EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

9 September 2024
EMA/CHMP/430688/2024
Committee for Medicinal Products for Human Use (CHMP)

Reflection paper on establishing efficacy based on single-arm trials submitted as pivotal evidence in a marketing authorisation application

Considerations on evidence from single-arm trials

Draft agreed by Drafting Group on single-arm trials	27 January 2023
Adopted by CHMP for release for consultation	17 April 2023
Start of public consultation	21 April 2023
End of consultation (deadline for comments)	30 September 2023
Agreed by the Methodology Working Party (MWP)	6 September 2024
Adopted by the CHMP	9 September 2024

Keywords	Single-arm trials, non-randomised trials, regulatory decision-making

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

Table of contents

1. Introduction and scope

Randomised controlled trials (RCTs) are the standard for providing confirmatory evidence on the efficacy of an investigational treatment. If the pivotal clinical data in a marketing authorisation application dossier are intended to stem from single-arm trials (SATs), it is the responsibility of the applicant to justify to regulators the reasons for deviating from the expected standard and the appropriateness of the SATs as alternative. As part of such justification, it must be substantiated that the SATs provide adequate pivotal evidence of efficacy.

The purpose of this reflection paper is to outline perspectives on SATs that are submitted as pivotal evidence for establishing efficacy in marketing authorisation applications. It aims to identify and critically reflect on specific and important features of SATs. Systematic assessment of these features supports applicants and regulators in evaluating the adequacy of SATs to provide pivotal evidence for regulatory decision-making. Such considerations strongly depend on the full clinical context and attributes of the investigational treatment. Obtaining scientific advice is therefore strongly recommended to discuss whether pivotal evidence from SATs could be considered acceptable for seeking marketing authorisation for a specific development programme.

The assessment of efficacy usually covers multiple endpoints and is an essential part of the benefit-risk assessment. Although this reflection paper is focused on establishing efficacy via SATs, also establishing safety via SATs is fraught with substantial shortcomings. Many of the critical considerations discussed in this reflection paper equally apply to the assessment of safety and need to be read in conjunction with other guidance on the assessment of safety. The assessment of a marketing authorisation application is based on the totality of evidence across the drug development programme which usually includes the conduct of multiple clinical trials. The key concepts described in this reflection paper apply to all objectives and in all contexts the SAT is used for. Many of the considerations described in this document also translate to SATs which are not submitted as pivotal evidence, including those used for decision-making in early development. The general requirements for the design, planning, conduct, analysis, and reporting of clinical trials also apply to SATs and are not the focus of this reflection paper.

Sections 1.1. and 1.2. specify the type of trials discussed in this reflection paper as well as characteristics specific to SATs. Following a list of relevant regulatory guidelines (Section 2.), key concepts and definitions useful to articulate considerations for assessment and interpretation of SATs are described in Section 3. , whereas Section 4. translates these concepts into practical considerations.

1.1. Description of single-arm trials

In SATs, all subjects entering the trial are planned to receive the investigational treatment and to be followed prospectively for a period of time. To address the primary question of clinical interest, the trial design for a SAT does not include a formal comparison to data from an internal or external control group. Notably, this paper does not address specific considerations for the situations where results of a SAT are contrasted against an external control that is constructed outside of the original trial protocol, e.g. as part of the totality of data submitted by applicants. In such settings, the present paper fully applies to the respective SAT. Designs that prospectively include a non-randomised external control (arm) in the trial protocol may not be considered SATs, but key considerations in this paper may still apply due to the lack of randomisation.

In general, the considerations in this reflection paper extend to trials that contain more than one arm, but do not randomise to a control for a formal comparison. This includes non-randomised trials, as well as trials in which only investigational treatment arms are randomised, but without formal comparisons

between the arms. An example for such a trial would be a particular kind of platform trial where several investigational treatment arms are included but which are not formally compared, and which can be viewed as a series of SATs. All these designs are considered SATs for the purpose of this reflection paper.

1.2. Specific characteristics of single-arm trials

Relative to double-blind RCTs, SATs lack the following key design features: a concurrent control arm, randomised allocation to treatment, enrolment of subjects without knowledge of their subsequent assignment, and blinding of participants, investigators and outcome assessors to treatment assignment. Consequently, SATs lack features that are instrumental to avoid bias (see Section 4.5). Due to the lack of randomisation, the design does not support an inherent causal interpretation as an effect of the investigational treatment and must rely on knowledge external to the SAT about the outcomes of subjects that would be observed had they not been treated with the investigational treatment. In addition, it does not include a randomised comparison against a control arm that allows to directly quantify the size of the treatment effect and the associated sampling variability. Thus, statistical methods to quantify the size of the treatment effect and corresponding precision and interpretation of results must rely on non-testable assumptions about the population distribution of the outcomes without the investigational treatment as well as on patient selection. As a consequence, the derived magnitude of effects is more difficult to interpret, and less reliable.

If results from SATs are intended to be used as pivotal evidence for approval, it is essential that their adequacy is systematically addressed in terms of their characteristics, limitations and uncertainties. This assists in establishing whether proof of efficacy could be based on SATs at all, and if so how to characterise the effect of the investigational treatment and understand remaining uncertainties to best inform and communicate benefit-risk assessment.

2. Legal basis and relevant guidelines

Part 4 from Directive 2001/83/EC states that: 'In general, clinical trials shall be done as 'controlled clinical trials' and if possible, randomized; any other design shall be justified.' This document is based on applicable EU regulation and should be read in conjunction with all other relevant EU and ICH guidelines. The following documents are of particular relevance:

- ICH guideline E8 (R1) on general considerations for clinical studies (EMA/CHMP/ICH/544570/1998)
- ICH E9 Statistical Principles of Clinical Trials (CPMP/ICH/363/96)
- ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials (EMA/CHMP/ICH/436221/2017)
- ICH E10 Choice of control group in Clinical Trials (CPMP/ICH/364/96)
- Guideline on clinical trials in small populations (CHMP/EWP/83561/2005)
- Points to consider on application with 1. Meta-analyses; 2. One pivotal study (CPMP/EWP/2330/99)
- Methodological issues in confirmatory clinical trials planned with an adaptive design (CHMP/EWP/2459/02)
- Guideline on adjustment for baseline covariates in clinical trials (EMA/CHMP/295050/2013)

- Guideline on the investigation of subgroups in confirmatory clinical trials (EMA/CHMP/539146/2013)
- Guideline on registry-based studies (EMA/426390/2021)
- Points to consider on multiplicity issues in clinical trials (CPMP/EWP/908/99)

In particular, ICH E10 offers general considerations and recommendations on SATs. This reflection paper expands on these, providing further articulation and details on key concepts relevant for the design, reporting and regulatory assessment of SATs.

3. Key definitions and terminology

To articulate key points for the design, planning, conduct, analysis and interpretation of SATs it is deemed important to more precisely define the following concepts and terminology.

Outcome

The individual outcome of a patient refers to the measurement(s) of an endpoint for said patient, e.g. cure. Statistical summary measures combine a set of individual outcomes for a group of patients or a population, e.g. 50% cured. Outcome and summary measure are key elements for the estimand.

Estimand and treatment effect of interest

The concept of estimand, defined as 'a precise description of the treatment effect reflecting the clinical question posed by the trial objective' (ICH E9(R1)), is equally important for SATs as for RCTs. However, due to the uncontrolled nature of SATs, some concepts from the estimand framework are more difficult to apply.

Regulatory assessment requires addressing whether there is an effect attributable to treatment and to estimate the magnitude of the treatment effect. These questions require to assess 'how the outcome of treatment compares to what would have happened to the same subjects under alternative treatment (i.e. had they not received the treatment, or had they received a different treatment)' (ICH E9(R1)). These regulatory questions of interest underline the shortcomings of submissions with pivotal clinical data from SATs. In particular, the estimand(s) which can be directly targeted by SATs are limited by observing results only under the investigational treatment. The estimand(s) from the SAT therefore do not directly address the regulatory question of interest. In the regulatory assessment, knowledge external to the SAT about what would have happened to subjects had they entered into the trial in absence of the investigational treatment (i.e. the counterfactual) is necessary to bridge from the estimand(s) of the SAT to the regulatory question of interest. This reflection paper addresses pre-requisites that may allow drawing conclusions based on a SAT about the primary questions of interest for marketing authorisation. As with all medicinal products, additional research questions of interest and decisions may need to be addressed. The same pre-requisites outlined in this reflection paper with respect to marketing authorisation may however not be sufficient to draw conclusions with sufficient certainty for these questions of interest and decisions based on the SAT.

Isolation of treatment effect

There is no general statistical or methodological definition for the concept of isolating a treatment effect. For the purpose of this reflection paper, the following definition is adopted: If in a SAT individual outcomes for the planned endpoint are observed that could not occur without effective treatment within the designated follow-up period for any trial participant, the SAT is able to isolate the treatment effect on that specific endpoint. Conceptually, this can allow a causal interpretation of the effect of the treatment, despite the limitations in trial design.

This is a theoretical concept which requires qualitative reasoning that leaves no doubt about the causal relationship between the treatment and outcome measured by the chosen endpoint. This will only be perfectly satisfied in exceptional cases. However, in general this concept enables systematic assessment of the uncertainties involved in attributing observed outcomes to the investigational treatment. This systematic assessment ultimately aids to determine whether causal conclusions can be drawn with sufficient certainty from the SAT on the effect of the treatment.

In practice, observed individual outcomes are subject to bias and variability, for example, in terms of errors in measurement or assessment. In contrast to RCTs, measurement errors or less stringent conduct of the SAT may lead to observed outcomes that would support an erroneous assessment that there is a treatment effect. There can also be uncertainty about which outcomes are truly impossible without treatment (such as the level of motor function in spinal muscular atrophy patients). In other situations it may be clear that the uncertainty and concerns on bias are always such that they do not allow isolation of treatment effects in a particular setting. In general, isolation of a treatment effect can only be assessed subject to a degree of uncertainty.

Depending on the therapeutic area and the development programme, the primary objective of the SAT may be the isolation of a treatment effect on an endpoint or the estimation of the size of the treatment effect. In general, SATs submitted as pivotal evidence should target an estimand and associated endpoint that allows causal attribution of the treatment effect as well as provide an estimate of clinically relevant benefit, as is usually the case for RCTs. This poses a specific challenge for SATs, as the clinically most relevant endpoint may not be suited to adequately isolate the treatment effect, while surrogacy of another endpoint that isolates the treatment effect may not be established (see Section 4.2.

Treatment effect estimate

In addition to establishing that an investigational treatment has an effect, a high degree of confidence in the estimates of the size of the (favourable and unfavourable) treatment effects is necessary for regulatory assessment. Statistical summary measure estimates from SATs have the limitation that only results under the investigational treatment (such as percentage of responders) are observed. Conceptually, the treatment effect estimate of interest for regulatory assessment is the contrast to an unobserved counterfactual (such as 0% responders). In specific applications, it may be justified that observing results from the investigational treatment arm suffices to draw conclusions about this comparative treatment effect size with an acceptable degree of uncertainty. In the following, the concept of treatment effect estimate refers to the estimate from the SAT contrasted to the unobserved estimate for the counterfactual.

Treatment effect estimates based on SATs are directly impacted by the selection of subjects included in the trial. Even though the observed individual outcomes in an RCT may be equally subject to the selection of subjects as in SATs, the RCT allows assessment of treatment efficacy through the direct comparison of treatment arms that are on average equally affected by selection. If a SAT is intended to provide pivotal evidence, confidence in the estimated size of treatment effects must be gained by addressing multiple potential sources of bias that could arise throughout the design, conduct, analysis and reporting of SATs (see Section 4.6.).

External validity

The external validity of SATs is characterised by the systematic difference between the treatment effect estimate based on the SAT and the true treatment effect in the target population. This type of bias also applies to treatment effect estimates from RCTs if the treatment effect differs between subgroups that are defined by a predictive factor and the distribution of these subgroups in the trial population is not representative of the target population. For example, if the treatment effect is larger

in biomarker positive subjects and the proportion of biomarker positive subjects in the trial population is higher than in the target population, this will bias the treatment effect estimate from the RCT compared to the treatment effect in the target population. Treatment effect estimates based on SATs are equally impacted by heterogenous treatment effects. In addition, treatment effect estimates based on SATs are biased if there is heterogeneity in disease prognosis and the trial population is not representative of the target population. For example, if biomarker positive subjects have a better disease prognosis regardless of treatment and the proportion of biomarker positive subjects in the trial population is higher than in the target population, this will bias the treatment effect estimate from the SAT compared to the treatment effect in the target population. Hence, external validity is more likely compromised in SATs.

Quantification of uncertainty

For regulatory decision-making, uncertainty in treatment effect estimates needs to be properly quantified, e.g. in the form of confidence intervals with adequate coverage probability. In RCTs this is done based on the statistical properties induced by randomisation and directly includes the uncertainty of the estimates under the control condition. Quantifying the uncertainty of treatment effect estimates based on SATs requires additional consideration. This is because only the variability of individual outcomes for the investigational arm is directly observed, but not for the for hypothetical control (see Section 4.5).

4. Methodological considerations for single-arm trials

4.1. General principles

Introduction

While SATs are used in all drug development phases, exploratory and confirmatory trials have different requirements regarding methodological and executional rigour. If SATs are submitted as pivotal evidence, best practice and strict criteria should be followed at the planning stage and during the entire conduct of the trial, and the assessment will need to follow standards that apply to the confirmatory setting. Due to the lack of safeguarding mechanisms like randomisation and blinding, any changes impacting the trial design or conduct (e.g. in terms of endpoints, trial population or statistical analysis approach) after trial initiation can have a large impact on the reliability of the results and the assessment of SATs.

Prespecification

From the confirmatory perspective, SATs which are submitted as pivotal evidence are expected to have an a priori definition of a clear success criterion. The success criterion needs to be justified based on suitable existing information, including knowledge about the disease and its natural course, uncertainties around the variability of outcomes and heterogeneity of the patient population. One way to define the success criterion in a SAT is through a threshold which the lower or upper bound of the confidence interval for the summary measure from the SAT must exceed (see Section 4.5.). It is recommended to seek scientific advice on the appropriateness of the proposed SAT success criterion in the specific clinical context.

Predefinition and adherence to the trial protocol when the trial is ongoing are always important, and this is even more pronounced in SATs. Due to the unblinded nature of SATs, claims on existing firewalls may not be sufficient to overcome concerns about potential data knowledge, and any amendment may need to be considered potentially data driven. This includes unplanned interim analyses, changes in endpoints, changes in or deviations from the planned number of participants (sample size changes), changes in the dosing regimen, changes in eligibility criteria, subgroup

selection, or treatment arm selection. In the context of regulatory decision-making, an especially critical unplanned change is the post hoc designation of a trial planned as exploratory trial (e.g. phase II) to a pivotal trial once trial data were available and to submit this as primary confirmatory evidence. Due to the unblinded nature of SATs, planned data-dependent modifications can also be considered critical. While predefinition is necessary, it is not the only condition to enable a SAT as confirmatory evidence.

4.2. Choice of endpoints

In general, the primary efficacy endpoint for the main trial(s) aiming to establish efficacy should reflect the variable capable of providing the most clinically relevant and convincing evidence directly related to the primary objective of the trial (ICH E9). This choice requires a fine balance between methodological aspects and a range of endpoint characteristics like validity, reliability, feasibility, and accepted norms and standards in the relevant field of research. From a regulatory standpoint the primary endpoint of a SAT must also be objectively measurable and able to isolate treatment effects (see Section 3). This means that observations of the desired outcome for the primary endpoint should be known to occur only to a negligible extent (in number of subjects or size of the effect) in the absence of an effective treatment. Consequently, a SAT cannot generate adequate pivotal evidence if there is no suitable endpoint with the ability to isolate treatment effects.

Any uncertainty whether observed individual outcomes are undoubtedly caused by the treatment complicates the interpretation of results from a SAT. In particular, these uncertainties can lead to concerns that results appear favourable due to a potential bias in the SAT. For example, if the probability of remission in the absence of treatment is small but not zero, it may be unclear to what extent selection bias could lead to false positive conclusions on efficacy (due to overrepresentation of patients in the trial with higher likelihood of remission in absence of an effective treatment). In addition, measurement error or misclassification may lead to recording erroneous individual outcomes in the SAT and, due to the lack of an equally affected control in the same trial, unduly favour the investigational treatment. In other situations, the disease may be episodic, characterised by a waxing and waning course. In such cases, any relevant primary endpoint would be affected by the natural course of disease in a way that would not permit isolating a treatment effect via a SAT.

Whether or not a specific endpoint is acceptable in a therapeutic area or allows establishing a clinically relevant treatment effect needs to be discussed on clinical grounds. The acceptability of a SAT and its primary endpoint depends on the clinical context and is therefore a disease area specific decision. In the following, some of the challenges with the most common types of outcome measures are discussed, without being exhaustive.

Time-to-event endpoints

Time-to-event endpoints such as time to death, progression-free survival, or time to first stroke measure time to events that can occur in the absence or presence of an effective treatment. For this reason, observed individual outcomes for such endpoints generally cannot be attributed to treatment and therefore time-to-event endpoints are usually not suitable in SATs to isolate a treatment effect. A major problem with time-to-event endpoints relates to the starting point of being at risk for a specific endpoint (‘time 0’), which is usually different from the start of the trial, and which cannot be determined with reasonable certainty except for very few experimental settings. In RCTs, the control arm provides an internal calibration for the trial participants’ history at risk prior to enrolment in the trial. This calibration is however lacking in SATs.

The impact of the course of a disease on time-to-event endpoints is usually unpredictable, particularly based on how prognostic factors impact the time until an event occurs. For time-to-event endpoints,

this amplifies the general problem that disentangling between prognostic factors (i.e. differences in expected outcomes irrespective of the investigational treatment) and predictive factors (i.e. differences in treatment effects caused by the investigational treatment) cannot be achieved based on the results from a SAT (see Section 4.3. 4.3.).

Continuous endpoints

Continuous endpoints are often expressed as change from baseline or are analysed in (repeated measures) models that are conceptually close to change from baseline analysis. Continuous endpoints allow for a precise and sensitive measurement of the changes that participants experience during the trial. However, when individual outcomes can change due to within-subject variability (random fluctuation over time), the natural course of a disease (systematic change over time), or measurement error, this change cannot be attributed to treatment. Therefore, a causal attribution of a treatment effect and the size thereof is difficult for continuous endpoints in SATs. A common phenomenon is 'regression to the mean' which may result from a combination of measurement error, within–subject variability and trial participant selection at baseline (see Section 4.6. 4.6.). For example, in case patients are selected based on disease severity as expressed by an extreme value of an endpoint at the time of inclusion in the trial (eligibility criterion), the measurements of these patients will have a tendency for improved values at a later point in time, irrespective of being treated with an effective treatment or not. Similar considerations apply for ordinal endpoints.

Binary endpoints / dichotomised endpoints

Binary endpoints are also not free of the problems described for time-to-event and continuous endpoints, but there may be diseases where a specific state usually does not change without intervention, e.g. being infected with hepatitis C. If after treatment intervention a 'cure' is achieved that would not be achievable without treatment, it may be plausible to conclude on a treatment effect. This can also apply to cases where patients are alive at a time point that substantially exceeds what patients would achieve without treatment or for continuous endpoints which cross a prespecified cut-off which cannot be achieved without treatment and is well beyond measurement uncertainty. In these cases, the binary endpoint can be considered to isolate the treatment effect with sufficient certainty. However, it should be emphasised that making wrong assumptions on such cut-offs at the trial planning stage makes the SAT results difficult to interpret with respect to a treatment effect (or the size thereof), even in the case of objective endpoints.

In principle, the issues of the underlying endpoint, regardless of its nature, are transferred to a version of that endpoint that is dichotomised by means of a cut-off. In specific cases it may, however, be possible to set the cut-off in advance in such a way that crossing it is not possible without effective treatment for any subject, even after accounting for potential sources of bias (as discussed above and in Section 4.6.).

4.3. Target and trial population

Recruiting an adequate trial population is required to ensure that conclusions about the effects of the investigational treatment are indeed valid for the intended target population, i.e., those subjects that will receive the treatment in routine practice. As described in Section 3, concerns about external validity are larger for SATs as compared to RCTs, because the treatment effect is not estimated relative to a control and therefore the composition of the trial population is especially relevant for estimates from a SAT. The trial population determines the plausibility of assumptions about the disease course in the hypothetical control group (counterfactual). It is important that the trial population is predefined and discussed with sufficient detail in relation to prognostic variables that can impact the natural disease course, as well as potential predictive variables.

The assumptions on the natural course of the disease necessarily need to hold for the trial population as included in the SAT. In practice, this means that the included trial population should not only share the known, but also the unknown characteristics of the patient population the assumptions were based on (the hypothetical control group), a requirement that is impossible to verify. As a consequence, the interpretation of results from SATs is more challenging in settings with high patient or disease heterogeneity.

In addition to inclusion and exclusion criteria defined in the protocol, less tangible and not easily documented selection mechanisms associated with prognosis do occur at the point of recruiting trial participants; both due to investigator decisions as well as subjects' choices, or even due to criteria related to selection of clinical trial sites. Such selection mechanisms may particularly impact SATs as they lack a control group providing a reference for the course of disease, which in turn can be related to previous trial experiences or epidemiological information about the target population. Consequently, a selection and understanding of the trial population that allows assessment of the benefit-risk balance for the target population is an essential prerequisite for a SAT to serve as pivotal evidence. To provide reassurance that the magnitude of an observed positive effect is not the result of a favourable selection of the trial population, prespecification and documentation of the subject selection process are of utmost importance to the assessment. In addition to well justified inclusion and exclusion criteria this includes details about the screening process, the decision for trial inclusion or exclusion, and clinical characteristics of the subjects included as well as about the subjects who were excluded.

In RCTs randomisation provides the basis for statistical inference by balancing in expectation the distribution of known and unknown prognostic or predictive variables over the treatment arms. Even if balance in important prognostic variables is not precisely achieved in the actual randomisation, including known prognostic variables appropriately into the prespecified confirmatory analysis will reduce the impact on treatment effect estimates. In contrast, in SATs the potential impact of unknown prognostic or predictive variables cannot be controlled. Furthermore, the estimation or control for the impact of known prognostic variables might not always be feasible. In particular, it is not possible to disentangle prognostic from predictive effects based on results from SATs (see Section 3.).

Biomarker-defined target and trial populations are one important example where interpretation of results from a SAT is challenging. Here, additional complications arise because the biomarker may not only be predictive for the treatment effect, but also be prognostic for the natural course of the disease. As the association between the biomarker and the endpoint measured is typically part of the development programme with limited or no historical data available, often no reliable estimate of the natural disease course in the targeted biomarker-defined subgroup is available. This limits applicability of SATs to provide pivotal clinical data for biomarker-based drug development to situations where very strong external knowledge about the (prognostic or predictive) role of the biomarker exists.

More generally, exploration of heterogeneity of treatment effects across subgroups is important, but also a particular challenge in SATs. This is because the lack of a control makes it impossible to clearly differentiate between subgroup heterogeneity caused by prognostic or by predictive factors based on the data from the SAT. In consequence, there should be strong biological plausibility for predictive effects and the associated expectations should be prespecified and justified before conducting the trial. Unexpected subgroup findings may cast doubt on the assumption that the course of the disease and the mechanism of action of the investigational treatment are well understood. Furthermore, strong prognostic factors may raise concern regarding selection bias and strengthen the need for a randomised concurrent control.

4.4. Target of estimation

Treatment condition

In RCTs both the treatment condition of interest as well as the alternative treatment condition to which a comparison will be made (ICH E9(R1)) are defined within the trial. In SATs, only the investigational treatment is administered and there is no alternative treatment condition to which a direct comparison can be made with the data from the SAT. In order to interpret the results from a SAT in the context of isolation and estimation of the size of a treatment effect, it is essential that the treatment condition that is considered the counterfactual, i.e. the condition in absence of the investigational treatment, is precisely and fully defined a priori. This definition includes the assumed absence of an effective treatment as the counterfactual, but it also needs e.g. sufficient detail on the supportive care that is (assumed to be) implemented for both the investigational and counterfactual treatment condition. Interpretation of the results from a SAT critically depends on a priori knowledge of the natural course of disease, and this knowledge not only needs to specifically apply to the trial population, but also to the counterfactual treatment condition.

The timing of treatment initiation in a SAT needs to be clearly defined, and it should be discussed upfront with regulators which events are to be considered intercurrent events. In some drug development programmes, it may be appropriate to define the start of treatment as a time point prior to first actual administration of the investigational treatment.

Intercurrent events

Intercurrent events are defined as 'Events occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest' (ICH E9(R1)). In SATs, intercurrent events are only observed for the investigational treatment which poses an additional challenge in relation to their interpretation and handling.

Strategies of handling intercurrent events in the SAT should be applied that ideally are aligned with the regulatory question of interest (see Section 3.). However, the lack of information on the occurrence of intercurrent events in the absence of the investigational treatment may complicate assessment of the impact of the chosen strategy on the conclusions of the SAT. Strategies that could lead to anti-conservative conclusions would be of largest regulatory concern.

Population summary measure

In a SAT the population summary measure is necessarily not comparative. Inferring the isolation of a treatment effect and its clinical relevance are therefore dependent on the full consideration of principles outlined in this reflection paper.

4.5. Statistical principles

Threshold setting

For endpoints that unambiguously (i.e. without uncertainty) isolate the treatment effect, establishing that an investigational treatment is effective can be straightforward based on results from SATs. In other cases, the trial success criterion to demonstrate efficacy can be formalised through a justified threshold which the lower or upper bound of the confidence interval of the summary measure at the trial population level must exceed. This is to ensure that clinical benefit of the investigational treatment for the target population can be established with sufficient certainty.

Such a threshold can be based on external clinical information of the natural course of the disease and heterogeneity in endpoint assessment. This bears the inherent risk of erroneous conclusions about the efficacy of the investigational treatment due to comparing results across different data sources.

Consequently, the choice of threshold should adequately reflect any uncertainties associated with the external information as well as its applicability to the SAT. In any case, the threshold needs to be prespecified and its clinical validity in the therapeutic context needs to be carefully justified.

In the case of endpoints for which the desired individual outcomes can occur in the absence of treatment, even if thought to a small extent only, variability may be observed such that the effect of the investigational treatment cannot be unambiguously isolated at the individual level. For such endpoints, it can be challenging to establish that there is a treatment effect and to quantify the size of the treatment effect. It is important to distinguish between the observed summary measure and the treatment effect (see Section 3.). Conceptually, if a threshold was specified which was thought to represent the corresponding summary measure under the hypothetical scenario of absence of treatment, the size of the effect attributable to the treatment would only be the difference between this threshold and the summary measure observed in the trial. However, a crude comparison does not account for the uncertainty in the point estimates which also needs to be considered by comparing confidence intervals against this threshold (see below 'Multiplicity'). In addition, the defined threshold will usually not be a known constant, but will be derived from external information that is prone to uncertainty. Therefore, treating this as a fixed constant does not properly reflect the underlying uncertainty that is inherent to its definition and a conservative threshold should be chosen. For example, the choice of threshold might be informed by (depending on the clinical context) the lower or upper limit of the confidence interval instead of the point estimate as derived based on external clinical information. Moreover, results depend on the selection mechanism in the trial. Overall, this makes the choice of a threshold difficult to justify. Therefore, scenarios without unambiguous isolation of the treatment effect on the individual level represent a critical risk for the interpretation of a SAT.

Analysis and Estimation

All analyses should be prespecified in a detailed statistical analysis plan before the SAT starts, i.e. before inclusion of the first patient. Importantly, this includes prespecification of the analysis in terms of the population of subjects and handling of intercurrent events (see Sections 4.3. and 4.4.). This is a fundamental part of achieving appropriate estimates from the SAT and bias due to inclusion or exclusion of subjects in the analysis set based on observed individual outcomes should be avoided.

In general, all subjects who initiated treatment should be used for the primary analysis. Situations may exist, however, which justify the exclusion of subjects to avoid biased estimates from a SAT towards a larger effect and thus towards overestimating clinical benefit. An example would be a situation where some subjects who are not diseased at trial entry and would therefore by definition be free of the disease at the end of the trial, were incorrectly included into the SAT. This situation can also occur when measurement methods to select patients for presence of a disease are different to those that are used for assessing the changes of the disease during the trial (e.g. response or resolution), and this situation is comparable to measurement error as discussed in Sections 3 and 4.1. Such cases should be avoided by trial design and conduct. If the number of subjects affected by this is relatively large, this may also question the validity of the endpoint and the trial. In particular, an individual outcome cannot be attributed as a response to treatment if a patient who was selected based on a measurement method at baseline would have been considered disease-free at baseline using the outcome measurement method. A further exemption may be required when an analysis is done before all subjects have reached a prespecified analysis timepoint (see below 'Missing data').

The statistical analysis used for estimation of the treatment effect should be fully prespecified. This should include a justification for which and how potential prognostic or predictive factors are to be incorporated, as well as a discussion on how the results will need to be interpreted. In SATs the method of estimation is of utmost importance, as the distribution of covariates is by design not

calibrated against a control that shares the same (randomised) characteristics (see Section 4.3.), and the handling of prognostic factors will impact the estimates for the targeted endpoint.

Factors that are predictive of the treatment effect can impact the estimation of the treatment effect both in RCTs as well as in SATs. However, in SATs, there is an additional problem for prognostic factors that does not generally exist in RCTs. This problem is related to how the levels of a factor are dealt with in a statistical analysis model for estimation. Due to the comparison against a randomised control this is not a problem in RCTs when using linear models, although it is a known problem in nonlinear models. In SATs however, the lack of calibration against a control makes the estimates calculated from a linear model dependent on how factor levels (i.e. their distribution observed in the trial sample) are treated in the analysis model. Consequently, if the distribution of the trial population does not per se resemble the distribution in the target population, estimation of the effect in the target population is a particular challenge. Investigating several distribution scenarios may provide additional analyses of interest, however, the exact distribution of the target population is usually unknown.

Overall, it is strongly encouraged to present sensitivity analyses to support the robustness of the estimates (see 'Sensitivity Analyses'). This issue is also related to selection bias, and the operational handling of data can never fully resolve selection in the common case that the selection mechanism is unknown (and may be broader than represented by few measured factors), or that patients are not adequately represented. If a sensitivity analysis with different handling of covariates leads to inconsistent results, this may question the overall reliability of the trial result.

Missing Data

Missing data across all relevant endpoints should be avoided through trial design and conduct. In the event of missing data, analysis methods should be applied that ideally provide unbiased estimates and as a necessary criterion do not overestimate the response to treatment. For example, if the endpoint is treatment failure, subjects who did not complete the pre-planned individual end-of-trial time point, but for whom it is already known that they failed should be included as failures in the analysis. In case of an interim analysis, those subjects who have not yet reached their individual end-of-trial time point and have not failed by then should not be included in the primary analysis, because these subjects should not be counted as non-failures. Sensitivity analyses are encouraged (see below 'Sensitivity Analysis').

Sensitivity analysis

Sensitivity analysis for the main estimator of the targeted estimand is a necessary, albeit not sufficient, criterion for assessing the influence of assumptions in the SAT, e.g. in relation to the handling of missing data. Of particular relevance to the interpretation of results from a SAT is the potential sensitivity to assumptions that cannot be tested based on the data generated in the SAT. These include assumptions about the natural course of the disease for the patients that were included in the SAT.

Multiplicity

While p-values from formal hypothesis testing are conceptionally of subordinate relevance compared to estimation (point estimates and confidence intervals) for the assessment of a given endpoint in SATs, it is still relevant for a SAT to control the probability of false positive conclusions at the trial level. As usual, multiplicity is present in case of several treatment arms, several endpoints or timepoints, interim analyses, or subgroup assessment. As outlined in Section 1, the general principles for (randomised) clinical trials apply also for SATs, and methods to address multiplicity should be pre-planned and adhered to.

Sample size

As for any other trial design, the sample size chosen for a SAT should be large enough to provide a reliable answer to the questions addressed, taking into consideration the planned analysis and the trial success criteria. While a SAT permits to allocate more subjects to an investigational treatment, uncertainty with respect to bias may outweigh any gains in precision compared to a randomised controlled design.

4.6. Sources of bias and potential mitigation

As described in Section 3. , unbiased estimates are difficult to obtain from SATs. Consequently, multiple potential sources of bias need to be addressed throughout the design, conduct, analysis and reporting of results from a SAT. Table 1 summarises potential sources and mitigation strategies for bias in SATs, some of which also apply to (open-label) RCTs. While these strategies may be considered necessary to reduce the risk for bias, they cannot be considered sufficient to fully remove bias and formal proof that treatment effect estimates are unbiased is impossible. Demonstration that the mitigation strategies were applied may thus not be sufficient to alleviate concerns about biased results from a SAT.

Table 1: Measures aiming to reduce potential bias in single-arm trials

Type of bias	Description	Potential bias reduction measures
Assessment bias	Knowledge of the treatment can influence the outcome assessment.	Endpoints in SATs should be objectively measurable and, if possible, assessments should be made independently and unaware of timing in relation to treatment (i.e. blinding of assessors to timing of measurement).
Attrition bias	Withdrawal of trial participants and missing data in general constitute an additional source of confounding that is difficult to resolve.	Avoid missing data by means of adequate trial design and conduct. Prespecify methods for handling of missing data that do not overestimate the response to treatment. Conduct suitable sensitivity analyses.
Bias due to lack of pre-planning	Any post trial-initiation changes in design, conduct and planned reporting (e.g. changes to the statistical analysis plan, protocol amendments on inclusion or exclusion criteria, adaptations in treatment, follow-up or allowed concomitant treatment) carries the risk of introducing bias.	Pre-planning is essential for all confirmatory trials, but the standard needs to be set even higher for SATs (e.g. statistical analysis plan needs to be finalised before trial initiation, absolutely minimise changes to the protocol and statistical analysis plan after trial initiation, if interim analyses are planned it is more problematic if they are flexible or not carried out at the pre-planned information level).
Bias due to regression to the mean	Subjects selected for inclusion in the trial based on extreme values are expected to show improved outcomes due to regression to the mean.	Define target population independently of disease severity at the time of or prior to inclusion. Avoid patient selection based on outcome measures that are subject to measurement error or fluctuation.
Bias due to variability in disease history	Patients can have substantial variability in their disease history before the investigational treatment is administered. This is especially (but not only) concerning for time-to-event endpoints where the disease history is usually strongly prognostic of the individual outcome.	Analysis of time-to-event endpoints is usually more difficult to assess without bias. Endpoints and analysis methods that do not directly rely on a time scale should be chosen.

Intercurrent event bias	Failure to identify relevant intercurrent events at the trial planning stage usually results in implicitly handling them with a treatment policy strategy, which is not necessarily the relevant strategy and likely results in a biased estimate for the regulatory question of interest.	Follow ICH E9 (R1), anticipating the intercurrent events at trial planning stage and ensuring the definition of estimand(s) as well as detailed collection of information on intercurrent events.
Selection bias in relation to the hypothetical control group	Subjects enrolled in a SAT may systematically differ from the hypothetical control group in ways that impact their prognosis.	Precisely prespecify inclusion and exclusion criteria such that the enrolled trial population matches the external information that assumptions are based on is a minimum pre-requisite. The clinical characteristics of the trial population as enrolled determines the actual risk of bias, hence prevention of additional selection and detailed documentation of selection processes during trial conduct is essential.
Selection bias in relation to the target population	Subjects enrolled in a SAT may systematically differ from the target population in ways that impact their prognosis.	Limit the number and extent of inclusion and exclusion criteria. Precisely prespecify expected prognosis in terms of the primary endpoint of the target population, including the external information this is based on.
Selection bias in relation to biomarker-defined subgroups	Subjects selected for targeted treatment based on a prespecified biomarker may differ in prognosis compared to the overall population (i.e. independent of the biomarker status).	Ensure prognosis of the biomarker-defined subgroup is sufficiently known prior to trial start.
Trial bias	Subjects in a SAT may have systematically different outcomes (independent of the investigational treatment) compared to clinical practice, e.g. due to different care in the trial setting.	Assure and show that the auxiliary care reflects the current standard.