HTA CG | MEMBER STATE COORDINATION GROUP ON HEALTH TECHNOLOGY ASSESSMENT

Guidance on Validity of Clinical Studies

V1.0

04 July 2024

Adopted on 19 September 2024 by the HTA CG pursuant to Article 3(7), point (d), of Regulation (EU) 2021/2282 on Health Technology Assessment

Contents

List of tables

List of figures

List of abbreviations

Abbreviation	Definition
CI	Confidence interval
GRADE	Grading of Recommendations Assessment, Development and Evaluation
HTA	Health Technology Assessment
HTAb	Health Technology Assessment body
HTAR	HTA Regulation (EU) 2021/2282
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 controlled 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 of clinical study results in an objective, reproducible, and transparent way. According to the HTA Regulation (HTAR, [1]), "*the degree of certainty of the relative effects, taking into account the strengths and limitations of the available evidence"* must be described in a Joint Clinical Assessment (JCA). However, a few aspects that determine the overall certainty of clinical study results are context-dependent. Therefore, the assessment of certainty and its contributing domains has to remain descriptive at the European level, also because overall conclusions or recommendations might interfere with the independent decision-making at the national level. As stated in the HTAR, "*The joint clinical assessment report should be factual and should not contain any value judgement, ranking of health outcomes, conclusions on the overall benefit or clinical added value of the assessed health technology, any position on the target population in which the health technology should be used, or any position on the place the health technology should have in the therapeutic, diagnostic or preventive strategy."*

Therefore, methodological systems that entail not only assessment, but also appraisal, can not (or only partially) be applied in European HTA, since appraisal includes value judgement and (pre-)determines a decision on pricing and reimbursement of a health technology. However, valid scientific principles are still required, not only to guide the development of JCAs at the European level, but also to support the understandability and usability of these results for national decision-making.

1.2 Scope/Objective(s) of the Guidance

This guidance 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 which synthesize evidence 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 guidance. Other HTAR guidance provides recommendations and guidance for the classification and assessment of these evidence syntheses [2,3]. Finally, the present guidance 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 [4].

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 (see other HTAR guidance). 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 relevant other HTAR guidance [5]. Additional considerations of the definition of patient-centred outcomes, and the assessment of their validity, reliability, and interpretability are discussed within another relevant HTAR guidance paper [6].

2 General considerations

2.1 Certainty of results

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 an 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 the HTAR [1], 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*".

Requirement for 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]); external validity (i.e., the extent to which study results provide a basis for generalisation to the target population, a concept which encompasses generalisability and applicability); 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 [7-9], 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. The three different dimensions of the certainty of results overlap with domains of different frameworks for the assessment of the quality of evidence. For instance, the GRADE (Grading of Recommendations Assessment, Development and Evaluation) domains of risk of bias, indirectness and imprecision correspond to internal validity, external validity, and statistical precision, respectively [10].

Although HTA usually requires a high target certainty of results (e.g. RCT evidence for new medicinal products), 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, external validity, or statistical precision in detail, if the PICO question can be sufficiently answered on the basis of higher-certainty results. For example, a detailed assessment of non-randomised controlled studies,

short-term data, or studies below a minimum sample size may be unnecessary in the presence of RCTs, long-term data, or large-sample studies. 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.

2.2 Internal validity

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 [8,11]. The classical hierarchy of evidence [12] includes several types of studies, 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 [13,14], but has practical value for a preliminary sorting between higher- and lower-quality evidence and for selecting a suitable RoB assessment tool.

Requirement for reporting

- In a JCA, the study design must be stated explicitly for all relevant studies.

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 in a given study). If the type of evidence requires it, the assessment of RoB needs to be level specific; however, the scope of the present HTAR guidance 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. For example, recall bias affects only retrospective designs. Therefore, different tools have been developed for RoB assessment in different study designs [15,16]. It is essential to use these standard tools.

Requirement for reporting

- Standard study design-specific tools should be used to assess RoB (or internal validity).

2.3 External validity

The terms ‘external validity’, ‘applicability’, ‘transferability’, ‘generalisability’, and ‘(in)directness’ are often used interchangeably. In the context of an HTA report, it is most appropriate to use the term ‘external validity’ [17]. A 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 [18]. In statistical terms, a lack of external validity of clinical evidence threatens the overall certainty of results if, because of relevant effect modification, the effect in the population of interest could be different from the effects in the clinical studies.

Limitations to the external validity 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 intervention, control interventions and/or subsequent interventions differ from the target setting; or (iii) the study outcomes (e.g., surrogate outcomes) do not give information about the outcomes of interest. For the external validity of clinically relevant evidence, effect modification has to be taken into account [18]. Details on effect modification and subgroup analysis can be found in other HTAR guidance documents [5].

Given that lack of external validity as compared with internal validity is usually 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. Many HTA bodies found this approach to be preferable and do not use a specific instrument or checklist to judge the external validity of clinical evidence. The external validity 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 external validity can only be made at the national (or even regional) level by each member state itself. Accordingly, the HTAR mentions that 'external validity' should be assessed in a JCA, but without forestalling any national judgement on applicability. To support national decision-making, specific issues in relation to external validity 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 external validity because of differences in patient population or control intervention) will only be commented on and briefly analysed, but without providing a conclusion on external validity or national applicability.

Issues with regard to surrogate outcomes usually require specific attention in HTA. However, surrogacy is outside the scope of this guidance, and will be addressed in future HTAR guidance.

Requirement for reporting

- Different aspects of external validity (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.

2.4 Statistical precision

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 specified 1-α level of confidence [19,20]. To increase the transparency and understandability of results, HTAR dossiers and JCA reports should contain counts and other types of descriptive statistics.

Statistical hypothesis testing in single studies can be used to draw causal inferences about intervention effects. Statistical testing in a clinical study requires transparent and clear prespecification of hypotheses, adequate handling of eventual multiplicity issues [21], full reporting of results [22], and careful interpretation [23,24] (see also other HTAR guidance). 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 [25]. However, the probabilities 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 [26].

Most comparative studies on interventions examine superiority hypotheses, but, depending on the medical context, non-inferiority and equivalence are also tested [27]. 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%). In a JCA, it should be specified whether statistical test results originated from confirmatory or exploratory hypothesis testing in a study. A p-value neither represents nor predetermines a conclusion of the added value of the assessed technology [28,29]. 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 [30], needs to be judged at the national context.

The certainty of a positive or negative effect will be higher if a very large effect size was found and the accompanying CI and p value safely exclude the possibility of no effect [31-33]. Which effect sizes can be considered very large and which p values can be accepted as sufficiently low is an unresolved scientific question [34,35]. 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 larger than 5 (or smaller than 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) for classifying effect estimates as very

large [31]. 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.

Requirement for 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 [36]. This guidance aims to describe classification and labelling of study design for use in JCA.

Studies are classified into two categories: interventional studies and observational studies (Figure 1). For consistency, synonyms, such as 'clinical trials' or 'experimental studies' for interventional studies or 'non-interventional' or 'non-experimental' for observational studies are not used in this guidance.

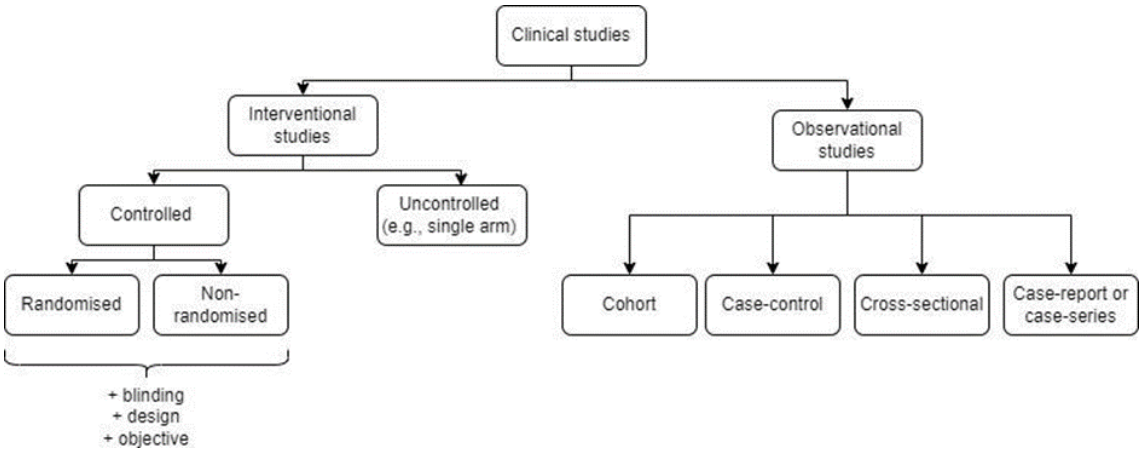

Figure 1: Classification of clinical studies

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

Classification of interventional studies is determined by study characteristics [37]. These characteristics have already been fully defined [38,39], but some terminology inconsistencies prevailed. Therefore, it is valuable to establish definitions of design characteristics for this guidance. Study characteristics are summarised in Table 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 [40], the EU Clinical Trials Register [41], and pertinent dictionaries [36,42]. Note that Table 1 is intended to combine definitions and not to be used as a reporting template.

Table 1: Key study characteristics

Characteristic	Definition
1. Control	
Controlled (or comparative)	A study in which the effect of one or multiple interventions of interest is compared to one or multiple comparator interventions.
Uncontrolled (or single-arm)	A study (observational or interventional) in which all participants receive the same intervention.
2. Randomisation	
Randomised controlled	An interventional study in which participants are randomly assigned to one of the intervention groups.
Non-randomised controlled	An interventional study in which participants receive one of the interventions without randomisation.
3. Blinding	
Blind	A study in which patients, healthcare providers, or outcome assessors are not aware, which study intervention was applied in a given study participant.
Unblinded (or open-label)	A study in which patients, healthcare providers, and outcome assessors are aware of the study intervention.
4. Additional design aspects	
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 intervention after a specified period (therefore, in most cases, each patient receives each intervention).
Factorial	Two or more interventions are evaluated simultaneously through the use of varying combinations of those interventions.
Intra-individual	Two or more interventions are evaluated simultaneously in the same patient ('split body', e.g. right versus left eye).
Cluster-based	A study in which each unit of analysis comprises more than just one patient (e.g. randomization of hospital wards).
5. Objective	
Superiority	Trial with the primary objective of showing that the response to the intervention is clinically superior to that of a comparator.
Non-inferiority	Trial with the primary objective of showing that the response to the intervention is not clinically inferior to that of a comparator. This requires definition of a non-inferiority margin.
Equivalence	Trial with the primary objective of showing that the response to two or more interventions differs by an amount that is clinically negligible. This requires definition of a lower and an upper equivalence margin of clinically acceptable differences.

With regard to blinding, the classical terminology of 'single-blinded' or 'double-blinded' trials should no longer be used because of the ambiguity of these descriptors [22,43-45]. In a JCA, the group of persons who are blinded (participants, care providers, outcome assessors) should

be explicitly specified, e.g. by labelling a trial as being 'patient- and observer-blinded', if treatment staff was not blinded.

The term 'cross-over' is sometimes used incorrectly to describe treatment switching. While cross-over means switching to the alternative intervention after a fixed, protocol-defined period, treatment switching in a broader sense encompasses any switch from the randomized intervention to any other intervention. In a stricter sense, treatment switching only refers to control group patients, who, e.g. after protocol-defined disease progression, receive the experimental intervention. This kind of treatment switching is common in oncology trials and can influence the estimated intervention effects (e.g., on survival). Statistical adjustment can be applied in an effort to reduce this bias, but these methods have their limitations [46-48].

3.1.2 Observational studies

In observational studies, the participants are allowed to follow routine care and are observed (except where the protocol requires visits at specific timepoints). Thus, the decision for or against an exposure[1] is not affected by the 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 [49].

Descriptive or analytical

Observational studies can be either descriptive, that is, without a control group (case-series and cross- sectional studies) or analytical with a control group (case-control and cohort studies) [39,50,51]. 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 intervention 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 [11,52]. Analytical studies, such as cohort and case-control design, can be useful when randomisation is deemed unethical or unfeasible [53].

Prospective and retrospective

The collection of the data in an observational study can be done prospectively or retrospectively. Prospective studies measure exposures before the occurrence of the outcome

[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.

of interest, whereas retrospective studies measure exposure after the occurrence of the outcome of interest [54].

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 [55]. Furthermore, there can be a high risk of recall bias if the determination of exposure status relies on recall or records only [56]. 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 [57].

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 verified.

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 [58]. 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 guidance paper, ‘cohort studies’ are always considered as comparative in that a cohort study follows up two or more groups from exposure to outcome [58].

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 exposures assessed if, at a proper time, they meet the eligibility criteria for the interventional study) [59]. This design is called a ‘trial within a cohort’ (TWIC) [60,61].

Case-control study

Case-control studies are retrospective studies that enrol 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’) [62,63].

The aim of this study design is to compare the exposure between case and controls to identify factors that might be associated with the occurrence of an outcome. Case-control designs are often chosen to address drug safety questions.

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 aim of this study design is to assess outcome and/or exposure prevalence in a population [8,64].

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 intervention of interest [7].

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 [65,66]. 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 [64]. However, they can help to detect new safety signals [66].

Requirement for reporting

- Classify and describe design characteristics for each study submitted as evidence for a JCA.

4 Specific strengths, weaknesses and recommendations regarding different designs

The JCA will report the certainty of the relative effects of the health technology of interest, taking into account the strengths and limitations of the available evidence [Article 9(1)]. As previously described, the certainty of relative effects as measured in a clinical study is determined by internal validity, external validity, and statistical precision.

Study design or conduct can lead to bias [67], impacting internal validity. Several standardised tools have been developed to evaluate RoB in various clinical study designs [68]. They are helpful for assessing the strengths and limitations of the available evidence and should be used when performing a JCA. RoB tools usually consist of several domains, which cover the main types of bias (i.e. bias domains), and a summary rating.

Of course, RoB assessment must address only that part of a study, that has relevance for the PICO question of interest. For example, if two interventions were compared in a cohort study, but only results for one of the treatments are used for HTA purposes (e.g. in an indirect comparison), it is not necessary to assess the validity of the irrelevant comparison.

4.1 Randomised clinical trials: gold standard for intervention 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 [69]. 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 intervention 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 additive causal effect of an intervention of interest over comparator intervention 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 [70-73].

To allow proper evaluation by member states, RoB should be assessed using the Cochrane Collaboration's RoB 1 instrument [74,75]. The RoB 1 tool covers confounding bias[2] (biased allocation to interventions), performance bias, detection bias, attrition bias, selective reporting bias, and other bias. The assessment of bias has to be made separately for different

[2] The term 'selection bias' was chosen by the developers of the RoB 1 instrument, but this type of bias is referred to as 'confounding bias' in the present Guideline.

outcomes, if specific types of bias (mainly due to lack of blinding or incomplete outcome data) differ between these outcomes. The developers of the RoB 1 tool nevertheless stress that assessors should group outcomes in order to limit the number of assessments [74]. For each bias domain, it is necessary to make the assessors' judgment transparent by providing a concise description of the relevant trial characteristic (ideally using a verbatim transcription from the trial report). The category of 'other bias' should be applied only for specific other domains; it should not be used to assess baseline comparability of groups, the trialists' potential conflicts of interests, poor reporting, or small sample size [76]. A summary judgment of RoB should be reached and reported as 'low risk', 'unclear risk' (or 'some concerns'), or 'high risk'. If necessary, this judgment again has to be made separately for each group of outcomes. Study results (or complete studies, if all outcomes have the same RoB) can alternatively be labelled as 'high certainty', 'moderate certainty', and 'low certainty' evidence, even though the overall certainty of study results is also influenced by other aspects, namely statistical precision and external validity. The overall judgment should be based on an explicit logic, whereupon a low RoB overall rating requires that RoB was judged to be low in all key domains.

In 2019, a new revised instrument for RoB assessment was published [16]. The RoB 2 tool was intended to replace RoB 1. It covers the same five domains of bias, but requires a more detailed assessment [77]. Ratings of 'probably yes' and 'probably no' are possible in RoB 2, in order to avoid the all-too-common rating of 'unclear' in RoB 1. In addition, RoB 2 requires that assessors' specify whether they are interested in the effect of assignment rather than adherence to an intervention. Both tools, RoB 1 and RoB 2, do not interfere with the principles of the estimand framework as for example outlined by the European Medicines Agency [40,78]. With regard to its practical use, RoB 2 was found to be challenging due to time requirements [79]. Because the application of the RoB 1 tool is well established and less time-consuming [80], this instrument currently appears preferable over the RoB 2 tool. It is nevertheless advisable that assessors familiarise themselves with RoB 2, so they can pay attention to more subtle mechanisms of bias and describe these in a JCA, if required.

Requirements for reporting

- For outcomes with evidence coming from RCTs, assess RoB using the Cochrane RoB 1 tool. RoB should be assessed for each outcome required in the assessment scope.
- For outcomes sharing the same characteristics (e.g., data collection, blinding aspects, or evaluation) RoB can be assessed on an overall level (i.e., grouped).
- RoB judgement should be provided for both, each individual domain (i.e. type of bias) 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. Non-RCTs 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 thus will differ from its true causal effect [71,81-83].

There are different methods that can be used to control for confounding within the trial, for example design-based methods, such as stratification or matching, or modelling-based methods, such as adjustment or models of causal inference (e.g., propensity scores or g-computation) [84]. 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.

Assessing the RoB of Non-RCTs is highly complex, as both, clinical and methodological expertise are necessary to judge the quality of adjustment methods [85-88]. A large number of RoB instruments were published, but many of these fail to assess internal validity exclusively, i.e. without including any other concepts, such as quality of reporting [89,90]. To allow proper evaluation by member states, RoB should be assessed using ROBINS-I [15,91]. Full guidance documents for ROBINS-I can be found on the Cochrane resource website (https://sites.google.com/site/riskofbiastool/welcome/home/current-version-of-robins-i). As for RoB 1, RoB assessment using ROBINS-I must be performed at the outcome level (if possible, within groups of similar outcomes).

Requirements for reporting

- For outcomes with evidence coming from comparative 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 (i.e. type of bias) 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 intervention 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 intervention of interest is delivered as part of a study intervention [92]. Therefore, patients in a single-arm trial can receive an intervention 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 intervention 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 [93-96], 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 from a single-arm trial can (at least theoretically) be used coupled with an external source of data as a control to allow for a comparative statistical analysis. When this is done, internal and external validity of the resulting indirect comparison, rather than of the uncontrolled study in its' own right, determines certainty of evidence for the assessment of relative effectiveness. In the context of a JCA, the assessment of such external comparisons is explicated in other HTAR guidance papers [2,3]. 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 [97].

Requirement for reporting

- Since uncontrolled trials per se are of very limited value for performing relative effectiveness assessment, RoB assessment is in general not required.
- When data from uncontrolled trials are combined with an external source of data, certainty of evidence for the resulting indirect comparison must be assessed and reported in the JCA.

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 [58]. 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 external validity.

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 [97]. Even with optimal methods, however, residual confounding cannot be ruled out.

Requirements for 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 (i.e. type of bias) 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, individuals are enrolled based on the occurrence of outcome and exposures are investigated in a retrospective manner [63]. Thus, they are at high risk of selection bias [98]. The selection of a control group is very likely to not allow verification of the exchangeability assumption [62]. It leads to the same issues as described before for non-RCTs and cohort studies regarding confounding bias (see Section 4.2) [99]. 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.

Requirements for 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.

Requirement for reporting

- Since 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 [64-66,93,100].

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.

Requirement for reporting

- Since evidence coming from case-series and case-reports is of very limited value for performing relative effectiveness assessment, no RoB assessment using a standardised tool is required for case-series and case-reports.

4.8 Additional design aspects

In addition to the aforementioned aspects, it is useful to examine and describe, whether specific design aspects affect RoB.

If a cross-over design was used, the rationale for this design choice and a justification of the duration of the wash-out phase is essential [101,102]. In this context, the risks of a carry-over effect or a period effect should be described [101,103-105].

If a factorial design was applied, no additional directly bias-related issues arise, but it is important to justify and to examine the assumption of no interaction between the study interventions [106,107].

If an intra-individual comparison (within-person or 'split-body' design) was used (e.g. for the evaluation of a topical drug), the risk of a potential carry-across effect should be paid attention to [108].

In a cluster-based study (e.g. a cluster-randomized trial), selection bias can occur, if participants are recruited after cluster randomization [109-111]. This specific bias due to the differential recruitment across treatment arms needs to be assessed for such studies [112-114]. In a cluster-based stepped-wedge design, secular trends have to be examined [115,116].

Unit of analysis issues should be specifically addressed in all aforementioned study designs, because data dependencies require adequate statistical analyses [109,117,118]. Conventional statistical analyses can lead to overestimated precision of effect estimates.

Requirements for reporting

- Non-standard study designs, such as crossover or factorial designs and intra-individual or cluster-based comparisons require consideration of additional design-specific aspects for RoB assessment.
- It should be verified that data dependencies are adequately addressed within the statistical analyses of non-standard study designs.

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) [119]. 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 protocol allowing the logistically efficient investigation of multiple hypotheses or interventions in one or multiple diseases [120,121]. 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 [122].

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 [122]. 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 [123,124]. 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) [120]. 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 [122].

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 interventions 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 relevant HTAR guidance [5]. Finally, it is imperative that the rules for adding new interventions into the trials are explicit and justified.

Requirements for reporting

- A platform trial is not a design per se; thus, the design of the clinical study should be described and classified according to the principles already described in this guidance. The same applies to RoB assessment.
- 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 should be considered. If the platform trial starts as a phase 2 trial, it is important to record whether the patients who were retained from phase 2 to phase 3 met the eligibility criteria for phase 3.
- If the interventions were modified (e.g. change in dosage or duration), it is important to consider the impact of these modifications on the study results. If interim analyses and multiple comparisons are reported, consult the appropriate HTAR guidance addressing multiplicity issues.

5.1.2 Basket trials

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 [120]. Basket trials aim to assess a targeted intervention across multiple diseases [120,125]. Eligibility of patients is based on a unifying criterion, which is a specific mechanism of action of the intervention 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 [120,125].

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 [125]. Nonetheless, the ability of basket trials to provide such certainty of results relies on multiple assumptions and conditions [126].

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 [120]. 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 [126]. 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 [126]. 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 [126].

Requirements for reporting

- A basket trial is not a design per se; thus, the design of the clinical study should be described and classified according to the principles already described in this guidance. The same applies to RoB assessment.
- Specific attention should be given to the performance of an eventual companion test, the plausibility of the hypothesis of homogeneity of effects, the interaction test for homogeneity of effects, and the study results within each 'cohort' of patients.

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 [120,125]. 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 intervention. 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 [122].

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 [120]. 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 [120]. 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 [126].

Requirements for reporting

- An umbrella trial 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 guidance. The same applies to RoB assessment.
- Specific attention should be given to the performance of an eventual companion test.

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 that 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] [127-130]. 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 studies are sometimes considered part of what are called 'pragmatic trials' [131]. 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 [61]. 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' [132]. 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 guidance, they can be used as sources of data for indirect comparisons (see the relevant HTAR

guidance [2,3]), 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 [133]. Second, their use can enhance the level of external validity and/or the level of statistical precision by facilitating the conduct of clinical studies on large samples 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 [131].

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 [134,135]. 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 [136]. For example, data about the dispensation of pharmaceuticals 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 [135]. Finally, particular attention to the assessment of outcomes and how those outcomes were adjudicated on (e.g., investigator versus central review, differences between sites) as well as timing of assessments is required [137].

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 intervention effectiveness [138]. 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.

Requirements for 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 guidance. The same applies to RoB assessment.

- 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 outcomes should be assessed.

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 [139]. 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' [137,140], and their RoB should be assessed as for any other RCT.

The strengths that were outlined for RWD-based clinical studies can be found in registry-based studies [141]. 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 [142], 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 [142].

Requirements for 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 guidance. The same applies to RoB assessment.
- For a given clinical study, it should be reported if a registry was 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. European Parliament and the Council of the European Union. 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. Official Journal of the European Union 2021; 64: L 458/451.

2. Health Technology Assessment Coordination Group (HTACG). Methodological Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons [online]. 2024 [Accessed: 04.07.2024]. URL: https://health.ec.europa.eu/publications/methodological-guideline-quantitative-evidence-synthesis-direct-and-indirect-comparisons_en.

3. Health Technology Assessment Coordination Group (HTACG). Practical Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons [online]. 2024 [Accessed: 04.07.2024]. URL: https://health.ec.europa.eu/publications/practical-guideline-quantitative-evidence-synthesis-direct-and-indirect-comparisons_en.

4. Whiting PF, Rutjes AW, Westwood ME et al. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. J Clin Epidemiol 2013; 66(10): 1093-1104. https://doi.org/10.1016/j.jclinepi.2013.05.014.

5. Health Technology Assessment Coordination Group (HTACG). Guidance on reporting requirements for multiplicity issues and subgroup, sensitivity and post hoc analyses in joint clinical assessments [online]. 2024 [Accessed: 04.07.2024]. URL: https://health.ec.europa.eu/publications/guidance-reporting-requirements-multiplicity-issues-and-subgroup-sensitivity-and-post-hoc-analyses_en.

6. Health Technology Assessment Coordination Group (HTACG). Guidance on outcomes for joint clinical assessments [online]. 2024 [Accessed: 04.07.2024]. URL: https://health.ec.europa.eu/publications/guidance-outcomes-joint-clinical-assessments_en.

7. Feinstein AR. Clinical epidemiology; the architecture of clinical research. Philadelphia: Saunders; 1985.

8. Fletcher RH, Fletcher SW, Wagner EH. Clinical epidemiology - the essentials (3rd edition). Baltimore: Williams & Wilkins; 1996.

9. Guyatt G, Rennie D, Meade MO, Cook DJ. User's guides to the medical literature - Essentials of evidence-based clinical practice (2nd Edition). New York: McGraw-Hill; 2008.

10. Guyatt GH, Oxman AD, Vist GE et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008; 336(7650): 924-926. https://doi.org/10.1136/bmj.39489.470347.AD.

11. Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet 2002; 359(9302): 248-252. https://doi.org/10.1016/S0140-6736(02)07451-2.

12. Oxford Centre for Evidence-Based Medicine. The Oxford Levels of Evidence 2 [online]. 2011. URL: https://www.cebm.ox.ac.uk/resources/levels-of-evidence/ocebm-levels-of-evidence.

13. Atkins D, Eccles M, Flottorp 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; 4(1): 38. https://doi.org/10.1186/1472-6963-4-38.

14. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. Lancet 2017; 390(10092): 415-423. https://doi.org/10.1016/S0140-6736(16)31592-6.

15. Sterne JA, Hernán MA, Reeves BC et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016; 355: i4919. https://doi.org/10.1136/bmj.i4919.

16. Sterne JAC, Savović J, Page MJ et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019; 366: l4898. https://doi.org/10.1136/bmj.l4898.

17. Murad MH, Katabi A, Benkhadra R, Montori VM. External validity, generalisability, applicability and directness: a brief primer. BMJ Evid-Based Med 2018; 23(1): 17-19.

18. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet 2005; 365(9453): 82-93. https://doi.org/10.1016/S0140-6736(04)17670-8.

19. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed) 1986; 292(6522): 746-750. https://doi.org/10.1136/bmj.292.6522.746.

20. Guyatt GH, Oxman AD, Kunz R et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. J Clin Epidemiol 2011; 64(12): 1283-1293. https://doi.org/10.1016/j.jclinepi.2011.01.012.

21. Li G, Taljaard M, Van den Heuvel ER et al. An introduction to multiplicity issues in clinical trials: the what, why, when and how. Int J Epidemiol 2017; 46(2): 746-755. https://doi.org/10.1093/ije/dyw320.

22. Moher D, Hopewell S, Schulz KF et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010; 340: c869. https://doi.org/10.1136/bmj.c869.

23. Greenland S, Senn SJ, Rothman KJ et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016; 31(4): 337-350. https://doi.org/10.1007/s10654-016-0149-3.

24. Rawlins M. De testimonio: on the evidence for decisions about the use of therapeutic interventions. Lancet 2008; 372(9656): 2152-2161. https://doi.org/10.1016/s0140-6736(08)61930-3.

25. Guyatt GH, Briel M, Glasziou P et al. Problems of stopping trials early. BMJ 2012; 344: e3863. https://doi.org/10.1136/bmj.e3863.

26. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995; 311(7003): 485. https://doi.org/10.1136/bmj.311.7003.485

27. Friedman LM, Furberg CD, DeMets DL. Fundamentals of clinical trials (3rd edition). New York: Springer; 1998.

28. Sedgwick P. Understanding P values. BMJ 2014; 349: g4550. https://doi.org/10.1136/bmj.g4550.

29. van Zwet E, Gelman A, Greenland S et al. A New Look at P Values for Randomized Clinical Trials. NEJM Evid 2024; 3(1): EVIDoa2300003. https://doi.org/10.1056/EVIDoa2300003.

30. Guyatt GH, Juniper EF, Walter SD et al. Interpreting treatment effects in randomised trials. BMJ 1998; 316(7132): 690-693. https://doi.org/10.1136/bmj.316.7132.690.

31. Guyatt GH, Oxman AD, Sultan S et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol 2011; 64(12): 1311-1316. https://doi.org/10.1016/j.jclinepi.2011.06.004.

32. Hultcrantz M, Rind D, Akl EA et al. The GRADE Working Group clarifies the construct of certainty of evidence. J Clin Epidemiol 2017; 87: 4-13. https://doi.org/10.1016/j.jclinepi.2017.05.006.

33. Institute for Quality and Efficiency in Health Care. General Methods; Version 7.0 [online]. 2023 [Accessed: 03.06.2024]. URL: https://www.iqwig.de/methoden/general-methods_version-7-0.pdf.

34. 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; 145: 101-111. https://doi.org/10.1016/j.jclinepi.2022.01.016.

35. Nagendran M, Pereira TV, Kiew G et al. Very large treatment effects in randomised trials as an empirical marker to indicate whether subsequent trials are necessary: meta-epidemiological assessment. BMJ 2016; 355: i5432. https://doi.org/10.1136/bmj.i5432.

36. Day S. Dictionary for clinical trials (2nd edition). Chichester: Wiley; 2007.

37. Seo HJ, Kim SY, Lee YJ 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; 70: 200-205. https://doi.org/10.1016/j.jclinepi.2015.09.013.

38. National Library of Medicine. ClinicalTrials.gov Glossary Terms [online]. 2024 [Accessed: 01.05.2024]. URL: https://www.clinicaltrials.gov/study-basics/glossary.

39. Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. Boston: Little, Brown & Co; 1985.

40. European Medicines Agency. ICH Topic E 9: Statistical Principles for Clinical Trials [online]. 2019 [Accessed: 01.05.2024]. URL: https://www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials-scientific-guideline.

41. European Medicines Agency. Glossary of Terms used in EU Clinical Trials Register [online]. 2010 [Accessed: 01.05.2024]. URL: https://www.clinicaltrialsregister.eu/.

42. Meinert CL. Clinical trials dictionary: Terminology and usage recommendations (2nd edition). Chichester: Wiley; 2012.

43. Devereaux PJ, Manns BJ, Ghali WA et al. Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. JAMA 2001; 285(15): 2000-2003. https://doi.org/10.1001/jama.285.15.2000.

44. Haahr MT, Hróbjartsson A. Who is blinded in randomized clinical trials? A study of 200 trials and a survey of authors. Clin Trials 2006; 3(4): 360-365. https://doi.org/10.1177/1740774506069153.

45. Montori VM, Bhandari M, Devereaux PJ et al. In the dark: the reporting of blinding status in randomized controlled trials. J Clin Epidemiol 2002; 55(8): 787-790. https://doi.org/10.1016/s0895-4356(02)00446-8.

46. Latimer NR, Dewdney A, Campioni M. A cautionary tale: an evaluation of the performance of treatment switching adjustment methods in a real world case study. BMC Med Res Methodol 2024; 24(1): 17. https://doi.org/10.1186/s12874-024-02140-6.

47. Morden JP, Lambert PC, Latimer N et al. Assessing methods for dealing with treatment switching in randomised controlled trials: a simulation study. BMC Med Res Methodol 2011; 11: 4. https://doi.org/10.1186/1471-2288-11-4.

48. Latimer NR, Henshall C, Siebert U, Bell H. Treatment switching: Statistical and decision-making challenges and approaches. Int J Technol Assess Health Care 2016; 32(3): 160-166. https://doi.org/10.1017/s026646231600026x.

49. Soni PD, Hartman HE, Dess RT et al. Comparison of Population-Based Observational Studies With Randomized Trials in Oncology. J Clin Oncol 2019; 37(14): 1209-1216. https://doi.org/10.1200/jco.18.01074.

50. Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. Lancet 2002; 359(9300): 57-61. https://doi.org/10.1016/S0140-6736(02)07283-5.

51. Bonita R, Beaglehole R, Kjellström T, World Health Organization. Basic epidemiology (2nd edition). Geneva: World Health Organization; 2006.

52. Sackett DL, Straus SE, Richardson WS et al. Evidence-based medicine: How to practice and teach EBM. London/UK: Churchill Livingstone; 2000.

53. Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. BMJ 2007; 334(7589): 349-351. https://doi.org/10.1136/bmj.39070.527986.68.

54. Vandenbroucke JP. Prospective or retrospective: what's in a name? BMJ 1991; 302(6771): 249-250. https://doi.org/10.1136/bmj.302.6771.249.

55. Nagurney JT, Brown DF, Sane S et al. The accuracy and completeness of data collected by prospective and retrospective methods. Acad Emerg Med 2005; 12(9): 884-895. https://doi.org/10.1197/j.aem.2005.04.021.

56. Sackett DL. Bias in analytic research. J Chronic Dis 1979; 32(1-2): 51-63. https://doi.org/10.1016/0021-9681(79)90012-2.

57. Feinstein AR, Horwitz RI. Double standards, scientific methods, and epidemiologic research. N Engl J Med 1982; 307(26): 1611-1617. https://doi.org/10.1056/nejm198212233072604.

58. Grimes DA, Schulz KF. Cohort studies: marching towards outcomes. Lancet 2002; 359(9303): 341-345. https://doi.org/10.1016/s0140-6736(02)07500-1.

59. Nickolls BJ, Relton C, Hemkens L et al. Randomised trials conducted using cohorts: a scoping review. BMJ Open 2024; 14(3): e075601. https://doi.org/10.1136/bmjopen-2023-075601.

60. Kessels R, May AM, Koopman M, Roes KCB. The Trial within Cohorts (TwiCs) study design in oncology: experience and methodological reflections. BMC Med Res Methodol 2023; 23(1): 117. https://doi.org/10.1186/s12874-023-01941-5.

61. Relton C, Torgerson D, O'Cathain A, Nicholl J. Rethinking pragmatic randomised controlled trials: introducing the "cohort multiple randomised controlled trial" design. BMJ 2010; 340: c1066. https://doi.org/10.1136/bmj.c1066.

62. Grimes DA, Schulz KF. Compared to what? Finding controls for case-control studies. Lancet 2005; 365(9468): 1429-1433. https://doi.org/10.1016/s0140-6736(05)66379-9.

63. Schulz KF, Grimes DA. Case-control studies: research in reverse. Lancet 2002; 359(9304): 431-434. https://doi.org/10.1016/s0140-6736(02)07605-5.

64. Grimes DA, Schulz KF. Descriptive studies: what they can and cannot do. Lancet 2002; 359(9301): 145-149. https://doi.org/10.1016/s0140-6736(02)07373-7.

65. Albrecht J, Meves A, Bigby M. Case reports and case series from Lancet had significant impact on medical literature. J Clin Epidemiol 2005; 58(12): 1227-1232. https://doi.org/10.1016/j.jclinepi.2005.04.003.

66. Nissen T, Wynn R. The clinical case report: a review of its merits and limitations. BMC Res Notes 2014; 7: 264. https://doi.org/10.1186/1756-0500-7-264.

67. Savović J, Turner RM, Mawdsley D et al. Association Between Risk-of-Bias Assessments and Results of Randomized Trials in Cochrane Reviews: The ROBES Meta-Epidemiologic Study. Am J Epidemiol 2018; 187(5): 1113-1122. https://doi.org/10.1093/aje/kwx344.

68. Viswanathan M, Ansari MT, Berkman ND et al. Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions. Agency for Healthcare Research and Quality Methods Guide for Comparative Effectiveness Reviews [online]. 2012 [Accessed: 01.02.2021]. URL: https://www.ncbi.nlm.nih.gov/books/NBK91433/pdf/Bookshelf_NBK91433.pdf.

69. Collins R, Bowman L, Landray M, Peto R. The Magic of Randomization versus the Myth of Real-World Evidence. N Engl J Med 2020; 382(7): 674-678. https://doi.org/10.1056/NEJMsb1901642.

70. Mansournia MA, Higgins JP, Sterne JA, Hernán MA. Biases in Randomized Trials: A Conversation Between Trialists and Epidemiologists. Epidemiology 2017; 28(1): 54-59. https://doi.org/10.1097/EDE.0000000000000564.

71. Savović J, Jones H, Altman D et al. Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. Health Technol Assess 2012; 16(35): 1-82. https://doi.org/10.3310/hta16350.

72. Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. Lancet 2002; 359(9305): 515-519. https://doi.org/10.1016/s0140-6736(02)07683-3.

73. Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet 2002; 359(9306): 614-618. https://doi.org/10.1016/s0140-6736(02)07750-4.

74. Higgins JP, Altman DG, Gøtzsche PC et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011; 343: d5928. https://doi.org/10.1136/bmj.d5928.

75. Higgins JPT, Savović J, Page MJ et al. Assessing risk of bias in a randomized trial. In: Higgins JPT, Thomas J, Chandler J et al (Ed). Cochrane handbook for systematic reviews of interventions. Hoboken: Wiley-Blackwell; 2019. p. 205-228.

76. Babic A, Pijuk A, Brázdilová L et al. The judgement of biases included in the category "other bias" in Cochrane systematic reviews of interventions: a systematic survey. BMC Med Res Methodol 2019; 19(1): 77. https://doi.org/10.1186/s12874-019-0718-8.

77. Flemyng E, Moore TH, Boutron I et al. Using Risk of Bias 2 to assess results from randomised controlled trials: guidance from Cochrane. BMJ Evid Based Med 2023; 28(4): 260-266. https://doi.org/10.1136/bmjebm-2022-112102.

78. European Medicines Agency. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials [online]. 2020 [Accessed: 01.05.2024]. URL: https://www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials-scientific-guideline.

79. Crocker TF, Lam N, Jordão M et al. Risk-of-bias assessment using Cochrane's revised tool for randomized trials (RoB 2) was useful but challenging and resource-intensive: observations from a systematic review. J Clin Epidemiol 2023; 161: 39-45. https://doi.org/10.1016/j.jclinepi.2023.06.015.

80. Savović J, Weeks L, Sterne JA et al. Evaluation of the Cochrane Collaboration's tool for assessing the risk of bias in randomized trials: focus groups, online survey, proposed recommendations and their implementation. Syst Rev 2014; 3: 37. https://doi.org/10.1186/2046-4053-3-37.

81. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ 1998; 317(7167): 1185-1190. https://doi.org/10.1136/bmj.317.7167.1185.

82. Page MJ, Higgins JP, Clayton G et al. Empirical Evidence of Study Design Biases in Randomized Trials: Systematic Review of Meta-Epidemiological Studies. PLoS One 2016; 11(7): e0159267. https://doi.org/10.1371/journal.pone.0159267.

83. Wood L, Egger M, Gluud LL et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ 2008; 336(7644): 601-605. https://doi.org/10.1136/bmj.39465.451748.AD.

84. D'Agostino RB Jr. Tutorial in Biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998; (17): 2265-2281. https://doi.org/10.1002/(sici)1097-0258(19981015)17:19<2265::aid-sim918>3.0.co;2-b.

85. Losilla JM, Oliveras I, Marin-Garcia JA, Vives J. Three risk of bias tools lead to opposite conclusions in observational research synthesis. J Clin Epidemiol 2018; 101: 61-72. https://doi.org/10.1016/j.jclinepi.2018.05.021.

86. Minozzi S, Cinquini M, Gianola S et al. Risk of bias in nonrandomized studies of interventions showed low inter-rater reliability and challenges in its application. J Clin Epidemiol 2019; 112: 28-35. https://doi.org/10.1016/j.jclinepi.2019.04.001.

87. Jeyaraman MM, Rabbani R, Copstein L et al. Methodologically rigorous risk of bias tools for nonrandomized studies had low reliability and high evaluator burden. J Clin Epidemiol 2020; 128: 140-147. https://doi.org/10.1016/j.jclinepi.2020.09.033.

88. Kalaycioglu I, Rioux B, Briard JN et al. Inter-rater reliability of risk of bias tools for non-randomized studies. Syst Rev 2023; 12(1): 227. https://doi.org/10.1186/s13643-023-02389-w.

89. Luijendijk HJ, Page MJ, Burger H, Koolman X. Assessing risk of bias: a proposal for a unified framework for observational studies and randomized trials. BMC Med Res Methodol 2020; 20(1): 237. https://doi.org/10.1186/s12874-020-01115-7.

90. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol 2007; 36(3): 666-676. https://doi.org/10.1093/ije/dym018.

91. Schünemann HJ, Cuello C, Akl EA et al. GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. J Clin Epidemiol 2019; 111: 105-114. https://doi.org/10.1016/j.jclinepi.2018.01.012.

92. European Medicines Agency. Reflection paper on establishing efficacy based on single-arm trials submitted as pivotal evidence in a marketing authorisation [online]. 2023 [Accessed: 01.05.2024]. URL: https://www.ema.europa.eu/en/establishing-efficacy-based-single-arm-trials-submitted-pivotal-evidence-marketing-authorisation.

93. Carey TS, Boden SD. A critical guide to case series reports. Spine (Phila Pa 1976) 2003; 28(15): 1631-1634. https://doi.org/10.1097/01.BRS.0000083174.84050.E5.

94. Munn Z, Barker TH, Moola S et al. Methodological quality of case series studies: an introduction to the JBI critical appraisal tool. JBI Evid Synth 2020; 18(10): 2127-2133. https://doi.org/10.11124/jbisrir-d-19-00099.

95. Murad MH, Sultan S, Haffar S, Bazerbachi F. Methodological quality and synthesis of case series and case reports. BMJ Evid Based Med 2018; 23(2): 60-63. https://doi.org/10.1136/bmjebm-2017-110853.

96. Slim K, Nini E, Forestier D et al. Methodological index for non-randomized studies (minors): development and validation of a new instrument. ANZ J Surg 2003; 73(9): 712-716. https://doi.org/10.1046/j.1445-2197.2003.02748.x.

97. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol 2016; 183(8): 758-764. https://doi.org/10.1093/aje/kwv254.

98. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004; 15(5): 615-625. https://doi.org/10.1097/01.ede.0000135174.63482.43.

99. Lanza A, Ravaud P, Riveros C, Dechartres A. Comparison of Estimates between Cohort and Case-Control Studies in Meta-Analyses of Therapeutic Interventions: A Meta-Epidemiological Study. PLoS One 2016; 11(5): e0154877. https://doi.org/10.1371/journal.pone.0154877.

100. Dalziel K, Round A, Stein K et al. Do the findings of case series studies vary significantly according to methodological characteristics? Health Technol Assess 2005; 9(2): iii-iv, 1-146. https://doi.org/10.3310/hta9020.

101. Senn S. Cross-over trials in clinical research. Chichester: Wiley; 2002.

102. Brown BW, Jr. The crossover experiment for clinical trials. Biometrics 1980; 36(1): 69-79.

103. Dwan K, Li T, Altman DG, Elbourne D. CONSORT 2010 statement: extension to randomised crossover trials. BMJ 2019; 366: l4378. https://doi.org/10.1136/bmj.l4378.

104. Lathyris DN, Trikalinos TA, Ioannidis JP. Evidence from crossover trials: empirical evaluation and comparison against parallel arm trials. Int J Epidemiol 2007; 36(2): 422-430. https://doi.org/10.1093/ije/dym001.

105. Fleiss JL. A critique of recent research on the two-treatment crossover design. Control Clin Trials 1989; 10(3): 237-243. https://doi.org/10.1016/0197-2456(89)90065-2.

106. Freidlin B, Korn EL. Two-by-Two Factorial Cancer Treatment Trials: Is Sufficient Attention Being Paid to Possible Interactions? J Natl Cancer Inst 2017; 109(9). https://doi.org/10.1093/jnci/djx146.

107. McAlister FA, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: a systematic review. JAMA 2003; 289(19): 2545-2553. https://doi.org/10.1001/jama.289.19.2545.

108. Leducq S, Dugard A, Allemang-Trivalle A et al. Design and Methodological Issues of Within-Person (Split-Body) Randomized Controlled Trials Evaluating a Topical Treatment: A Systematic Review. Dermatology 2023; 239(5): 720-731. https://doi.org/10.1159/000530149.

109. Campbell MK, Piaggio G, Elbourne DR et al. Consort 2010 statement: extension to cluster randomised trials. BMJ 2012; 345: e5661. https://doi.org/10.1136/bmj.e5661.

110. Bolzern J, Mnyama N, Bosanquet K, Torgerson DJ. A review of cluster randomized trials found statistical evidence of selection bias. J Clin Epidemiol 2018; 99: 106-112. https://doi.org/10.1016/j.jclinepi.2018.03.010.

111. Donner A, Klar N. Design and analysis of cluster randomization trials in health research. London: Arnold; 2000.

112. Hemming K, Taljaard M. Key considerations for designing, conducting and analysing a cluster randomized trial. Int J Epidemiol 2023; 52(5): 1648-1658. https://doi.org/10.1093/ije/dyad064.

113. Hemming K, Taljaard M, Moerbeek M, Forbes A. Contamination: How much can an individually randomized trial tolerate? Stat Med 2021; 40(14): 3329-3351. https://doi.org/10.1002/sim.8958.

114. Bolzern JE, Mitchell A, Torgerson DJ. Baseline testing in cluster randomised controlled trials: should this be done? BMC Med Res Methodol 2019; 19(1): 106. https://doi.org/10.1186/s12874-019-0750-8.

115. Hemming K, Taljaard M. Reflection on modern methods: when is a stepped-wedge cluster randomized trial a good study design choice? Int J Epidemiol 2020; 49(3): 1043-1052. https://doi.org/10.1093/ije/dyaa077.

116. Nevins P, Ryan M, Davis-Plourde K et al. Adherence to key recommendations for design and analysis of stepped-wedge cluster randomized trials: A review of trials published 2016-2022. Clin Trials 2024; 21(2): 199-210. https://doi.org/10.1177/17407745231208397.

117. Altman DG, Bland JM. Statistics notes. Units of analysis. BMJ 1997; 314(7098): 1874. https://doi.org/10.1136/bmj.314.7098.1874.

118. Sauerland S, Lefering R, Bayer-Sandow T et al. Fingers, hands or patients? The concept of independent observations. J Hand Surg Br 2003; 28(2): 102-105. https://doi.org/10.1016/s0266-7681(02)00360-1.

119. Lange S, Sauerland S, Lauterberg J, Windeler J. The Range and Scientific Value of Randomized Trials. Dtsch Arztebl Int 2017; 114(38): 635-640. https://doi.org/10.3238/arztebl.2017.0635.

120. Park JJH, Siden E, Zoratti MJ et al. Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols. Trials 2019; 20(1): 572. https://doi.org/10.1186/s13063-019-3664-1.

121. Woodcock J, LaVange LM. Master Protocols to Study Multiple Therapies, Multiple Diseases, or Both. N Engl J Med 2017; 377(1): 62-70. https://doi.org/10.1056/NEJMra1510062.

122. Park JJH, Harari O, Dron L et al. An overview of platform trials with a checklist for clinical readers. J Clin Epidemiol 2020; 125: 1-8. https://doi.org/10.1016/j.jclinepi.2020.04.025.

123. Bhatt DL, Mehta C. Adaptive Designs for Clinical Trials. N Engl J Med 2016; 375(1): 65-74. https://doi.org/10.1056/NEJMra1510061.

124. Griessbach A, Schonenberger CM, Taji Heravi A et al. Characteristics, Progression, and Output of Randomized Platform Trials: A Systematic Review. JAMA Netw Open 2024; 7(3): e243109. https://doi.org/10.1001/jamanetworkopen.2024.3109.

125. Park JJH, Hsu G, Siden EG et al. An overview of precision oncology basket and umbrella trials for clinicians. CA Cancer J Clin 2020; 70(2): 125-137. https://doi.org/10.3322/caac.21600.

126. Lengliné E, Peron J, Vanier A 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; 22(10): e430-e434. https://doi.org/10.1016/S1470-2045(21)00337-5.

127. Arlett P, Kjaer J, Broich K, Cooke E. Real-World Evidence in EU Medicines Regulation: Enabling Use and Establishing Value. Clin Pharmacol Ther 2022; 111(1): 21-23. https://doi.org/10.1002/cpt.2479.

128. Concato J, Stein P, Dal Pan GJ et al. Randomized, observational, interventional, and real-world - What's in a name? Pharmacoepidemiol Drug Saf 2020; 29(11): 1514-1517. https://doi.org/10.1002/pds.5123.

129. U.S. Food and Drug Administration. Framework for FDA’s Real-World Evidence Program [online]. 2018 [Accessed: 01.05.2024]. URL: https://www.fda.gov/media/120060/download.

130. Council for International Organizations of Medical Sciences (CIOMS) Working Group XIII. Real-world data and real-world evidence in regulatory decision making [online]. 2024 [Accessed: 03.06.2024]. URL: https://doi.org/10.56759/kfxh6213.

131. Zuidgeest MGP, Goetz I, Groenwold RHH et al. Series: Pragmatic trials and real world evidence: Paper 1. Introduction. J Clin Epidemiol 2017; 88: 7-13. https://doi.org/10.1016/j.jclinepi.2016.12.023.

132. Nicol GE, Piccirillo JF, Mulsant BH, Lenze EJ. Action at a Distance: Geriatric Research during a Pandemic. J Am Geriatr Soc 2020; 68(5): 922-925. https://doi.org/10.1111/jgs.16443.

133. Huml RA, Dawson J, Lipworth K et al. Use of Big Data to Aid Patient Recruitment for Clinical Trials Involving Biosimilars and Rare Diseases. Ther Innov Regul Sci 2020; 54(4): 870-877. https://doi.org/10.1007/s43441-019-00009-1.

134. van Walraven C, Austin P. Administrative database research has unique characteristics that can risk biased results. J Clin Epidemiol 2012; 65(2): 126-131. https://doi.org/10.1016/j.jclinepi.2011.08.002.

135. Meinecke AK, Welsing P, Kafatos G et al. Series: Pragmatic trials and real world evidence: Paper 8. Data collection and management. J Clin Epidemiol 2017; 91: 13-22. https://doi.org/10.1016/j.jclinepi.2017.07.003.

136. Welsing PM, Oude Rengerink K, Collier S et al. Series: Pragmatic trials and real world evidence: Paper 6. Outcome measures in the real world. J Clin Epidemiol 2017; 90: 99-107. https://doi.org/10.1016/j.jclinepi.2016.12.022.

137. Karanatsios B, Prang KH, Verbunt E et al. Defining key design elements of registry-based randomised controlled trials: a scoping review. Trials 2020; 21(1): 552. https://doi.org/10.1186/s13063-020-04459-z.

138. Nørgaard M, Ehrenstein V, Vandenbroucke JP. Confounding in observational studies based on large health care databases: problems and potential solutions - a primer for the clinician. Clin Epidemiol 2017; 9: 185-193. https://doi.org/10.2147/clep.S129879.

139. European Medicines Agency. Guideline on registry-based studies [online]. 2021 [Accessed: 12.10.2022]. URL: https://www.ema.europa.eu/documents/scientific-guideline/guideline-registry-based-studies_en-0.pdf.

140. Lauer MS, D'Agostino RB, Sr. The randomized registry trial--the next disruptive technology in clinical research? N Engl J Med 2013; 369(17): 1579-1581. https://doi.org/10.1056/NEJMp1310102.

141. Gliklich R, Dreyer N, Leavy M. Registries for evaluating patient outcomes: a user´s guide; 3rd edition; AHRQ pub. no. 13(14)-EHC111 [online]. 2014 [Accessed: 12.10.2022]. URL: https://effectivehealthcare.ahrq.gov/sites/default/files/related_files/registries-guide-3rd-edition-vol-2-140430.pdf.

142. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Concepts for the generation of routine practice data and their analysis for the benefit assessment of drugs according to §35a Social Code Book V (SGB V) [online]. 2020 [Accessed: 11.07.2023]. URL: https://www.iqwig.de/download/a19-43_routine-practice-data-for-the-benefit-assessment-of-drugs_extract-of-rapid-report_v1-0.pdf.