positive therapeutic effects of a medicinal product against its own risks as regards patients' or public health. This assessment is carried out by a Competent National Authority (e.g. INFARMED, I.P.) or by the European Medicines Agency (EMA), depending on the applicable assessment procedure, which depends on the type of product/therapeutic area. This authorisation is the sole requirement for marketing the medicine in the jurisdiction in which it is valid.

Reimbursement

Following this MA, decisions on the reimbursement and pricing of a medicine can be taken at national, regional or local level in each EU Member State. Portugal has a National Health Service (SNS) that reimburses health technologies in part or in their entirety. In the case of medicines, only those that have

obtained the respective MA can be reimbursed. To support the reimbursement decision, a Health Technology Assessment (HTA) is carried out. In Portugal, this assessment is also carried out by INFARMED, I.P. as the HTA Agency (organically through DATS and CATS), regardless of the body that assessed the MA.

Difference between Marketing Authorisation and Reimbursement

Although briefly, and because it is (still) a frequent cause of confusion, it seems important to note in this document the differences in the assessment of medicines and other health technologies (hereinafter referred to as health technologies) between Regulatory Agencies and HTA Agencies, which result from the existence of different objectives, perspectives and assessment methodologies. While regulators assess the quality, efficacy and safety of health technologies, in the perspective of a positive relationship between the therapeutic effects of that health technology and the respective risks in the therapeutic indication under assessment, HTA Agencies make a recommendation as to the reimbursement of the medicine or other health technology, taking into account the existence of other therapeutic alternatives already reimbursed and in use in clinical practice, through a comparative analysis of the efficacy and safety of the new medicine or other health technology compared to the alternatives commonly used in national clinical practice.

It is therefore essential for applicants to plan and develop in advance the evidence needed to provide adequate information. This information should make it possible to answer both the questions of the Regulatory Agencies and the questions of the HTA Agencies, which pursue different objectives and assess different perspectives of health technologies through different assessment methodologies. The existence of gaps in the quantity and quality of available clinical evidence brings additional challenges to these agencies' assessments and leads to decisions being made with greater uncertainty and can be an obstacle to access health technologies or other health technologies for people who need them.

Health Technology Assessment (HTA)

As mentioned above, HTA aims to support the decision of use and reimbursement (co-payment and/or prior assessment) of health technologies in the SNS. This decision is based not only on the quality, safety and efficacy criteria required of all medicinal products, but also on comparative efficacy and safety criteria in order to optimise the use of available resources.

Generally speaking, in Portugal, the public reimbursement process can be divided into the following phases: application, pharmacotherapeutic assessment, pharmacoeconomic assessment, negotiation and decision. In the pharmacotherapeutic phase of the HTA, to which this document refers, it is intended to ensure that no SNS reimbursement of health technologies that are not useful and/or necessary is recommended.

HTA has a long tradition in Portugal and in 2015 was updated with the creation of the National Health Technology Assessment System (SiNATS), through Decree-Law No. 97/2015, of 1 June. SiNATS is constituted by the set of entities and means that carry out the HTA, and its management is entrusted to INFARMED, I. P.. This legal framework established the creation of CATS, an advisory committee of INFARMED, I. P. to support SiNATS, under the terms and conditions provided for in article 8 of Decree-Law No. 46/2012, of 24 February, as amended by Decree-Law No. 97/2015, of 1 June, in its current wording.

1.6 Scope of application of the methodology

As set forth in paragraphs 2 and 4 of article 7 of Administrative Rule no. 195-A/2015, of 30 June, in its current wording, the pharmacotherapeutic assessment of health technologies is subject to an opinion/deliberation by CATS in the case of a medicine with an International Nonproprietary Name (INN) or therapeutic indication that is not yet been co-paid or authorised for use in the institutions and services under the responsibility of the Government member responsible for health and/or whenever requested by the competent services of INFARMED, I.P.. This document presents the assessment methodology generically used by CATS to issue these pharmacotherapeutic recommendations.

It should be noted that the methodology provided herein may undergo adaptations to allow for the assessment of specific health technologies/therapeutic areas, and its application to medicines already in well-established clinical use at a national level or frontier products (e.g. medicated medical devices) may not be justified. Some of these situations are foreseen in section 5. Additionally, fixed-dose combinations to replace single components already financed in the same doses and respective therapeutic indications, prophylactic vaccines and medicines derived from human plasma (or their recombinant or modified versions) are, in principle, excluded from this methodology.

This document should be read in conjunction with the legal framework of Health Technology Assessment in Portugal and other normative documents on this subject.

1.7 Main changes in the new version

The current version of the Pharmacotherapeutic Assessment Methodology (version 3.0), has undergone a major revision compared to version 2.0 of 23 November 2016, with the structure of the document having been changed, and new sections introduced, while other sections have undergone profound changes that have generally entailed new content and greater development and detail.

In Chapter 2 (Operationalisation of the assessment), a new section has been created (2.2. Definition of the assessment matrix), which includes the criteria for selection of comparators, where the reasons that led to the revision of these criteria are developed in detail.

In Chapter 3 (General methodology), a new section was created (3.3. Outcome measures) detailing the different types of outcome measures and describing the requirements for validation of a surrogate outcome measure (3.3.3.3.). A new section has been added (section 3.4. Systematic Reviews) which makes recommendations on how to conduct a systematic literature review process and stresses its importance to the assessment process.

A new chapter has been created (4. Comparison methods), where comparison methods are described in detail, including conventional meta-analysis and network meta-analysis, and adjusted indirect comparison methods (MAIC, STC) for use in the context of rare and ultra-rare diseases (Chapter 5).

Chapter 6 (Subgroup analysis) has been expanded, now including greater detail, notably on the principles and criteria to be checked for the specification of subgroups within the initial assessment matrix.

Chapter 10 (Added therapeutic value) has been completely rewritten, with the new version being more detailed regarding the process for recognising added therapeutic value with the aim of clarifying the CATS recommendation regarding the synthesis of results.

2 OPERATIONALISATION OF THE ASSESSMENT

2.1 Introduction

Health technology assessment begins with the definition of the assessment matrix, that is, by the definition of PICO. PICO is an acronym used in evidence-based practice (and specifically in Evidence-Based Medicine) to structure and answer a clinical or medical care question. The PICO structure is also used to develop search strategies in the literature, for example, in systematic reviews. The acronym PICO means: P – Population; I – Intervention; C - Comparison, control or comparator; O – Outcome measures. It is this matrix that will define the terms in which the assessment will be carried out.

2.2 Definition of the assessment matrix

Definition of population(s)

The population (patients) should be defined taking into account the clinical characteristics of the patients included in the therapeutic indication under assessment. In case the indication under assessment includes different populations distinguished by the presence of effect modifiers, or usually receiving different treatment, it should be considered to divide the population included in the approved indication into two or more populations, and to assess the treatment effect separately for each one (see also 6 Subgroup analysis section).

Intervention

The intervention should include only the intervention under assessment and should not include other products that are not part of the indication of interest. However, if the technology under assessment is to be used in combination with other technologies, these should be part of the definition of the intervention.

Selection of comparators

Comparators are all therapeutic alternatives commonly used in clinical practice in Portugal to treat the indication for which the medicinal product under assessment has marketing authorisation (MA). Comparators are options against which the new medicinal product is compared with the objective of assessing whether it has additional benefit and is cost-effective.

The selection of a given intervention as a comparator does not translate into a judgement on its efficacy, the only inclusion criterion being its usual use in clinical practice in Portugal. Therefore, the relevant comparators should not be constrained by the comparators used for the control group in clinical trials of the medicinal product under assessment.

Comparators may include:

- medicines with MA for the indication;
- inactive therapeutic options (e.g. best supportive care, monitoring), if commonly used in Portuguese clinical practice for the indication;
- non-pharmacotherapeutic active options (e.g. surgery), if used in Portuguese clinical practice for the indication;

- therapeutic sequences in which the medicinal product under assessment is used in a second line or another subsequent line, if applicable to the indication and if permitted in its MA and, if relevant, permitted in the comparators' MA.

In exceptional cases, medicines without MA may be included as comparators for the indication provided that their use is well established in Portuguese clinical practice for the indication.

At the stage of defining the assessment matrix, all relevant comparators for the indication should be identified.

Justification for the identification of all relevant comparators

The pharmacotherapeutic assessment aims to determine the efficacy, safety and added therapeutic value of the medicinal product under assessment in relation to each of its comparators. As pharmacotherapeutic assessment is comparative, the results and conclusions necessarily depend on the efficacy and safety of all relevant comparators because the additional benefits of the medicinal product under assessment depend on the evidence on various measures of efficacy and safety, namely the magnitude of differences from comparators and the uncertainty about these differences.

Justification for the inclusion of comparators that are medicines without an MA for the indication, but that are commonly used in Portuguese clinical practice for the indication

There are specific clinical situations where clinical practice includes the use of a non-MA medicinal product for an indication. For the therapeutic assessment to reflect the comparators in Portuguese clinical practice, it is necessary to include these medicinal products as comparators. If these medicinal products are not included, therapeutic and cost-effectiveness assessments may conclude that the added value of the medicinal product is greater than it actually is.

Justification for inclusion of inactive therapeutic options

Inactive therapeutic options, such as improved supportive care or monitoring, are relevant for indications where Portuguese clinical practice includes them as therapeutic options. Even if there are medicinal products with an MA for an indication where better supportive care or monitoring are options, it is necessary to include these in the medicinal product assessment. The exclusion of these options may lead to the additional benefits of the medicine being overestimated or its costs underestimated because these options may offer advantages in terms of adverse effects or costs. As such, their exclusion could have adverse consequences for the conclusion on their added value and the price at which they will be reimbursed by the SNS.

Justification for inclusion of active non-pharmacotherapeutic options

Non-pharmacotherapeutic active options are treatments that do not involve medicinal products, such as surgery, physiotherapy, psychological counselling, etc. As discussed above, the exclusion of these options, when they are part of the therapeutic stockpile in Portuguese clinical practice, may lead to the additional benefits of the medicine being overestimated or its costs underestimated, with adverse consequences for the reimbursement decision.

Justification for consideration of therapeutic sequences

Therapeutic sequences are relevant when patients can be treated with one of the pre-existing first-line treatment options, with the new medicinal product being reserved if the first-line treatment is not effective. In cases where the efficacy of the new second-line medicinal product is high, and pre-existing

therapeutic options have some efficacy, high safety and the duration of treatment is short, the therapeutic sequence may have similar efficacy to the new medicinal product. Therapeutic sequences are relevant when the MA does not restrict the medicine to a particular therapeutic line.

Definition of efficacy and safety measures and classification of their importance

A set of outcome measures should be proposed, related to the efficacy and safety of the intervention. These measures should allow for a comprehensive view of the effect of treatment and should include measures relevant to the patient. To this end, measures that assess the patient's symptoms, functional capacity or life expectancy, i.e. mortality, morbidity (symptoms and complications), duration of illness, and health-related quality of life (HRQoL) are considered relevant to the patient.

Therapeutic efficacy and safety measures must be classified, according to the degree of importance attributed to them, as critical and important but not critical, according to the CATS assessment methodology. Therapeutic efficacy and safety measures should be considered critical when, from the evaluator's perspective, they may influence the direction of the assessment. As a general rule, measures that assess symptoms, functional capacity or patient life expectancy, i.e. mortality, morbidity (symptoms and complications), duration of illness, and health-related quality of life, should be considered critical.

The classification is made on a scale of one to nine: measures classified as important should be quantified with a score between four and six and measures classified as critical between seven and nine. The final score assigned to the outcome measures should be the average of the scores assigned by each of the Assessment Group members during the meeting to discuss the assessment matrix. The score is rounded to the nearest whole number, whereby decimal places are disregarded and, where the number after the comma is five or more, this number is increased by one (1).

2.3 Conclusions

Health technology assessment starts by defining the structure of PICO.

All therapeutic alternatives that are commonly used in clinical practice in Portugal to treat the indication for which the medicinal product under assessment has marketing authorisation and the applicant has applied for funding should be selected as comparators. The selection of a given intervention as a comparator does not translate into a judgement on its efficacy, the only inclusion criterion being its usual use in clinical practice in Portugal.

A set of outcome measures should be proposed, related to the efficacy and safety of the intervention. These measures should allow for a comprehensive view of the effect of treatment and should include measures relevant to the patient. To this end, measures that assess the patient's symptoms, functional capacity or life expectancy, i.e. mortality, morbidity (symptoms and complications), duration of illness, and health-related quality of life, are considered as relevant to the patient.

3 GENERAL METHODOLOGY

3.1 The relevance of the certainty of results

The aim of HTA is to inform decision-makers, with as much confidence as possible, about whether there is available evidence that proves the benefits or harms of a specific intervention compared to alternatives already used in clinical practice.

For the assessment of the added therapeutic value, the Evidence-Based Medicine methodology is used. Note, that in medicine the benefit of interventions is assessed in terms of probability: benefit is demonstrated when the intervention increases the probability of a given beneficial outcome or reduces the probability of a non-beneficial outcome.

Evidence-Based Medicine allows us to assess the extent to which the available evidence is reliable. To this end, it uses a set of internationally accepted rules and instruments that form the basis of benefit assessments. Assessments include analysing a range of details about how studies were planned, conducted, analysed and published.

Comparative randomised clinical trials are considered the most appropriate method for estimating measures of relative treatment effect. These should be integrated into a systematic review and synthesised through meta-analysis, conventional or network (see section 4 Comparison methods). Uncertainty in the evidence should be identified and explored in sensitivity analyses. Non-randomised evidence can only be used in specific situations, which should be adequately justified (see section 5 Comparison methods in exceptional situations).

It is recommended that HTA should be based only on studies with sufficient certainty of results. It is the responsibility of the company that holds the marketing authorisation to submit the dossiers for assessment by INFARMED, I.P. including all the evidence it considers relevant. If it is apparent from the review of this evidence that the studies included in the submission process do not answer the research questions, it may be concluded that, with the documentation submitted, there is no evidence available to prove the additional benefit of a specific intervention.

3.2 The link between certainty of results and proximity to everyday conditions

It is often stated that studies with high certainty of results (for the purposes of this paper, ‘certainty of results’ means high confidence in effect estimates) have high internal validity, but do not always represent the population in current practice, which is usually more heterogeneous. In other words, the results have low external validity and are therefore not generalisable.

However, this criticism does not result from the methodology used in the study but from the fact that the eligibility criteria for such studies are generally very restrictive, often excluding elderly patients or patients with multiple comorbidities; or the randomised study protocol does not reflect clinical practice (e.g. differences in dosage, differences in stop-restart rules, differences in patients’ previous therapy; differences in subsequent therapy, differences in monitoring intensity). Thus, increasing external validity does not imply reducing the certainty of results, but rather including those patient groups considered relevant.

Thus, high certainty of results and proximity to everyday conditions are not mutually exclusive. Comparative, randomised studies with high internal and external validity (e.g. pragmatic studies) are preferable.

3.3 Outcome measures

There must be a prior definition of which therapeutic efficacy and safety measures will be used in the assessment. The therapeutic efficacy and safety measures used should be relevant to the patient. The following measurements are recommended for this purpose: mortality, morbidity (symptoms and complications), duration of illness, and health-related quality of life.

3.3.1. Clinical effect measures

Clinical effect measures are characteristics or variables that reflect a patient's symptoms, functional capacity or life expectancy. In the context of the assessment of an intervention (e.g. a medicinal product), the treatment effect on these measures is a change that is detectable by the patient, such as an improvement in symptoms, an improvement in functional capacity, a decrease in the likelihood of developing a disease or a complication of that disease, or an increase in survival.

In health technology assessment, and in the definition of the assessment matrix, measures of clinical effect should preferably be valued. Therefore, and as a general rule, only measures of clinical effect should be given maximum importance (measures whose importance should be rated as critical).

3.3.2. Surrogate effect measures

For the purposes of this document, a biomarker is defined as a characteristic that is objectively measured and assessed as an indicator of pharmacological response to a therapeutic intervention.

A surrogate effect measure is a biomarker that is intended to replace a clinical outcome measure, i.e. a biomarker that is expected to be able to predict clinical benefit, harm, or lack of benefit or harm. This expectation should be supported by robust evidence (‘validation’).

In health technology assessment, and in the definition of the assessment matrix, surrogate effect measures should be valued less than clinical effect measures. Thus, and as a general rule, surrogate effect measures should not be accorded maximum importance (i.e. critical importance). These measures should generally be classified as ‘important but not critical’.

3.3.3. Validation of surrogate effect measures

3.3.3.1. Introduction

The purpose of replacing a clinical outcome measure with a surrogate outcome measure in a randomised trial is to allow valid statistical inference regarding the efficacy of an intervention on a clinical outcome measure, without the effect on that clinical outcome measure having been observed. Consequently, the use of a surrogate outcome measure requires extrapolation beyond the observed data to estimate the true benefits to be expected for patients.

In medical research it is common to use surrogate outcome measures as substitutes for patient-relevant therapeutic efficacy measures, with the aim of obtaining conclusions about the efficacy of interventions on clinical outcome measures earlier and at lower cost. The use of surrogate outcome measures in clinical trials allows for a reduction in the number of participants and the duration of trials compared to the use of clinical outcome measures.

In the Committee's view, the use of surrogate outcome measures has the potential advantage of accelerating access to innovative technologies that offer added value for patients. However, surrogate outcome measures are often not able to reliably predict the overall effect on clinical outcome measures. The aim of this section is to recommend the use of a methodology that ensures that the surrogate outcome measures used are able to reliably predict the overall effect of the intervention on clinical outcome measures.

If the evidence submitted by the Marketing Authorisation Holder (MAH) uses surrogate outcome measures, it should also contain information on which clinical outcome measure the surrogate measure replaces and include demonstration of the validation of the surrogate measures used, using the methodology recommended herein.

Studies using a surrogate outcome measure as the basis for medicinal product co-payment/funding decisions

Studies using surrogate outcome measures often overestimate the effect of treatment (2). In recent years, different countries have approved a substantial number of medicinal products based on surrogate outcome measures (3)(4).

The need to assess studies using surrogate outcome measures may be of particular relevance in the context of early assessment of the beneficial effects of medicinal products.

3.3.3.2. Requirements of a surrogate outcome measure

For a surrogate outcome measure to be an effective substitute for a clinical outcome measure, the effects of the intervention on the surrogate outcome measure must be able to reliably predict the overall effect on the clinical outcome measure, but in practice this condition is often not observed. Among other explanations for this fact, there is the possibility that the pathological process affects the clinical outcome measure through various causal mechanisms not mediated by the surrogate, and the effect of the intervention on these causal mechanisms is different from its effect on the surrogate (5). Even more likely, the intervention may affect the clinical outcome measure by unrecognised, unanticipated, and unintended mechanisms of action that operate independently of the pathological process (5).

Importantly, in some cases, the biomarker is strongly predictive of survival but does not predict the effect of treatment on survival. CD4+ counts, used in HIV studies, are an example of such markers.

Thus, in assessing the additional benefit of an intervention, surrogate measures of therapeutic efficacy may be considered as substitutes for measures of clinical efficacy provided they have been previously validated.

3.3.3.3. Validation of a surrogate outcome measure

There are no standardised procedures to validate a surrogate outcome measure. The methodological literature often advocates the use of correlation methods for surrogate validation, recommending that correlations be estimated at the individual level and at the study level (3). Thus, in their assessments of benefit, preference should be given to validations using these procedures. These procedures generally require the conduct of meta-analyses of randomised trials, reporting the surrogate and final outcomes, in which the effect of the intervention on the surrogate outcome measure and the clinical outcome measure is assessed. Only in exceptional cases are alternative methods considered.

Thus, validation goes through three stages. First, assess the biological plausibility of the relationship between the surrogate outcome measure and the clinical outcome measure (level three). Second, assess whether there is a strong correlation between the surrogate outcome measure and the clinical outcome measure in different cohorts or at the individual patient level (this correlation does not validate the surrogate measure but may identify good prognostic markers) [level two]. Third, assess whether there is demonstration of a relationship between the treatment effect in the surrogate and the effect on the clinical outcome measure, preferably in multiple randomised trials (level one). In the case of new health technologies, which commonly use surrogate outcome measures, evidence should be sought from other studies evaluating the same or similar health technologies (including drugs from the same class or, if this evidence is not available, including drugs from different classes) (6). While the second criterion is easily met, the third is not. There is no consensus on the correlation values (thresholds) required for validation of a surrogate, but often correlation coefficient values (R_{study} or $R_{individual}$) between 0.85 and 0.955 are given. If there is not a high correlation, the surrogate threshold effect (STE) can still be used. This parameter is also based on the analysis of several randomised trials and defines what is the minimum absolute value of the effect on the surrogate that has to be observed to deduce an effect on the clinical outcome measure (3). Thus, the STE at which a certain level of variation in the biomarker turns into clinical benefit can be calculated. In both cases, certainty in the conclusions depends on pre-specified levels of significance.

To validate a surrogate, the correlation between the treatment effect in the surrogate and the treatment effect on the clinical outcome measure, assessed at the study level using the thresholds defined above, should be used primarily.

Note that correlation estimators are sensitive to small changes in the data and the calculation of their confidence interval is problematic when the samples of the included studies are small or moderate (7). On the other hand, correlation measures only reflect (approximately) linear relationships between treatment effects on the surrogate and the clinical outcome measure and cannot be used to demonstrate relationships with other forms. It is therefore important to consider the 3 stages of validation of a surrogate measure, as using correlation alone may exaggerate or dilute the importance of the relationship between the measures under consideration (8).

It may be considered acceptable that, in exceptional situations, non-validated surrogate outcome measures may be accepted in cases where there is a reasonable likelihood that the marker is capable of predicting clinical benefit, and provided that the practical impossibility of validating the surrogate outcome measure is demonstrated, for example because the time required to observe the event (clinical outcome measure) is excessively long. A practical example of this is the use of sustained virologic response as a surrogate outcome measure of mortality or hepatocellular carcinoma in chronic hepatitis C. For the purposes of this 'reasonability' there must be at least biological plausibility (level three validation), and a correlation must be observed between the surrogate and the clinical outcome measure (level two validation).

It should be noted that the existence of a correlation between the treatment effect on the surrogate and the treatment effect on the clinical outcome measure in an intervention with a specific mode of action does not necessarily mean that this correlation is observed with other interventions used to treat the same disease that have a different mechanism of action. Thus, validation of a surrogate outcome measure is usually done in a specific population, and for a specific intervention, i.e. validation is disease-specific, population-specific, and therapeutic area-specific.

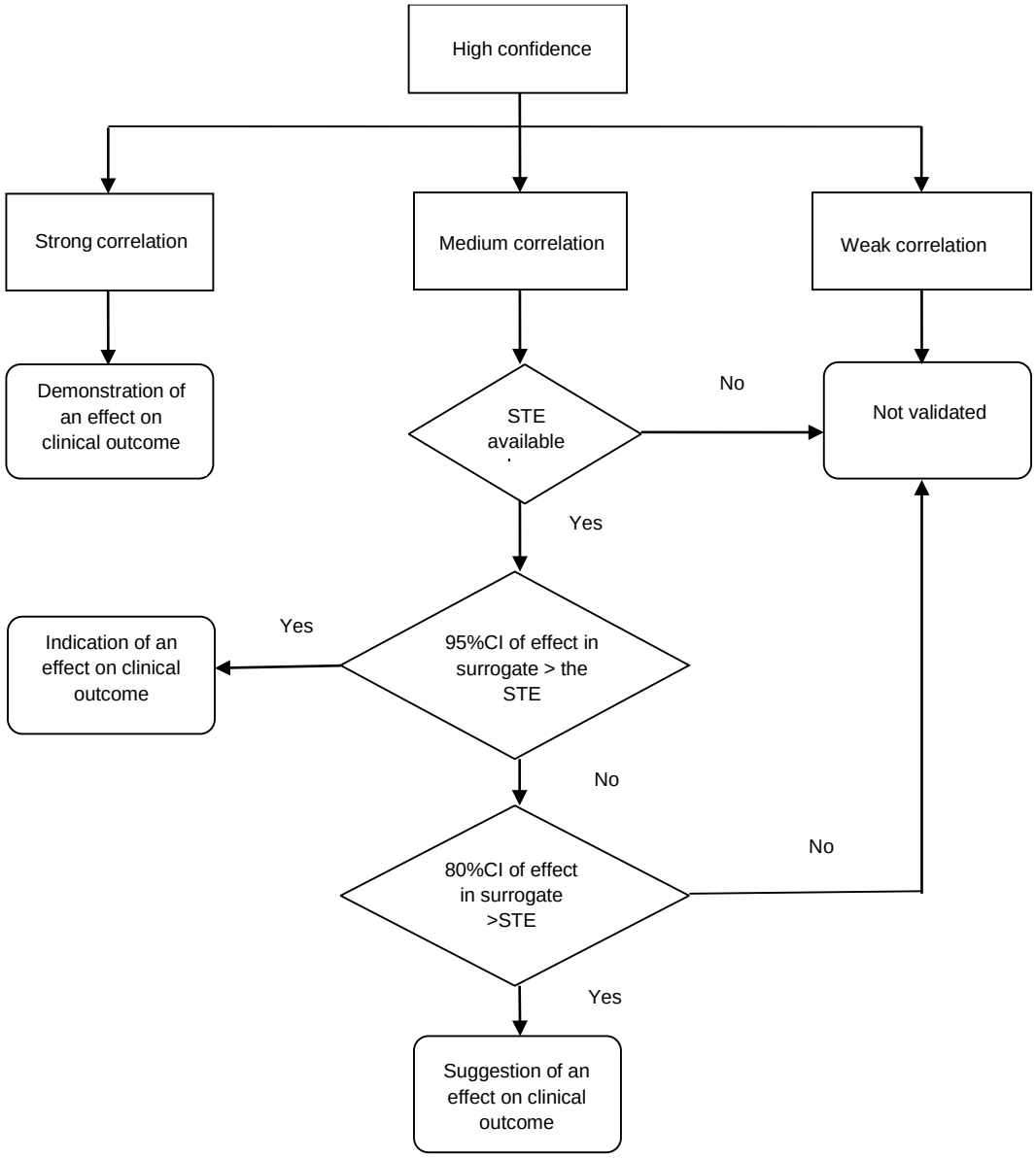
Since new health technologies (using new/different mechanisms of action) are initially evaluated in studies using surrogate outcome measures, there may be no evidence based on clinical outcome measures. Thus, validation of the surrogate can only originate from studies with drugs with different mechanisms of action/different pharmaceutical classes. It is recommended that the use of surrogate outcome

measures that have only been validated for medicinal products used for the same indication, but with different mechanisms of action, only take place when there are no treatment alternatives for the indication under assessment or when there is an indication, with reasonable probability, that the new medicinal product may present additional benefit in relation to existing alternatives and the disease is severe or life-threatening.

The conclusion about the validation of a surrogate depends on two factors: the quality of the evidence supporting validation and the strength of the correlation between the effect of the intervention on the surrogate and the effect of the intervention on the clinical outcome measure.

In case the validation study is classified as high quality, the conclusion about the validation of the surrogate depends on the strength of the correlation between the treatment effect in the surrogate and in the clinical outcome measure or STE value. The flow diagram in Figure 1 describes the classification process in detail (9).

Figure 1: ***Conclusions on the validation of the surrogate in the case of high-quality evidence***



Adapted from Ref. 9

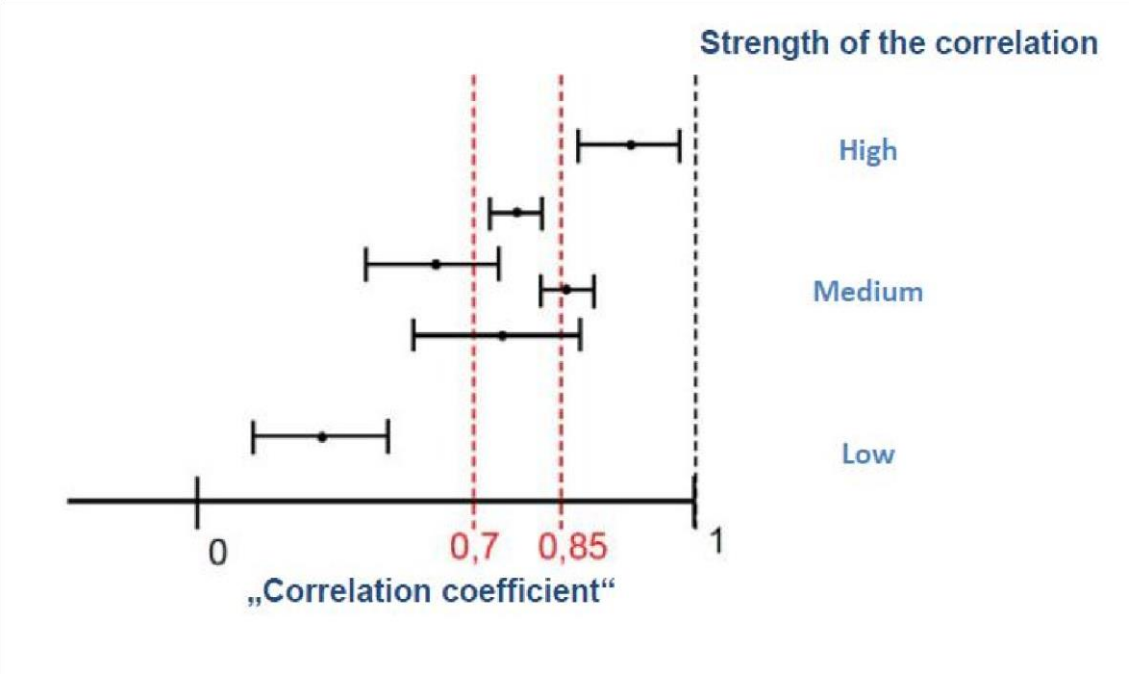
CI: confidence interval; STE: Surrogate threshold effect

Regarding correlation strength, a correlation is classified as strong if the lower limit of the confidence interval of the correlation coefficient R is ≥0.85, it is classified as weak if the upper limit of the confidence interval of R is ≤0.70, and it is classified as medium if the confidence interval of R overlaps, even partially, the interval between <0.85 and >0.70 (Figure 2) (9).

In case the validation study is classified as high quality, it is considered that there is demonstration of validation of the surrogate if there is a strong correlation between the effect of the intervention on the surrogate and the effect on the clinical outcome measure; that there is no demonstration of validation of the surrogate if a weak correlation was observed between the effect of the intervention on the surrogate and the effect on the clinical outcome measure; and it is considered unclear whether the surrogate is validated if there is a medium correlation (Figure 1) (9).

In this case, the surrogate threshold effect (STE) and the confidence interval of the effect of the intervention on the surrogate are used to reach a conclusion on validation (Figure 2).

Figure 2*: Classification of correlation strength as a function of the correlation between treatment effect on the surrogate and the clinical outcome measure*



In case the validation study is classified as of moderate or low quality, it is considered unclear whether the surrogate is validated (Table 1).

Table 1: *Validity of the surrogate outcome measure as a function of the quality of evidence (validation study) and the correlation between the treatment effect on the surrogate and on the clinical outcome measure*

Quality of evidence	**Correlation**	**Validity**
High	Strong	Yes
	Medium	Not clear - use STE
	Weak	No
Moderate		Not clear - use STE
Low		Not clear
Very Low		

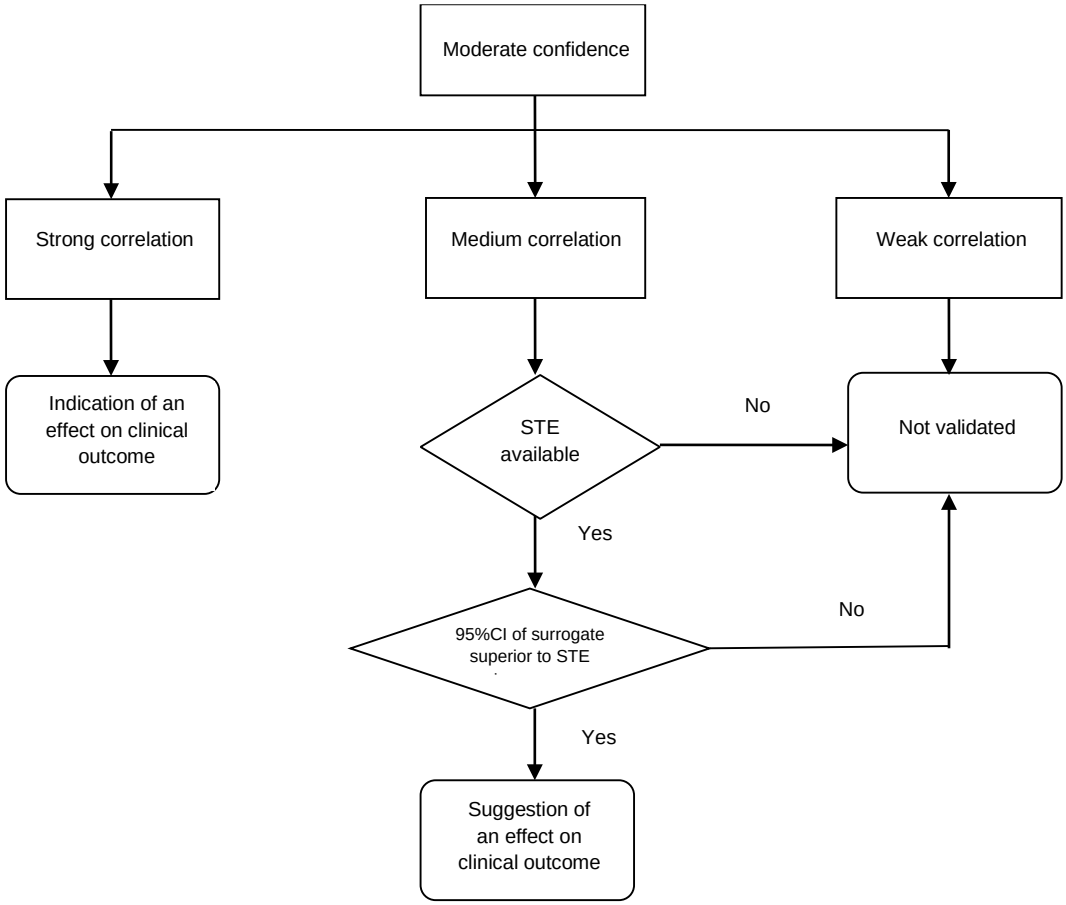
Source: adapted from IQWiG Reports – Commission No. A10-05. Validity of surrogate endpoints in oncology. Version 1.1. 21.11.2011. (9)

In case the validation study is classified as of moderate quality, the conclusion on the validation of the surrogate depends on the strength of the correlation between the treatment effect in the surrogate and

on the clinical outcome measure or the STE value. The flow diagram in Figure 3 describes the classification process in detail (9).

In case the application of this methodology leads to different results in different surrogate validation studies, the validation outcome is considered to be the one resulting from the majority of high-quality studies.

Figure 3**:** *Conclusions on the validation of the surrogate in the case of moderate-quality evidence*



Source: Modified from Ref. (IQWiG Reports – Commission No. A10-05. Validity of surrogate endpoints in oncology. Version 1.1. 21.11.2011. (9)

CI: confidence interval; STE: Surrogate effect threshold

3.3.3.4. Conclusions

The validation of a surrogate outcome measure goes through three steps. First, assess the biological plausibility of the relationship between the surrogate outcome measure and the clinical outcome measure (level three validation). Second, assess whether there is a strong correlation between the surrogate outcome measure and the clinical outcome measure in different cohorts or at the individual patient level (level two validation). Third, assess whether there is demonstration of a relationship between the treatment effect in the surrogate and the effect on the clinical outcome measure, preferably in multiple randomised trials (level one validation).

There is no consensus on the correlation values (thresholds) required for validation of a surrogate, but often correlation coefficient values (R_{study} or $R_{individual}$) between 0.85 and 0.955 are given. If there is not a high correlation, the surrogate threshold effect (STE) can still be used. This parameter defines what is the minimum absolute value of the effect on the surrogate that must be observed to infer an effect on the clinical outcome measure.

Since new health technologies (using new/different mechanisms of action) are initially evaluated considering studies using surrogate outcome measures, there may be no evidence based on clinical outcome measures. Thus, validation of the surrogate can only originate from studies with drugs with differ-

ent mechanisms of action/different pharmaceutical classes. In this case, and in order to assess transferability, validation studies that include several interventions in the same indication should at least include data on heterogeneity. However, the use of surrogate outcome measures that have only been validated for drugs used in the same indication but with different mechanisms of action is only justified when there are no treatment alternatives for the indication under assessment or when there is an indication, with reasonable probability, that the new drug may present additional benefit compared to existing alternatives and the disease is severe or life-threatening.

Non-validated surrogate outcome measures may be accepted where there is a reasonable likelihood that the marker is capable of predicting clinical benefit, provided that the practical impossibility of validating the surrogate outcome measure is demonstrated, for example because the time required to observe the event (clinical outcome measure) is excessively long. For the purposes of this 'reasonability' there must be at least biological plausibility (level three validation), and a correlation must be observed between the surrogate and the clinical outcome measure (level two validation).

If the evidence submitted by MAH uses surrogate outcome measures, it should also contain information on which clinical outcome measure the surrogate measure replaces and include demonstration of the validation of the surrogate measures used, using the methodology recommended here.

3.4 Systematic reviews

3.4.1. Introduction

The systematic review is a methodological process specifically devolved to identify, select and critically assess the available studies on a clearly formulated question. It is fundamental to the assessment process that the evidence base considered is comprehensive and complete. The systematic review should be transparent and objective in order to reduce bias and ensure that the most valid answer is obtained.

Thus, the scientific evidence to be considered to inform the assessment and determine the comparative efficacy of treatment includes secondary studies, namely systematic review of clinical trials on the intervention under analysis, and primary studies. Within primary studies, randomised trials provide the highest standard of evidence regarding the comparative efficacy of a treatment and should be preferred whenever possible. However, data from non-randomised studies may be required to supplement the available data or to provide information on other assessment parameters, such as adverse effects and cost. Data from the included studies can be synthesised using meta-analysis (see section 4 Methods of comparison).

3.4.2. Research protocol

It is the responsibility of the MA holder to systematically review relevant evidence to inform the health technology assessment process. This process should be conducted according to the best internationally established standards for systematic reviews, namely those defined by PRISMA (10) (Preferred Reporting Items for Systematic Reviews) and MOOSE (11) (Meta-analysis of Observational Studies*)* consensuses*.* Also mentioned are the EUnetHTA (European Network for Health Technology Assessment) recommendations for the information gathering process for systematic reviews and clinical efficacy assessments of health technologies (12).

This imposes the need to formulate a protocol for the conduct of the systematic review, in which the inclusion and exclusion criteria, effect measurement, search strategy and planned analyses are stated in advance. These aspects should be guided by what was previously defined as population, intervention,

comparators and outcome measures of interest for the assessment, and should be documented in sufficient detail to ensure the transparency and reproducibility of the systematic review performed.

3.4.3. Databases

With regard to the sources of information to be searched, a wide range of bibliographic databases should be included to ensure the identification of all relevant studies for the topic under assessment. The existence of completed or ongoing clinical trials of relevance to the intervention under assessment should also be systematically searched and reported, including the clinicaltrials.gov and EUDRACT (European Union Drug Regulating Authorities Clinical Trials Database) interfaces. For different aspects of the assessment, sources of information other than published literature, such as registers, may also be considered, as appropriate.

3.4.4. Research strategy and study selection

The research strategy must be reproducible, established in line with the framework and the final objective of the assessment. Literature selection should be based on explicit inclusion and exclusion criteria, using standardised and recognised methodologies. Ineligible studies should be listed, along with the justification for their exclusion and a flow chart summarising this information. Every effort should be made to include all relevant evidence, regardless of language.

3.4.5. Assessment of the quality of evidence

The review should reveal the best and most up-to-date evidence on the clinical efficacy of the intervention in relation to its comparators. It is therefore fundamental to critically evaluate the scientific evidence used to make the assessment with regard to its validity, quality and applicability. Any potential biases that result from the design of the studies used in the assessment should be explored and documented. The external validity of the results of the studies included in the review should also be considered, as well as their applicability to Portuguese clinical practice.

Clinical efficacy estimates of comparator treatments shall be based on data from the best quality studies available and shall apply, within the indication under assessment, to the typical patient in normal clinical circumstances, assessing relevant clinical effect measures and making comparison with appropriate comparators, using relative and absolute measures of efficacy and appropriate measures of uncertainty.

Many factors can affect the overall estimate of relative treatment effect that is obtained from the systematic review. Differences between included studies may result from differences in patient characteristics (e.g., age, gender, severity of illness) or other factors, such as differences in the measurement of effect measures or in the context of care delivery, for example. Potential modifiers of treatment effect should be identified prior to data analysis, through extensive review of the topic and discussion with experts in the clinical discipline concerned.

Where sufficient valid and relevant data expressed in comparable measures of effect are available, quantitative synthesis by meta-analysis is possible and appropriate. Similarly, where comparator treatments have not been evaluated in the same randomised clinical trial, consideration should be given to performing network meta-analysis as appropriate. These methodologies and their application are detailed in section 4.4 Network meta-analysis.

4 METHODS OF COMPARISON

4.1 Introduction

Health technology assessment evaluates the additional benefit and cost-effectiveness of an intervention in relation to comparators of interest. For this purpose, it uses different methods of comparison.

Randomised clinical trials summarised in a meta-analysis (conventional or network) are preferred for estimating comparative effects of the intervention under study and its comparators. Non-randomised evidence may be accepted in specific situations, which should be adequately justified (see section 5 Methods of comparison in exceptional situations).

4.2 Direct and indirect comparisons: definitions

Direct comparison between two specific treatments is understood as the comparison in a study of these two treatments, or the combination of multiple studies of these same treatments, to generate a combined estimate (meta-analysis) of the relative efficacy of the two treatments.

Indirect comparison is the estimate of the relative efficacy between two or more treatments in the absence of studies that directly compare them. Mixed treatment comparison is defined as estimating the relative efficacy of 3 or more treatments using both direct and indirect comparisons simultaneously. The term network meta-analysis encompasses direct, indirect and mixed comparisons.

When the available evidence includes several studies that compare treatments directly, sometimes the results of these studies are combined using meta-analytic techniques to generate a pooled estimate of the relative efficacy of the two treatments.

However, sometimes there is insufficient data to reliably estimate the relative efficacy of two treatments or there may be a need to compare more than two treatments simultaneously, in which case it is necessary to use multiple treatment comparison methods, for example network meta-analysis.

Thus, multiple treatment comparison methods can be used to infer the relative efficacy of two or more treatments in the absence of studies comparing them directly or by combining direct and indirect comparisons.

It is important to emphasise that, the method of meta-analysis (conventional or network) used must keep intact the original randomisation of the included primary studies. Comparisons that do not maintain randomisation have the same value as comparisons using observational studies and are not recommended.

Conventional or network meta-analyses (including indirect comparisons) should only be conducted if the available studies are comparable, homogeneous and consistent, so that the results obtained can be reliable. These issues are discussed in more detail in subsequent sections.

4.3 Conventional meta-analysis

4.3.1. Introduction

Meta-analysis consists of the application of a set of statistical methodologies that allow the aggregation of results from a set of primary studies in order to generate one or more meta-analytic summary

measures. It is also possible to analyse the presence, magnitude and potential moderators of heterogeneity in the identified evidence base (13)(14)(15). This methodology of synthesis and quantitative analysis of the evidence base follows naturally from a systematic review and is usually its last phase.

When the available evidence includes two or more studies making a direct comparison of the interventions of interest, the outcomes of these studies can be combined using meta-analytic techniques to generate a pooled estimate of the relative efficacy of the two treatments. However, conventional meta-analysis only allows two treatments to be compared with each other. Aggregation of distinct interventions to form a single 'treatment' for the purposes of meta-analysis is not advisable and must be clinically justified. When multiple treatments are under consideration, network meta-analysis methods should be used (section 4.4 Network meta-analysis).

Once the measures of effect have been identified and data extracted from each primary study allowing their calculation and associated measures of precision (standard error or confidence intervals), it will be possible to calculate meta-analytic summary measures that represent in aggregate form the quantitative results of the included primary studies. These measures result from the aggregation of the effect measures of the various primary studies included, considering specific and distinct weightings for each study. These different weightings take into account the precision of each study included (different sample sizes, different variability) and the heterogeneity between studies. Naturally, for example, it will be expected that if a study has a larger number of individuals analysed, its outcome will carry more 'weight', at the time of the calculation of the summary measure, than the results of other smaller studies.

The advantages of a meta-analysis include increased statistical power and precision, the possibility of answering questions not directly asked in the individual studies, and the resolution of controversies when individual studies reach contradictory conclusions.

However, it should be kept in mind that the results of a meta-analysis may be affected by differences in the design and characteristics of the included studies and by different types of bias.

When performing a meta-analysis it is necessary to specify the effect measure used to describe the efficacy of the intervention (e.g. relative risk), the statistical method of weighting (e.g. inverse variance weighting), model used (fixed-effects, random-effects), and the method of statistical inference (frequentist or Bayesian). The variability, magnitude and relevance of differences between studies (heterogeneity) should also be assessed. The consistency of the results of the individual studies may influence the decision to combine them through a meta-analysis and the decision on the analytical model to be used.

4.3.2. Factors affecting precision

Sometimes individual studies are too small to estimate the relative efficacy of two treatments with sufficient precision. The use of meta-analytic techniques to combine the results of several studies generates a combined estimate of the relative efficacy of the two treatments and may result in an increase in the precision of the estimate of treatment effect.

The inverse variance weighting method is a simple and common method of conducting meta-analyses, used for both dichotomous and continuous variables. It is so called because the weight given to each study is the inverse of the variance of the effect estimate (i.e. one over the square of its standard error). More weight is given to larger studies, with smaller standard errors, than to smaller studies, which have larger standard errors. This harnesses the evidence from all included primary studies and minimises the inaccuracy (uncertainty) of the combined estimate of treatment efficacy.

4.3.3. Fixed-effect and random-effects models

Two statistical models are often used in meta-analyses:

- the fixed-effect model, assumes that all estimates of the intervention effect from different primary studies estimate the same (true) effect in the population of interest, and that differences observed between different studies included in the meta-analysis reflect only random variation (due to being from one sample);

- the random-effects model assumes that there is variation in the estimates of the intervention effect between the included studies, in addition to random variation due to sampling. This model assumes that each study estimates the true effect value in the study population and that these true effect values follow a particular distribution. In general, this distribution of true effects of each study is assumed to be normal (Gaussian) and the statistical estimation of the mean and variance of this distribution is sought in the context of a hierarchical model with two distinct levels. In this model, the included studies are considered to represent a random sample from a theoretical population of studies that answer the question of interest.

The random-effects model can deal with heterogeneity between studies that cannot be explained by other factors, incorporating it in the calculation of the meta-analytic summary. In a set of heterogeneous studies, the random-effects model gives more weight to smaller study results than the fixed-effects model. The existence of a large heterogeneity among the included studies may cause problems in interpreting the results of the meta-analysis, in particular, by giving more weight to the results of smaller studies, the random-effects model may cause problems in interpreting the results of the meta-analysis (small studies bias).

The fixed-effects model only considers variability within each study, while the random-effects model also considers variability between studies. Consequently, the fixed-effect model gives rise to narrower confidence intervals (i.e. better precision). In the absence of heterogeneity between studies, the results obtained with both models coincide.

In general, meta-analytic measures can be seen as the best available answers to the research question at hand, provided that they result from the appropriate synthesis of the best available evidence, the selection of which has been made in a comprehensive and unbiased manner.

In cases where the hypothesis of homogeneity between studies is not plausible, the random-effects model should be used. However, when we are faced with severe heterogeneity, there is a suggestion that the included studies estimate measures of effect apparently from very different realities. In this case, meta-analytic measures should be interpreted with particular care. Additionally, it is important to try to identify the causes of heterogeneity - that is, it is fundamental to identify the clinical and/or methodological differences between the studies that may explain the observed heterogeneity.

If there are clinical or statistical reasons to assume homogeneity in the relative effects estimated by the different studies, the fixed-effects method can be used. If possible, a sensitivity analysis using the random-effects model should also be presented. However, it is generally not possible to estimate heterogeneity between studies with sufficient precision in meta-analyses with few studies, so a minimum of 3 studies should be considered for a meta-analysis with a random-effects model.

4.3.4. Heterogeneity

Heterogeneity can be defined as any type of variability between studies:

- variability in participants, interventions and outcome measures (clinical heterogeneity);

- variability in study design, measurement of effect estimation and risk of bias (methodological heterogeneity);

- variability in the effect of the intervention to be assessed between different studies or in the baseline risk of different populations (statistical heterogeneity). This may be a consequence of both clinical and methodological heterogeneity.

A meta-analysis should only be performed when a study group is sufficiently homogeneous in terms of participants, interventions and outcome measures.

In case there are studies classified as having high risk of bias, these should be excluded from the meta-analysis and only studies with low risk of bias should be included in the main analysis.

Statistical heterogeneity will be referred to from this point on as just heterogeneity. Some variation (inconsistency) in the results of different studies is expected due to chance alone. Variability that cannot be attributed to chance, reflects real differences in study results, i.e. heterogeneity. If the confidence intervals of the results of the different studies show little overlap, it is an indication of the presence of heterogeneity. This can be assessed more formally through a statistical test.

Heterogeneity can be assessed using the X^2 (Chi-square) test, based on Cochran's Q statistic. This test assesses whether differences between results are due to chance alone. A low p-value (or a high X^2 test relative to degrees of freedom) is evidence of heterogeneity in the estimates of the intervention effects. However, caution is needed in interpreting the results from this test as it has low power when there are few studies included in the meta-analysis or when the sample size is small and excess power when there are many studies included in the meta-analysis.

In contrast, the I^2 statistic quantifies the percentage of the total variation in the estimated effects of the different studies included in the meta-analysis that is due to heterogeneity and not to random variability of a sampling nature and, therefore, should always be considered in addition to the hypothesis test based on Cochran's Q statistic. Some authors consider an I^2 value of less than 25% to be low. There are no universally accepted cut-off points, however, it is generally considered that an I^2 value higher than 40-50% configures a situation of moderate to severe heterogeneity, which should deserve particular attention and exploration. However, the I^2 statistic also suffers from large uncertainty when only a few studies are available and is sensitive to the accuracy of the included studies. Reporting the degree of uncertainty of I^2 (with a 95% confidence interval) is recommended. The heterogeneity estimate, τ (tau) or τ^2 and its confidence interval should also be taken into account (16). When few studies exist, inferences about heterogeneity should be cautious. As a rule of thumb, when there is severe heterogeneity the interpretation of meta-analytic measures should always be done with extreme caution, as they may then not exactly correspond to the best estimate of the treatment effect that is intended to be assessed.

When considerable or severe heterogeneity is observed, it is important to consider the reasons that may explain it. In particular, heterogeneity may be due to differences between subgroups of studies. Also, errors in the execution of the systematic review and data extraction are a common cause of heterogeneity in the results.

4.3.5. Subgroup analysis and meta-regression

When the estimate of the intervention effect varies with different populations or with characteristics of the intervention such as dose or duration of treatment, this variation is known as an interaction or modification of effect. Subgroup analysis and meta-regression are methods used to determine whether there is interaction or whether the results are robust. The rules for defining and assessing the credibility of subgroup analysis are detailed in the section 6 Subgroup Analysis.

Adjustments for meta-regression are considered observational results and should be used for exploratory analyses to identify effect modifiers, or to which results are sensitive, rather than main outcomes.

Subgroup analysis is performed to investigate heterogeneous outcomes and to answer specific questions about a particular patient group, type of intervention or type of study. The results of subgroup analyses can be misleading as they are not based on randomised comparisons.

Meta-regression is an alternative method to test for differences between subgroups. In this case a random-effects model is preferable, due to the risk of false-positive results when a fixed-effects model is used to compare subgroups.

Meta-regression is an extension of subgroup analysis that allows the effect of both continuous and categorical characteristics to be investigated, and the effect of multiple comparators to be investigated simultaneously (if an adequate number of studies exist). Meta-regression should not be considered if there are fewer than 10 studies in the meta-analysis.

In a meta-regression, the outcome measure is the estimate of the effect of the intervention and the explanatory variables are characteristics of the studies that may influence the effect size of the intervention. To avoid risk of bias meta-regression with individual patient data is preferable but rarely possible in the context of pharmacotherapy assessment. The risks of potential aggregation bias should be taken into account when interpreting meta-regression results based on aggregate data.

4.3.6. Meta-analysis with individual data

Meta-analysis of individual patient or participant data (IPD) is a type of meta-analysis that involves obtaining and synthesising individual participant data from several related clinical trials (17). It is considered the benchmark methodology for meta-analyses and is particularly relevant for determining the efficacy of interventions given the specific characteristics of the participants. However, this IPD approach is not often used in practice, as obtaining data from individual participants in each study is an operationally complicated task and often impossible to perform. When it is not possible to obtain individualised data for all studies in the meta-analysis, methods that combine individual participant data with aggregated data can be used (18).

This type of meta-analysis should be performed when conventional meta-analysis is not adequate to answer the pre-defined clinical question. In this case, the use of individual participant data across the different randomised trials allows for increased statistical power to detect distinct treatment effects. The availability of individual participant data facilitates the standardisation of statistical analysis across studies and the direct retrieval of the desired information, regardless of statistical significance or how it was reported in the individual studies. It is possible to analyse data in more detail, obtain results with longer follow-up, include more participants, and investigate hypotheses different from those of the original studies. It also decreases the risk of bias associated with the use of aggregate data in meta-regression. Meta-analysis of individual data is considered more reliable than conventional meta-analysis and may lead to different conclusions. However, it is an organisationally more complex, costly and time-consuming technique.

Statistical methods used in meta-analysis of individual data should preserve the clustering of participants in each study (19). The clustering of participants is maintained during the analysis, and two possible approaches can be used, one with a single stage and one using two stages. In the two-stage approach, first, individual participant data are analysed independently in each individual study using the statistical method appropriate for the type of data being analysed, which produces aggregate results for each study. Then, in a second stage, these data are synthesised using a model suitable for aggregated data analysis, in a similar way to conventional meta-analysis. Thus, fixed-effect and random-effect models can be used to estimate the effect of the intervention. Alternatively, a one-stage technique can be used, i.e. in a single model where individual data from participants in the various studies can be analysed simultaneously using specific techniques to maintain the grouping of patients in each study (typically a regression with a separate term for each study or one that varies between studies via random effects).

Again, it is necessary to use a specific model for the type of data to be analysed and to respect the assumptions of meta-analysis. These two one- or two-stage meta-analysis techniques usually yield similar results. However, when the included studies are small and/or the effect is large or the events are rare, there is a risk of bias with the two-stage technique because some pre-specifications of the second stage may not be appropriate (20).

It is also important to recognise that the quality of the individual data for meta-analysis is dependent on the quality of the original studies as well as a properly performed systematic review. Thus, a meta-analysis of individual data should follow a pre-defined protocol and include an assessment of the quality of the original studies (20). If appropriate, it should be made clear how the inclusion of lower quality data affects the conclusions.

4.3.7. Effect measures and their interpretation

For binary outcome measures, the most common measures to estimate the effect of the intervention include hazard ratio, odds ratio and risk difference.

For continuous outcome measures, the measure used to estimate the effect of the intervention is the mean difference. Its use is appropriate when the estimation of the intervention effect is done on the same scale for the different studies.

When the studies to be combined use different scales the effect of the intervention in each study can be divided by the standard deviation to form a standardised mean difference that reflects the magnitude of the effect of the intervention in each study relative to the variance of the scale. The choice of the standard deviation to be used is not consensual and may cause bias and heterogeneity, especially in the case where the included studies are small - the standard deviation observed in each study is usually used. However, this assumes that this standard deviation is identical across all included studies which is rarely realistic (13)(21). The use of externally estimated standard deviations for each scale can be a solution and facilitates the conversion of the estimated relative effect to one of the original scales which facilitates interpretation (13).

For time-to-event outcome measures, the hazard ratio is the most common measure to estimate the effect of the intervention. The logarithm of the hazard ratio and its standard error for each study must be included in the meta-analysis. Risk ratio and odds ratio (related to events occurring at a given time) are not equivalent to hazard ratio, and median survival times should not be used in a meta-analysis.

The selection of the measure used to describe the combined efficacy of the intervention depends on 3 factors:

- consistency: the measure chosen should originate similar estimates in all meta-analysis studies and subpopulations in which the intervention will be applied. The more consistent the measure chosen, the more robust is the justification for describing the combined estimate of the effect of the intervention with a single value. Relative measures are usually more consistent than absolute measures and therefore, meta-analyses using the risk difference should be avoided. Odds ratios and risk ratios, are usually equivalent in terms of consistency; meta-analysis of odds ratios has better statistical properties, but risk ratios are easier to interpret (22)(23);
- mathematical properties: the chosen measure must have the necessary mathematical properties to perform a valid meta-analysis where the most relevant property is the existence of an adequate and easily applicable estimator of its variance (22)(23);

- the measure used to describe the efficacy of the intervention should be easily interpretable and applicable to the purpose of the meta-analysis. It may be suitable to perform a sensitivity analysis to determine whether the measure chosen to estimate the combined efficacy of the intervention influences the conclusions of the meta-analysis.

4.3.8. Ways to report the results of a meta-analysis

The results of a meta-analysis are usually illustrated using a forest plot:

- a forest plot includes the intervention effect estimate and confidence intervals for individual studies and from the meta-analysis;
- each study is represented by a block with the estimated intervention effect and a horizontal line extending to either side of the blocks;
- the block area indicates the point estimate of each study included in the meta-analysis, while the horizontal line represents the confidence interval (usually a 95% confidence interval);
- the confidence interval represents the intervention effect interval compatible with the study result;
- the block size may also be indicative of the weight of the study (when the block size is larger, the confidence intervals are usually narrower) for the combined estimate of the effect of the intervention;
- the combined estimate of the effect of the intervention is normally represented by a diamond at the end of the forest plot;
- for each meta-analysis, a measure of consistency of the results of the included studies, such as I^2 and τ (measures of heterogeneity), with the respective confidence intervals should also be presented.

4.4 Network meta-analysis

4.4.1. Introduction

When there is insufficient data to reliably estimate the relative efficacy of two interventions or when there is a need to compare more than two interventions simultaneously, network meta-analysis methods need to be used. For example, an indirect comparison may be necessary when two interventions have not been directly compared in clinical trials but have a common comparator (e.g. placebo). Indirect comparisons are special cases of network meta-analysis where there are no studies that directly compare any of the interventions concerned and conventional meta-analysis is a special case that only includes two interventions.

It should be noted that network meta-analysis methods are particularly relevant in cases where in current clinical practice, several (more than one) interventions are used for the same indication, and where there may therefore be several comparators selected for assessment. When there are multiple comparators with relevant evidence, they should be compared simultaneously in a network meta-analysis to take into account direct and indirect evidence and to ensure consistency of conclusions for all comparators.

The method for choosing the interventions to be included in the network should be specified in advance, be reproducible and ensure a unique network of interventions to be compared (24) (25).

All studies comparing two or more interventions mentioned in PICO (I or C) should be included provided they meet the other pre-specified criteria (see section 2.2). Network meta-analyses (including indirect comparisons) should only be conducted when the available studies are comparable (i.e. sufficiently homogeneous) and of sufficient quality (see section 9.3) so that the results obtained can be relied upon.

The network of comparisons should be designed so that the interventions included define vertices (nodes) and the existence of studies comparing two interventions directly define edges (note that studies with more than 2 arms define multiple comparisons and all should be included in the network representation). Network meta-analysis produces relative effects for all comparisons of pairs of interventions included in the network, provided they form a connected network, i.e. a network in which it is possible to establish a path (using the edges) from one intervention (vertex) to any other. Interventions that are not connected by edges cannot be compared. In a network with loops (i.e. where it is possible to define a path that starts and ends at the same vertex) the comparisons included in each loop are composed of direct and indirect evidence which increases accuracy but requires assessment of the consistency of the evidence (see section 4.4.2 and 4.4.3).

In case the network is disconnected, the inclusion criteria can be extended to include randomised trials comparing additional interventions with one (or more) of the interventions under assessment, which can connect the network. These additional studies should meet all other inclusion criteria, including having a population comparable to that specified in PICO. When there is more than one additional intervention that may link the interventions under consideration, all linking interventions should be included so that all relevant evidence is considered (24).

The inclusion of additional interventions in a connected network may be justified to increase the accuracy of the results. In this case, the same principles of including all interventions and additional studies that may increase the accuracy of the network should be followed. A sensitivity analysis should be presented with the results of the meta-analysis from the original network (including only the interventions mentioned in PICO).

When it is not possible to link all the network interventions under consideration using randomised trials conducted in the same population, the evidence base can be extended to other populations where the relative effects can be assumed to be comparable or can be adjusted, for example by meta-regression. The use of non-randomised or observational studies to form edges connecting disconnected networks is discouraged because of their high potential for bias, which could contaminate any comparison that uses this edge to connect interventions. Given the high potential for bias, the use of population adjustment methods to link disconnected networks and to include data from non-comparative (one-armed) studies is not recommended, except in exceptional situations. These exceptional situations are described in the section 5 and must be adequately justified.

There are several methods of multiple comparisons, generically called network meta-analysis, which include (among others) the Bucher method (26) for indirect comparisons, and frequentist and Bayesian methods for mixed comparisons of interventions (27). The choice of method should be individualised for each case taking into account the type of evidence and the structure of the network. A detailed description of these methodologies is beyond the scope of this document.

Indirect comparison methods can be used to infer the relative efficacy of two interventions in the absence of studies that directly compare them, if there is another intervention (e.g. placebo) to link the two interventions under consideration. Bucher's method (28) should be applied exclusively to situations of indirect comparisons between two interventions with a single linking intervention, where only one study is available for each comparison. In case there are multiple studies available for each comparison, the

method can only be used when these are combined using fixed-effect meta-analysis models. If multiple studies are pooled using meta-analysis with random-effects, Bucher's method is not appropriate. Network meta-analysis methods (Bayesian or frequentist) should be used to make indirect comparisons with random-effects.

Indirect comparison methods are also applicable to the comparison of multiple interventions connected to a single common comparator forming a star network (network of interventions only connected by a common comparator). The use of methods for mixed comparisons (of network meta-analysis) is recommended since they are more efficient in the case of fixed-effect models and allow better estimation of heterogeneity in random-effect models. In a star network, the effect estimates for the comparison of any two interventions via the common comparator are only affected by the studies that make up that indirect comparison. When a fixed-effect model is used, component studies of the star network that do not involve these two interventions do not affect the relative effect estimate. However, if a random-effects model is used, all studies contribute to the estimation of heterogeneity common to the network, which may affect the confidence intervals for all comparisons.

The choice of network meta-analysis method should be separate for each case. A detailed description of these methodologies is beyond the scope of this document, however, we note the following points (see also section 4.4.3):

- only methods that ensure consistency of outcomes and that use meta-analytical principles, that is, that combine relative effects (and not absolute effects) of interventions, should be used. The network meta-analysis models proposed by Lu & Ades (29) and described in the NICE DSU documents (30) are recommended and can be estimated using Bayesian (31) (32) or frequentist (33) methods;
- the network meta-analysis model proposed by Rücker et al (34) is also suitable, but estimation is performed differently, so the models described in the previous section are preferable for models with random effects;
- the models proposed by Lumley (35) and Hong (36) include different assumptions and should not be used.

4.4.2. Assumptions of a network meta-analysis

Network meta-analysis methods are extensions of direct meta-analysis methods for comparisons of more than two interventions. Consequently, all the assumptions underlying the validity of conventional direct meta-analyses (see section 4.3) also apply to network meta-analyses and indirect comparisons.

Participants should be comparable across included studies and should be relevant to the ongoing assessment. The studies included should be sufficiently homogeneous and not differ substantially in characteristics that could alter the relative effects of the interventions. The studies should include the interventions specified in PICO. The extent of interventions included to connect the network or increase the accuracy of the results should be adequately justified (see section 4.4.1).

The extension of the assumptions mentioned in section 4.3 to network meta-analysis implies the following additional points:

- there should be no differences between participants included in studies comparing different interventions, i.e., in principle any participant could have been randomised to any intervention and included in any of the studies;
- direct and indirect comparisons estimate the same relative effect in the included population. That is, for any pair of interventions, the effect of intervention X compared to

intervention A and the effect of intervention Y compared to A is the same as would be observed in a study that included interventions A, X and Y.

By ensuring that the studies included in the network meta-analysis are sufficiently homogeneous with respect to clinically relevant aspects (and that all the assumptions mentioned above are fulfilled) the consistency of the results of the network meta-analysis or indirect comparison to be performed is, in theory, ensured. However, this consistency should be checked statistically whenever possible, i.e. in networks with loops.

4.4.3. Technical aspects in network meta-analysis

Models for network meta-analysis with fixed-effects or random-effects may be used, depending on the clinical assumptions and the variability of the studies included. In the case of random-effects models, the most commonly used models assume the same level of heterogeneity for all comparisons, that is, they estimate a heterogeneity common to all comparisons. Models that estimate different levels of heterogeneity for different comparisons are more complex and, in most cases, there is not enough information to estimate them (37). Models estimating a common heterogeneity parameter are therefore acceptable.

To ensure stable results when computing network meta-analysis results, the following points should be taken into consideration:

- the intervention network must be connected (see section 4.4.1);
- in networks with small studies or that investigate rare binary events, it is common to observe zero events in one or more study arms. Studies with zero events in all arms should be removed since they do not contribute to the estimation of relative effects. Note that the network may be disconnected when these studies are removed (see section 4.4.1);
- the reference treatment should be chosen so that it is one of the vertices with more links to the other treatments included in the network, i.e. the treatment with the highest number of edges and *in the centre* of the network. Normally control interventions or placebos satisfy this condition and should be used as a reference. Theoretically the choice of reference intervention does not affect the results of the network meta-analysis, since all comparisons are estimated. However, the choice of a reference treatment with few comparisons or more distant from others in the network (i.e. with more edges to go through to perform comparisons) can lead to computational problems causing difficulties in estimation, for example a reduced speed of convergence of the algorithms used and high autocorrelation between estimates.

The choice of relative effects scale for network meta-analysis should take into account the type of outcome measure and the statistical properties of relative effects appropriate for that type of outcome measure (see section 4.3.7).

In most situations, the choice of effect estimation method (Bayesian or frequentist) does not affect the outcome of the network meta-analysis, provided that:

- the network is connected;
- appropriate software is used for the model to be estimated, for example WinBUGS, OpenBUGS, JAGS or Stan with appropriate code for Bayesian and Stata or R methods using appropriate functions for frequentist methods;
- there is a sufficient number of studies to estimate the level of heterogeneity in the case of using a random-effects model;

- there are no studies measuring discrete outcomes that have not observed events in one or more arms (i.e. with the presence of outcomes with zero events);
- no external evidence is used in a Bayesian analysis with informative prior distributions.

Bayesian estimation methods should be preferred when:

- there are studies with zero events – the Bayesian estimation models accept results with zeros and are not subject to introducing bias with the need to add 0.5 to the cells of the arms with zeros (28). When the only connection of an intervention to the network is made by one or more studies with zero events observed in one of the arms, the network may be disconnected. In these cases, network connection should be reassessed with these excluded studies. If the network is disconnected, methods that add 0.5 to the cells of the studies with zeros required to connect the network can be considered. However, it should be noted that the results will be slightly biased (28). Studies in which there are no events in either arm should be excluded;
- there is an insufficient number of studies to estimate heterogeneity, but it is necessary to consider a random-effects model due to the characteristics of the included studies – in this case the use of informative prior distributions for the heterogeneity parameter is recommended (38);
- there is relevant external information that should be used as aprior distribution to estimate the relative effects, for example for connection of disconnected networks. In the context of pharmacotherapeutic assessment, the existence of validated external information is rare, so this scenario should only be applied in special situations and with adequate justification.

In the case of using Bayesian estimation methods, the choice of prior distributions should be justified for all parameters to be estimated and subject to sensitivity analysis. Typically, non-informative prior distributions should be chosen for parameters that estimate relative (and absolute) effects of interventions.

In models with random effects it is not possible to define truly uninformative prior distributions for the heterogeneity parameter, so distributions considered to be weakly informative should be used. Sensitivity analyses using different distributions for this parameter should be performed. Informative prior distributions for the heterogeneity parameter (39) can be used, but their impact should be explored in sensitivity analyses.

The quality of the model fit to the data shall be assessed. This is particularly important when a fixed-effect analysis is used, since it is essential to validate the common-effect assumption of the included studies, but it should also be done for random-effects models. A model with a poor fit to the included data indicates the possible failure of one, or more, of the assumptions inherent to the synthesis (e.g. excess heterogeneity or inconsistency in the evidence network). The method used to assess fit depends on the type of model used and must be justified in view of the synthesis method used. In general, methods of residual analysis or variance analysis are recommended. Akaike's information criterion (AIC) or the deviance information criterion (DIC) can be used for model comparison, respectively for frequentist or Bayesian synthesis methods.

Network meta-analysis models assume homogeneity and consistency of estimates from direct and indirect evidence. When the network contains loops, this means that there is direct and indirect evidence for the comparisons involved in that loop. In these cases, the assumption of consistency can, and should, be assessed statistically.

There are methods to assess consistency between direct and indirect evidence in one loop at a time (locally) or in the whole network. The most appropriate method for assessing consistency depends on the structure of the network and the number of loops, and should be determined on a case-by-case basis:

- for consistency assessment locally, the Bucher method (40) is suitable for networks with independent loops that are estimated with a fixed-effect model, and the node-splitting method is suitable for more complex networks (41) (42);
- for overall inconsistency assessment, inconsistency models can be used (43) (44);
- assessment of the fit of the consistency model should be inspected and compared with the original network meta-analysis model;
- in random effect models the heterogeneity parameter should be evaluated. A reduction in its estimate in the inconsistency model in relation to the original network meta-analysis is informative and suggests the existence of inconsistency between direct and indirect evidence.

If inconsistency between direct and indirect evidence is detected, the inclusion of all studies should be reviewed to ensure that they meet the requirements of the systematic review and are relevant; the data extracted and included in the model should be checked to exclude the possibility of error; and the existence of risk of bias or the presence of effect-modifying variables should be explored. Methods used to explain heterogeneity between studies can also be used to explain inconsistency between direct and indirect evidence, for example meta-regression methods and subgroup consideration (see sections 4.3.5 and 4.4.4).

Results estimated from networks with inconsistency have a reduced confidence level and are subject to bias. In some cases it is possible to isolate parts of the network that are not affected by the inconsistency so that some comparisons may be of higher quality - this depends on the structure of the network and the results of the inconsistency analysis and should be investigated and justified in each case.

For example, if a comparison in the network is formed solely by one (or more) small study(ies) with zero (or 100%) events in one arm, extreme and implausible relative effects can be estimated by direct evidence in that comparison. However, indirect evidence can estimate more realistic relative effects if it is based on larger studies and with more (or fewer) events. Inconsistency can be detected in this case when direct evidence is compared with indirect evidence and is only caused by the extreme size of the direct outcomes (i.e. both types of evidence show effects in the same direction, but the size of the direct effect is implausible). In these cases, the acceptance of the relative effect estimated by the network meta-analysis can be considered credible if it is adequately justified.

If the network structure includes a subnetwork where the inconsistency is located, but this subnetwork does not include the comparisons of interest for the assessment, this inconsistency can be ignored if the rest of the network shows no evidence of inconsistency. In this case, the main analysis should only include the subnetwork consistent with the full network included as a sensitivity analysis (in a random-effects model the full network can increase the accuracy of the results by allowing more precision in estimating heterogeneity).

The full results of the network meta-analysis should be reported (31) (45) (46), including:

- diagram with the network structure and table including the data used in the synthesis;
- table with all estimated relative effects and their confidence intervals, accompanied by a forest plot with these results (if possible);

- measure of heterogeneity and its confidence interval;
- measures of the ranking of interventions and their uncertainty (note that the probability of an intervention being 'the best' or the SUCRA measure (47) are not sufficient to characterise the uncertainty in the ranking of interventions, so the rank of each intervention and its confidence interval should also be presented);
- full details of the statistical model used, including software used, the reference treatment, model fit statistics and prior distributions used in Bayesian models.

4.4.4. Meta-regression and bias adjustment

High heterogeneity (clinical or statistical) among the included studies indicates the presence of effect-modifying variables that interact with the treatment effect. These variables may reflect two types of variation: clinical variation between treatment effects due to variability of populations, protocols or context in the included studies; or variation due to different quality of studies and their risk of bias. Studies classified as having high risk of bias should be excluded from the network and only studies with low risk of bias should be included in the main analysis (48). The inclusion of additional studies can be presented in a sensitivity analysis.

Meta-regression methods can be used to obtain results adjusted for observable effect-modifying variables. Its simplest example is subgroup analysis, but continuous variables, for example baseline risk level, can also be considered (see section 4.3.5).

In a network meta-analysis, excess variability due to effect-modifying variables can cause both heterogeneity and inconsistency. Meta-regression methods can be used to explain (and eliminate) heterogeneity and inconsistency in results (49, 50).

Although the reasons for heterogeneity are equivalent in network and conventional meta-analysis, due to the inclusion of a larger number of studies and interventions that may cover a longer time horizon, network meta-analysis may potentially include more heterogeneous studies, for example in terms of the absolute baseline risk of the included patients, which may be an important effect modifier. The investigation of potential absolute risk variation in the included studies, while not conclusive, is an indication of potential heterogeneity or inconsistency between direct and indirect evidence. In that case, it should be assessed whether the baseline risk level is a potential effect modifier (51).

Various meta-regression models are possible in a network meta-analysis (51)(52). Models that assume a common interaction for all comparisons are the most relevant for pharmacotherapeutic assessment, provided they have clinical validity. However, their results are considered observational and should only be used for exploratory or sensitivity analyses.

4.5 Conclusions

When the available evidence includes two or more studies, the results of these studies can be combined using meta-analytic techniques to generate a combined estimate of the relative efficacy of the interventions with greater statistical power and precision.

Conventional meta-analysis allows two treatments investigated in multiple studies to be compared with each other. Network meta-analysis allows the comparison of multiple interventions compared across multiple studies, provided they form a connected network. The method for choosing the treatments to

be included in the network must be specified in advance, be reproducible and ensure a unique network of treatments to be compared.

Network meta-analysis methods are particularly relevant in cases where several treatments are used in current clinical practice for the same indication, and where there may therefore be several comparators selected for assessment. These methods should be preferred, provided that a linked treatment network can be formed, based on relevant randomised trials and without high risk of bias. In case the network is disconnected, the inclusion criteria may be extended to include randomised trials comparing additional treatments with one (or more) of the treatments under assessment, which may connect the network.

Conventional or network meta-analyses should only be conducted when the available studies are sufficiently homogeneous, i.e. comparable (not differing substantially in characteristics that could alter the relative effects of treatments) and of such quality that the results obtained can be relied upon.

5 METHODS OF COMPARISON IN EXCEPTIONAL SITUATIONS

5.1 Anchored Adjusted Indirect Comparison (MAIC, STC)

Any meta-analysis, conventional or network, has as its main assumption that there are no differences in the distribution of effect-modifying variables in the included studies. However, this assumption is not always verified, namely when there is a high level of clinical or statistical heterogeneity between studies. Indirect comparisons, and meta-analyses that include few studies, are particularly vulnerable to the existence of these differences.

Using models that adjust results based on effect-modifying variables can produce more relevant and credible relative effects. Meta-analysis (conventional or network) with meta-regression using individual participant data from all studies is the preferred method (see section 4.3.6).

In the context of the pharmacotherapy assessment, the submitting MAH usually only has access to the individual data of its studies. The matching adjusted indirect comparisons (MAICs) (53) and simulated treatment comparisons (STCs) (54) methods were developed to deal with situations where:

- an indirect comparison between two treatments is required;
- there are differences in one or more effect-modifying characteristics between the population of studies that will form the indirect comparison;
- the company has access to the individual data from its study, but not from the other studies.

The MAIC method uses inverse propensity score weighting to weight the effect of treatments used in the population for which individual data are available, to the effect that would be observed in the study population for which individual data is not available. Methods typically used for this weighting give equivalent results (55). The STC method uses regression to adjust the treatment effect in the population for which individual data are available, for the effect that would be observed in the study population for which individual data are not available. Random samples of the joint distribution of covariates in the study with aggregate data are used to calculate the predicted effect in that population using a regression model. However, these methods:

- generally, do not make comparisons on the scale of relative effects that would be preferred in a conventional (simple or network) meta-analysis. It is recommended that comparisons be made on the scale chosen for the relative effects (56,57).
- produce relative effects applicable to the population of one of the studies (the study with no individual data available), but do not ensure that the results are applicable to the population most relevant to the assessment (the population defined in the PICO).
- assume a distribution for effect modifiers in the comparator study based only on statistical summaries described in publications.
- Only allow adjustment of variables with summaries described in the publications.

Note that, as the adjustment is made on the basis of randomised comparisons, it is not necessary (and is even discouraged) to adjust for purely prognostic variables.

Furthermore, the MAIC method can only be applied when there is sufficient overlap in the distributions of the variables in the study population with individual data and the comparator study. When there is

little overlap, the method does not produce credible results (56,57). However, when there is a large overlap in the distributions of variables in the study population with individual data and the comparator study, it is not necessary to use adjustment methods since the populations are expected to be comparable.

The use of simulation in the STC method introduces additional variation. The estimated effect is not the average effect in the population with aggregate data, but the predicted effect in a randomly selected individual from that population (i.e. from the predictive distribution), which leads to overestimation of uncertainty in the final indirect comparison.

The need for the use of MAIC or STC should be justified with reference to the characteristics of the included studies and the existing evidence to support effect modification. This type of analysis uses methods and assumptions that represent a departure from the methods typically used in pharmacotherapeutic assessment and should therefore be considered less credible than meta-analysis (simple or network) based on randomised trials without evidence of effect modifiers. Note that the estimates obtained refer to the population of the comparator study. It will be necessary to compare this population with the population under evaluation.

The MAIC and STC methods do not extend to the situation where there are multiple possible indirect comparisons using different comparators and their assumptions are difficult to validate (56,57). There is therefore the possibility of heterogeneity of results in assessments of different products for the same therapeutic area, which derives from the choice of studies used for adjustment and which had individual data available in each circumstance.

The multi-level network meta-regression (ML-NMR) method (58) is an alternative that allows for the adjustment of the effects of a treatment network to the study population, avoids the risk of clustering bias and produces directly interpretable results despite having a more complex implementation. The choice of population for which effects are adjusted should be adequately justified.

5.2 Use of non-randomised studies

When there is quality evidence from randomised comparative trials, non-randomised evidence can be used to complement, but not replace, the evidence from randomised trials, for example to validate its application to the Portuguese context.

However, in exceptional situations, due to lack of randomised studies, it may be necessary to consider non-randomised, or ‘real world’ evidence. Note that this type of evidence has a high risk of bias. However, non-randomised studies may be acceptable to inform specific endpoints (e.g. long-term safety), provided it is based on the type of study most appropriate for the purpose and minimises the risk of bias (e.g. where there is a history of controlled studies, appropriately adjusted). This type of evidence can be used to define the patient's clinical history in the absence of treatment, to compare safety data, and to inform efficacy endpoints in cases where randomised trials are clearly not feasible, for example in the case of rare or ultra-rare diseases.

The MAIC and STC methods can also be used to make comparisons between single-arm studies or to connect disconnected networks. Given the high potential for bias, the use of population adjustment methods to link disconnected networks and to include data from non-comparative (one-armed) studies is not recommended except in exceptional situations that should be justified in detail (56,57). In particular, unanchored indirect comparison methods should not be used when an anchored comparison is possible as they have a higher potential for bias and less precision than anchored comparisons.

Of these exceptional situations, the following stand out:

- rare diseases, defined by a prevalence of less than five in 10,000 people, where there are no therapeutic alternatives, or where the effect of these alternatives is unproven or uncertain, or where the treatment includes medicines with well-established use.
- ultra-rare diseases, defined as a disease with a prevalence of ≤ one patient per 100,000 people.

In these cases, where there is no evidence from randomised trials, the use of unanchored adjusted indirect comparisons (MAIC and STC) from single-arm studies is considered acceptable as demonstration of additional proof of benefit.

Both MAIC and STC should be used simultaneously. Demonstrating proof of additional benefit will imply that the two methods (MAIC and STC) give concordant results. The use of alternative methods should be adequately justified.

An unanchored comparison assumes that the absolute effect of the intervention can be predicted based on population characteristics, i.e. it assumes that all prognostic and effect-modifying variables are included in the prediction model. This assumption is stronger than the assumption used in anchored comparisons in which it is not necessary to consider predictor variables, and virtually impossible to verify. The failure of this assumption implies a level of bias in the comparisons made which is difficult to quantify. When effect measures based on unanchored comparisons are used, it is necessary to demonstrate the plausible error size due to the lack of inclusion of variables in the adjustment of the estimated relative effect (56).

6 SUBGROUP ANALYSIS

6.1 Introduction

Patients in a particular indication may vary in characteristics that affect the magnitude of the benefits of the new health technology, as well as the costs associated with its treatment. This variation in characteristics is known in the literature as heterogeneity. Heterogeneity may influence the choice of treatment, as it is possible to select the treatment that most benefits the patient (or is most cost-effective) given their characteristics.

The existence of heterogeneity may be due to different reasons, the most frequent being heterogeneity in relative treatment effect (i.e. modification of therapeutic effect) and heterogeneity in baseline risk, such as in risk of progression or risk of events. There may also be situations in which the baseline risk is correlated with the relative treatment effect (59).

In most clinical trials and systematic reviews, treatment effects are not homogeneous across the included population. Subgroup analyses allow these differences in response to treatment to be assessed, as well as other sources of heterogeneity, to enable greater personalisation in health decisions.

These recommendations are intended to inform the pharmacotherapeutic assessment, and do not preclude the assessment of other subgroups in the pharmacoeconomic assessment, in line with the recommendations of the pharmacoeconomic guidelines.

6.2 Definition/ specification of subgroups

The specification of subpopulations should comply with the criteria established by the assessment matrix (PICO) (see section 2.2).

Ideally, these subpopulations identified in the initial matrix would be assessed in separate studies, or studies designed to have adequate statistical power to study subpopulations within the same study. However, sometimes these subpopulations are included in the same study, and a subgroup analysis is performed by the MAH and/or research team to detect different treatment effects. When this happens, it is important to conduct an assessment of the credibility of the subgroup analysis in the evidence submitted (see section 6.4).

Separation into subpopulations should be proposed if there are characteristics that are potential treatment effect modifiers, preferably documented in previous studies in the same pathology (interaction term). Any definition of a subpopulation which is not explicitly provided for in the approved indication must be justified by the proponent of that subpopulation - either CATS or the applicant. If there is doubt about a potential effect modification that has not been studied previously, one may choose to note the subgroup effect that is intended to be checked in a footnote in the table of the initial assessment matrix.

One should not propose to separate the population into subgroups just because there are different clinical characteristics or prognostic factors, if this does not predictably influence the treatment effect. Division into subgroups solely for clinical heterogeneity may even raise equity and/or ethical issues. For example, the use of age or age class may be appropriate if the effect of age is a modifier of treatment effect (e.g. less effective treatment in older patients) or disease progression. On the other hand, if age does not reflect an effect on treatment (or disease) its use may not be equitable and such considerations should be explicitly made when defining subgroups.

Subgroup analyses have important methodological limitations and often do not meet the necessary methodological criteria, leading to erroneous results (60). Randomised controlled trials are the preferred type of evidence for identifying treatment effect modifiers. Non-randomised evidence (e.g. large longitudinal observational studies), may, however, be appropriate for the identification of other types of subgroups mentioned above, such as baseline risk and prognostic heterogeneity information. Regardless of the study design, sample size in subgroups is often small, and without statistical power to detect differences between groups in the same population, even if these differences exist.

6.3 Recommendations for subgroup analysis - MAH perspective

The use of a set of rules in subgroup analysis is recommended from the perspective of the MAH:

- subgroup analyses should be defined before the study starts and should be limited to a small number of clinically relevant issues;
- the study protocol should include information on how the subgroups were selected, and why they were selected;
- the exact definitions and categories of the subgroup variables shall be defined explicitly from the outset. For continuous or categorical variables, the thresholds for analysis must be pre-defined;
- the direction and magnitude of the expected effect on the subgroup should be defined *a priori*;
- in the study design, consideration should be given to stratification of randomisation by important subgroup variables;
- where important subgroup - treatment effect interactions are expected, the study should have sufficient statistical power to reliably detect such interactions;
- the rules for stopping the study should take into account expected subgroup - treatment effect interactions and not just the overall treatment effect;
- if the relative treatment effect is likely to be related to baseline risk, the analysis plan should include a stratification of outcomes according to the predicted risk. The model or risk score should be pre-selected so that the relevant baseline data are recorded;
- the significance of the treatment effect in individual subgroups should not be reported, as the percentages of false positives and false negatives are extremely high. The only reliable statistical approach is to test the subgroup - treatment effect interaction, i.e. the correct analysis is not the statistical significance of the treatment effect in one or another particular subgroup, but whether the effect differed significantly between subgroups (subgroup - treatment effect interaction test). In epidemiological terms, interaction means effect modification. It is important to note that interaction testing should use relative measures of effect (relative risk or hazard ratio or odds ratio) and not absolute risk reduction. This is because the intrinsic property of the treatment is represented by the relative measures, which tend to be constant across different strata of absolute risk;
- the statistical significance of treatment effect - subgroup interactions must be adequately adjusted for when multiple subgroup analyses are performed;
- subgroup analyses should be reported as relative risk reductions and absolute risk reductions.
- Ideally, only one outcome measure should be studied, preferably the primary outcome measure of the study;

- the comparability of prognostic factors between treatment groups should be confirmed in subgroups;
- where multiple subgroups-treatment effect interactions are identified, additional analyses are required to verify that their effects are independent;
- descriptions of the statistical significance of the treatment effect in individual subgroups should be ignored, especially reports of lack of benefit in a particular subgroup in a study in which overall benefit was observed, unless there is a significant subgroup-treatment effect interaction;
- unexpected genuine subgroup treatment effect interactions are rare. Therefore, apparent interactions that are discovered *post hoc* should be interpreted with care. In this case, no significance test is reliable;
- *pre-hoc* subgroup analyses are not intrinsically valid and should be interpreted with caution. The probability of false positives increases with the number of tests and can be evaluated by the formula 1-(1-p)^c, where p is the significance level and c, the number of tests;
- the best test of validity of subgroup interactions - treatment effect is its reproducibility in other studies;
- few studies have the statistical power to detect subgroup effects, so the percentage of false negative interaction tests is high. If a genuine subgroup - treatment effect interaction exists, the probability of a false negative result with a formal test for interaction will be much higher than the 5% false positives seen in a study where there is no true interaction;
- the uncertainty in the identified subgroups should be properly quantified and expressed in an appropriate manner (e.g. confidence interval, standard deviation). An adequate quantification of uncertainty in a subgroup analysis is usually achieved via analysis of patient-level data. Formal modelling (e.g. via regression models, meta-regression) facilitates establishment of subgroup-treatment effect interactions, with estimation of parameter and between-parameter uncertainty.

6.4 Assessing and rating the credibility of subgroup analyses

Given the numerous limitations inherent in subgroup analysis, it is crucial to critically evaluate subgroup analyses in order to infer their degree of credibility. Thus, it is recommended to evaluate the credibility of subgroup analyses in two stages:

- analysis of compliance with the criteria for assessing the credibility of a subgroup analysis;
- rating the degree of credibility of the analysis from ‘extremely unlikely’ to ‘extremely plausible’.

Table 2: *Criteria for assessing the credibility of subgroup analysis*

Criterion to be evaluated	**Y/N/U**
Design	
1. Is the subgroup variable a characteristic measured after randomisation or at baseline?	
2. Is the effect suggested by comparisons within the study more than between studies?	
3. Was the hypothesis specified *a priori*?	
4. Has a small number of hypotheses been tested?	
5. Was the direction of effect in the subgroup specified *a priori*?	
Analysis	
6. Does the test for interaction suggest a low probability that the apparent effect of the subgroup is explained by chance?	
7. Is the effect of the subgroup independent?	
Context	
8. Is the magnitude of the subgroup effect large?	
9. Is the interaction consistent between studies?	
10. Is the interaction consistent in the closely related outcome measures of the study?	
11. Is there indirect evidence to support the hypothetical interaction (biological rational)?	

Yes: Y=Yes; N=No; U=Unknown

We suggest checking the 11 criteria in Table 2 to assess the credibility of a subgroup analysis. Note that the credibility assessment of subgroup analysis is not a dichotomous question, but one that results in a spectrum of credibility from ‘extremely unlikely’ to ‘extremely plausible’. We suggest adherence to the credibility assessment considered by the "User's Guide to the Medical Literature" (61) (62) considering that criteria 1, 3 and 9 are the most relevant for a correct assessment of the credibility of the subgroup analysis.

If subgroup analysis is not credible, it is suggested to analyse the total study population if there is confidence that the characteristic of the subgroup studied does not appear to be a modifier of the treatment effect. If this analysis is not feasible or if there is doubt about a potential effect modification, it will not be possible to assess the additional benefit of the drug in the subpopulation in question. This conclusion may or may not lead to a restriction of the indication under assessment.

6.5 Conclusions

- Patients in an indication may vary in characteristics that affect the magnitude of the benefits of the new medicinal product as well as the costs associated with its treatment;
- From the perspective of the MAH, rules should be followed to define subgroup analyses in order to present more credible analyses;

- From the perspective of the evidence assessment group, the subgroup analysis submitted should be analysed and graded as to its credibility according to the 11 criteria set out in Table 2, on a continuum between ‘extremely unlikely’ and ‘extremely plausible’.
- Subgroup analyses are exploratory analyses to identify effect modifiers, or sensitivity of outcomes, conditioned on their degree of credibility.

7 PARTICULAR ASPECTS IN BENEFIT ASSESSMENT

7.1 Impact of study outcomes not published in the conclusions

An essential prerequisite for the validity of a benefit assessment is the full availability of the results of studies conducted on a topic. An evaluation based on incomplete or possibly even selectively compiled data may produce biased results.

Additionally, the biases resulting from publication bias and outcome reporting bias have been comprehensively described in the literature. To minimise the consequences of this problem, it is recommended that the information search, in addition to including a search in bibliographic databases, should also include, for example, searching international platforms of trial records (see section 3.4).

7.2 Dramatic effect

If the course of a disease is certainly or almost certainly predictable, and no treatment options are available to influence this course, proof of a benefit of a medical intervention can also be provided by the observation of a reversal of the (more or less) deterministic course of the disease in well-documented case series of patients. If, for example, it is known that a disease is highly likely to lead to death within a short period of time after diagnosis, and it is described in a case series that, after the application of a specific intervention, most of those affected survive for a long period of time, this ‘dramatic effect’ may be sufficient to provide proof of a benefit. An example of this effect is the replacement of vital hormones in diseases with a lack of hormone production (e.g. insulin therapy in patients with Type 1 diabetes *mellitus*). An essential prerequisite for classification as a ‘dramatic effect’ is sufficiently reliable documentation of the fatal course of the disease in the literature and of its diagnosis in the patients included in the study being assessed. In this context, possible harm from the intervention should also be taken into consideration. Empirical data suggest that an observed relative risk of five to ten cannot be explained by confounding factors alone. If, in the period leading up to the assessment, there is sufficient information available indicating that a dramatic effect caused by the intervention being assessed can be expected (e.g. due to a preliminary literature search), the assessment should include those studies that demonstrate greater certainty in the results due to their design.

7.3 Duration of study

The duration of the study is an essential criterion in the selection of studies relevant to benefit assessment. In evaluating a therapeutic intervention for acute diseases, where the primary objective is, for example, to reduce the duration of the disease and alleviate acute symptoms, it makes no sense to require long-term studies unless late complications are expected. On the other hand, in the assessment of therapeutic interventions for chronic diseases, short-term studies are usually not adequate to obtain a complete assessment of the benefits of the intervention. This applies especially if the treatment is necessary for several years, or even lifetime. In such cases, studies covering a treatment period of several years are particularly relevant and desirable. As benefits and harm can be distributed differently over time in long-term interventions, comparing the benefits and harm of an intervention is only possible with sufficient certainty if studies of sufficient duration are conducted.

8 SUPERIORITY, NON-INFERIORITY, AND EQUIVALENCE STUDIES: DEFINITIONS AND CRITERIA FOR CHANGING OBJECTIVES

8.1 Introduction

Evidence of efficacy can be obtained from different types of controlled studies. Superiority studies seek to show that an intervention is superior to a control (placebo, no treatment, lower dose of intervention). Another type of study is one that compares the intervention with an active treatment (active control). Although this type of study can also aim to demonstrate superiority, it often aims to show that the difference between the new treatment and the active control is small and that, based on its performance in previous studies and the assumed efficacy of the active control in the current study, it is possible to conclude that the new intervention is also effective. However, the design and interpretation of the results of these studies pose specific challenges, so it is necessary to set out some considerations in this regard.

8.2 Demonstration of equivalence

One of the most frequent serious errors in interpreting medical data is to classify a non-significant result of a significance test as evidence that the null hypothesis is true.

To demonstrate 'equivalence', methods that allow the hypothesis of equivalence to be tested must be used, i.e. the study must have an equivalence design that allows the absence of a significant difference between treatments to be confirmed (for example, that the mean value of the difference between 2 groups is exactly zero).

This objective is obtained through the calculation and observation of confidence intervals, since the use of statistical tests is not possible. At the time of protocol development, it is necessary to define a margin (Δ) of clinical equivalence by defining the largest difference that is clinically acceptable, such that a larger difference would be relevant in clinical practice. The two treatments are considered equivalent if the 95% confidence interval (two-sided), which defines the range of plausible differences between the two treatments, is within the interval - Δ to + Δ. In practice, what is shown is not the existence of exact equivalence (that the difference between the mean values of the 2 groups is exactly zero), but that the difference between the two groups is irrelevant. As with superiority studies, the sample size in equivalence studies at the time of protocol preparation must be estimated.

In the case of bioequivalence studies, 90% confidence intervals of the difference between treatments has been accepted as the standard, in the evaluation of the mean values of pharmacokinetic parameters. In the case of an inhaled generic or a topically applied product, where bioequivalence studies are impossible, it is acceptable to carry out clinical bioequivalence studies using 95% confidence intervals.

When the 95% confidence interval defining the range of plausible differences between the two treatments is one-sided, the study is referred to as non-inferiority.

8.3 Demonstration of non-inferiority

Non-inferiority studies aim to demonstrate that the effect of the new treatment is not inferior (meaning that it may have the same or more efficacy) to the effect of the existing treatment, by a specified value, called the margin (Δ) of non-inferiority. As mentioned, the confidence intervals approach also applies here, but now we are only interested in assessing the possible difference in a single direction. Thus, the 95% confidence interval (two-sided) of the difference between treatments should be completely to the

right of the value - Δ. The statistical test is given by comparing the upper bound of the (two-sided) confidence interval for the comparison of the two treatments with the margin previously specified. If the upper bound of the confidence interval is less than the margin, non-inferiority is established.

As noted above, non-inferiority studies are sometimes erroneously referred to as equivalence studies, representing a source of confusion.

8.4 One-sided and two-sided confidence intervals

According to the *ICH E9 Note for Guidance* (63,64), two-sided 95% confidence intervals should always be used in all clinical studies regardless of their objective. If one-sided confidence intervals are used, they should be used to cover a 97.5% probability. In the special case of bioequivalence studies, the use of bilateral 90% confidence intervals have been recommended.

One possibility for setting the margin (Δ) is to establish a value equal to the known effect of the existing treatment compared to placebo, based on previous randomised trials. With this choice of margin, and assuming that the drug under assessment reaches this level of efficacy in the non-inferiority study, non-inferiority means that the test drug has an effect greater than 0. However, a usual choice is to establish that the margin (Δ) is equivalent to a clinically relevant portion of the known existing treatment effect relative to placebo, namely the portion of the control treatment effect that is important to preserve in the drug under assessment, based on clinical judgement.

8.5 Superiority studies

Superiority studies are designed to detect a difference between treatments, and the demonstration of superiority is made through the use of a test of statistical significance. The test of statistical significance tests the null effect, i.e. hypothesises that there are no differences in clinical effects between two treatments. The degree of statistical significance (p-value) indicates the probability that the observed difference occurred by chance in the hypothesis that, in reality, there is no difference between treatments. However, results should not simply be reported as having or not having ‘statistical significance’ but should be interpreted in the context of the type of study and the associated risk of bias.

Once the hypothesis of non-difference between treatments is found to be untenable as unlikely (a $p < 0.05$ indicates that this probability is less than 5%), it is important to estimate the magnitude of the difference to assess whether this magnitude is clinically relevant. To this end, it is necessary to calculate the best estimate of the magnitude of the difference between treatments, usually called point estimate, which, in normally distributed data, corresponds to the difference in mean values of the efficacy measure used. The confidence interval should also be calculated, which corresponds to the range of plausible values of the true difference. This interval should not include the null effect, which is zero if the effect size is a continuous variable (e.g. the standardised mean difference), or one in the case of ratios (risk ratio, odds ratio or hazard ratio), because the null hypothesis (zero difference between treatments) has already been rejected. Thus, the following statements are considered equivalent: the 95% confidence interval of the difference between treatments excludes the null value and the two values are statistically different at the 5% two-sided significance level ($p < 0.05$).

In superiority studies, assessing whether a difference between treatments with statistical significance is clinically relevant requires *a posteriori* value judgement. In contrast, in equivalence and non-inferiority studies, clinical relevance is defined in advance, at the time of designing the study protocol, by defining the Δ (non-inferiority margin [-Δ] or equivalence margin [±Δ]).

8.6 Relevance of Δ pre-definition in non-inferiority and equivalence studies

The demonstration of ‘equivalence’ or ‘non-inferiority’ depends on the Δ value selected which should represent the maximum acceptable difference for the objective of interest. It is important to note that, upon inspection of the data, it is always possible to select a value of Δ that leads to the conclusion of ‘equivalence’ or ‘non-inferiority’.

Since the choice of Δ always requires clinical judgment, there is always a risk of bias in this choice, but this risk increases exponentially if Δ is selected after inspection of the data. Thus, the selection of Δ should always be made at the time of designing the study protocol and justified based on plausible arguments.

8.7 Relevance of pre-definition of the study as superiority, non-inferiority or equivalence

According to the guidelines set out by the European Medicines Agency (EMA) in the Committee for Proprietary Medicinal Products (CPMP) document *"Points to consider on switching between superiority and non-inferiority"* (65), pre-definition of the study as superiority, non-inferiority, or equivalence is essential for the following reasons:

- ensures that comparators, doses of medicinal products, populations to be included, and outcome measures are appropriate;
- allows the estimation of the sample size to be based on appropriate calculations of statistical power;
- ensures that the criteria of equivalence or non-inferiority are pre-defined;
- allows the protocol to describe in detail the appropriate statistical analysis;
- ensures that the study has sufficient sensitivity to achieve its objectives.

8.8 Is it possible to change the purpose of a comparison?

The change between superiority and non-inferiority is the only change with practical relevance. Equivalence studies are so specific that there is no possibility of change, either between equivalence and superiority or between equivalence and non-inferiority.

8.9 Interpreting a non-inferiority study as a superiority study

Where the entire 95% confidence interval of the difference between treatments is not only to the right of - Δ but also to the right (above) of the null value (zero in the case of continuous variables, and one in the case of ratios), there is evidence of superiority in terms of statistical significance at the 5% level ($p < 0.05$). In this case it is acceptable to calculate the p-value associated with a superiority test (statistical significance test) and assess whether it is statistically significant. The interpretation of this test is not affected by the multiplicity problem (type I error), since it corresponds to a single statistical test of significance.

Thus, according to the EMA CPMP it is possible to change the objective from a non-inferiority study to a superiority study provided that:

- the study has been designed and conducted in accordance with the requirements of a non-inferiority study;
- the p-values for superiority are given;
- the study has been analysed according to the intention-to-treat principle.

8.10 Interpretation of a superiority study as a non-inferiority study

If a superiority study does not show a statistically significant difference between treatments, there could be interest in a lower endpoint for establishing non-inferiority. If the difference between treatments in a superiority study is presented as a confidence interval, the lower bound of the confidence interval provides an estimate of the minimum effect of the new treatment relative to the comparator. If the study protocol defines a non-inferiority margin deemed acceptable, changing the study objective from superiority to non-inferiority is possible and acceptable.

In superiority studies where the non-inferiority margin was not defined at the time of protocol design, it is not acceptable to change the study objective since, upon inspection of the data, it is always possible to select a value of Δ that leads to the conclusion of equivalence or non-inferiority. Thus, a *post hoc* definition of Δ is associated with a high risk of bias and is therefore not acceptable.

Thus, a superiority study where the test of significance for the difference between treatments showed no statistical significance, if you have not defined the non-inferiority margin at the time of protocol design, should be considered only as a negative superiority study, and it is not acceptable to change the objective from superiority to non-inferiority.

However, changing the purpose of a study from superiority to non-inferiority may be feasible provided the following requirements are met, as recommended by EMA:

- the margin of non-inferiority in relation to the control treatment has been predefined or can be justified;
- analyses according to the intention-to-treat principle and according to protocol, with confidence intervals and p-values for the null hypothesis of inferiority, show similar results;
- the study was properly designed and conducted in accordance with the requirements for non-inferiority studies;
- the sensitivity of the test is sufficiently high to ensure that it is capable of detecting relevant differences, if any;
- there is direct and indirect evidence that the control treatment demonstrates its usual level of efficacy.

9 ASSESSMENT OF THE QUALITY OF EVIDENCE

9.1 Assessment of risk of bias by study

The assessment of the quality of evidence should begin by evaluating the risk of bias for each of the studies included in the intervention assessment. Six domains are used to assess the risk of bias for each study: lack of random sequence generation, lack of allocation concealment, lack of blinding (participants, investigators and adjudicators of outcome measures), incomplete accounting of patients and outcome events, selective reporting of outcome measures and other sources of bias (e.g. stopping early for benefit, use of unvalidated outcome measures, carryover effects in crossover trial).

The rating of the risk of bias of each study should be explained in detail for each of the six domains mentioned above and, for the set of studies assessed, it is recommended that it be summarised using a table and/or a bar chart (66).

9.2 Assessment of quality of evidence (certainty of evidence) in conventional meta-analysis

The methodology described here is essentially applicable in the context of conventional meta-analysis and is not generally applicable in the context of network meta-analysis. However, in situations of simple networks, as is the case with Bucher's method (40), this methodology can also be applied.

The quality of evidence for each comparison should be assessed, for each outcome measure, but the final rating should refer to the set of studies assessed. 'Quality' reflects our confidence that the effect estimates are correct.

The quality of the evidence should be rated into 4 levels: high, moderate, low or very low. This rating applies not to individual studies, but to each comparison (certainty of evidence) and to each outcome measure. In the initial assessment, evidence from randomised trials included in each comparison starts as high-quality evidence, but this initial rating may be reduced by five factors (risk of bias, imprecision, heterogeneity, indirectness and publication bias). Details on the methodology to be used for assessing the quality of evidence will be described in points 9.2.2 to 9.2.6. The rating of the quality of evidence allows the certainty of results to be ranked:

- high quality means high certainty of results (meaning: we are very confident that the true effect is very close to the effect estimates);
- moderate quality means moderate certainty of results (meaning: we are moderately confident in effect estimation. The true effect is likely to be close to the estimated effect, but there is a possibility that it may be substantially different);
- low quality means low certainty of results (meaning: our confidence in effect estimates is limited. The true effect may be substantially different from the effect estimate);
- very low quality means very low certainty of results (meaning: our confidence in effect estimates is very limited. The true effect may be very different from the effect estimate).

9.2.1. Rating the overall quality of evidence

The previous conclusions obtained separately for each outcome measure are then summarised in an overall rating of confidence in estimates of effect.

In this final assessment, in general, the rating assigned to the quality of the overall evidence is the same as that assigned to the 'critical' outcome measure that provides the lowest confidence.

9.2.2. Rating the quality of evidence: risk of bias

The quality of evidence based on randomised studies is initially rated as high but may be reduced by the five factors mentioned above. Note that the evidence quality rating does not refer to individual studies, but to each outcome measure used in each comparison (assessment for each outcome measure), this being the basis for a subsequent overall quality assessment (for all outcome measures).

The first factor (risk of bias) results from methodological problems in the design or conduct of the study and includes a set of five methodological problems:

- Lack of allocation concealment: the investigators have prior knowledge of the group to which the next included patient will be allocated;
- lack of blinding (participants, investigators and adjudicators of outcome measures): patients, investigators, those recording outcome measures, those adjudicating outcome measures, and/or those analysing the data are aware of the arm to which patients are allocated (or the medication they are currently receiving in the case of a crossover study);
- incomplete accounting of patients and outcome events: loss of patients for follow-up and non-adherence to the intention-to-treat principle in superiority studies. Historically, methodologists have suggested arbitrary thresholds for acceptable loss to follow-up (e.g. less than 20%). However, the significance of a loss to follow-up depends on the relationship between loss to follow-up and number of events. As a general rule, the greater the difference between the percentage loss to follow-up and the percentage of events in the intervention and control groups, the greater the risk of bias. For example, if the events are 2% and 4% in the intervention and control groups, a loss to follow-up of 5% is of concern;
- selective reporting of outcome measures: incomplete or absent reporting of some outcome measures influenced by results. To assess this domain, protocols recorded in clinical trial databases (e.g. https://clinicaltrials.gov) should be analysed and it should be assessed whether all recorded outcome measures were analysed;
- other limitations: early termination of the study for benefit, use of non-validated surrogate outcome measures, etc.). Empirical evidence suggests that studies stopped early for benefit overestimate the treatment effect (67).

In the initial rating of the quality of evidence, it should be kept in mind that this is a specific assessment for each outcome measure, so the impact of these methodological problems on each of these measures may vary substantially. The risk of bias due to lack of treatment blinding or lack of allocation concealment is higher in studies with subjective outcome measures. For example, the lack of treatment blinding is not a serious problem if the outcome measure to be assessed is overall mortality, so in this case the quality rating should not be reduced.

For each comparison, and for each outcome measure, the quality of the evidence should be reduced by one level (to moderate), or two levels (to low), if it is considered that there are serious or very serious limitations, respectively.

However, if the serious limitations are not at the level of the outcome measure, but at the level of an individual study, the possibility of excluding that study from the assessment process should be considered. Note that in the case of network meta-analysis comparisons this exclusion may disconnect the network.

9.2.3. Rating the quality of evidence: imprecision

For each outcome measure, the main criterion for assessing precision is the 95% confidence interval around the estimate of the relative treatment effect. Conceptually, the 95% confidence interval can be interpreted as the interval within which, in 95% of cases, the true value lies. In assessing the quality of evidence, the question is whether the confidence interval around the estimate of the relative effect of the intervention is sufficiently narrow.

Randomisation allows prognostic variables to be balanced in the intervention and control groups. However, this equilibrium is only achieved if the sample size is large enough. Large treatment effects in the presence of a small sample size may simply be the result of an imbalance of prognostic variables between treatment groups even in randomised trials with results with narrow confidence intervals.

Thus, to assess whether the results are sufficiently precise, the following two criteria should be used cumulatively:

- the 95% confidence interval is sufficiently narrow and excludes the null effect;
- the number of participants included in the studies under review is equal to or greater than the 'optimal information size (OIS)'. The OIS is obtained by calculating the number of patients needed to be included in a study with sufficient statistical power. It is basically a question of estimating the sample size of a study with sufficient statistical power.
- If the criteria defined in the two previous paragraphs are not cumulatively fulfilled, the quality rating of the evidence should be reduced due to imprecision.

9.2.4. Rating quality of evidence: inconsistency

The criteria used here to establish the existence of 'inconsistency' refer to relative measures (relative risk, risk ratio or odds ratio), not absolute measures, and apply in the context of a pairwise meta-analysis.

The rating of the quality of evidence may be reduced by the presence of inconsistency but is not increased by its absence.

A set of four criteria should be used to assess inconsistency in the context of a pairwise meta-analysis:

- effect estimates vary substantially between studies;
- the confidence intervals do not show any overlap or only a minimal overlap;
- statistical tests for heterogeneity (usually the Q-test) – which test the null hypothesis that all studies included in a meta-analysis show the same effect magnitude – show a significant P value;
- I^2 has a high value. The value of I^2 can vary between 0 and 100%. I^2 tells us what proportion (%) of the variance of the observed effects reflects the variance of the true effects [and therefore does not result solely from sampling error, where I^2 would be 0%].

If using the previous criteria, one comes to the conclusion that the results present problems of inconsistency, the rating of the quality of evidence should be reduced by one or two levels.

9.2.5. Rating the quality of evidence: indirectness

The evidence may not be directly relevant in three ways:

- the study population is different from the population of interest;
- the intervention tested is different from the intervention of interest;
- the outcome measures are different from the outcome measures of interest, for example, the use of surrogate outcome measures.

The possible impact of non-directly relevant evidence (different populations or interventions and use of surrogates) on outcomes should be assessed and a decision made on whether to downgrade the quality of evidence.

9.2.6. Rating quality of evidence: selective reporting of outcome measures

Selective reporting of outcome measures is defined as the selection of only a portion of the initially defined variables, based on the results, for inclusion in the published report of the study. Selective reporting of outcome measures can arise in a number of ways, some affecting the study as a whole and others related to specific outcome measures:

- selective omission of some outcome measures in the report: in this case, only some of the outcome measures analysed are included in the report. If the choice is based on the outcomes and in particular on the statistical outcome, the corresponding (meta-analytical) estimates are likely to be biased as well;
- selective choice of data for an outcome measure: for each specific outcome measure, there may be different times when that outcome measure was observed, or different instruments may have been used to measure the outcome measure at a particular time (e.g. different scales);
- selective reporting of analyses using the same data: there are multiple different ways to analyse the effect of treatment on an outcome measure. Changing the analysis from that initially planned to other forms of analysis may bias the results;
- incomplete reporting of data: sometimes data are reported incompletely, for example, in a way that does not allow inclusion in a meta-analysis.

The effect measures reported should be compared with the effect measures provided in the study protocol. The failure to report the treatment effect of pre-specified outcome measures in the study protocol in the study report indicates the existence of selective data reporting. The absence of treatment effect data on outcome measures considered key in the context under assessment should also be considered an indication of selective data reporting.

9.3. Quality of evidence assessment in network meta-analysis

The assessment of the quality of evidence based on a network meta-analysis, due to its complexity, requires specific assessment methods, which take into account the fact that the estimates for each pair

of interventions may be based on direct and indirect evidence and the complexity of the network structure.

The CiNeMA (Confidence in Network Meta-Analysis) (68,69) and Threshold analysis (70) (71) methods take into account the mixed nature (direct and indirect) of the evidence and incorporate the influence of each study on the final estimate. The quality of each study is not directly related to its contribution to the final outcome. For example, a high-quality study may have little influence on the final estimates of the network meta-analysis or vice versa (72).

CiNeMA or threshold analysis methods should be used to describe the confidence in the results of network meta-analyses.

The CiNeMA method considers six criteria: risk of within-study bias, risk of selective outcome measure reporting bias, indirect evidence, inaccuracy, heterogeneity and incoherence/ inconsistency.

It is therefore important to mention the following:

- the contributions matrix, which describes the percentage of information that each study contributes to the results of the network meta-analysis, is used to produce (semi-automatic) rankings of the risk of bias and indirect evidence (69);
- unlike the GRADE methodology, classifications are not used to judge imprecision, heterogeneity and inconsistency criteria, but the impact of these fields on the clinical decision is assessed;
- the CiNeMA method is implemented using the R software (package netmeta) (73) so it is only applicable to network meta-analyses performed using the Rücker (2012) frequentist method (74).

Threshold analysis quantifies the extent to which the evidence could be changed (for example, due to bias adjustments or sampling variation) without changing the recommendation and identifies what the new recommendation is if the evidence falls outside the calculated thresholds.

The impact of threshold analysis is highlighted below:

- threshold analysis should be performed for each study included in the meta-analysis, and for each relative effect calculated by the meta-analysis;
- threshold analysis is implemented in R software (nmathresh) (75) and can be used to evaluate frequentist or Bayesian analyses;
- the result of the network meta-analysis is considered robust if it is considered unlikely that the evidence might fall outside the calculated thresholds; otherwise, the result is sensitive to likely changes in the evidence;
- if there are studies identified as likely to alter the recommendations of the network meta-analysis, they should be inspected in detail to determine the plausibility of changes to their estimated effect beyond the calculated thresholds, taking into account the risk of bias and relevance of the study to the population under assessment;
- the assessment group is normally only interested in comparisons of the technology under assessment with the comparators in use. The thresholds calculated for these comparisons should be inspected in detail to determine the plausibility of changes in these effects beyond the calculated thresholds, taking into account the quality of the studies making up the network.

10 ADDED THERAPEUTIC VALUE

10.1. Introduction

According to Decree-Law No. 97/2015, in its current wording, paragraph 1 of Article 14 and paragraph 3 of Article 25, the co-payment/reimbursement of medicines requires, cumulatively, the technical-scientific demonstration of therapeutic innovation or therapeutic equivalence and the demonstration of their economic advantage, for the therapeutic indications claimed.

According to paragraph 6 of Article 14 and paragraph 7 of Article 25 of the same statute, the MAH for the medicinal product has the burden of proving its efficacy, added therapeutic value or therapeutic equivalence and economic advantage.

The assessment of a health technology used to treat a given indication includes the assessment of the additional benefit of that health technology compared to therapeutic alternatives commonly used in clinical practice to treat that same indication. When added benefit is demonstrated, it is considered to be a technical-scientific demonstration of added therapeutic value.

10.2. Criteria for demonstrating added therapeutic value

The assessment process compares the treatment effect of the health technology under evaluation, with the treatment effect of the comparators, on a set of outcome measures that were defined in the scoping phase (see section 2.2).

The outcome measures used should be relevant to the patient. For this purpose, patient-relevant refers to how a patient feels, functions or survives, meaning mortality, morbidity (symptoms and complications), duration of illness, quality of life, and safety.

Non-clinical outcome measures may be used as substitutes for clinical outcomes (surrogate measures), provided they have been previously validated. In the case of using surrogate measures, the company must submit evidence to demonstrate their validation.

Measures selected in the scoping phase to evaluate the effect of treatments are scored between one and nine, depending on the degree of importance assigned to them by the assessors, with a score of one to three for measures that are not important, four to six for measures that are important but not critical, and seven to nine for measures that are critical. This score makes it possible to prioritise efficacy measures and safety measures according to the importance attributed to them. It is recommended that clinical effect measures [mortality, morbidity (symptoms and complications), duration of illness, quality of life] be rated as critical, and that surrogate effect measures be rated as non-critical (score less than seven).

For each comparison (for this purpose ‘comparison’ means comparison of treatment effect between two drugs or therapeutic regimens), the treatment effect on outcome measures should be assessed, using the measures selected in the scoping phase. Thus, it follows from this assessment that, for each comparison, there will be an estimate of the relative treatment effect on each of the selected outcome measures. For each outcome measure, it is thus possible to determine whether the treatment effect of the drug under assessment presents superiority or not, in relation to each comparator.

It is then necessary to summarise these results in order to express the overall effect.

For each comparison, the conclusion of superiority of one intervention over the other, is based on the relative treatment effect on the outcome measure to which greater importance has been attributed. In

case there are several outcome measures with the same score of importance, the treatment effect estimate whose result is most reliable should be used, among the measures with the highest importance score. In case there are several outcome measures that meet these criteria (equal in score of importance and credibility of estimates of effect), the determination of the existence or not of superiority in global terms, is made using the estimates of the relative treatment effect observed on these outcome measures, being the existence or not of superiority presented descriptively (treatment effect on the outcome measures with the highest importance score, and that are equal in importance score and credibility of estimates of effect) in case of divergence.

Estimates of the treatment effect on the other outcome measures rated as critical, are valued in order to strengthen or weaken the conclusions on superiority, determined by the treatment effect on the highest scoring and most credible outcome measure, but are not used, on their own, to determine the existence or non-existence of superiority.

It may happen that a surrogate effect measure, initially rated as important but not critical, is found to be determinant for the direction of the recommendation because of the absence of treatment effect data on clinical outcome measures. In this case, this surrogate measure may be rerated and given the appropriate score and degree of importance provided that this surrogate measure has been previously validated.

10.3. Drafting of conclusions on added therapeutic value

In order to determine the added therapeutic value, and based on the scientific analysis of the available data, it is recommended that conclusions be expressed based on the degree of certainty of the results: ‘proof’ (high certainty of results when the quality of evidence is high), ‘indication’ (moderate certainty of results when the quality of evidence is moderate), ‘hint’ (low certainty of results when the quality of evidence is low), or none of the above when no data is available or the quality of evidence is very low. The outcome of the assessment of whether there is added therapeutic value should be expressed in one of the following ways: there is proof, indication or hint of added therapeutic value of an intervention.

10.4. Criteria for determining 'therapeutic equivalence'

The following scenario may result from the assessment described above: for each comparison, the overall treatment effect shows that the treatment under assessment is not superior to the comparator, but the Committee was convinced of the beneficial effect of the drug, so, using this criterion alone, it would recommend its funding. In these cases, ‘therapeutic equivalence’ is considered to exist for pricing purposes.

10.5. Criteria for not recommending co-payment / funding

The following scenario may result from the assessment described above: for each comparison, the overall treatment effect shows that the treatment under assessment is not superior to the comparator, and the Committee was not convinced of the beneficial effect of the drug. In such cases, the Committee recommends that health technology should not be funded.

If the described assessment does not provide data to conclude that the treatment is superior to the comparator and the Committee is not convinced of the beneficial effect of the drug, the Committee will, also in these cases, recommend not to fund the health technology.

10.6. Rating the magnitude of added therapeutic value

To rate the magnitude of the added therapeutic value it is recommended that the estimate of the overall treatment effect and its confidence interval be taken into account, using the upper limit of the 95% confidence interval and the thresholds defined in the Table below, and be classified in one of the following ways:

- substantial added therapeutic value (major);
- moderate added therapeutic value;
- marginal added therapeutic value (minor);
- non-quantifiable added therapeutic value.

Binary outcome measure: The determination of the extent of added therapeutic value should take into account the quality of evidence regarding the effect of treatment on the outcome measure, and be based on relative risk taking into account Table 3 and Table 4.

Time to event: The 95% confidence interval of the hazard ratio is required to determine the extent of the treatment effect in the case of outcome measures assessed by 'time to event'. If there is a meta-analysis of several studies in which the outcome measure is time to event, the hazard ratio should be used. The same thresholds as Table 3 and Table 4 shall be used to determine the extent of the added therapeutic value. If there is no risk ratio or it is not calculable, the possibility of calculating a relative risk should be considered. Where appropriate, the relative risk shall be calculated on a given date.

Magnitude of Added Therapeutic Value: it is not always possible to quantify the magnitude of the added therapeutic value at the level of the outcome measure or at the overall level. For example, if a statistically significant effect is observed in a surrogate measure, but there is no confidence data on the treatment effect on a patient-relevant outcome measure, it will not be possible to quantify the magnitude of the treatment effect. In this case, the added therapeutic value will be considered as 'non-quantifiable'. This same classification applies to cases of immature analyses, in which the treatment effect estimate may be overestimated.

Table 3: *Rating the magnitude of added therapeutic value (qualitative)*

		***Outcome* category**			
		Overall mortality	**Symptoms (or late complications) and serious (or severe) adverse events**	**Health-related quality of life**	**Symptoms (or late complications) and non-serious (or non-severe) adverse events**
Extension categories	**Substantial (major)** **Major improvement in a sustained way** in the measure of benefit assessment, which was not previously achieved in relation to the appropriate comparator	Important (major) increase in survival time	Suppression or prolonged avoidance	Major improvement	Not applicable
	Moderate **Marked improvement** in the measure of benefit assessment, which was not previously achieved relative to the appropriate comparator	Moderate increase in survival time	Suppression or prolonged avoidance	Important improvement	Important avoidance
	Marginal (minor) **Moderate and not only marginal improvement** in the measure of benefit assessment, which was not previously achieved relative to the appropriate comparator	Any increase in survival time	Any reduction	Relevant improvement	Relevant avoidance

Table 4: *Rating the magnitude of added therapeutic value (quantitative)*

		***Outcome* category**		
		Overall mortality	**Symptoms (or late complications) and serious (or severe) adverse events**	**Symptoms (or late complications) and non-serious (or non-severe) adverse events**
Extension categories	**Substantial (major)** **Major improvement in a sustained way** in the measure of benefit assessment, which was not previously achieved in relation to the appropriate comparator	0.85	0.75 and risk ≥5%	Not applicable
	Moderate **Marked improvement** in the measure of benefit assessment, which was not previously achieved relative to the appropriate comparator	0.95	0.90	0.80
	Marginal (minor) **Moderate and not only marginal improvement** in the measure of benefit assessment, which was not previously achieved relative to the appropriate comparator	1.00	1.00	0.90

The values in the Table refer to an upper threshold below which the 95% confidence interval of the relative risk should lie. The estimated 95% confidence interval of the relative risk should be below (furthest from 1) the defined threshold, i.e. the upper limit of the confidence interval should be below the defined threshold. For example, in relation to overall mortality, for the added therapeutic value to be considered disruptive, the upper bound of the 95% confidence interval of the relative risk must be at least 0.84, i.e. the reduction in relative risk of death relative to the appropriate comparator must be included in a 95% confidence interval whose upper bound is 16% or less (0.84).

Source: Modified from Ref. (IQWiG General Methods – Benefit assessment. Available at https://www.iqwig.de/en/about-us/methods/methods-paper/

11 REFERENCES

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 Lisbon, Portugal

www.infarmed.pt
infarmed@infarmed.pt

ANNEX

Deliberation of the Executive Board of INFARMED, I.P. which approved the technical-scientific criteria for the pharmacotherapeutic evaluation of health technologies

REPÚBLICA PORTUGUESA
SAÚDE

SNS SERVIÇO NACIONAL DE SAÚDE

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

Deliberação n.º 76 /CD/2022

O Sistema Nacional de Avaliação de Tecnologias de Saúde (SiNATS), criado através do Decreto-Lei n.º 97/2015, de 1 de junho, surgiu com o objetivo de dotar o Serviço Nacional de Saúde (SNS) de um instrumento único que melhorasse o seu desempenho, introduzindo neste âmbito a experiência já existente em Portugal e as melhores práticas ao nível europeu, no que se refere à avaliação e reavaliação de tecnologias de saúde.

Nos termos do referido diploma, o financiamento pelo SNS de medicamentos e dispositivos médicos depende da demonstração do seu valor terapêutico acrescentado ou equivalência terapêutica e da sua vantagem económica

Nos termos do disposto no artigo 5º n.º 9, do Decreto-Lei nº 97/2015, de 1 de junho, alterado Decreto-Lei n.º 115/2017, os critérios técnico-científicos para a avaliação das diferentes tecnologias de saúde são estabelecidos em regulamento aprovado pelo Conselho Diretivo do INFARMED, I. P.

Neste sentido, tendo por base o trabalho desenvolvido por um grupo de peritos constituído no âmbito da Comissão de Avaliação das Tecnologias de Saúde (CATS) do INFARMED — Autoridade Nacional do Medicamento e de Produtos de Saúde, I. P. (INFARMED, I. P.), foi elaborada a Metodologia de Avaliação Farmacoterapêutica, versão 3.0, que atualiza a metodologia anterior, de forma a fornecer a melhor análise da evidência de apoio à decisão, pretendendo-se, assim, dar resposta aos desafios metodológicos encontrados nos últimos tempos, atualizando-se o conhecimento e desenvolvimento científicos em matéria de avaliação farmacoterapêutica de tecnologias de saúde.

Assim, nos termos e ao abrigo do disposto no n.º 9 do artigo 5.º, do Decreto-Lei n.º 97/2015, de 1 de junho, alterado pelo Decreto-lei n.º 115/2017, de 7 de setembro, o Conselho Diretivo do INFARMED, I. P., delibera:

1. São aprovados em anexo os critérios técnico-científicos para a avaliação farmacoterapêutica (Metodologia de Avaliação Farmacoterapêutica, versão 3.0) das diferentes tecnologias de saúde que consta do anexo à presente Deliberação.

2. A presente Deliberação é publicada na página eletrónica do INFARMED, I.P. e produz efeitos a partir de 01 de novembro de 2022.

Lisboa, 29 JUL. 2022

O Conselho Diretivo

Rui Santos Ivo, Presidente

Carlos Lima Alves, Vice-Presidente

Erica Viegas, Vogal