EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

24 October 2025
EMA/301654/2025
Committee for Medicinal Products for Human Use (CHMP)/
Methodology Working Party (MWP)

Guideline on non-inferiority and equivalence comparisons in clinical trials

Draft

Draft agreed by Methodology Working Party	24 October 2025
Adopted by CHMP for release for consultation	3 November 2025
Start of public consultation	13 November 2025
End of consultation (deadline for comments)	31 May 2026

This guideline replaces the 'Guideline on the choice of the non-inferiority margin' (EMEA/CPMP/EWP/2158/99) and 'Points to consider on switching between superiority and non-inferiority' (CPMP/EWP/482/99).

Comments should be provided using this EUSurvey form**.** For any technical issues, please contact the EUSurvey Support.

Keywords	***Non-inferiority, equivalence, biosimilarity***

Official address Domenico Scarlattilaan 6 ● 1083 HS Amsterdam ● The Netherlands
Address for visits and deliveries Refer to www.ema.europa.eu/how-to-find-us
Send us a question Go to www.ema.europa.eu/contact **Telephone** +31 (0)88 781 6000 An agency of the European Union

Guideline non-inferiority and equivalence comparisons in clinical trials

Table of contents

1. Introduction

This guideline lays out general principles for the design and analysis of confirmatory clinical trials that include non-inferiority or equivalence comparisons. The terms 'non-inferiority *comparison'* and 'equivalence *comparison'* are used instead of the terms 'non-inferiority *trial*' and 'equivalence *trial'* to acknowledge that a trial may have different objectives for the same endpoint or for different endpoints.

This guideline replaces the *Guideline on the choice of the non-inferiority margin* from 2005 (EMEA/CPMP/EWP/2158/99) and the *Points to consider on switching between superiority and non-inferiority* from 2000 (CPMP/EWP/482/99). It addresses all the topics that were addressed in the two previous guidelines, with updated recommendations to reflect current EMA positions and the concepts introduced in the estimand framework (ICH E9 R1).

2. Scope

In scope of this guideline are design, conduct and analysis of confirmatory randomised controlled trials that aim to demonstrate:

- efficacy over a putative placebo (absolute efficacy),
- non-inferior efficacy versus an active comparator (relative efficacy),
- non-inferiority of risk profiles,
- biosimilarity in clinical efficacy endpoints (Comparative Clinical Efficacy Studies),
- therapeutic equivalence,
- equivalence in pharmacodynamic properties.

Out of scope of this guideline are bioequivalence, pharmacokinetics and quality assessments.

3. Legal basis

This guideline has to be read in conjunction with the introduction and general principles (4) and the Annex I to 2001/83 as amended.

The following regulatory guidelines refer to the design and conduct of clinical trials also with respect to non-inferiority and equivalence trials and should be read and followed in conjunction with this guideline:

- ICH Note for Guidance E9 (Statistical Principles for Clinical Trials) (CPMP/ICH/363/96).
- ICH E9(R1) addendum on estimands and sensitivity analysis (EMA/CHMP/ICH/436221/2017).
- ICH Note for Guidance E10 (Choice of Control Group) (CPMP/ICH/364/96).
- EMA Guideline on clinical trials in small populations (CHMP/EWP/83561/2005).
- EMA Reflection paper on a tailored clinical approach in biosimilar development (EMA/CHMP/BMWP/60916/2025).

- EMA guidelines on biosimilarity (CHMP/437/04 Rev.1) EMA guideline on missing data in confirmatory clinical trials (EMA/CPMP/EWP/1776/99 Rev. 1).
- EMA Guideline on adjustment for baseline covariates in clinical trials (EMA/CHMP/295050/2013).
- EMA Points to consider on multiplicity issues in clinical trials (CPMP/EWP/908/99).
- EMA Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design (CHMP/EWP/2459/02).
- EMA clinical efficacy and safety guidelines in the relevant disease area.

4. Trial objectives

Non-inferiority and equivalence comparisons can be conducted to address different objectives, which are described below. These objectives are not mutually exclusive, as a single trial can be designed to address more than one objective. Whenever a trial aims to address multiple objectives, multiplicity issues need to be considered, see section 10 for more detail.

Objectives of non-inferiority comparisons

Absolute efficacy

The typical aim of a non-inferiority comparison in the regulatory context is to demonstrate absolute efficacy. This means that an active treatment (the *test treatment*) is compared to another active treatment (the *comparator*) for the purpose of showing that the test treatment is superior to no treatment or placebo, without necessarily being as effective as the active comparator. This situation arises when a trial would ideally be placebo controlled but an active comparator is selected for ethical reasons. The analysis is essentially an indirect comparison between the test treatment and a putative placebo arm.

Relative efficacy

Another objective of a non-inferiority comparison that is sometimes relevant in the regulatory context is to show relative efficacy. This means that a test treatment is compared to an active comparator for the purpose of showing that the test treatment is not worse by more than a clinically acceptable amount. This is commonly expected in situations where a defined amount of the efficacy of the comparator needs to be retained to avoid serious, long-term, or irreversible harm.

From a scientific perspective, the best way to study the relative efficacy of a test treatment and its relative efficacy is to have a three-arm trial in which participants are randomised to the test treatment, an active comparator, or to placebo. This design makes it possible to show absolute efficacy by directly comparing the test treatment to placebo, after which relative efficacy can be shown by comparing the test treatment to the active comparator.

Although the three-arm design is ideal from a scientific perspective, it is often difficult to justify ethically because some patients will receive placebo even though the active comparator is known to be effective. A situation in which a three-arm design might be justified is when the disease is transient or slowly progressive and when giving placebo for a limited period does not lead to a significant loss of

chance in the form of long-term or irreversible harm. The placebo-treated patients can be switched to an active treatment later in the trial. A three-arm design can also be suitable or even necessary when the efficacy of the active comparator is heterogeneous, questioning whether the trial can have sufficient assay sensitivity and making it difficult to justify any non-inferiority margin.

Non-inferior safety

Another objective of a non-inferiority comparison is to show non-inferior safety. Non-inferior safety means that a test treatment is compared to an active comparator, to placebo, or to no treatment for the purpose of showing that the test treatment does not lead to an unacceptably large increase in the risk of an adverse drug reaction.

Objectives of equivalence comparisons

Equivalence comparisons aim to show that two active treatments are similar enough to be considered equivalent. This guideline focuses on equivalence in terms of pharmacodynamics or clinical efficacy, which are typically studied in biosimilarity trials.

Equivalence in terms of clinical efficacy is often referred to as clinical equivalence or therapeutic equivalence.

5. Basic analysis concepts

This section explains how trials with non-inferiority or equivalence comparisons are analysed using the *fixed-margin approach* (see section 8).

To explain how non-inferiority comparisons work, it is useful to start with the better-known case of demonstrating superiority. Figure 1 shows 95% confidence intervals for the difference between the test treatment and the comparator. Superiority is demonstrated with a type-I error rate of 2.5% when the two-sided 95% confidence interval lies above 0 (above 1 for ratio effects). This approach ensures that the probability of falsely concluding superiority is 2.5% (the type-1 error rate). Here, we are assuming that larger values reflect a better result of the test treatment, but the reasoning can easily be adjusted for endpoints where smaller values reflect a better result.

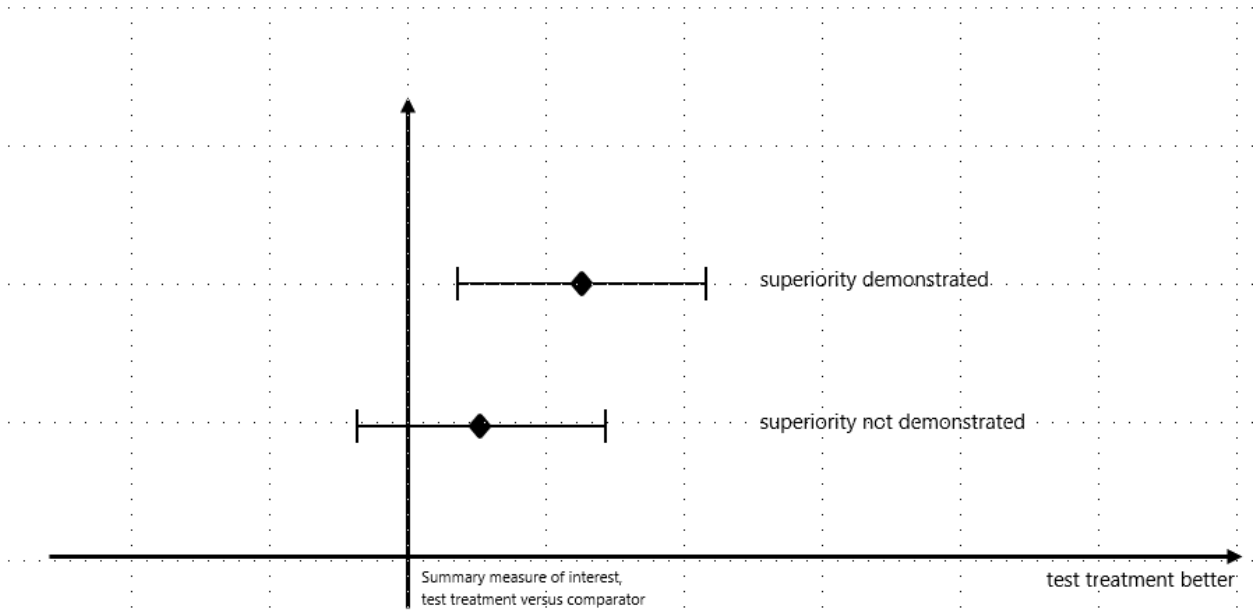

Figure 1: Superiority testing using 95% confidence intervals.

Non-inferiority testing is similar to superiority testing (Figure 2). The difference is that, instead of checking that the two-sided 95% confidence interval lies above 0 (1 for ratios), the confidence interval should lie above a number *less than 0* (less than 1 for ratios) called the non-inferiority margin, which is denoted by -Δ (pronounced 'negative delta' or 'minus delta'). This approach ensures that the probability of falsely concluding non-inferiority is 2.5% (the type-1 error rate).

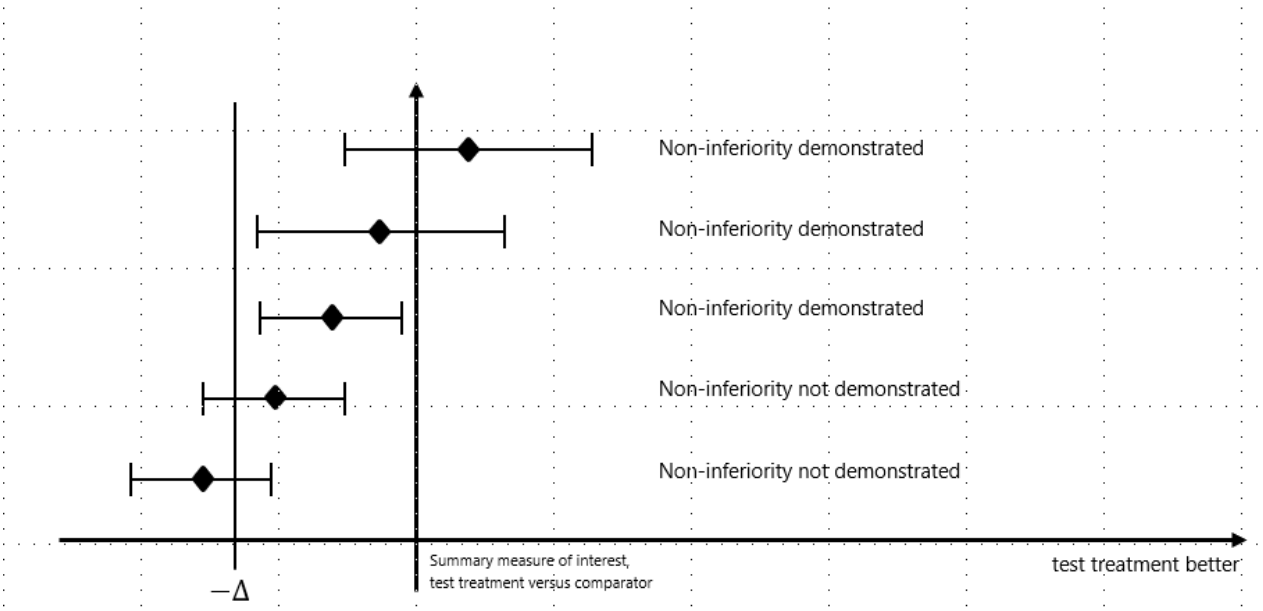

Figure 2: Non-inferiority testing using 95% confidence intervals.

Equivalence comparisons are similar to non-inferiority comparisons (Figure 3). The difference is that the two-sided 95% confidence interval should lie within both a lower bound (-Δ) and an upper bound (+Δ). Both bounds together are called the equivalence margin (±Δ). This approach ensures that the probability of falsely concluding equivalence is 2.5% (the type-1 error rate).

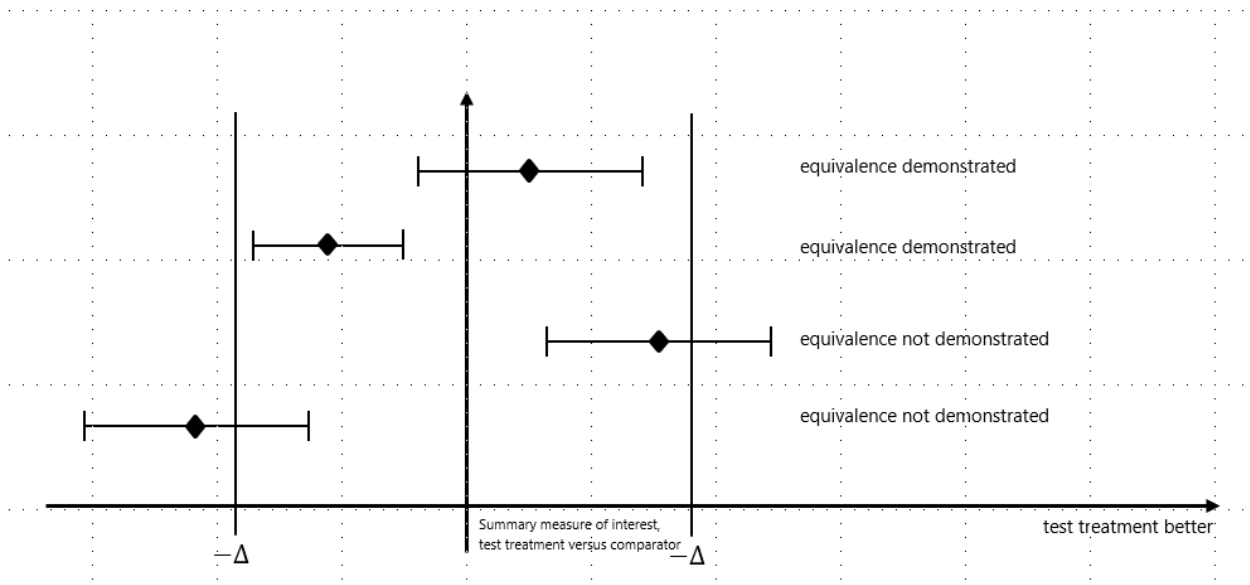

Figure 3: Equivalence testing using 95% confidence intervals.

As a general remark, regardless of the trial objectives, rejecting the null hypothesis based on a pre-specified non-inferiority margin does not automatically lead to a positive decision on the benefit-risk. An overall assessment of the trial results on efficacy and safety, including an evaluation of the methodological strengths and limitations from study planning to trial analysis, will be the basis for the regulatory decision-making.

6. Assay sensitivity and trial quality

In the present context, *assay sensitivity* is the ability of a clinical trial to detect differences between the investigational treatment and reference product / comparator, if they exist. For non-inferiority and equivalence comparisons, assay sensitivity is critical for the internal validity. Lack of assay sensitivity would make the test and reference treatments appear more similar than they really are, which increases the probability of falsely concluding equivalence or non-inferiority. This is a crucial difference to superiority comparisons, where lack of assay sensitivity does not increase the probability of falsely concluding superiority.

Assay sensitivity rests upon an adequate estimand (target population, comparator, endpoint, summary measure and handling of intercurrent events), good trial design, good trial conduct, suitable study population and adequate statistical analysis. Even unsystematic deficiencies can lead to a biased estimate of the difference between the new treatment and the active comparator or a putative placebo, increasing the likelihood of falsely concluding non-inferiority or equivalence. The size of the bias or extent of underestimation of the variability associated with deficiencies is generally unknown and often cannot be quantified with sufficient certainty. Therefore, questionable assay sensitivity may render a non-inferiority or equivalence comparison uninterpretable.

In absence of a placebo arm, the assay sensitivity of a trial is strongly linked to the credibility of the *constancy assumption* that the current effect of the active control is similar to that observed in the past studies used to estimate the active control effect in the same population.

If the assay sensitivity of clinical trials in a certain disease area is questionable a priori, it is not recommended to design a non-inferiority trial without a placebo arm. If randomisation to placebo is not considered feasible in these settings, a superiority trial against an active comparator might be the only feasible option for demonstrating absolute efficacy compared to placebo or relative efficacy compared to the active comparator.

Assessment of assay sensitivity

The assessment of assay sensitivity is critical for the interpretation of a non-inferiority or equivalence comparison, and the trial should be planned to allow an adequate assessment. The assessment of the assay sensitivity should correspondingly be pre-specified in the trial protocol and included in the clinical study report.

In a three-arm trial including placebo, a comparison between the active comparator and placebo allows for a direct assessment of the assay sensitivity. The active comparator would be expected to be superior with a treatment effect compared to placebo of a similar magnitude as compared to previous trials.

The assessment of assay sensitivity in a two-arm trial without a placebo arm is more complicated because it cannot be based on a treatment effect estimate against placebo. In a two-arm trial without a placebo arm, applicants are expected to contextualise the analysis with data from previous trials of the active comparator for justifying that the non-inferiority and equivalence comparison has sufficient assay sensitivity. Applicants are expected to analyse and compare not only the treatment response and its variability in the active arm with previous trials but also compared the distributions of prognostic and predictive factors in the study populations, the frequencies and patterns of protocol deviations, type and number of intercurrent events and extent of missing data. The impact of observed differences on the treatment responses and trial interpretation should be critically assessed.

A difference in the observed performance of the active comparator compared to placebo in the new trial as compared to historical results, or other differences in the aforementioned elements, could indicate a lack of assay sensitivity and question the constancy assumption. This would be of special relevance in borderline cases where the confidence interval of the non-inferiority comparison is close to the margin, or the statistical test is only borderline significant. The pre-defined non-inferiority margin may no longer be appropriate at all if the comparator performs differently from what was assumed when defining the non-inferiority margin.

7. Estimands

The primary estimand(s) of a clinical trial describes the primary scientific question(s) of interest that the clinical trial aims to answer, using the framework developed in ICH E9(R1). The objective(s) of the non-inferiority or equivalence comparison(s) (see section 4) and the relevant regulatory question(s) should inform the choice of estimand. In contrast to superiority trials, in the context of non-inferiority and equivalence comparisons, the assay sensitivity of the experiment must be an additional consideration for the choice and justification of attributes of the primary and secondary / supplementary estimands' attributes.

In most situations, a single primary estimand is expected to be used. However, usually the primary estimand alone cannot address all regulatory and clinical questions of interest for a given endpoint. Supplementary estimands will be needed to address additional key questions, and justification and pre-specification are needed on their impact on the interpretation of the clinical trial and on the conclusion whether the trial objective was met. Particularly, depending on the clinical setting and regulatory requirements, each objective (e.g., showing absolute efficacy against placebo or relative efficacy against the active comparator) might require a dedicated estimand.

Generally, the available scientific guidelines on designing and analysing clinical trials in specific disease areas should be followed when considerations on the preferred estimand for demonstrating superiority or non-inferiority are included. It should be noted that some guidelines only include a discussion on the preferred estimand for a superiority comparison, which should not be understood as the default preference of the same estimand for a non-inferiority setting.

The following sections outline specific considerations for the different estimand attributes, depending on the type of non-inferiority or equivalence comparisons and the underlying objective(s) (see section 4).

7.1. Population

One of the estimand attributes, as defined in ICH E9 (R1), is the population of patients targeted by the clinical question. The appropriate target population of a non-inferiority or equivalence comparison depends on the underlying objective. It is important to distinguish the target population from the study population, comprising the included patients in the trial (operationalised with the in- and exclusion criteria, see section above on assay sensitivity), and the analysis data set (see section 9 on statistical considerations). For the constancy assumption to hold, the target population as well as the participant sample (the study population as included in the trial) need to be considered.

For demonstrating absolute efficacy of a new treatment by means of a non-inferiority comparison, the constancy assumption is critical. Correspondingly, the intended indication and the target population of the pivotal trials of the active comparator against placebo should guide the choice of population. Additionally, the planned study population of the pivotal trials of the active comparator need to be considered in the light of the constancy assumption. If the target populations or (planned) study populations differ between the pivotal trials of the active comparator and the planned non-inferiority comparison, a justification is required, and an assessment is needed of the impact on the constancy assumption. In those cases where violations of the constancy assumption are deemed acceptable at the study planning stage, the non-inferiority margin should be more stringent to compensate for expected violations of the constancy assumptions. Correspondingly, the study populations should be compared after the trial conduct to assess possible violations of the constancy assumptions and their expected impact on the conclusions (see section on Assay sensitivity above).

For demonstrating relative efficacy of the new treatment against the active comparator, a population relevant for the intended conclusion should be chosen, which may differ from the population of the pivotal trials of the active comparator. However, in any case the efficacy of the active comparator in

the selected population should have been robustly demonstrated and the magnitude of effect be known with sufficient precision.

For demonstrating therapeutic equivalence, in principle a more sensitive trial population (as compared to the intended indication) for detecting any differences in treatment effect can be chosen, i.e. a more homogeneous trial population with less variability leading to a better signal to noise ratio. When a different target population for the equivalence trial is proposed, the following points need to be considered:

- The target population should only be restricted for potential prognostic factors of the outcome that are not predictive for the treatment effect (test treatment versus comparator) to ensure that these results can be extrapolated from the subpopulation to the target population of interest. Since many prognostic factors can also be predictive, the trial sponsor needs to justify at the study planning stage and provide sufficient evidence that the variables used for restricting the population are not predictive.

As an exception to the above, choosing a sub-population with a known higher treatment effect size of the active comparator (against placebo) for the study of interest and using the margin derived from the effect of the active comparator in the (broader) target population would usually be acceptable.

- The clinical justification of the margin plays a more important role than the statistical justification for equivalence comparisons and usually leads to a stricter margin (see the section on selecting a margin). Nevertheless, it needs to be ensured that the effect size of historical trials of the active comparator in the historical target population has still been adequately considered in the margin derivation.

7.2. Treatment

For all types of efficacy non-inferiority and equivalence comparisons, the efficacy of the active comparator should be established in the sought indication. Therefore, (i) sufficiently robust evidence demonstrating that it is an effective treatment in the intended setting, and (ii) a sufficiently precise quantification of the treatment effect should be available. If there are no data from randomised controlled trials for the active comparator available and the estimation of the effect of the comparator against placebo in the intended patient population is not possible, demonstrating absolute efficacy at the expected level of certainty is more challenging and would require a thorough justification.

Comparators with older pivotal trials carry the risk that the underlying standard of care and other factors have changed over time, questioning the comparability of the pivotal trials and thereby the constancy assumption.

For demonstrating relative efficacy or non-inferiority of risk profiles, the most appropriate treatment options (usually the best available treatment option based on current scientific evidence) is of highest interest as a comparator.

7.3. Variable

The choice of an appropriate endpoint for demonstrating the trial objectives (see section 4) depends on the clinical setting and the treatment purpose (curative, preventive, maintenance, palliative).

For demonstrating absolute efficacy, the most relevant clinical endpoint for demonstrating efficacy should be used and scientific guidelines from the specific disease areas should be followed. Usually, it is expected that this endpoint was also used in the pivotal trial(s) of the active comparator. In case the chosen primary endpoint has changed over time (e.g., evolution of measurement scale, change in measurement technique, different endpoint altogether or different measurement timepoints between the pivotal trial of the active comparator and the new trial), it is important that sufficient evidence on the superiority of the active comparator against placebo and the magnitude of the treatment effect for the chosen endpoint is available (see also the section on selecting a margin).

For demonstrating relative efficacy or therapeutic equivalence, a sensitive endpoint for showing any difference between treatment arms is required, which is not necessarily the most clinically relevant endpoint or the primary endpoint in the pivotal trials of the active comparator. In general, quantitative measures are seen as more sensitive than qualitative measures and continuous endpoints should not be dichotomised to form a responder endpoint.

7.4. Population-level summary

The summary measure must correspond to the non-inferiority margin and vice versa.

A potential issue arises when historical summary data for the comparator are available only for a different summary measure than what is considered relevant for the definition of the non-inferiority margin, without individual patient data. See section 8 for more detail on the derivation of the non-inferiority margin.

7.5. Strategies for addressing intercurrent events

The frequency and pattern of relevant intercurrent events can be informative about relevant differences between the compared treatments but can also render clinical trials uninterpretable (e.g. unexpected high treatment discontinuation rates). Especially in an equivalence trial, differences in the frequency or pattern of relevant intercurrent events are not expected a priori between the treatment arms and can in themselves indicate a difference between the investigated treatments. For all types of non-inferiority comparisons, a priori expectations about the patterns of the relevant intercurrent events should be formulated in the trial protocol. At the trial analysis stage, frequencies and patterns of the intercurrent events should be compared between treatment arms and their impact on the trial's conclusion should be understood. The comparison of the frequencies and patterns of relevant intercurrent events should be reported for all trials.

There is no general recommendation for the strategies to be used to handle intercurrent events that is valid for every situation. The primary estimand of the trial must be tailored to the specific situation to address the scientific and regulatory questions of interest. The granularity for defining intercurrent events and selecting strategies for handling them must be balanced against the increased complexity of many different intercurrent events that are proposed to be handled with different strategies.

For demonstrating absolute efficacy, the constancy assumption for the effect of the active comparator is important for valid conclusions. At the same time, the primary estimand for demonstrating absolute efficacy should address the most important regulatory question. When fulfilling the constancy assumption and addressing the most relevant regulatory question conflict, it is generally considered more important to use a primary endpoint that addresses the most relevant regulatory question. This might lead to violations of the constancy assumption when the same intercurrent event, e.g., use of an additional medication, is handled differently in the pivotal trial(s) of the comparator and the trial comparing the new treatment to the active comparator. Violating the constancy assumption by choosing a different strategy from the pivotal trial of the active comparator requires an assessment of the impact on the possibility to conclude absolute efficacy for the estimand of interest. Under a range of clinically plausible scenarios, the bias and increase in probability for falsely concluding absolute efficacy need to be assessed. This assessment should (i) inform the decision whether a change in the initially intended estimand is necessary and (ii) be considered in the selection and justification of the margin to compensate for possible bias due to the violation of the constancy assumption. In this situation, supplementary estimands are likely needed.

A similar exercise is expected when the estimand framework was not used in the pivotal trial(s) of the active comparator, and it might not be possible to reconstruct the (implicit) estimand of the pivotal trial(s) with confidence. For demonstrating absolute efficacy, the availability of any additional or alternative medication in the pivotal trial(s) for the active comparator should be considered for the choice of strategy.

For some estimand strategies the true value of the estimand changes when the rate of intercurrent events change (treatment policy and composite). Whenever these strategies are used, it is of particular importance in the process of checking the constancy assumption to compare the rates and patterns of intercurrent events between the historical pivotal trial of the active comparator and the new trial. A higher / lower rate of the intercurrent event in one or both treatment arms of the new trial can bias the comparison in favour of the new treatment.

While general guidance for how to handle all intercurrent events cannot be given, the below considerations are expected to apply widely.

In a non-inferiority or equivalence trial, particular attention should be given to intercurrent events that might make the investigated treatments appear more similar or where participants are expected to do considerably better after the intercurrent event, for example initiating / changing additional medications (e.g., the background medication), or switching from, modifying, or discontinuing the assigned treatment.

The intercurrent event of *switching from the assigned treatment to the active comparator or new investigational product* should not be handled with a treatment policy strategy.

For intercurrent events related to additional or alternative medication intake, it is important to understand whether the main regulatory interest lies in comparing treatment regimens that include the respective additional or alternative medication or comparing treatment regimens excluding them. A treatment policy strategy could be adequate in the former but not in the latter case. As a cautionary remark, an estimand using a treatment policy strategy for additional medications might not detect

important differences between the efficacy of the arms (e.g., because additional medications were used more often in the less efficacious arm compensating for the lower efficacy), which could result in products with relevant inferior efficacy being more likely to (falsely) demonstrate non-inferiority or equivalence. This is especially important when the rates of intercurrent events differ between arms, and such events are more common on the test arm. In these settings, a hypothetical strategy for this class of intercurrent events may be preferable.

8. Selecting a margin

The non-inferiority margin or equivalence range should be selected based upon a combination of statistical reasoning and clinical judgement dependent on the trial objectives (see section 4). The study protocol should clearly describe the objectives of the comparisons and explain for each objective separately the role of the clinical justification or statistical justification in choosing the margin, see following subsections. The estimands for the different study objectives and the estimands used in historic trials of the active comparator should be considered when selecting and justifying the corresponding non-inferiority margin or equivalence range (see Annex *Reviewing existing evidence*).

The non-inferiority margin and equivalence range are design features and correspondingly need to be pre-specified. Post-hoc definitions of, or changes to, the non-inferiority margin are strongly discouraged because it is difficult to convincingly demonstrate that they were not data-driven (see also section 10).

For demonstrating absolute efficacy, the statistical considerations for the chosen approach and justification (see section 8.1) of the margin are critical for ensuring superiority over placebo. For subsequent tests, a clinically justified margin might become relevant for additional conclusions, e.g., relative efficacy compared to the active comparator.

For demonstrating relative efficacy to the active comparator, a clinical justification (see section 8.2) of the margin is needed to define what difference in clinical outcomes versus the active comparator would be considered acceptable. The clinical justification is based on knowledge of disease characteristics and patients' experiences. Equally for relative comparisons in the setting of safety concerns (e.g., cardiovascular outcome trials), the margin should be based on a clinical justification, e.g., an acceptable increase in risk.

For equivalence comparisons an equivalence range is chosen instead of a margin. However, due to the usual requirement of symmetric equivalence margins, this range can be defined by a single margin $\pm\Delta$ on the additive scale of evaluation or upper bound being the inverse of the lower bound on the multiplicative scale of evaluation. The clinical justification of the margins is highly relevant as the aim is to show relative efficacy, and not only absolute efficacy.

The chosen margin for the appropriate outcome measures influences the required sample size and thereby affects study feasibility. The sample size of the trial should be chosen to achieve all the relevant trial objectives.

When available, the recommendations on acceptable margins in the therapeutic area specific guidance of the EMA should be followed.

8.1. Demonstrating absolute efficacy

For demonstrating absolute efficacy, the chosen statistical approach, including the margin or fraction of the effect of the comparator over placebo to preserve, should provide robust evidence that the new treatment is superior to (a putative) placebo. The effect of the active control compared to placebo should typically be estimated based on past randomised controlled trials (see Annex *Reviewing existing evidence*). Preserving a defined fraction of the effect of the active control over placebo ensures superiority of the new treatment compared to placebo. Usually, it is expected that the to be preserved fraction of the treatment effect is clearly larger than 0% and acts as a safeguard against violations of the constancy assumption that might bias the indirect comparison of the new treatment against placebo in favour of the new treatment. Clinical and statistical input will be needed to define the adequate safeguard to compensate for violations of the constancy assumption that might plausibly occur (e.g. slight differences in the study populations of the old and new trial, changes in standard of care treatment, higher/lower rates of important intercurrent events, etc.).

There are different statistical methods for demonstrating absolute efficacy:

1) In *the fixed margin approach,* the margin is pre-specified based on historical studies of the active comparator. Following the 95%-95% method, two 95% confidence intervals are calculated to infer superiority of the test drug over placebo:

    a) The 95% confidence interval from the historical studies for the effect of the comparator relative to placebo, and

    b) the 95% confidence interval from the current trial for the effect of the test drug relative to the comparator.

    To calculate the 95% confidence interval of the effect of the active comparator relative to placebo from point a), results of all available placebo-controlled trials of the active comparator could be pooled by means of an appropriate meta-analytic method (see section on review of existing evidence in the annex for details). This confidence interval is used for defining the non-inferiority margin. To take the uncertainty in the true effect of the comparator against placebo into account, the maximal possible margin for testing absolute efficacy needs to be smaller or equal to the lower limit of this 95% confidence interval. Usually, a fraction of the maximal margin is expected to be used for demonstrating absolute efficacy as a safeguard against lower assay sensitivity as compared to the pivotal trials of the comparator against placebo and violations of the constancy assumption. Here, the fraction is to be calculated as (lower limit of 95% confidence interval from the historical study minus the margin) / lower limit of the 95% confidence interval from the historical study. For example, assuming a 95% confidence interval of [10, 16] for the difference between active comparator and placebo, a non-inferiority margin of 4 would preserve at least (10-4)/10 = 60% of the estimated benefit of the active comparator under ideal circumstances.

    As a second step, the 95% confidence interval from the current trial estimating the effect of the test drug relative to the comparator from point b) is compared to the margin derived in the

previous step. If the new treatment is shown to be non-inferior to the active comparator by that margin, it is concluded that the new treatment is superior to a putative placebo.

The two-step fixed margin method controls the type I error of the non-inferiority comparison, conditional on the pre-specified margin. However, this conditional error rate of the non-inferiority comparison does not quantify the error rate of the indirect inference. Data from historical trials are used for defining the margin and thereby contribute to the indirect comparison of the new treatment to placebo.

2) The *synthesis approach* combines the data from historical trials with the data from the current trial to derive an estimate and confidence interval for the comparison of the new product (T) to a putative placebo (P) in a single step. Like for superiority trials, the null hypothesis $\mathrm{T} - P < 0$ against the alternative $\mathrm{T} - P \geq 0$ can be tested by showing that the lower bound of the confidence interval is greater than 0. The indirectly estimated effect of the new product against placebo can be compared to the effect of the active comparator against placebo (effect of the new product divided by effect of active comparator) to test whether a pre-defined fraction of the effect of the active comparator is preserved for the new product. By design, in the synthesis method, only a to-be preserved fraction can be pre-defined, but no margin on the scale of the endpoint. Therefore, no clinical justification of the margin independent of effect retention is possible.

   As such, the synthesis approach is in essence an evidence synthesis approach using external/historical information with the reference treatment as an anchor for the indirect comparison. Based on the point estimate and the confidence interval for the effect of the new treatment compared to placebo, the effectiveness of the new drug can then be compared to the risks in the benefit risk assessment. The synthesis method is intended to control the (unconditional) type I error associated with the 'meta-analytic' inference integrating the new trial data and the historical data.

Both the fixed margin approach and the synthesis approach use evidence from historical studies for indirect inference about the effect of the new treatment compared to placebo. Therefore, the selection of studies to inform this comparison is a critical element of both approaches and should be carefully planned to avoid (selection) bias in favour of the new treatment, see Annex *Comparison of fixed margin and synthesis approaches* for more details. The historical studies used for the fixed margin approach and synthesis approach should be described in sufficient detail in the study protocol. Adequate operating characteristics of both approaches (lack of bias, type-I error rate control) crucially depend on the constancy assumption. Since (small) violations of the constancy assumption cannot be excluded a priori and are not always easy to detect, a safeguard against these violations by selecting a large enough fraction of the effect of the active comparator against placebo to preserve is usually necessary.

When the same fraction of the effect of the active comparator compared to placebo is used in both the fixed margin approach (translated into a margin on the outcome scale) and the synthesis approach (used directly as a proportion), the standard error of the fixed margin approach is larger compared to the synthesis approach. While the inflated standard error provides some safeguard against violations of

the constancy assumption, the magnitude of the inflation cannot be known at the study planning stage (as it depends on the standard error of the treatment effect estimate in the new trial) and therefore might not be a sufficient compensation for possible violations of the constancy assumption. Therefore, it might still be necessary to use a fraction of the comparator's effect to preserve larger than zero to ensure sufficient robustness of the conclusion on absolute efficacy against violations of the constancy assumption. For the synthesis approach, that does not have an inbuild safeguard against violations of the constancy assumption, it is necessary to preserve a fraction of the comparator effect notably larger than zero. Early interaction with regulators is encouraged when the synthesis approach is planned to be used for demonstrating absolute efficacy. Regardless of the statistical approach chosen, a thorough discussion of plausible violations of the constancy assumption is expected and a justification for the chosen statistical approach, including, if relevant, the margin and the fraction of effect to preserve, should be included in the study protocol.

For both approaches, a confidence interval for the comparison against putative placebo should be reported that is consistent with the chosen statistical approach.

For demonstrating absolute efficacy, the synthesis approach might be acceptable if sufficiently justified.

8.2. Demonstrating relative efficacy or equivalence

In case the relative efficacy of the test treatment compared to the active comparator or showing equivalence/biosimilarity is of interest, an appropriate choice of margin should provide assurance that the test product is not inferior to the comparator by more than a clinically acceptable amount. A clinical justification of the margin is only compatible with a fixed margin approach (see section 8.1). A margin selected based on clinical considerations about excluding a meaningful loss of efficacy should logically also be small enough to demonstrate absolute efficacy. Consequently, the sample size required for a test to demonstrate relative efficacy of the test product to the comparator is expected to be higher compared to when demonstrating absolute efficacy.

The discussion on which difference between the new treatment and the active comparator is clinically relevant should include reference to other trials in the same therapeutic area where such results were seen, and just as importantly, trials where a specific magnitude of difference was concluded to be clinically irrelevant. Consultation with experts in the field or patient representatives are recommended, especially if clinical evidence is sparse in the targeted clinical area. The definition of a non-inferiority margin needs to be distinguished from minimal clinically important difference (MCID; available in some therapeutic areas) or hypotheses for power calculations for superiority trials. Nonetheless, the clinically irrelevant difference can be expected to be smaller than a MCID or the underlying hypothesis for a relevant difference in a superiority trial.

The clinically acceptable amount of inferiority of the new treatment can be established by relating the chosen margin to outcomes that are relevant to patients. However, the interpretation of clinically relevant differences is not always straightforward when the correlation of different outcomes is not high or of unknown magnitude.

The non-inferiority margin should be considerably lower than the effect size of the least effective treatment (which is still generally accepted to be effective).

Furthermore, it is not appropriate to clinically justify the non-inferiority margin as a fraction of the difference between active comparator and placebo without considering what the loss of that proportion of effect means clinically. While this approach provides a safeguard for a statistical justification as a basis for demonstrating absolute efficacy, see section 8.1, it does not provide a clinical justification per se. For example, if the active comparator has a large effect compared to placebo, it does not follow that a large reduction in efficacy would be clinically acceptable.

In situations where a clinically relevant difference or even lack of absolute efficacy cannot be excluded with a feasibly sized trial, it is not good practice to define an arbitrary achievable margin and use that to demonstrate relative/absolute efficacy. Rather, it should be acknowledged that it is not possible to demonstrate relative/absolute efficacy at the expected standard in this case.

For the benefit–risk assessment, a modest loss in efficacy may be acceptable if the product demonstrates a clinically meaningful advantage in another important domain (e.g. improved safety/tolerability or patient convenience). Such aspects may be well defined before a trial, for example having an easier form of administration than the comparator, while other need to be confirmed by data, such as reduced toxicity, or reduced need of supplementary medications.

While these additional aspects can play an important role in characterising the new treatment and comparing it to the active comparator at the stage of benefit/risk assessment after they have been demonstrated in the pivotal trial, at the planning stage of the trial it is not considered acceptable to use any assumed advantages in other domains as arguments for using a larger than usually acceptable non-inferiority margin.

A trial where the results support a sound pre-specified hypothesis increases its evidentiary weight. Therefore, when such additional benefits are planned to be demonstrated, it is advisable to pre-specify them in the study protocol and incorporate them in the testing strategy as appropriate, e.g. by having two dedicated endpoints to show absolute efficacy in the efficacy parameter and superiority to the active comparator for the other important parameter. In case both treatments are assumed to have different advantages, it is advised to plan the trial for demonstrating absolute efficacy followed by demonstrating the advantages of the new treatment over the comparator, instead of increasing the margin for demonstrating relative efficacy.

9. Statistical considerations

The statistical analysis should be aligned with, and sufficiently *conservative* for, the estimand of interest.

Confidence intervals

Confidence intervals comparing the test treatment to the comparator should always be provided. These confidence intervals should typically be two-sided (even in non-inferiority comparisons) and have a two-sided coverage probability of 95%.

The two-sided coverage probability of 95% should be symmetric. This means that the lower confidence limit has a 97.5% probability of being below the parameter of interest and that the upper limit has a 97.5% probability of being above the parameter of interest. This property ensures that the 95% confidence interval corresponds to a two-sided significance test at 5% or a one-sided significance test at 2.5%. Most standard methods for calculating confidence intervals have symmetric coverage probabilities.

The method used to calculate confidence intervals needs to be pre-specified to ensure that the type-1 error is controlled.

When the objective of a non-inferiority comparison is to demonstrate absolute efficacy, having a confidence interval for the indirect comparison of the test treatment to a putative placebo arm can be helpful in the assessment of the clinical relevance of the efficacy of the test treatment and in the benefit-risk evaluation. This confidence interval should be consistent with the chosen statistical approach (see Annex for how the confidence interval for the indirect effect of the new treatment compared to placebo is calculated under the fixed margin and the synthesis approach).

The demonstration of absolute efficacy using a fixed margin approach, relative efficacy, and equivalence should typically be based on the pre-specified non-inferiority or equivalence margin and not on the observed confidence interval of the new treatment compared to the reference treatment. Basing conclusions on the observed confidence interval is a special case of reducing the non-inferiority margin, where repeated tests are conducted with gradually smaller margins until the test is no longer statistically significant (at the edge of the confidence interval). Doing so can inflate the type-1 error if the trial has secondary endpoints which are included in the confirmatory testing scheme. Therefore, if conclusions based on different margins are planned, a suitable statistical testing approach should be prespecified and included in a multiple testing procedure (see also section 10).

Analysis sets

All randomised patients should typically be included in the analysis. Some modifications to this rule can be acceptable, such as excluding patients who never received study treatment if the trial was double blind or who were wrongly included despite not meeting in- or exclusion criteria.

The *default use* of per-protocol sets, which exclude patients with missing data, protocol deviations or intercurrent events, should generally be avoided in primary and secondary analyses.

Missing data

Missing data should be handled in a way that is appropriately *conservative* (under a range of plausible clinical scenarios) and the principles from the EMA guideline on missing data in confirmatory clinical trials should be followed.

In equivalence comparisons, the difference between the test treatment and the comparator should not be biased towards similarity (e.g., a bias towards 1 when a ratio is used as a summary measure, or a bias towards 0 when a difference is used as a summary measure).

In contrast, in non-inferiority comparisons, a bias towards similarity can be conservative in some situations, but not in others. One factor that can affect the likelihood of a statistical method to be

conservative for a non-inferiority comparisons is whether the new treatment performs better or worse than the comparator. For example, when the test treatment is superior to the comparator treatment, a bias towards equality could be conservative in both a non-inferiority comparison and a superiority comparison, as it attenuates the estimated benefit of the test treatment. However, if the test treatment is worse than the comparator (even if by a small amount), a bias towards similarity is not conservative. In general, in a non-inferiority comparison, the effect estimate should not be biased in favour of the test treatment (by overestimating the benefit of the test treatment, by underestimating the benefit of the comparator, or both).

At the study planning stage, it is unknown which of these two scenarios is true. However, the risk of falsely concluding non-inferiority is typically larger in the scenario where the test treatment is worse than the active comparator. Therefore, unless otherwise justified, it is expected that this scenario drives the selection of statistical methods.

The study protocol should include a discussion about the assumption on whether the treatment is assumed to be better or worse than the comparator. Sensitivity analyses should be planned for both scenarios, but the focus should be on sensitivity analyses that address the setting of highest likelihood for falsely concluding non-inferiority. While pre-planning is critical, the final evidentiary weight for the study conclusions of the pre-specified primary analysis and the sensitivity analyses will only be evaluable at the analysis stage when there are fewer uncertainties about which statistical approach can be considered conservative.

Some missing-data methods, such as reference-based imputation, use the same distribution for imputing missing data in both study arms. This leads to an attenuation of the treatment effect if the amount, reason and pattern of missing data are similar in both study arms. Consequently, these methods are rarely conservative in equivalence comparisons, so they should not be used in the primary analysis. In non-inferiority comparisons, these methods will be conservative if the test treatment is superior to the comparator, but not if the test treatment is worse than the comparator. Therefore, these methods should typically not be used in the primary analysis of non-inferiority comparisons.

Other methods assume that the missing data come from a distribution similar to the data seen in patients who prematurely discontinued the study treatment but continued follow up in the same arm (this is also known as using off-treatment information). These methods can be appropriate in both equivalence and non-inferiority comparisons when a treatment policy strategy is being used to handle treatment discontinuation, but care must be taken to ensure that they are conservative if the reasons for discontinuation of treatment or follow-up differ between the arms or cannot be reliably determined. Additionally, care is needed if it is plausible that participants with and without observed outcome differ regarding factors that could influence the outcome. Adequate performance of these types of approaches also requires that a sufficient number of patients are followed up after the intercurrent event, which underlines the importance of good trial conduct.

The same considerations as above are also important when observed data are excluded from the analysis because a hypothetical strategy is used to handle an intercurrent event.

If patients with missing data are handled differently depending on the reason for missingness, a sufficiently clear and reliable categorisation of the reasons is critical and should be pre-specified.

In general, the chosen method of analysis should not incentivise and cannot compensate for poor trial conduct (e.g., high proportions of missing data or higher than expected rates of intercurrent events like treatment discontinuation).

Sample size re-estimation

Both blinded and unblinded sample size re-estimation could substantially inflate the type-1 error in non-inferiority and equivalence comparisons. Therefore, blinded sample size re-estimation should be avoided unless special statistical methods, which reliably control the type-1-error rate, are used.

10. Multiple objectives or changing the objective

Multiple objectives

Clinical trials can be designed to address multiple objectives. For example, the main objective can be to demonstrate absolute efficacy compared to placebo, and a secondary objective can be to demonstrate relative efficacy or superiority compared to the reference treatment in the same or a different endpoint.

Demonstrating superiority to the reference treatment not only provides evidence of greater efficacy, but it can also provide reassurance of the absolute or relative efficacy of the test treatment because, in contrast to non-inferiority comparisons, superiority comparisons within the same trial do not require the constancy assumption or assay sensitivity for the conclusions to be valid.

Having multiple objectives is acceptable when the following aspects have been addressed:

- Objectives should be pre-specified.
- The objectives should be included in a multiple testing procedure if they will be used to make conclusions about efficacy or safety. This is needed even when the same confidence interval is used for multiple tests (for example, when first testing for absolute efficacy compared to placebo and then for superiority compared to the reference treatment on the same primary endpoint).
- The objectives might require different estimands even if the endpoint is the same.
- The suitability of the trial design needs to be evaluated critically for each objective.
- A test of absolute efficacy should be performed before a test of relative efficacy, as having absolute efficacy is a prerequisite for having relative efficacy. Both objectives can, of course, be tested simultaneously if the non-inferiority margin is strict enough to ensure both objectives.

Note that testing for superiority after absolute or relative efficacy have been demonstrated is a special case of reducing the non-inferiority margin to zero (or 1 for a ratio effect). As explained in Section 10, reducing the non-inferiority margin should only be done if it is pre-specified and type-1-error controlled.

Changing the objective

In contrast to having pre-specified multiple objectives, sponsors sometimes want to change the objective of an ongoing or completed trial. This is often an attempt to rescue a superiority trial that has failed or is likely to fail by changing the objective to a non-inferiority comparison for the purpose of demonstrating absolute or relative efficacy.

Changing the objective from a superiority comparison to a non-inferiority comparison after start of a trial is rarely acceptable for several reasons:

- The change can be data-driven (an attempt to rescue a failed superiority trial). This reduces the credibility of the results and undermines the confirmatory nature of the trial.
- Changing the objective can create a multiplicity problem in case there are secondary endpoints tested in the confirmatory testing scheme.
- An appropriate non-inferiority margin will probably not have been prespecified, and post-hoc definitions of the margin can be data driven.
- The comparator may not be appropriate for non-inferiority testing.
- The estimand used in the superiority comparison may not be appropriate for a non-inferiority comparison, and an estimand designed for a non-inferiority comparison will likely not have been pre-specified.
- Protocol deviations (such as non-adherence to treatment) can be a bigger problem when evaluating non-inferiority than superiority and the trial conduct might not have been planned for a non-inferiority objective (see Section 6).
- Assay sensitivity might be questionable, and the constancy assumption may not be fulfilled.
- Demonstrating absolute efficacy via a non-inferiority comparison may not be sufficient for marketing authorisation. If demonstrating absolute efficacy is indeed sufficient, it is recommended to pre-specify both objectives (as explained above).

Definitions

Absolute efficacy: The efficacy of a treatment (on top of other treatments) compared to placebo (on top of other treatments).

Assay sensitivity is the ability of a clinical trial to distinguish an effective treatment from a less effective or ineffective treatment.

Conservative: A characteristic of a statistical approach for estimation and testing under an assumed data generating mechanism. A statistical approach is considered *conservative* for estimating an estimand of interest and for hypothesis testing, if it is not expected to favour the study's working (alternative) hypothesis (e.g., demonstration of non-inferior efficacy to an active control) under a range of circumstances expected to occur in the study because of a bias in the point estimate or because of an inflation of the type-I error rate.

Constancy assumption: the assumption that the effect of the active control in the non-inferiority or equivalence trial is similar to the effect in the past studies used for deriving the margin.

Relative efficacy: The efficacy of a treatment (on top of other treatments) compared to an active comparator treatment (on top of other treatments).

Annex

1. Reviewing existing evidence

A systematic review should be conducted before planning the trial to identify studies relevant to the comparison of the active comparator with placebo in the condition being considered. Selected studies should be documented in the study protocol before initiation of the study and can be used for estimating the difference between the comparator and placebo in the intended patient population.

There are several issues regarding the literature search and the historical studies considered relevant for justification of the treatment difference between active comparator and placebo that will need to be discussed by the applicant:

- Selection bias. The criteria used for selecting which of the available studies to include should be thoroughly documented so that, as far as is possible, an unbiased selection of studies was made, including a discussion of a potential publication bias.
- Constancy assumption: Consideration should be given to potential differences between the current trial in comparison to the previous trials regarding changes that may affect treatment outcome. Some of the studies may be of little relevance because clinical practice may have changed, or the criteria or methods for measuring the comparator's effect have changed, e.g., inclusion criteria, method of diagnosis, concomitant treatments allowed, dosing regimen of comparator, endpoints measured, timing of assessments, etc. This may also include changes in the treatment difference seen over time, as the event rates in some conditions may have decreased over time because of general improvements in healthcare. In the latter, it might be appropriate to include only the more recent studies in the estimations. The durability of historical data concerning the estimand (population, handling of intercurrent events, endpoint, summary measure (scale), standard of care, trial setting, etc) must be judged on a case-by-case basis.

The constancy assumption emphasizes the importance that the design/estimand of the current trial matches the previous trials on which the statistical assumptions are based on (incl. margin justification), as changes regarding the design may affect treatment outcome. In case of a difference between the estimand of interest for the non-inferiority comparison and the historical estimands, the margin derived from the statistical justification needs to be sufficiently small to guarantee superiority to putative placebo (e.g., by retaining a larger proportion of the treatment effect, see section 8.2).

Too strict criteria for the selection of studies limit the number of studies and a balance is needed between estimation precision/sample size and potential for bias. As certain design aspects very often differ due to between-study heterogeneity, including/excluding certain studies might be at the cost of introducing potential bias. Like for meta-analyses in general, the following factors must be considered to evaluate whether the studies are appropriate to be included in the review and whether a common

treatment effect estimate would still be meaningful: are the study populations in the different studies sufficiently representative of the population of the non-inferiority/equivalence trial; to what extent can dosage, treatment duration, sample size and other study conditions differ among studies. If the active comparator is part of a class where individual products can be assumed to be equally effective and safe, it might be acceptable to use the overall class difference from placebo/no treatment. It is assumed that adequate methods are used for the estimation of the overall effect as well as for the evaluation of heterogeneity. If publication bias is considered possible or constancy assumption seems questionable, this should be discussed and adequately compensated for either when producing the historical confidence interval for the effect of the active comparator versus placebo or when selecting the non-inferiority margin.

2. Comparison of fixed margin and synthesis approaches

For demonstrating absolute efficacy by means of a non-inferiority comparison, the non-inferiority comparison synthesises information from the current trial and the historical evidence on the active comparator's efficacy compared to placebo. Two statistical approaches are discussed in this guideline for demonstrating absolute efficacy: the synthesis approach and the fixed-margin approach.

For outlining and comparing the fixed margin and the synthesis approaches step by step, consider the following situation. A new treatment T is compared to an active comparator C in the new trial. The active comparator C was compared to placebo P in historical trials. Since the effect of the active comparator in the current trial could be different from the performance in the historical placebo-controlled trials, we denote the true means of a relevant outcome measure of the active comparator in the current trial as $\mu_{C_{new}}$ and in the historical evidence in placebo-controlled trial as $\mu_{C_{hist}}$ and correspondingly the true effects of the active comparator over placebo as $\mu_{C_{new}} - \mu_P$ and $\mu_{C_{hist}} - \mu_P$. In the new trial, the true effect of the new treatment compared to the active comparator is denoted as $\mu_T - \mu_{C_{new}}$. Throughout this example, we assume that positive values in the outcome correspond to an improvement in disease state, and that negative values correspond to a worsening of the disease state. Further, we will use the notation $\overline{\mathrm{T}}$ and $SE(\overline{\mathrm{T}})$ to notate the estimator and its standard error for estimating μ_T and $\bar{t}$ and $SE(\bar{t})$ for the observed estimate.

Synthesis approach

As the name suggests, the results from the new trial and the historical evidence is synthesised to estimate the effect $\mu_{\mathrm{T}} - \mu_{\mathrm{P}}$. The constancy assumption corresponds to the assumption that the active comparator would have the same effect over placebo in the new trial as it had in the historical trial, $\mu_{C_{new}} - \mu_P = \mu_{C_{hist}} - \mu_P$. If the constancy assumption holds, the effect of the new treatment compared to placebo can be written as $\mu_{\mathrm{T}} - \mu_{\mathrm{P}} = (\mu_{\mathrm{T}} - \mu_{\mathrm{C_{new}}} + \mu_{C_{hist}} - \mu_P)$. The null hypothesis $\mu_{\mathrm{T}} - \mu_{\mathrm{P}} < 0$ can be tested against the alternative $\mu_{\mathrm{T}} - \mu_{\mathrm{P}} \geq 0$ by deriving the test statistic or confidence interval assuming a normal distribution for the difference in sample means. The effect $\mu_{\mathrm{T}} - \mu_{\mathrm{P}}$ is estimated via $\overline{\mathrm{T}} - \overline{\mathrm{P}} = \overline{\mathrm{T}} - \overline{\mathrm{C_{new}}} + \overline{\mathrm{C_{hist}}} - \overline{\mathrm{P}}$ and the standard error of the distribution of $\overline{\mathrm{T}} - \overline{\mathrm{P}}$ can be calculated based on standard errors of $\overline{\mathrm{T}} - \overline{\mathrm{C_{new}}}$ and $\overline{\mathrm{C_{hist}}} - \overline{\mathrm{P}}$ as $(\overline{\mathrm{T}} - \overline{\mathrm{C_{new}}})$ and $(\overline{\mathrm{C_{hist}}} - \overline{\mathrm{P}})$are independent random variables:

$$\mathrm{SE}(\overline{\mathrm{T}}-\overline{\mathrm{P}}\ )=\sqrt{\mathrm{Var}(\overline{\mathrm{T}}-\overline{\mathrm{P}}\ )}\ =\ \sqrt{\mathrm{Var}(\overline{\mathrm{T}}-\overline{\mathrm{C}_{\mathrm{new}}}+\ \overline{\mathrm{C}_{\mathrm{hist}}}-\overline{\mathrm{P}}\ )}\ =\ \sqrt{\mathrm{Var}(\overline{\mathrm{T}}-\overline{\mathrm{C}_{\mathrm{new}}})+\ \mathrm{Var}(\overline{\mathrm{C}_{\mathrm{hist}}}-\overline{\mathrm{P}}\ )}$$

$$=\ \sqrt{SE(\overline{\mathrm{T}}-\overline{\mathrm{C}_{\mathrm{new}}})^2+\ \mathrm{SE}(\overline{\mathrm{C}_{\mathrm{hist}}}-\overline{\mathrm{P}})^2}$$

The Null hypothesis would be rejected if the lower bound of the 95% confidence interval for $\mu_{\mathrm{T}}-\mu_{\mathrm{P}}$ is larger than zero:

$$\bar{\mathrm{t}}-\overline{c_{\mathrm{new}}}\ +\ \overline{c_{\mathrm{hist}}}-\bar{p}\ -1.96\,SE(\bar{\mathrm{t}}-\overline{c_{\mathrm{new}}}\ +\ \overline{\mathrm{c}_{\mathrm{hist}}}-\bar{\mathrm{p}}\ )>0$$

Due to the above proven equality for the standard errors, this is equivalent to

$$(1)\qquad \bar{\mathrm{t}}-\overline{c_{\mathrm{new}}}\ +\ \overline{c_{\mathrm{hist}}}-\bar{p}\ -1.96\sqrt{SE(\bar{\mathrm{t}}-\overline{c_{\mathrm{new}}})^2+SE(\overline{\mathrm{c}_{\mathrm{hist}}}-\bar{\mathrm{p}}\ )^2}>0$$

In the synthesis approach, no explicit fixed margin M for the comparison between new treatment and active control can be defined. However, it can be specified that a certain percentage q of the treatment effect of the active comparator over placebo must be retained. This tests the null hypothesis $\mu_{\mathrm{T}}-\mu_{\mathrm{P}}<\ \mathrm{q}\ (\mu_{C_{hist}}-\mu_P)$, which corresponds under the constancy assumption to $\mu_{\mathrm{T}}-\mu_{\mathrm{C_{new}}}+(1-\mathrm{q})(\mu_{C_{hist}}-\mu_P)<\ 0$. Similarly to $\mathrm{q}=0$, the Null hypothesis would be rejected if the lower bound of the 95% confidence interval for $\mu_{\mathrm{T}}-\mu_{\mathrm{C_{new}}}+(1-\mathrm{q})(\mu_{C_{hist}}-\mu_P)$ is larger than zero:

$$\bar{\mathrm{t}}-\overline{c_{\mathrm{new}}}\ +(1-q)(\overline{c_{\mathrm{hist}}}-\bar{p})-1.96\,SE(\bar{\mathrm{t}}-\overline{c_{\mathrm{new}}}\ +(1-q)(\overline{\mathrm{c}_{\mathrm{hist}}}-\bar{\mathrm{p}})\ )>0$$

Similar to choosing M in the fixed margin approach, see below, q could be chosen sufficiently large as a safeguard for violations of the constancy assumption.

Fixed margin approach

The fixed margin approach aims to demonstrate that the mean of the new treatment is not worse than the mean of the active control by a fixed margin M, i.e. $\mu_{\mathrm{T}}-\mu_{\mathrm{C_{new}}}\geq\ -\mathrm{M}$. The margin M is selected based on the historical evidence for the effect $\mu_{C_{hist}}-\mu_P$ and should be no larger than the effect the active control is expected to have in the NI study. Practically, $-M$ is selected at most as large as the lower bound of the 95% confidence interval for the effect estimate for $\mu_{C_{hist}}-\mu_P$, i.e. $M\leq\overline{\mathrm{c}_{\mathrm{hist}}}-\bar{\mathrm{p}}\ -1.96\,SE(\overline{\mathrm{c}_{\mathrm{hist}}}-\bar{\mathrm{p}}\ )$ using a normal approximation and denoting the observed estimate from the trial as $\bar{\mathrm{t}}-\overline{c_{\mathrm{new}}}$. The observed estimate and standard error from the historical evidence are $\overline{\mathrm{c}_{\mathrm{hist}}}-\bar{\mathrm{p}}$ and $SE(\overline{\mathrm{c}_{\mathrm{hist}}}-\bar{\mathrm{p}}\ )$. The Null hypothesis $\mu_{\mathrm{T}}-\mu_{\mathrm{C_{new}}}<\ -\mathrm{M}$ is rejected if the lower bound of the 95% confidence interval for $\mu_{\mathrm{T}}-\mu_{\mathrm{C_{new}}}$ lies above $-\mathrm{M}$

$$\bar{\mathrm{t}}-\overline{c_{\mathrm{new}}}\ -1.96\,SE(\bar{\mathrm{t}}-\overline{c_{\mathrm{new}}})>-M$$

$$\bar{\mathrm{t}}-\overline{c_{\mathrm{new}}}\ -1.96\,SE(\bar{\mathrm{t}}-\overline{c_{\mathrm{new}}})>-\big(\ \overline{\mathrm{c}_{\mathrm{hist}}}-\bar{\mathrm{p}}\ -1.96\,SE(\overline{\mathrm{c}_{\mathrm{hist}}}-\bar{\mathrm{p}}\ )\big)$$

This expression is equivalent to

$$(2)\qquad \bar{\mathrm{t}}-\overline{c_{\mathrm{new}}}\ +\overline{\mathrm{c}_{\mathrm{hist}}}-\bar{\mathrm{p}}\ -1.96\big(\ SE(\bar{\mathrm{t}}-\overline{c_{\mathrm{new}}})\ +SE(\overline{\mathrm{c}_{\mathrm{hist}}}-\bar{\mathrm{p}}\ )\big)\ >0$$

As a safeguard against violations of the constancy assumptions, the margin M can be selected smaller than the complete observed benefit of the active comparator compared to placebo based on the historical evidence. Like for the synthesis approach, M can be selected to ensure that the effect of the new treatment over placebo is at least as large as a pre-defined fraction q of the entire effect of the

active control over placebo. Under these circumstances $M = (1-q)\left(\overline{c_{\text{hist}}} - \bar{p} - 1.96\, SE(\overline{c_{\text{hist}}} - \bar{p})\right)$ and the tested null hypothesis is $\mu_T - \mu_{C_{\text{new}}} < -(1-q)\left(\overline{c_{\text{hist}}} - \bar{p} - 1.96\, SE(\overline{c_{\text{hist}}} - \bar{p})\right)$.

Comparison of the fixed margin and synthesis approach

Comparing equations (1) and (2) shows that the difference between the fixed margin approach and the synthesis approach for testing $\mu_T - \mu_P$ is the standard error used for the inference. The fixed margin approach uses $SE(\bar{t} - \overline{c_{\text{new}}}) + SE(\overline{c_{\text{hist}}} - \bar{p})$ whereas the synthesis approach uses $\sqrt{SE(\bar{t} - \overline{c_{\text{new}}})^2 + SE(\overline{c_{\text{hist}}} - \bar{p})^2}$ as the standard error for deriving the confidence interval. From the triangle inequality it follows that $SE(\bar{t} - \overline{c_{\text{new}}}) + SE(\overline{c_{\text{hist}}} - \bar{p}) \geq \sqrt{SE(\bar{t} - \overline{c_{\text{new}}})^2 + SE(\overline{c_{\text{hist}}} - \bar{p})^2}$, i.e. the fixed margin approach uses a larger standard error in the confidence interval calculation and thus is more conservative for demonstrating absolute efficacy over placebo, all else being equal. With both approaches, a confidence interval for the effect of the new treatment compared to placebo can be derived. While this confidence interval is a natural output of the synthesis approach (for q=0), for a fixed margin approach, such a confidence interval would need to be calculated in addition to the test statistic.

For both approaches, a defined fraction of the effect of the active comparator over placebo could be pre-specified to be preserved to safeguard against violations of the constancy assumption. By retaining a certain percentage of the treatment effect, the amount of discounting would be known a priori. In contrast, the amount of variance inflation of the fixed margin approach can vary and is not known precisely before the conduct of the trial (a plausible range can be derived, though, based on the expected standard error in the new trial).

The fixed margin approach can be interpreted as a conservative version of the synthesis approach by inflating the variance component and therefore, the same limitations apply for both approaches for testing absolute efficacy. Both approaches rely strongly on an adequate selection of the historical data and bias could result from the fact that the historical part of the data is already known at the design stage of the NI trial. While using the same margin for several drugs compared to the same comparator can be achieved with the fixed margin approach, in the synthesis approach consistency can be achieved by using the same historical data.

However, consistency in the above sense might not always be desirable. Relevant changes in the evidence for the efficacy of the active comparator could justify using a different margin in the fixed margin approach or a different estimate of the active comparator's effect against placebo in the synthesis approach.

Example

To illustrate the above-described approaches, consider a setting with a continuous endpoint and an estimated effect $\widehat{c_{\text{hist}} - p}$ of 13 on the relevant scale with a 95% confidence interval of [10, 16] and approximate standard error of 1.53.

In the fixed margin approach, the lower bound of the 95% confidence interval of 10 would be the maximal acceptable margin M for demonstrating absolute efficacy. However, to safeguard against violations of the constancy assumption, it might be decided that at least 60% of the benefit of the active comparator should be preserved, therefore the margin M would be calculated as $-(1-0.6) *$

$10 = -4$. Correspondingly, the non-inferiority comparison would be expected to demonstrate that the effect $\mu_T - \mu_{C_{new}}$ is larger than -4 by showing that the lower bound of the 95% confidence interval for $\mu_T - \mu_{C_{new}}$ is larger than -4.

If the non-inferiority comparison results in an estimate of $\bar{t} - \overline{c_{new}}$ of 5 with a 95% confidence interval of [-1, 11] and an approximate standard error of 3.06, the lower bound of the confidence interval would be larger than -4 and it can be concluded that at least 60% of the effect of the active comparator over placebo is preserved for the new treatment over placebo.

Using the above-described formular for the standard error used in the fixed-margin approach, we can calculate the standard error as $SE(\bar{t} - \overline{c_{new}}) + SE(\overline{c_{hist}} - \bar{p}) = 3.06 + 1.53 = 4.59$. Using a normal approximation, we can calculate a 95% confidence interval for $\mu_T - \mu_P$ via

$$\bar{t} - \bar{p} +/- 1.96 * \left(SE(\bar{t} - \overline{c_{new}}) + SE(\overline{c_{hist}} - \bar{p})\right) = 13 + 5 +/- 1.96 * 4.59 = [9.0, 27.0]$$

In the synthesis approach, similar to the fixed margin approach, it is pre-specified that at least 60% ($q = 0.6$) of the benefit of the active comparator should be preserved, as a safeguard against violations of the constancy assumption. Consequently, the null hypothesis $\mu_T - \mu_P < q\,(\mu_{C_{hist}} - \mu_P)$ or, equivalently under the constancy assumption, $\mu_T - \mu_{C_{new}} + 0.4 * (\mu_{C_{hist}} - \mu_P) < 0$ is to be tested. For constructing a 95% confidence interval for $\mu_T - \mu_{C_{new}} + 0.4 * (\mu_{C_{hist}} - \mu_P)$, the point estimates and standard errors from the new trial and historical evidence are used in line with the formulas above as:

$$SE(\ (\bar{t} - \overline{c_{new}}) + 0.4 * (\overline{c_{hist}} - \bar{p})\ ) = \sqrt{SE(\bar{t} - \overline{c_{new}})^2 + (0.4)^2 SE(\overline{c_{hist}} - \bar{p})^2} = \sqrt{3.06^2 + (0.4)^2 1.53^2} = 3.12$$

The 95% confidence interval for $\mu_T - \mu_{C_{new}} + 0.4 * (\mu_{C_{hist}} - \mu_P)$ is calculated as

$$(\bar{t} - \overline{c_{new}}) + 0.4 * (\overline{c_{hist}} - \bar{p}) +/ -1.96 * SE(\ (\bar{t} - \overline{c_{new}}) + 0.4 * (\overline{c_{hist}} - \bar{p})\ ) = 5 + 0.4 * 13 +/- 1.96 * 3.12$$
$$= [4.1, 16.3]$$

Since the lower bound of the 95% confidence interval is larger than 0, we can conclude that the new treatment preserves at least 60% of the benefit of the active comparator.

Additionally, by selecting $q = 0$, we can calculate the point estimate and 95% confidence interval for the indirectly estimated effect $\mu_T - \mu_P$ via

$$\bar{t} - \bar{p} = \bar{t} - \overline{c_{new}} + \overline{c_{hist}} - \bar{p} = 5 + 13 = 18$$

$$SE(\bar{t} - \bar{p}) = SE(\ (\bar{t} - \overline{c_{new}}) + (\overline{c_{hist}} - \bar{p})\ ) = \sqrt{SE(\bar{t} - \overline{c_{new}})^2 + SE(\overline{c_{hist}} - \bar{p})^2} = \sqrt{3.06^2 + 1.53^2} = 3.42$$

Thereby, we can calculate the 95% confidence interval for the indirectly estimated effect of the new treatment over placebo as

$$\bar{t} - \bar{p} +/- 1.96 * SE(\bar{t} - \bar{p}) = [11.3, 24.7]$$