EUROPEAN NETWORK FOR HEALTH TECHNOLOGY ASSESSMENT

EUnetHTA 21 - Individual Practical Guideline Document

D4.6 VALIDITY OF CLINICAL STUDIES

Version 1.0, 16.12.2022

Template version 1.0, 03/03/2022

Document history and contributors

Version	Date	Description
V0.1	**23/03/2022**	First draft
V0.2	**25/05/2022**	Second draft
V0.3	**20/06/2022**	Draft for public consultation
V0.4	**28/09/2022**	Draft for CSCQ validation
V0.5	**02/11/2022**	Endorsed by CEB
V1.0	**16/12/2022**	Date of publication

Disclaimer

This Practical Guideline was produced under the Third EU Health Programme through a service contract with the European Health and Digital Executive Agency (HaDEA) acting under the mandate from the European Commission. The information and views set out in this Practical Guideline are those of the author(s) and do not necessarily reflect the official opinion of the Commission/ Executive Agency. The Commission/Executive Agency do not guarantee the accuracy of the data included in this study. Neither the Commission /Executive Agency nor any person acting on the Commission's/Executive Agency's behalf may be held responsible for the use which may be made of the information contained therein.

Participants

Hands-on Group	Gemeinsamer Bundesausschuss [G-BA], Germany Haute Autorité de Santé [HAS], France Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen [IQWiG], Germany
Project Management	Zorginstituut Nederland (ZIN), the Netherlands
CSCQ	Agencia Española de Medicamentos y Productos Sanitarios (AEMPS), Spain Austrian Institute for Health Technology Assessment (AIHTA), Austria Belgian Health Care Knowledge Centre (KCE), Belgium Gemeinsamer Bundesausschuss (G-BA), Germany Haute Autorité de Santé (HAS), France Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG), Germany Italian Medicines Agency (AIFA), Italy National Authority of Medicines and Health Products, I.P. (INFARMED), Portugal National Centre for Pharmacoeconomics, St. James Hospital (NCPE), Ireland National Institute of Pharmacy and Nutrition (NIPN), Hungary Norwegian Medicines Agency (NOMA), Norway The Dental and Pharmaceutical Benefits Agency (TLV), Sweden Zorginstituut Nederland (ZIN), the Netherlands
CEB	

The work in European Network for Health Technology Assessment (EUnetHTA) 21 is a collaborative effort. While the agencies in the Hands-on Group wrote the deliverable, the entire EUnetHTA 21 consortium was involved in its production throughout various stages. This means that the Committee for Scientific Consistency and Quality (CSCQ) reviewed and discussed several drafts of the deliverable prior to validation. Afterwards the Consortium Executive Board (CEB) endorsed the final deliverable prior to publication.

Associated HTAb & Stakeholders participating in public consultation

The draft deliverable was reviewed by associated HTAb and was open for public consultation between 04.07.2022 and 02.08.2022.

Associated HTA bodies who reviewed	Dachverband der Österreichischen Sozialversicherung, [DVSV], Austria Directorate for Pharmaceutical Affairs [DPA], Malta Evaluation and Planning Unit – Directorate of the Canary Islands Health Service, [SESCS], Spain Health Information and Quality Authority [HIQA], Ireland Norwegian Institute of Public Health, [NIPH], Norway Regione Emilia-Romagna, [RER], Italy Swedish Agency for Health Technology Assessment and Assessment of Social Services [SBU], Sweden
Stakeholders who reviewed during public consultation	European Organisation for Research and Treatment of Cancer (EORTC), Belgium International Association of Mutual Benefit Societies (AIM), Belgium European Union of General Practitioners/Family Physicians (UEMO), Belgium European Confederation of Pharmaceutical Entrepreneurs (EUCOPE), Belgium European Federation of Pharmaceutical Industries and Associations (EFPIA), Belgium Alliance for Regenerative Medicine (ARM), Belgium European Organisation for Research and Treatment of Cancer (EORTC) , Belgium The European Socienty for Paediartic Oncology (SIOPE), Belgium Takeda Pharmaceuticals International AG, Brussels, Switzerland, local operating companies across the European Union Edwards Lifesciences, Europe European Federation of Statisticians in the Pharmaceutical Industry (EFSPI) HTA SIG, Europe MedTech Europe (MTE), Europe - Belgium Lymphoma Coalition - Lymphoma Coalition Europe (LCE), France EHA, France EURORDIS, France Ecker + Ecker GmbH (E+E), Germany SKC Beratungsgesellschaft mbH (SKC), Germany Verband Forschender Arzneimittelhersteller (vfa) e.V, Germany GKV-Spitzenverband (GKV-SV), Germany Bayer AG & Bayer Vital GmbH, Germany German Medicines Manufacturer´s Association (BAH), Germany AstraZeneca (AZ) Global, (UK based) F. Hoffmann-La Roche Ltd (Roche), Switzerland Medtronic, Switzerland GSK, UK

Copyright

Table of Contents

LIST OF ABBREVIATIONS

CEB	Consortium Executive Board
CI	Confidence interval
CSCQ	Committee for Scientific Consistency and Quality
EUnetHTA	European Network of Health Technology Assessment
GRADE	Grading of Recommendations, Assessment, Development and Evaluation
HTA	Health Technology Assessment
HTAb	Health Technology Assessment body
HTAR	Health Technology Assessment Regulation
HTD	Health Technology Developer
ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
JCA	Joint Clinical Assessment
MAMS	Multi-arm, multi-stage trials
PICO	Population, intervention, comparator, outcome
RCT	Randomised clinical trial
RoB	Risk of bias
RWD	Real-world data
RWE	Real-world evidence
TWIC	Trial within a cohort

1 INTRODUCTION

1.1 Problem statement

One key element of Health Technology Assessment (HTA) is to assess and describe the certainty (and validity) of clinical study results in an objective, reproducible, and transparent way. In 2020, the European Network of Health Technology Assessment (EUnetHTA) Executive Board concluded that GRADE (1) (or any other system for rating the overall quality of evidence and developing healthcare recommendations) can only partially be applied within EUnetHTA because overall conclusions or recommendations might interfere with the independent contextualisation and decision-making at the national level (2). However, valid scientific principles are still required, not only to guide the development of Joint Clinical Assessments (JCAs) at the European level, but also to support the understandability and usability of these results for national decision-making.

1.1 Scope/Objective(s) of the Guideline

This Practical Guideline is dedicated to the definition, classification, and assessment of the certainty of results of studies leading to the statistical analysis of data considered as originating from or part of a single study (i.e., one sample of patients). Studies that consist in evidence synthesis by pooling the results of multiple already-analysed data sets from multiple samples of patients [e.g., pairwise meta-analysis, indirect comparison, or interventional studies such as single-arm trials coupled with an external source of data as a control group (including historical control)] are not included in this Guideline. The EUnetHTA 21 Methodological and Practical Guidelines *Direct and Indirect Comparisons* provide recommendations and guidance for the classification of these evidence syntheses. Finally, the present Guideline does not offer guidance on how to assess diagnostic accuracy studies, because these studies might have a conventional cross-sectional or cohort design, but still require specific assessment of internal validity (3).

The way in which the validity of clinical studies will be assessed and interpreted for drawing conclusions at a national level cannot be dissociated from the population, intervention, comparator, outcome (PICO) question that will be formulated by Health Technology Assessment bodies (HTAbs) (see the EUnetHTA 21 Practical Guideline *Scoping Process*). For complementary elements relating to the reporting and assessment of multiple hypothesis testing, subgroup, sensitivity, and post-hoc analyses, the reader is referred to the EUnetHTA 21 Practical Guideline *Applicability of Evidence - Practical Guideline on Multiplicity, Subgroup, Sensitivity, and Post-Hoc Analyses*. Additional considerations of the definition of clinically relevant outcomes and endpoints, and the assessment of their validity, reliability, and interpretability are discussed within the EUnetHTA 21 practical guideline *Endpoints*.

1.2 Relevant articles in Regulation (EU) 2021/2282

Articles from Regulation (EU) 2021/2282 directly relevant to the content of this practical guideline are:

- Recital (14);
- Recital (28);
- Article 8: Initiation of Joint Clinical Assessments;
- Article 9: Joint Clinical Assessment Reports and the Dossier of the Health Technology Developer.

2 GENERAL CONSIDERATIONS

HTA requires the relative effectiveness of an intervention to be determined as correctly and precisely as possible. Relative effectiveness is the quantification of the effect caused by the intervention relative to a comparator (e.g., standard of care) on an outcome of interest. Interventions can be medicinal products, medical devices, *in vitro* diagnostic medical devices, medical procedures, as well as measures for disease prevention, diagnosis, or treatment. For any effectiveness assessment, it is essential to examine and report the certainty of results systematically. Given that the certainty of results is fundamental, this needs to be communicated alongside the numerical results. According to Article 9 of EU-HTA Regulation (4), it is therefore essential that a JCA contains a description of both 'the relative effects of the health technology' and 'the degree of certainty of the relative effects, taking into account the strengths and limitations of the available evidence'.

Practical Guideline (requirement for JCA reporting)

Any effectiveness result in a JCA report must be accompanied by a description of its certainty.

The certainty of effectiveness results is determined by three concepts: **internal validity** [i.e., the extent to which a study is free from bias (also called systematic errors), a concept analogous to Risk of Bias (RoB)]; **applicability** (i.e., the extent to which study results provide a basis for generalisation to the target population, a concept close to external validity and generalisability); and **statistical precision** (i.e., the uncertainty associated with study results due to random sampling variability). These three concepts assess three different dimensions of the certainty of results, which, for example, means that shortcomings in internal validity cannot be remedied by higher statistical precision. Furthermore, evidence that has high internal validity does not necessarily have high external validity. Although HTA usually requires a high target certainty of results, it is necessary to assess all available data, as submitted by the Health Technology Developer (HTD). Nevertheless, there might be justification to not assess the evidence that ranges below a minimum level of internal validity, applicability, or statistical precision in detail, if the PICO question can be sufficiently answered on the basis of higher-certainty results. Furthermore, the certainty of results is independent of the medical context of the PICO question. It is methodologically inappropriate, for example, to take the rareness of a disease or the impossibility of blinding as a justification to ignore the resulting uncertainties in the clinical evidence.

Following international standards of evidence-based medicine, the internal validity of a study has a paramount role in determining the overall certainty of the study results (i.e., if study results have a low level of internal validity, the levels of statistical precision and external validity are irrelevant) (5,6). The classical **hierarchy of evidence** (7) includes several types of study, from case-reports and nonclinical data (level 5 evidence, the lowest level of evidence,), case-control studies (level 4), retrospective (or lower-quality) cohort studies (level 3), prospective (or higher-quality) cohort studies (level 2), up to randomised controlled trials (RCTs; level 1, the highest level of evidence). Classification of study design alone (see Section 3) is insufficient for a full assessment of internal validity (8,9), but has much practical value for distinguishing between higher- and lower-quality evidence and for selecting a suitable RoB assessment tool.

Practical Guideline

For internal validity, it is useful in a JCA to distinguish between different study designs.

RoB can be defined as any potential systematic error in clinical research that might lead to an incorrect estimate of the effect of interest. RoB can be present at different levels, including: (i) the meta level (e.g., publication bias in a systematic review or meta-analysis); (ii) the study level (e.g., confounding bias in a cohort study); and (iii) the outcome level (e.g., information bias caused by unblinded assessment of an outcome). If the type of evidence requires it, the assessment of RoB needs to be level specific; however, the scope of the present Guideline is limited to bias at the study and outcome levels. Furthermore, some types of bias can occur only in certain study designs, whereas other types can affect all types of study. Therefore, different tools have been developed for RoB assessment in different study designs (10,11). It is essential to use these standard tools (see other Guidance documents at https://www.eunethta.eu/methodology-guidelines/).

Practical Guideline

Standard study design-specific tools should be used to assess RoB.

The terms '**applicability**', 'external validity', 'transferability', 'generalisability', and 'directness' are often used interchangeably. In the context of an HTA report, it is most appropriate to use the term 'applicability' (12,13), although the term 'indirectness' can be chosen in the context of GRADE methodology. The key question is how well the evidence matches the elements of the PICO question and, therefore, whether it can be applied to answer that question (14). In statistical terms, a lack of applicability of clinical evidence threatens the overall certainty of results if, because of relevant effect modification, the effect in the population of interest is probably different from the effects in the clinical studies.

Limitations to the applicability of the evidence can occur if: (i) the study population (based on eligibility criteria or actual recruitment) differs from the intended target population; (ii) the experimental or control interventions were not performed in the way that they are applied or would be applied in the target setting; or (iii) the study outcomes (e.g., surrogate outcomes) fail to offer information about the outcomes of interest. For the applicability of clinically relevant evidence, effect modification has to be taken into account (15). For example, if the relative effectiveness of a drug was shown to vary substantially with age, the application of overall study results would be questionable. Instead, the subgroup results for the corresponding age groups or other analytical techniques could supplement information on relative effectiveness.

Given that lack of applicability as compared with internal validity is usually less relevant and more straightforward to detect, it might be sufficient to assess any issues with regard to patients and interventions on a case-by-case basis using qualitative descriptive methods. Most HTA agencies found this approach to be preferable and do not use a specific instrument or checklist to judge the applicability of clinical evidence (16). The applicability of a study can differ between European member states, not only because PICO questions are often different, but also because of different healthcare settings (e.g., organisational aspects). Therefore, a final judgment on applicability can only be made at the national (or even regional) level by each member state itself. Accordingly, the HTA Regulation (HTAR) mentions that ‘external validity’ (I.e., applicability) should be assessed in a JCA, but without forestalling any national judgement on applicability. To support national decision-making, specific issues in relation to applicability should be described and addressed in a JCA, where necessary. This primarily includes any potential mismatch between the PICO of interest and the PICO examined in a clinical study. However, in the JCA, each aspect (e.g., questionable applicability because of differences in patient population or control intervention) will only be commented on and briefly analysed, but without providing a conclusion on applicability.

Issues with regard to surrogate outcomes usually require specific attention in HTA (17). However, surrogacy is outside the scope of this Practical Guideline, and is addressed in the EUnetHTA 21 practical guideline *Endpoints*.

Practical guideline

Different aspects of applicability (primarily any PICO mismatch between assessment scope and clinical study) should be addressed in a JCA, but the final judgment on the applicability of study results must be left to the discretion of each member state.

Statistical precision is a quantitative concept that can be applied for each outcome of interest, at both the meta and study level. Variation, and the uncertainty that comes with it, can occur in both primary studies and evidence synthesis, and differentiation of both (within-study and between-study variability) is required to better understand the underlying sources of variation. Effect estimates and other key results should always be accompanied by the corresponding measures of statistical precision, preferably confidence intervals (CIs) at a 95% level (18,19). To increase the transparency and understandability of results, data submissions and analyses should contain counts and other types of descriptive statistics.

In a single study, statistical hypothesis testing can be used to decide whether an effect was proven. Statistical testing in a clinical study requires transparent and clear prespecification of hypotheses, adequate handling of eventual multiplicity issues (20), full reporting of results (21), and careful interpretation (22) (see also EUnetHTA 21 D4.5 Practical Guideline *Applicability of Evidence: Practical guideline on Multiplicity, Subgroup, Sensitivity, and Post-hoc Analyses*). Data-driven statistical tests provide results of low internal validity. Similarly, early unplanned stopping of clinical studies, deliberate extension of recruitment, and selective reporting of results all undermine the validity of study results (23). However, the rates of type I and type II errors in a clinical trial are not directly related to the validity of the observed treatment effects, because these errors are relevant only when interpreting the results of statistical tests (24).

Most comparative studies on interventions examine superiority hypotheses, but, depending on the medical context, non-inferiority and equivalence are also tested. Although the type of question (superiority, non-inferiority, or equivalence) is also important in HTA, common work on a JCA should consider the rejection of the null hypothesis of a statistical hypothesis test against a prespecified α level [which, in biomedical research, is usually set at 0.05 (5%)]. This neither represents nor predetermines a conclusion of the added value of the assessed technology. Similarly, the clinical relevance of an effect size, which can be assessed by comparing the effect size with a predefined threshold or by responder analyses (25), needs to be judged at the national context. This point is addressed in the EUnetHTA 21 Practical Guideline *Endpoints*.

The certainty of a positive or negative effect will be higher if a very large effect size was found and the accompanying 95% CI and *p* value safely exclude the possibility of no effect (26–28). Which effect sizes can be considered very large and which *p* values can be accepted as sufficiently low is an unresolved scientific question (29). Nevertheless, in the context of a JCA, it might be helpful to highlight such situations, especially when no RCT evidence is available. For effect sizes expressed as relative risks, the threshold of a relative risk superior to 5 (or inferior to 0.2) and a *p* value <0.01 (as an indicator of sufficient precision) was proposed as a ‘rule of thumb’ (i.e., an arbitrary rule based on expert opinion) (26,30). The JCA report will describe effect estimates, but without a conclusion on whether the certainty of results is increased, because this is best made at the national level.

Practical Guideline (requirement for JCA reporting)

To describe statistical precision accurately, effect estimates should always be accompanied by the corresponding measures of variation, preferably CIs at a specified 1-α level of confidence, which is 0.95 (95%) in most cases.

3 CLINICAL STUDY DESIGNS

3.1 Terminology

Classification and labelling of studies design can vary. For this Practical Guideline, we establish standardised definitions for classifying and labelling clinical studies. These are used without prejudice to the definitions that might be applied in national legislation and related regulatory guidance.

Studies are classified into two categories: **interventional studies** and **observational studies**. For consistency, synonyms, such as ‘clinical trials’ or ‘experimental studies’ for interventional studies or ‘non-interventional’, ‘non-experimental’, or ‘nonrandomised studies’ for observational studies are not used in this Guideline.

Distinction between interventional and observational studies depends on whether the intervention under assessment is assigned by the investigator(s) through the study protocol (interventional) or is given during routine clinical care (observational).

3.1.1 Interventional studies

In interventional study, the intervention(s) (one or several) under assessment are assigned by the investigator(s).

Classification of interventional studies could be established based on the study characteristics. These have already been fully defined but can sometimes vary (31,32). Therefore, it is valuable to establish definitions of design characteristics for this Guideline. Study characteristics are summarised in Table 3.1, based on the glossary from the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E9 Statistical Principles for Clinical Trials (33) or EU Clinical Trials Register (34). Note that Table 3.1 is intended to combine definitions and not to be used as a reporting template.

Table 3.1. Interventional study characteristics

Characteristic	Definition
1. Control	
Controlled (or comparative)	The study compares the effect of one or multiple treatments of interest to one or multiple comparators
2. Randomisation	
Randomised	A form of controlled allocation whereby patients are randomly assigned to one of the treatment groups
3. Blinding	
Blind	When people (patients and/or investigators and/or outcome assessors and/or statisticians) do not know which intervention is being given
4. Design	
Single arm	A trial in which all patients receive the same intervention
Parallel	Two or more interventions are evaluated concurrently in separate groups of patients
Cross-over	Comparison of two (or more) interventions in which patients are switched to the alternative treatment after a specified period (therefore, in most cases, each patient receives each treatment)
Factorial	Two or more treatments are evaluated simultaneously through the use of varying combinations of those treatments
5. Objective	
Superiority	Trial with the primary objective of showing that the response to the treatment(s) of interest is clinically superior to that of a comparator
Non-inferiority	Trial with the primary objective of showing that the response to the treatment(s) of interest is not clinically inferior to that of a comparator. This is usually demonstrated by showing that the true treatment difference is unlikely to cross a threshold of an acceptable non-inferiority margin
Equivalence	Trial with the primary objective of showing that the response to two or more treatments differs by an amount that is clinically negligible. This is usually demonstrated by showing that the true treatment difference is likely to lie between a lower and an upper equivalence margin of clinically acceptable differences

3.1.2 Observational studies

In observational studies, there is no forced change in routine care (except where the protocol requires visits at specific timepoints) and neither is the usual decision for exposure[1] affected by an observational study. Given that observational studies are performed based on routine healthcare, this suggests that they allow the assessment of relative effectiveness of only those interventions that are already used in medical practice, rather than of new ones.

Descriptive or analytical

Observational studies can be either **descriptive**, that is, without a control group (case-series and cross-sectional studies) or **analytical** (case-control and cohort studies) with a control group. Analytical studies provide a measure of the association between exposure (notably interventions) and outcome of interest. In a case-series, changes over time can be analysed (i.e., before and after the introduction of the treatment of interest); however, under usual circumstances, such before–after changes are unlikely to assess interventional effects. It is generally important to remember that association does not necessarily imply causality. Analytical studies, such as cohort and case-control design, can be useful when randomisation is deemed unethical or unfeasible.

[1] In the context of observational studies, the broader term ‘exposure’ is used to denote anything whose relationship with an ‘outcome’ is being explored. In HTA, exposures of interest are mainly medical interventions.

Prospective and retrospective

The collection of the data from those studies can be done prospectively or retrospectively. **Prospective** studies measure exposures before the occurrence of the outcome of interest, whereas **retrospective** studies measure exposure after the occurrence of the outcome of interest.

Retrospective data are usually collected from existing data sources. Thus, retrospective studies can be quicker to complete compared with prospective studies but are limited by the availability of the existing data. Furthermore, there can be a high risk of recall bias if the determination of exposure status relies on recall or records only. In that case, the fundamental assumption that cause precedes effect can be violated, which implies that the study of causality between exposure and outcome of interest is unfeasible.

By contrast, prospective observational studies might be more time consuming to perform, but the patient follow-up is standardised, and the availability of data that can be collected is not determined before the conduct of the study. Furthermore, if a prospective study is designed to ensure that exposition precedes outcome, the aforementioned fundamental assumption for causality can be assumed.

Cohort study

Cohort studies, also known as incidence studies, longitudinal studies, follow-up studies, or prospective studies, are studies following a group of subjects (a cohort) with a common exposure over time, but without having experienced the outcome of interest at enrolment. Patients are followed during a specified period, and data on outcomes of interest are collected in a prospective manner.

While the term “cohort” alone is sometimes used to define a longitudinal follow-up of patients irrespective of a comparison or not, in this Guideline, “cohort studies” are always considered as comparative in that a cohort study follows up two or more groups from exposure to outcome. As previously mentioned, this chosen classification/definition is used without prejudice to definitions that might be applied elsewhere.

Sometimes, a cohort study data set can serve as a basis for enrolling patients into an interventional study, which can be a RCT (i.e., a subset of newly included or already-included patients can be allocated to one of the exposuress assessed if, at a proper time, they meet the eligibility criteria for the interventional study). When it happens, this design is called a ‘trial within a cohort’ (TWIC) (35).

Case-control study

Case-control studies are retrospective studies that enroll patients who have experienced a particular outcome of interest (‘cases’), compared with patients who have not experienced the outcome of interest but who are representative of the study population on some controlled criterion (‘controls’).

The aim of this study design is to compare the exposure between case and controls to identify factors that might contribute to be associated to the occurrence of an outcome.

Cross-sectional study

Cross-sectional studies, also known as transversal studies, measure outcomes and exposure status simultaneously in a specified population to study the frequency and characteristics of an outcome at a particular point in time.

The main use of this study design is to assess outcome and/or exposure prevalence in a population.

Case study: case-report and case-series

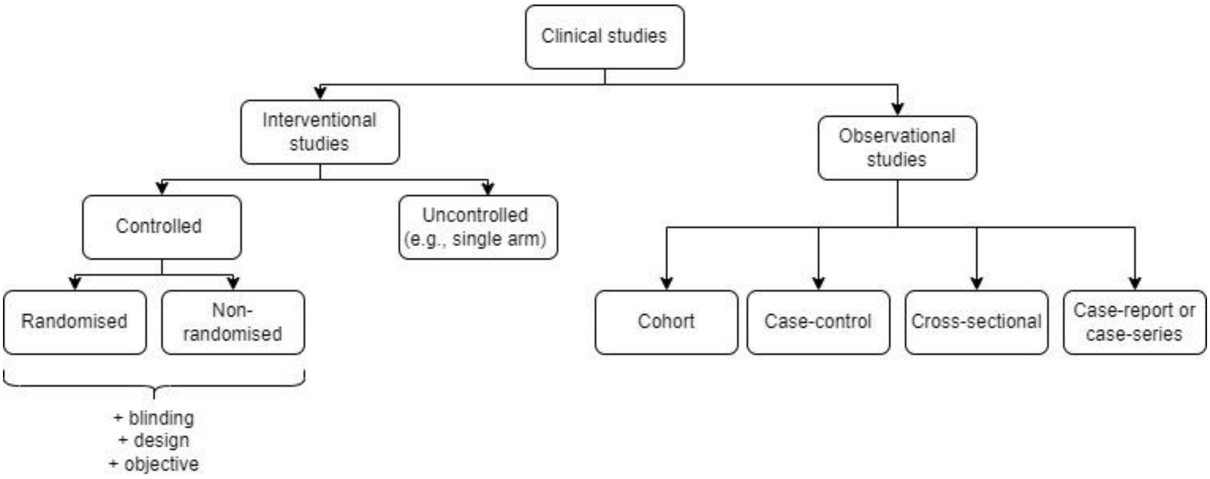
Case studies are descriptive studies of a single case (**case-report**) or a group of subjects with similar diagnoses or exposure (**case-series**) followed over time. It provides detailed descriptions of cases without the use of a control group. However, in a case-series, it is possible to compare the health status of participants over time, for example, to estimate the pre–post changes induced by an exposure. Given the characteristics of this design, such changes are unlikely to estimate the true effect of the treatment of interest.

Case studies can be used to describe rare events or early trends, such as unusual manifestations of a disease or unusual response to an exposure. Some case-reports in the medical literature are intended to prove the feasibility of an exposure. Those study designs cannot be used to assess the effectiveness of an exposure. However, they can help to detect new safety signals.

3.2 Classification

The classification of clinical studies is presented in Figure 3.1.

Figure 3.1. Classification of clinical studies



Practical Guideline (requirement for JCA reporting)

Classification and design characteristics for each study submitted as evidence.

As per its definition, this classification is used in this Guideline without prejudice to the definitions that might be applied elsewhere (36,37).

4 SPECIFIC STRENGTHS, WEAKNESSES, AND RECOMMENDATIONS REGARDING DIFFERENT DESIGNS

The JCA will report the certainty of results of the relative effectiveness of the treatment(s) of interest, taking into account the strengths and limitations of the available evidence [Article 9(1)]. As previously described, the certainty of results is determined by internal validity, applicability, and statistical precision.

Study design or conduct can lead to bias, impacting internal validity. Several standardised tools have been developed to evaluate RoB in various clinical study designs (13,14). They are helpful for assessing the strengths and limitations of the available evidence and should be used when performing JCA. This Guideline recommends the systematic use of Cochrane’s tools to assess RoB.

4.1 Randomised clinical trials: gold standard for treatment effect estimation

RCTs are the gold standard for evaluating causal relationships between interventions and outcomes because randomisation eliminates much of the bias inherent to other designs (38). In brief, a proper randomisation allows the trial to be conducted under the assumption of **exchangeability** (i.e., if patients from one group were substituted to the other, the same treatment effect would be observed). This underlying assumption implies the absence of confounding bias (both on known and unknown confounders and effect modifiers). Moreover, blinding alongside with identical and standardised follow-up between each group help to maintain exchangeability over time and prevent measurement bias. As a result of randomisation and blinding, relative effectiveness assessment allows estimation of the supplementary causal effect of an intervention of interest over comparator treatment effects. Finally, rigorous follow-up and analysis of the adequate population (e.g., intention-to-treat population for a superiority RCT) help control attrition. Nonetheless, depending on numerous factors, such as the quality of the design and conduct of the study, the certainty of results of a particular RCT can be questioned and biases can arise (39,40).

To allow proper evaluation by member states, RoB should be assessed using **ROB-2** (10). Full guidance documents for ROB-2 could be found on the Cochrane resource website (https://methods.cochrane.org/risk-bias-2). Given that ROB-2 assumes that overall RoB is performed at the outcome level, RoB should be performed for every outcome required in the assessment scope (i.e., 'O', from PICO). Indeed, although the occurrence of some biases can frequently impact internal validity at the study level (i.e., irrespective of the outcomes being assessed), other biases can be outcome dependent (e.g., nonblinding can affect the assessment of outcomes differently, such as overall survival and quality of life measured by patient-reported outcomes). Moreover, results in the JCA report will be presented primarily according to PICO questions. Therefore, it is valuable to have a RoB assessment at the outcome level. The use of ROB-2 does not exclude the possibility of assessing evidence with an analysis strategy that corresponds best to a given PICO question (e.g., for addressing the issue of the adequate management of intercurrent events and missing data), as defined according to the principles of the estimand framework outlined in ICH E9 and its addendum (E9(R1)) (41,42).

Practical Guideline (Requirement for JCA reporting)

For outcomes with evidence coming from RCTs, assess RoB using ROB-2.

RoB should be assessed for each outcome required in the assessment scope.

The RoB of outcomes sharing the same characteristics (e.g., data collection, blinding aspects, or evaluation) can be assessed on an overall level (i.e., grouped).

RoB judgement should be provided for both each individual domain level and overall.

4.2 Nonrandomised controlled trials

Non-RCTs are clinical trials in which participants are allocated to intervention under assessment or reference intervention using methods that are not random. Allocation could be based, for example, on investigator's choice, participant's choice, or calendar dates. They allow direct estimation of relative effects between interventions. However, such non-random allocation breaks the underlying assumption of exchangeability and, therefore, is likely to lead to confounding bias. Thus, the estimated association between intervention and outcome is likely to be biased and will differ from its true causal effect.

There are different methods that can be used to control for confounding (i.e., allowing if properly conducted, **conditional exchangeability**, e.g., design-based methods, such as stratification or matching, or modeling-based methods, such as adjustment or models of causal inference (e.g., propensity scores or g-computation)] within the trial. Any method for controlling confounding bias when allocation was not randomised requires exhaustivity (i.e., all relevant confounders and effect modifiers must be known and adequately measured within the trial), an unverifiable underlying assumption. By contrast, known and unknown, measured and unmeasured confounding factors and effect modifiers are fully controlled through randomisation.

To allow proper evaluation by Member States, RoB should be assessed using **ROBINS-I**. Full guidance documents for ROBINS-I could be found here using the Cochrane resource website (https://sites.google.com/site/riskofbiastool/welcome/home/current-version-of-robins-i). As for ROB-2, RoB assessment using ROBINS-I must be performed at the outcome level.

Practical Guideline (Requirement for JCA reporting)

For outcomes with evidence coming from non-RCTs, assess RoB using ROBINS-I.

RoB should be assessed for each outcome required in the assessment scope.

The RoB of outcomes sharing the same characteristics (e.g., data collection, blinding aspects, or evaluation) can be assessed on an overall level (i.e., grouped).

RoB judgement should be provided for both each individual domain level and overall.

4.3 Uncontrolled clinical trials (e.g., single-arm trials)

Unlike comparative clinical trials, uncontrolled trials, when they are the only source of data submitted as evidence, do not allow relative effectiveness assessment (i.e., supplementary effect over comparator treatment effect). In terms of strengths and weaknesses, they can be considered mostly akin to case-series. However, a difference with case-series is that the treatment is delivered as part of a study intervention. Therefore, patients in a single-arm trial can receive a treatment in a more-standardised manner and with a more-rigorous follow-up compared with those from a case-series. In the context of HTA, uncontrolled clinical trials are of very limited value for estimating treatment effectiveness.

Given the lower importance of uncontrolled trials for relative effectiveness assessment and HTA, it is deemed unnecessary to propose any formal rules for assessing RoB of single-arm trials. Some tools have been developed in the past (43–46), but RoB of uncontrolled studies appears to be affected by only a few specific aspects of internal validity, such as the consecutiveness of recruitment, the prespecification of sample size and analyses, and the blinded assessment of outcomes. Nevertheless, RoB of an uncontrolled study is very unlikely to be changed by formal RoB assessment; thus, this work appears dispensable.

Data of a single-arm trial can be used coupled with an external source of data as a control to allow for a comparative statistical analysis. In the context of a JCA, the assessment of such external comparisons is explicated in the EUnetHTA 21 methodological and practical guidelines *Direct and indirect comparisons*. Such a comparison requires an adequate use of a method of causal inference. In such a context, the framework of the emulation of a target trial can help to formulate the appropriate causal research question that is addressed. It allows defining the appropriate estimand, eligibility criteria as well as exposition and outcome(s) of the targeted (i.e., ideal) RCT the external comparison tries to emulate (47).

Practical Guideline

Uncontrolled trials per se are of very limited value for performing relative effectiveness assessment.

Although the (partial) use of some tools for RoB assessment is possible, the overall conclusion on the (very limited) internal validity of uncontrolled studies is very unlikely to be changed by RoB assessment. Therefore, RoB assessment is not required.

4.4 Cohort studies

Cohort studies can be used when allocation of an intervention in a controlled manner is deemed unethical or unfeasible. Compared with interventional studies, they can allow larger sample sizes and longer follow-up, improving statistical precision or the detection of long-term adverse events (48). They can also help to investigate the effectiveness of interventions when used in routine healthcare on a sample of patients with less-stringent eligibility criteria compared with an interventional study, which could enhance applicability.

Given that the intervention is not randomised between patients, the underlying assumption of exchangeability cannot hold, which is very likely to lead to confounding bias. Thus, without the proper use of an appropriate method for controlling for confounding (see Section 4.2), the estimated association between exposure and outcome of interest will most likely differ from its true causal effect. As described in Section 4.3, when controlling for confounding by using an appropriate method of causal inference, the framework of the emulation of a target trial can help to formulate the appropriate causal research question that is addressed (47).

Practical Guideline (Requirement for JCA reporting)

For outcomes with evidence coming from cohort studies, assess RoB using ROBINS-I.

RoB should be assessed for each outcome required in the assessment scope.

The RoB of outcomes sharing the same characteristics (e.g., data collection, blinding aspects, or evaluation) can be assessed on an overall level (i.e., grouped).

RoB judgement should be provided for both each individual domain level and overall.

4.5 *Case-control studies*

A case-control study design is useful to examine rare outcomes, and multiple factors affecting one outcome can be studied.

In case-control studies, patients are enrolled based on the occurrence of outcome and exposures are investigated in a retrospective manner. Thus, they are at high risk of selection bias. The selection of a control group is very likely to not allow verification of the exchangeability assumption. It leads to the same issues as described before for non-RCTs and cohort studies regarding confounding bias (see Section 4.2). Moreover, case-control studies are also likely to lead to a measurement bias, especially recall bias, because exposure is measured after the onset of the disease or outcome. Moreover, because data are collected in a retrospective manner, it is uncertain that the exposure of interest precedes the occurrence of the outcome of interest, which can lead to violation of a fundamental rule of causation (exposure must precede effect).

Finally, this study design is not suited for rare exposures and for studying more than one outcome.

Practical Guideline (Requirement for JCA reporting)

For each outcome with evidence coming from a case-control study, assess RoB using ROBINS-I.

RoB should be assessed for each outcome required in the assessment scope.

The RoB of outcomes sharing the same characteristics (e.g., data collection, blinding aspects, or evaluation) can be assessed on an overall level (i.e., grouped).

RoB judgement should be provided for both each individual domain level and overall.

4.6 *Cross-sectional studies*

A cross-sectional study design is useful to investigate multiple outcomes and exposures simultaneously.

This type of study estimates association but cannot be used to study the cause–effect relationship or causality because there is no temporality; thus, it is not possible to distinguish whether the exposure preceded or followed the outcome. Therefore, it is deemed unnecessary to propose any formal tool for assessing RoB of cross-sectional studies.

Practical guideline

Evidence coming from cross-sectional studies is of very limited value for performing relative effectiveness assessment.

No RoB assessment using a standardised tool is required for cross-sectional studies.

4.7 *Case-series and case-reports*

These studies allow the generation of hypotheses, such as identifying unexpected effects (adverse or beneficial) and describing unusual syndromes that could later be studied using study designs with a higher certainty of results.

These studies are only descriptive and are rarely used to test hypotheses or establish causal effects. Any effect estimate generated from a study lacking a control group is only a pre–post change, thus the interpretation of such change as a causal effect requires the very unlikely assumption that no change would have occurred without the intervention. Furthermore, case-reports generate selection bias and lack external validity because of low representativeness. Therefore, it is deemed unnecessary to propose any formal tool for assessing RoB of case-series and case-reports.

Practical Guideline

Evidence coming from case-series and case-reports is of very limited value for performing relative effectiveness assessment.

No risk of bias assessment using a standardised tool is required for case-series and case-reports.

5 PARTICULARITIES

Specific topics in clinical methodology that are of particular relevance for HTA will be introduced in this section. Indeed, although these topics are methodological concepts that are now prevalent when discussing the design of clinical studies, they cannot be strictly classified according to the principles described earlier in the document (see Section 3). These particularities can be compatible with many features of the aforementioned designs (e.g., some can be compatible with the principles of RCTs). Nonetheless, their definitions, strengths, and weaknesses need to be highlighted separately because they can justify looking for specific methodological points of attention.

5.1 Master protocols

'**Master protocol**' refers to the use of an overarching logistic, design protocol allowing the investigation of multiple hypotheses or interventions in one or multiple diseases (49,50). The master protocol proposes a common infrastructure establishing uniformity and standardisation of procedures in designing and assessing different interventions. Usually, the concept of a master protocol encompasses three subtypes: **platform trials** [also called **multi-arm, multi-stage trials (MAMS)**], **basket trials**, and **umbrella trials** (51).

5.1.1 Platform trials

Platform trials allow, for a particular disease, the comparison, either simultaneously and/or sequentially, of multiple interventions with a common control group (51). Sometimes the different interventions can also be compared with each other. The master protocol defines the overall infrastructure and sets the overarching principles of the design, but specific addendum protocols are created when a new intervention is assessed. Given that the assessment of certain interventions can be stopped or, alternatively, added to the trial, platform trials can be considered mainly as *adaptive trials* (52). The intervention that is used as a control can also evolve over time if the standard of care is updated following the start of the platform trial. Platform trials are compatible with the principles of RCT design and, when used for assessing the effectiveness of medicinal products, they are frequently phase 3 RCTs (i.e., a confirmatory assessment of effectiveness), but they sometimes start as phase 2 trials (i.e., an exploratory assessment of effectiveness, which can be uncontrolled), and the switch from phase 2 to phase 3 is conducted under the same master protocol (i.e., a platform trial with a '*seamless*' design) (50). In that case, the most promising interventions based on the results of the phase 2 trial are retained for the phase 3 trial. Therefore, the follow-up of some patients from a phase 2 trial can be extended to the phase 3 trial (providing they meet the phase 3 eligibility criteria).

Methodologically, the main strength of platform trials is their flexibility. Thus, they can be considered as more 'disease focused' compared with more commonly used traditional RCTs because they can provide a more efficient assessment of multiple interventions in a manner that can be potentially perpetual with the possibility to be adapted to both scientific discoveries provided by the trial and external discoveries (51).

In itself, platform trials are not a specific type of methodological design *per se*. Therefore, platform trials can provide the same certainty of results than more commonly used traditional RCTs providing they are conducted in conformity with the same methodological principles. Nonetheless, because of their flexibility, several specific points of attention must be considered. First, platform trials can sometimes start as phase 2 trials. Thus, it is important that the criteria to select interventions that are going to phase 3 are clearly defined (e.g., the criteria for defining sufficient presumption of effectiveness). Moreover, because patients from phase 2 can participate in phase 3 of the trial, it is necessary that these patients still meet the eligibility criteria for phase 3. Second, because the inclusion of new patients in the control group can occur over long periods, the contemporaneity of the control group in relation to the assessment of some interventions can be brought into question and the relevance of the intervention proposed within the control group must be scrutinised. Third, although blinding of patients and investigators is possible, it requires the use of multiple dummies, which can be difficult to achieve when there are multiple treatments with different pharmaceutical formulations that are assessed simultaneously. Thus, numerous platform trials are conducted in an open manner. Fourth, multiple interim analyses are usually performed as well as multiple comparisons between groups. Thus, there is a risk of an inflated type 1 error rate if not properly managed. Therefore, assessment of the quality of these analyses (interim analyses and multiples groups comparisons) should follow the guidelines proposed in the EUnetHTA Practical Guideline *Applicability of Evidence: Practical Guideline on Multiplicity, Subgroup, Sensitivity and Post-hoc Analyses*. Finally, it is imperative that the rules for adding new interventions into the trials are explicit and justified.

Practical Guideline (Requirement for JCA reporting)

If a platform trial is an RCT then, in general, RoB needs to be assessed according to the principles described in Section 4.2.

If a platform trial is not an RCT then, in general, RoB needs to be assessed according to the principles described in the Section herein corresponding to the design of the study.

Specific points for attention

If the platform trial starts as a phase 2 trial, the quality of the definition of the rules to select interventions that are going to phase 3 must be considered.

If the platform trial starts as a phase 2 trial, do the patients that were retained from phase 2 to phase 3 meet the eligibility criteria for phase 3? (Yes/No)

Design considerations when adding new intervention(s) (criteria, process, or timing) to the trial.

The potential modifications of the intervention of the control group.

The results of interim analyses and multiple comparisons in accordance with the EUnetHTA practical guideline *Applicability of Evidence: Practical Guideline on Multiplicity, Subgroup, Sensitivity and Post-hoc Analyses*.

5.1.2 Basket trials

Basket trials aim to assess a targeted intervention across multiple diseases (50,53). Eligibility of patients is based on a unifying criterion, which is a specific mechanism of action of the treatment of interest with prognostic value (e.g., a specific molecular alteration or a common pathological process). The targeted intervention is supposed to produce a beneficial effect for all patients because it targets a common process. Therefore, basket trials pool patients with diseases that are classified as different in terms of usual nosography (e.g., cancers from different primary organs or different cardiovascular diseases). Basket trials are currently mainly used in oncology for assessing the effectiveness of interventions designed to target specific molecular alterations, but other medical areas can be concerned by the use of such trials (50,53).

The main strength of basket trials is their potential ability to generate evidence of effectiveness regarding interventions targeting a specific mechanism of action with prognostic value, therefore generating evidence for multiple diseases in one trial (53). Nonetheless, the ability of basket trials to provide such certainty of results relies on multiple assumptions and conditions (54).

In itself, a basket trial is not a type of methodological design *per se*. Therefore, the certainty of results provided by such a trial is mainly dependent on its design. Although basket trials can be RCTs, most are currently uncontrolled trials and, therefore, do not provide a higher certainty of results compared with single-arm trials (50). Randomisation and relative effectiveness assessment in the context of basket trials can be difficult because they investigate multiple diseases and, therefore, can require multiple control interventions (50). Second, the hypothesis that the effect of the targeted intervention will be beneficial, on average, for each ‘cohort’ of patients (e.g., the first cohort is patients with breast cancer, the second one is patients with lung cancer, etc.) relies on the assumption of homogeneity of between-cohorts effects (54). This assumption cannot be proven by analysing the data of the conducted basket trial. There is the possibility of performing an interaction statistical hypothesis test between the intervention and the different cohorts (54). However, even if the test does not reject the null hypothesis of homogeneity of effects, it does not experimentally prove homogeneity because the test can be nonsignificant as a result of a lack of power. Thus, the plausibility of this assumption relies mainly on the basis of biological arguments of the mechanisms of actions or on the proximity to other situations in which the hypothesis of homogeneity has been accepted or proven. Third, the specific effect of the targeted intervention in a specific 'cohort' (e.g., patients with breast cancer only) can suffer from a lack of statistical precision because it can be expected that some cohorts will have a low number of patients given that the occurrence of the targeted mechanism of action can be rare. Finally, eligibility criteria often rely on the screening of a specific molecular alteration or biomarker. Inclusion within a basket trial often relies on the results of a companion test and, therefore, the performance of the test (sensitivity, specificity, predictive values, or probability reports, calibration, and discriminatory capacity for biomarkers measured on a continuum) must be known and must be of an acceptable level (54).

Practical Guideline (Requirement for JCA reporting)

If a basket trial is an RCT then, in general, RoB needs to be assessed according to the principles described in Section 4.2.

If a basket trial is not an RCT then, in general, RoB needs to be assessed according to the principles described in the Section corresponding to the design of the study.

Specific points for attention

Rationale for the plausibility of the hypothesis of homogeneity of effects.

If the eligibility of patients within the basket trial relies on the results of a companion test, its performance, availability, and methods used for detection (e.g., on which tumor sample the test is performed).

If an interaction test for homogeneity of effect was performed, how it was performed (statistical method) and an appropriate description of its result.

Results of effectiveness within each 'cohort' of patients with appropriate statistical estimates.

5.1.3 Umbrella trials

Umbrella trials, which are also mostly used in oncology, aim to assess multiple targeted interventions for what is considered a single disease according to usual nosography (50,53). Patients with a single disease are included (e.g., advanced breast cancer) and are stratified into subgroups based on the baseline value of a biomarker or risk factor with a prognostic value. Thus, the single disease is split into multiple subtypes with eligibility for each intervention group defined by the mechanism of action of each treatment. Each intervention group receives a different targeted intervention that is supposed to have a beneficial effect that is better suited for the specific subgroup of patients for which it is proposed.

The main strength of umbrella trials is their ability to propose targeted therapies that have the potential to be better suited for subgroups of patients of a same disease, which can ultimately enhance the development of stratified medicine (51).

As for any other types of master protocol, umbrella trials are not a type of methodological design *per se*. Therefore, the certainty of results provided by an umbrella trial is mainly dependent on its design. Akin to basket trials, although umbrella trials can be RCTs, most are currently uncontrolled trials and, therefore, do not provide a higher certainty of results compared with single-arm trials (50). Nonetheless, randomisation and relative effectiveness assessment can be considered easier to achieve in the context of an umbrella trial compared with a basket trial, because the existing standard of care (or placebo, if there is no established care) for the disease being studied can be used as a common control for all the subgroups (50). As for basket trials, inclusion often relies on the search for a specific molecular alteration or biomarker. Therefore, the performance of the test (sensitivity, specificity, predictive values, or probability reports, calibration, and discriminatory capacity for biomarkers measured on a continuum) must be known and must be of an acceptable level (54).

Practical Guideline (Requirements for JCA reporting)

If an umbrella trial is an RCT then, in general, RoB needs to be assessed according to the principles described in Section 4.2.

If an umbrella trial is not an RCT then, in general, RoB needs to be assessed according to the principles described in the Section corresponding to the design of the study.

Specific points of attention

If the eligibility of patients within the umbrella trial relies on the results of a companion test, its performances, availability, and methods used for detection (e.g., on which tumor sample the test is performed).

5.2 Real-world data and real-world evidence

Real-world data (RWD) is an umbrella term encompassing the use of various types of data that share the common property they have been generated in the context of routine healthcare [e.g., electronic health records, medical claims and billing data, administrative healthcare databases, patient-generated data (including in-home-use settings) and data produced from various sources (such as electronic devices) that can inform on health status] (55–57). Therefore, the term excludes data collected explicitly for experimental intervention research purposes. In relation to the concept of RWD, **real-world evidence** (RWE) is a term defining clinical evidence of a health technology or medical condition derived from the analysis of RWD for a given research question. RWD can be used to generate RWE for different purposes: for example generating hypotheses for testing in future RCTs, assessing trial feasibility, informing prior probability distributions for Bayesian statistical models, identifying patient baseline characteristics or prognostic and predictive factors, describe usage of a health technology in real-world setting, and assessing the effectiveness and/or safety of health technologies (e.g., for new indications of already-used technologies or for documenting long-term follow-up).

Although ‘RWD’ is used to describe data generated in the context of routine healthcare, such data can be used for various purposes in the context of clinical research. Thus, RWD can be coupled with data generated for clinical research purposes. Indeed, a specific source of RWD can be used as a basis for conducting a RCT in which the collection of necessary data can exclusively come from a set of RWD, or as a primary source complemented by data specifically collected for the clinical study (i.e., a secondary source). These types of study are sometimes considered part of what are called ‘pragmatic trials’ (58). When only a subset of newly included patients within the collection of a specific RWD (e.g., a cohort of patients with data collected from electronic health records) are randomised over time, the corresponding RCT can be considered a *TWIC* (35). When the secondary source of data is collected using fully remote pathways (e.g., electronic informed consent, digital assessment tools, or virtual study visits), the corresponding RCT is sometimes called a ‘*contactless trial*’ (59). RWD can also be used as the only or as the primary source of data for any type of other clinical trial (e.g., single-arm trial) or observational studies (e.g., cohort study). Although this is out of the scope of this Guideline, they can be used as sources of data for indirect comparisons (see the EUnetHTA 21 Methodological Guideline *Direct and Indirect Comparisons*), or as additional historical data borrowing for enriching data of a control group in an already existing clinical trial (e.g., when the trial concerns a rare disease).

The use of RWD in generating evidence can be useful in multiple ways. First, their use can enhance the recruitment of patients in clinical trials, especially for rare diseases (60). Second, their use can enhance the level of applicability of evidence (or external validity) and/or the level of statistical precision by facilitating the conduct of clinical studies on large sample of patients with less stringent inclusion criteria compared with a classical clinical trial, by assessing the effectiveness and/or safety of health technologies in ‘real-world’ settings, and by allowing studies with clinical trials with a longer follow-up than usual (58).

Potential weaknesses in using RWD when conducting clinical studies are mainly linked to the fact that a set of RWD was not primarily structured for conducting a clinical study. Thus, data validity, data integrity, and data monitoring are dependent on the quality of already-existing procedures before the conduct of a given clinical study (61). A related issue can be the use of certain variables from databases as proxies of the characteristics they are supposed to measure in a given clinical study, which can lead to measurement bias (62). For example, data about the dispensation of pharmaceutical drugs coming from administrative databases can be used as a proxy for usage even though the two concepts are not equivalent (even if correlated). Second, follow-up of patients included in a clinical study using RWD might not be as standardised as in *de novo* clinical studies (especially if RWD are the only source of data that will be used for analysis), which can result in a greater risk of attrition bias (61). Finally, particular attention to the assessment of endpoints and how those endpoints were adjudicated on (e.g., investigator *versus* central review, differences between sites) as well as timing of assessments is required (63).

To conclude, in itself, RWD does not define a type of clinical study design and RWE can be produced with varying certainty of results for a given research question. Therefore, the certainty of results that is produced, especially the level of internal validity, is mainly determined by the study design of a given clinical study based on the use of RWD. Especially because most clinical studies using RWD are currently not RCTs, controlling for confounding bias is one of the main issues when estimating treatment effectiveness. Indeed, the lack of randomisation requires the proper use of methods to control for confounding bias (see Section 4.2), which rely on assumptions (e.g., the assumption of exhaustivity on confounders and effect modifiers) that are, in part, unverifiable.

Practical Guideline (Requirement for JCA reporting)

RWD is not a design *per se*; thus, the design of a clinical study should be described and classified according to the principles already described in this guideline.

RoB should be assessed according to the principles already described in this guideline.

Specific points of attentions

For a given clinical study, it should be reported if RWD are the sole source of data, or a primary source of data complemented by a secondary source specifically collected for research purposes (and, if so, to which specific design it corresponds).

Given the at least partial use of data that were not initially structured for clinical research, the validity and reliability of RWD for adequately answering a given research question is of particular importance, especially the potential use of proxy variables, the risk of attrition bias, and the adequate measurement of endpoints.

5.3 Registries

Clinical registries are organised systems collecting data on a group of patients defined by a common characteristic or set of characteristics, which can be the occurrence of a particular disease, condition, exposure or use of a particular health technology or health-related service (64,65). After inclusion of a patient into the registry, follow-up data (i.e., outcomes) are collected. Data collected within the registry can then be used to conduct registry-based studies. Given that they are often a collection of observational data from routine healthcare practices, data from registries can be considered as RWD (57), but it could be advocated that some registries are organised systems that are explicitly devoted to research purposes. Nevertheless, registry data can be used in the same way (e.g., as the sole source of data or as a primary source of data) and for as many purposes as RWD. Furthermore, RCTs conducted using, exclusively or in part, data from registries are often called ‘registry-based RCTs (63,66).

The strengths that were outlined for RWD-based clinical studies can be found in registry-based studies (67). A particular point that can sometimes apply is the fact some registries aim toward an exhaustive coverage of a population of interest. This means that they aim to include the entire population of interest of patients presenting the characteristic leading to their inclusion in the registry (e.g., the diagnosis of a particular disease) within the boundaries of a specific geographical area (which can sometimes be at a national level). Therefore, some registry-based studies can have the ability to produce the true parameter value in the population of interest rather than an estimate (provided the population covered by the registry is the same as the target population)..

However, many weaknesses identified for RWD-based clinical studies are also present in registry-based studies, but some of these aforementioned weaknesses can be mitigated depending on the context. Indeed, first, registries are sometimes built around the idea of answering specific research questions. Thus, registries can produce data with a structure that is more adequately suited to answer specific research questions compared with other sources of RWD. Second, data validity, integrity, and monitoring can be primary concerns in well-structured registries (e.g., national-level registries for a particular disease) and, thus, registry-based studies can sometimes profit from data with a higher level of quality regarding these aspects compared with other types of RWD, especially regarding attrition bias. However, registry data should not be automatically assumed as presenting a high level of validity and reliability and procedures for collection and monitoring of data should be scrutinised anyway when assessing the validity of a registry-based study. Finally, the same remark can be made as for RWD: registry data, in themselves, do not define a clinical study design. Therefore, certainty of results that is produced using registry data, especially the level of internal validity, is mainly determined by the design of a given registry-based clinical study (68).

Practical guideline (Requirement for JCA reporting)

Registries are not a design *per se*; thus, the design of a clinical study should be described and classified according to the principles already described in this Guideline.

RoB should be assessed according to the principles already described in this Guideline.

Specific points for attention

For a given clinical study, it should be reported if a registry is the sole source of data, or a primary source of data complemented by a secondary source specifically collected for research purposes (and, if so, to which specific design it corresponds).

6 REFERENCES

1. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008 Apr 24;336(7650):924–6.

2. EUnetHTA. Partial Use of GRADE in EUnetHTA Framework [Internet]. Available from: https://www.eunethta.eu/wp-content/uploads/2021/05/EUnetHTA-GRADE-framework-paper.pdf?x16454

3. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, QUADAS-2 Steering Group. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. J Clin Epidemiol. 2013 Oct;66(10):1093–104.

4. Regulation (EU) 2021/2282 of the European Parliament and of the Council of 15 December 2021 on health technology assessment and amending Directive 2011/24/EU (Text with EEA relevance) [Internet]. OJ L Dec 15, 2021. Available from: http://data.europa.eu/eli/reg/2021/2282/oj/eng

5. Clinical Epidemiology: The Essentials [Internet]. [cited 2022 Mar 18]. Available from: https://www.wolterskluwer.com/en/solutions/ovid/clinical-epidemiology-the-essentials-2532

6. Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet Lond Engl. 2002 Jan 19;359(9302):248–52.

7. OCEBM Levels of Evidence — Centre for Evidence-Based Medicine (CEBM), University of Oxford [Internet]. [cited 2022 Mar 18]. Available from: https://www.cebm.ox.ac.uk/resources/levels-of-evidence/ocebm-levels-of-evidence

8. Atkins D, Eccles M, Flottorp S, Guyatt GH, Henry D, Hill S, et al. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group. BMC Health Serv Res. 2004 Dec 22;4(1):38.

9. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. Lancet Lond Engl. 2017 Jul 22;390(10092):415–23.

10. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019 Aug 28;366:l4898.

11. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016 Oct 12;355:i4919.

12. Murad MH, Katabi A, Benkhadra R, Montori VM. External validity, generalisability, applicability and directness: a brief primer. BMJ Evid-Based Med. 2018 Feb;23(1):17–9.

13. Levels of Evidence - Applicability of evidence for the context of a relative effectiveness assessment Amended JA1 Guideline Final Nov 2015 - EUnetHTA [Internet]. 2015 [cited 2022 May 23]. Available from: https://www.eunethta.eu/levels-of-evidence-applicability-of-evidence-for-the-context-of-a-relative-effectiveness-assessment-amended-ja1-guideline-final-nov-2015/

14. Windeler J. [External validity]. Z Evidenz Fortbild Qual Im Gesundheitswesen. 2008;102(4):253–9.

15. Rothwell PM. External validity of randomised controlled trials: “to whom do the results of this trial apply?” Lancet Lond Engl. 2005 Jan 1;365(9453):82–93.

16. EUnetHTA JA2. Applicability of evidence for the context of a relative effectiveness assessment [Internet]. Available from: https://www.eunethta.eu/wp-content/uploads/2018/01/Levels-of-Evidence-Applicability-of-evidence-for-the-context-of-a-relative-effectiveness-assessment_Amended-JA1-Guideline_Final-Nov-2015.pdf

17. EUnetHTA. Endpoints used in relative effectiveness assessment of pharmaceuticals - Surrogate Endpoints [Internet]. Available from: https://www.eunethta.eu/wp-content/uploads/2018/01/Surrogate-Endpoints.pdf

18. Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. J Clin Epidemiol. 2011 Dec;64(12):1283–93.

19. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J Clin Res Ed. 1986 Mar 15;292(6522):746–50.

20. Li G, Taljaard M, Van den Heuvel ER, Levine MA, Cook DJ, Wells GA, et al. An introduction to multiplicity issues in clinical trials: the what, why, when and how. Int J Epidemiol. 2017 Apr 1;46(2):746–55.

21. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. PLoS Med. 2010 Mar 24;7(3):e1000251.

22. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016 Apr;31(4):337–50.

23. Guyatt GH, Briel M, Glasziou P, Bassler D, Montori VM. Problems of stopping trials early. BMJ. 2012 Jun 15;344:e3863.

24. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ. 1995 Aug 19;311(7003):485.

25. Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials. BMJ. 1998 Feb 28;316(7132):690–3.

26. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol. 2011 Dec;64(12):1311–6.

27. Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, et al. The GRADE Working Group clarifies the construct of certainty of evidence. J Clin Epidemiol. 2017 Jul;87:4–13.

28. General Methods | IQWiG.de [Internet]. IQWIG. [cited 2022 Mar 18]. Available from: https://www.iqwig.de/en/about-us/methods/methods-paper/

29. Hozo I, Djulbegovic B, Parish AJ, Ioannidis JPA. Identification of threshold for large (dramatic) effects that would obviate randomized trials is not possible. J Clin Epidemiol. 2022 Jan 25;145:101–11.

30. Bross IDJ. Pertinency of an extraneous variable. J Chronic Dis. 1967 Jul 1;20(7):487–95.

31. Glossary [Internet]. NICE. NICE; [cited 2022 Feb 11]. Available from: https://www.nice.org.uk/glossary?letter=c

32. Glossary of Common Site Terms - ClinicalTrials.gov [Internet]. [cited 2022 Feb 10]. Available from: https://www.clinicaltrials.gov/ct2/about-studies/glossary

33. E 9 Statistical Principles for Clinical Trials. 2006;37.

34. EU Clinical Trials Register - Update [Internet]. [cited 2022 Feb 11]. Available from: https://www.clinicaltrialsregister.eu/

35. Relton C, Torgerson D, O’Cathain A, Nicholl J. Rethinking pragmatic randomised controlled trials: introducing the “cohort multiple randomised controlled trial” design. BMJ. 2010 Mar 19;340:c1066.

36. Seo HJ, Kim SY, Lee YJ, Jang BH, Park JE, Sheen SS, et al. A newly developed tool for classifying study designs in systematic reviews of interventions and exposures showed substantial reliability and validity. J Clin Epidemiol. 2016 Feb;70:200–5.

37. Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. Lancet Lond Engl. 2002 Jan 5;359(9300):57–61.

38. Collins R, Bowman L, Landray M, Peto R. The Magic of Randomization versus the Myth of Real-World Evidence. N Engl J Med. 2020 Feb 13;382(7):674–8.

39. Mansournia MA, Higgins JPT, Sterne JAC, Hernán MA. Biases in randomized trials: a conversation between trialists and epidemiologists. Epidemiol Camb Mass. 2017 Jan;28(1):54–9.

40. How to spot bias and other potential problems in randomised controlled trials | Journal of Neurology, Neurosurgery & Psychiatry [Internet]. [cited 2022 Feb 11]. Available from: https://jnnp.bmj.com/content/75/2/181

41. EMA. ICH E9 statistical principles for clinical trials [Internet]. European Medicines Agency. 2018 [cited 2022 Sep 27]. Available from: https://www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials

42. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials [Internet]. Available from: https://www.ema.europa.eu/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf

43. Carey TS, Boden SD. A critical guide to case series reports. Spine. 2003 Aug 1;28(15):1631–4.

44. Munn Z, Barker TH, Moola S, Tufanaru C, Stern C, McArthur A, et al. Methodological quality of case series studies: an introduction to the JBI critical appraisal tool. JBI Evid Synth. 2020 Oct;18(10):2127–33.

45. Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J. Methodological index for non-randomized studies (minors): development and validation of a new instrument. ANZ J Surg. 2003 Sep;73(9):712–6.

46. Institute of Health Economics | [Internet]. [cited 2022 Jun 20]. Available from: https://www.ihe.ca/advanced-search/development-of-a-quality-appraisal-tool-for-case-series-studies-using-a-modified-delphi-technique

47. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol. 2016 Apr 15;183(8):758–64.

48. Grimes DA, Schulz KF. Cohort studies: marching towards outcomes. Lancet Lond Engl. 2002 Jan 26;359(9303):341–5.

49. Master Protocols to Study Multiple Therapies, Multiple Diseases, or Both | NEJM [Internet]. [cited 2022 Mar 18]. Available from: https://www.nejm.org/doi/10.1056/NEJMra1510062?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub++0www.ncbi.nlm.nih.gov

50. Park JJH, Siden E, Zoratti MJ, Dron L, Harari O, Singer J, et al. Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols. Trials. 2019 Sep 18;20(1):572.

51. Park JJH, Harari O, Dron L, Lester RT, Thorlund K, Mills EJ. An overview of platform trials with a checklist for clinical readers. J Clin Epidemiol. 2020 Sep 1;125:1–8.

52. Adaptive Designs for Clinical Trials | NEJM [Internet]. [cited 2022 Mar 18]. Available from: https://www.nejm.org/doi/full/10.1056/nejmra1510061

53. An overview of precision oncology basket and umbrella trials for clinicians - Park - 2020 - CA: A Cancer Journal for Clinicians - Wiley Online Library [Internet]. [cited 2022 Mar 18]. Available from: https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.21600

54. Lengliné E, Peron J, Vanier A, Gueyffier F, Kouzan S, Dufour P, et al. Basket clinical trial design for targeted therapies for cancer: a French National Authority for Health statement for health technology assessment. Lancet Oncol. 2021 Oct 1;22(10):e430–4.

55. Arlett P, Kjær J, Broich K, Cooke E. Real-World Evidence in EU Medicines Regulation: Enabling Use and Establishing Value. Clin Pharmacol Ther. 2022 Jan;111(1):21–3.

56. US Food and Drug Administration. Framework for FDA’s Real-World Evidence Program. 2018.

57. Concato J, Stein P, Dal Pan GJ, Ball R, Corrigan-Curay J. Randomized, observational, interventional, and real-world—What’s in a name? Pharmacoepidemiol Drug Saf. 2020 Nov;29(11):1514–7.

58. Zuidgeest MGP, Goetz I, Groenwold RHH, Irving E, van Thiel GJMW, Grobbee DE. Series: Pragmatic trials and real world evidence: Paper 1. Introduction. J Clin Epidemiol. 2017 Aug;88:7–13.

59. Nicol GE, Piccirillo JF, Mulsant BH, Lenze EJ. Action at a Distance: Geriatric Research during a Pandemic. J Am Geriatr Soc. 2020 May;68(5):922–5.

60. Huml RA, Dawson J, Lipworth K, Rojas L, Warren EJ, Manaktala C, et al. Use of Big Data to Aid Patient Recruitment for Clinical Trials Involving Biosimilars and Rare Diseases. Ther Innov Regul Sci. 2020 Jul 1;54(4):870–7.

61. Meinecke AK, Welsing P, Kafatos G, Burke D, Trelle S, Kubin M, et al. Series: Pragmatic trials and real world evidence: Paper 8. Data collection and management. J Clin Epidemiol. 2017 Nov;91:13–22.

62. Welsing PM, Oude Rengerink K, Collier S, Eckert L, van Smeden M, Ciaglia A, et al. Series: Pragmatic trials and real world evidence: Paper 6. Outcome measures in the real world. J Clin Epidemiol. 2017 Oct;90:99–107.

63. Karanatsios B, Prang KH, Verbunt E, Yeung JM, Kelaher M, Gibbs P. Defining key design elements of registry-based randomised controlled trials: a scoping review. Trials. 2020 Dec;21(1):552.

64. European Medicines Agency. Guideline on registry-based studies. 2021.

65. Eunethta. Vision paper on the sustainable availability of the proposed Registry Evaluation and Quality Standards Tool (REQueST). 2019.

66. Lauer MS, D’Agostino RB. The Randomized Registry Trial — The Next Disruptive Technology in Clinical Research? N Engl J Med. 2013 Oct 24;369(17):1579–81.

67. Gliklich RE, Dreyer NA, Leavy MB, editors. Registries for Evaluating Patient Outcomes: A User’s Guide [Internet]. 3rd ed. Rockville (MD): Agency for Healthcare Research and Quality (US); 2014 [cited 2022 Mar 18]. (AHRQ Methods for Effective Health Care). Available from: http://www.ncbi.nlm.nih.gov/books/NBK208616/

68. [A19-43] Development of scientific concepts for the generation of routine practice data and their analysis for the benefit assessment of drugs according to §35a Social Code Book V – rapid report

[Internet]. IQWIG. [cited 2022 Mar 18]. Available from: https://www.iqwig.de/en/projects/a19-43.html