EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

17 April 2023
EMA/CHMP/564424/2021
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

Considerations on evidence from single-arm trials

Draft

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

Comments should be provided using this template. The completed comments form should be sent to RP-SATs@ema.europa.eu

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

Table of contents

1. Introduction and scope

Randomised controlled trials (RCTs) are the standard for providing confirmatory evidence on the efficacy of a new treatment. However, in a relevant proportion of marketing authorisation applications the pivotal clinical data stems from single-arm trials (SATs). This is observed across different therapeutic areas, including for rare diseases.

The purpose of this reflection paper is to outline the current thinking about SATs that are submitted as pivotal evidence for establishing efficacy in marketing authorisation applications. Defining general conditions under which SATs may be considered acceptable as pivotal evidence for marketing authorisation is outside the scope of this reflection paper. Such considerations strongly depend on the clinical context and other things such as the drug treatment modality. It is the responsibility of the applicant to adequately justify to regulators why a SAT, which deviates from the standard approach of providing pivotal evidence on efficacy through RCTs, can provide clear pivotal evidence of efficacy. Obtaining scientific advice is therefore strongly recommended to discuss whether pivotal evidence from SATs may be considered acceptable for seeking marketing authorisation for a specific development programme.

The assessment of efficacy is a relevant 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 and many of the critical considerations discussed equally apply to the assessment of safety.

Moreover, 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. Depending on the therapeutic area and the development programme, the primary objective of the SAT may differ (see Section 3). The key concepts described in this reflection paper apply to any of the objectives and contexts the SAT is used for. 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. Many of the considerations described also translate to SATs which are not submitted as pivotal evidence.

This reflection paper is structured as follows. 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 listing of relevant 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 experimental treatment and to be followed prospectively for a period of time. SATs can have specific design features, such as a monitoring period to obtain baseline data of the subjects before the start of treatment.

In general, the considerations in this reflection paper extend also 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 experimental arms are randomised, but without formal comparisons between the arms. An example for such a study would be a platform trial, where several treatment arms are included, but 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 patients 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 a causal interpretation as an effect of the treatment and must rely on knowledge external to the SAT to estimate the average outcome of the trial population if patients had not been treated with the experimental drug. In addition, it does not include a randomised comparison against a control arm that allows to directly quantify the effect of the treatment and the associated sampling variability. Thus, statistical methods to quantify the effect of the treatment and corresponding precision and interpretation of results must rely on assumptions about the population distribution of the outcomes without active treatment and on patient selection. As a consequence, the derived magnitude of effects is more difficult to interpret, and less reliable.

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

2. Relevant guidelines

This document 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)

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.

Estimands

The concept of estimands, 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 estimands framework are more difficult to apply, specifically in relation to the five estimand attributes:

- Treatment ('The treatment condition of interest and, as appropriate, the alternative treatment condition to which comparison will be made…', ICH E9(R1)): 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 derived from the SAT.
- Population: See Section 4.2.
- Variable (or endpoint): See Section 4.1.
- Handling of 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 arm which poses an additional challenge in relation to their interpretation and handling and even the timing of treatment initiation may be less clear than in RCTs.
- Population-level summary: See definition of treatment effect estimate in this section and Section 4.4.

Conceptually, appropriateness of a SAT depends on whether it can address the targeted estimand of interest. Specific problems associated with this are addressed in Section 4.

Treatment effect of interest

Following ICH E9, a treatment effect is 'an effect attributed to a treatment in a clinical trial. In most trials the treatment effect of interest is a comparison (or contrast) of two or more treatments'. For the purpose of this reflection paper, the term treatment effect of interest refers to the comparison (contrast) of the summary measure under the experimental treatment to the summary measure under the alternative of the trial population not being treated with the experimental treatment (counterfactual). This term is used in this reflection paper in the context of assessing whether there is an effect attributable to treatment and of (unbiased) estimation of the size of the treatment effect.

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 observed individual outcomes in a SAT for the defined endpoint within the designated follow-up could not have occurred without active treatment in any patient who entered the trial, 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 study design.

This is a theoretical concept which requires detailed knowledge of the clinical context. Specifically, there must be qualitative reasoning that leaves no doubt about the causal relationship between the treatment and outcome measured by the endpoint; which will only be fully satisfied in exceptional cases. In practice, observed individual outcomes are subject to bias and various sources of variability, for example, in terms of measurement or assessment. Hence, in contrast to RCTs, measurement errors or less stringent conduct of the SAT may erroneously lead to such observed outcomes and consequently the erroneous assessment that there is a treatment effect. There can also be residual uncertainty about which outcomes are truly impossible without treatment (such as level of motor function in spinal muscular atrophy patients). For other endpoints it is clear that they do not support the isolation of treatment effects in a particular setting.

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 (e.g., estimation of the Pearl index for contraceptives).

Treatment effect estimate

Statistical summary measures are used to estimate the treatment effect of interest. Some of the summary measures typically chosen for SATs (such as the percentage of responders) estimate the contrast to an assumed counterfactual (such as 0% responders). In other cases, treatment effect estimates defined for SATs may include contrasts to external control group data.

For regulatory decision making, a high degree of confidence in the estimates of the size of the treatment effects (favourable and unfavourable) is necessary. Treatment effect estimates based on SATs are directly impacted by the selection of patients included in the study. Even though the observed individual outcomes in an RCT may be equally prone to the selection of patients as in SATs, the RCT allows a direct comparison of treatment arms. Hence, multiple potential sources of bias throughout the design, conduct, analysis and reporting of SATs in the estimation of the treatment effect need to be addressed (see Section 4.5).

Internal validity

The internal validity of a SAT (compared to a well-designed RCT) can be conceptualised as the systematic difference between the treatment effect estimate from the SAT and the treatment effect estimate that would have resulted from the matching RCT had it been conducted in the same population and had the test treatment thereby been calibrated against a (placebo) control arm. This matching RCT can be understood as the target trial for the SAT. The absence of the randomised control arm substantially increases the risk of bias and thus reduces internal validity.

External validity

The external validity of SATs is characterised by the systematic difference between the treatment effect estimate from 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 and the trial population is not representative of the target population. For example, if the treatment effect is larger in biomarker positive patients and the proportion of biomarker positive patients 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 from SATs are equally impacted by heterogenous treatment effects. In addition, treatment effect estimates from 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 patients have a better disease prognosis regardless of treatment and the proportion of biomarker positive patients 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 appropriately known coverage probabilities. 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 special consideration. This is because only the variability of individual outcomes for the experimental arm is directly observed, but not for the for hypothetical control (see Sections 4.1 and 4.4).

4. General considerations for single-arm trial designs

In general, RCTs are the most suitable method to provide reliable estimates of clinical efficacy. However, in certain situations, evidence from SATs may be considered acceptable for marketing authorisation, and in such cases obtaining scientific advice is recommended. The following Sections describe important considerations related to the design, the conduct, the interpretation and the assessment of SATs that are presented as pivotal evidence for marketing authorisation.

4.1. 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 different endpoint characteristics like validity, reliability, feasibility, and accepted norms and standards in the relevant field of research. For a SAT the primary endpoint must also be able to isolate treatment effects (see Section 3), i.e. it is required that the primary endpoint is such that it is known that observations of the desired outcome would occur only to a negligible extent (in number of patients or size of the effect) in the absence of an active treatment.

Any uncertainty whether observed individual outcomes are undoubtedly caused by the treatment severely complicates the interpretation of results derived from a SAT. In particular, these uncertainties can lead to concerns that results just 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 not be clear to what extent selection bias due to overrepresentation in the trial population of patients with higher likelihood of remission in the absence of treatment could lead to false positive conclusions on efficacy. In addition, measurement error or misclassification might result in erroneously recording a certain individual outcome in the SAT and, due to the lack of a comparator within the trial that is equally affected, unduly favour the experimental treatment. In other situations, the disease may be episodic, characterised by a waxing and waning course. In such cases, all relevant primary endpoints 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 of a clinically relevant treatment effect needs to be discussed on clinical grounds. The acceptability of a SAT and its primary endpoint strongly depend on the clinical context and mechanism of action of the drug and are therefore a case-by-case and 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 active treatment. For this reason, observed individual outcomes on such endpoints generally cannot be attributed to treatment and therefore time-to-event endpoints are usually not suitable to be used in SATs. Exceptions could be endpoints that measure time to positive events that cannot occur at all without treatment. 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 comparator arm provides an internal calibration for the patients' history at risk prior to enrolment in the trial, which is however lacking in SATs.

The impact of the course of a disease on time-to-event endpoints is usually highly 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 experimental treatment) and predictive factors (i.e. differences in treatment effects for the experimental treatment) cannot be achieved based on the results derived from a SAT (see Section 4.2).

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 assessment. Continuous endpoints allow for a precise and sensitive measurement of the changes that patients experience during the trial. However, when the individual outcome can change due to within-patient 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. A common phenomenon is 'regression to the mean' which may result from a combination of measurement error, within–patient variability and patient selection at baseline (see Section 4.5). For example, in case patients are selected based on disease severity as expressed by a low value of a specific endpoint at the time of inclusion in the trial (eligibility criterion), the measurements of the same patients will have a tendency for improved values at a later point in time, irrespective of being treated with an effective treatment or not.

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 specific diseases where a certain state does usually 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, then 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 pre-specified threshold 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 thresholds at the trial planning stage might make the SAT results difficult to interpret regarding the representation of 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 threshold. In specific cases it may, however, be possible to set the threshold in advance in a way that crossing it is not possible without treatment for any patient, even after accounting for potential sources of bias (as discussed exemplarily above and in Section 4.5).

4.2. Target and trial population

Recruiting an adequate trial population is required to ensure that conclusions about the effects of the experimental 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 (external validity), concerns about external validity are in general larger for SATs as compared to RCTs, because the treatment effect is not directly estimated relative to a control and the composition of the trial population is especially relevant for estimates from a SAT. In this regard, importantly, the trial population determines the plausibility of assumptions about the disease course of a hypothetical control group or the comparability with an external data source (see Section 4.3).

The assumptions on the natural course of the disease must apply for the trial population in the SAT. In practice, this means that the 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 validate. As a consequence, the interpretation of results derived from SATs becomes even 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 patients; both due to investigator decisions as well as patients' choices, or even criteria related to selection of study 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, specification 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, and about the subjects who were not selected.

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 pre-defined 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 derived from SATs.

Biomarker-defined populations are one important example for the choice of target and trial population, where interpretation of results derived 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 specific 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 within the targeted subgroup is available. The role of the biomarker and its contribution to an effect of the treatment cannot be established within the SAT.

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 drug are well understood. Furthermore, strong prognostic factors may raise concern regarding selection bias and strengthen the need for a randomised concurrent control.

4.3. Role of external information

Due to the lack of a comparator within the trial, the role of relevant external (extra-study)information is critical for the interpretation of the results derived from a SAT. External information may take the form of (i) general knowledge about the natural course of the disease, e.g. that an endpoint will not change without active treatment, or (ii) external clinical data. Use of external information in the analysis or interpretation of a SAT is a crucial design element and should be pre-specified in the study protocol. Most importantly, any external information used to describe the hypothetical control condition (counterfactual) of the SAT should include a precise and a priori definition and description of the control condition(s) to be covered. It is strongly recommended to seek scientific advice on the use and the choice of external information before the study protocol of the SAT is finalised.

In some cases, when the treatment effect is clinically dramatic, occurs rapidly following treatment, and is unlikely to have occurred spontaneously, this may be sufficient to consider that isolation of the treatment effect as well as clinical relevance are demonstrated. Assessment in the context of use may then be based on general knowledge about the disease and the target population. In other cases, external information may be used to establish a threshold for efficacy that can be demonstrated to fulfil the conditions that support isolating a treatment effect (see Section 4.4).

In exceptional cases, the assessment of efficacy is envisaged to be informed by a direct comparison against external clinical data (i.e. an external control). Guidance on the choice of and comparison with external data is beyond the scope of this reflection paper. While methods that directly incorporate external data into the analysis come with a promise to provide useful insights and potentially reduce bias, they add complexity to pre-specification and rely on additional assumptions that are often not transparent. Consequently, approaches that directly incorporate external data should be carefully evaluated on a case-by-case basis.

4.4. Statistical principles

General Principles

Exploratory and confirmatory trials have different requirements regarding statistical rigour, and it is acknowledged that SATs are used for all treatment development phases. If SATs are submitted as pivotal evidence, best practice and strict criteria should be followed at planning and conduct of the trial, and the assessment will need to follow standards that apply to the confirmatory setting (ICH E9). Due to the lack of safeguarding mechanisms like randomisation and blinding (see Sections 3 and 4.2), choices in the statistical analysis approach after trial initiation can have a larger impact on the reliability of the results and the assessment of SATs that are used for pivotal evidence as compared to RCTs.

Predefinition

As with confirmatory RCTs, SATs which are submitted as pivotal evidence are expected to have an a priori definition of a clear success criterion. Such a criterion needs to be justified based on suitable external information (see Section 4.3) such as knowledge about the disease, uncertainties around the variability of (primary) outcomes and treatment effect estimates and are ideally pre-agreed with regulators.

Unplanned changes to trials are always problematic. Therefore, predefinition and adherence to the study protocol when the trial is ongoing are critical. This is even more pronounced in SATs. Due to the unblinded nature of SATs, claims on existing firewalls can hardly overcome concerns on potential data knowledge and any amendment is considered potentially data driven. This includes unplanned interim analyses, changes in endpoints, changes in or deviations from the planned number of patients (sample size changes), changes in the dosing regimen, changes in eligibility criteria, subgroup selection, or treatment arm selection (e.g. in a platform trial). In the context of regulatory decision making, an especially critical unplanned change is the post-hoc designation of a trial planned as exploratory phase II trial to a pivotal trial once trial data were available and to submit this as primary confirmatory evidence. Due to the unblinded nature of SATs, also planned data-dependent actions can be considered critical.

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 study 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.

Analysis Set

Predefinition of the primary analysis set is of utmost importance and bias due to inclusion or exclusion of patients in the analysis set based on observed individual outcomes should be avoided. Therefore, the full analysis set, i.e. all subjects that entered the SAT upon providing informed consent, should be used as the primary analysis set. Situations may exist, however, where the analysis based on the full analysis set may bias 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 respective disease at study end, were incorrectly included into the SAT. This situation can also occur, when measurements to select patients for presence of a disease (state) are different to those that are used for assessing the changes of the disease (state) during the trial (e.g. response or resolution), and this situation is comparable to measurement error as discussed in Section 3 and 4.1. Such cases should be avoided by study 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 study. In particular, an individual outcome cannot be attributed as a response to treatment if the patient, who was selected based on a baseline measurement, would have been considered disease-free at baseline if the outcome measurement had been used. Only for cases where the inclusion of the patient results in an optimistic estimate of the response, it should be predefined that such subjects are excluded from the primary analysis set. A further exemption may become necessary when an analysis is done before all patients have reached a pre-defined analysis timepoint (see below 'Missing data').

Missing Data

With respect to missing data, 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, patients who did not complete the pre-planned individual end-of-study timepoint, but for whom it is already known that they fail should be included as failures in the analysis. On the other hand, in a study with an interim analysis, patients who have not yet reached their individual end-of-study time point without having failed by then should not be included in the primary analysis because they should not be counted as non-failures. Sensitivity analyses are encouraged (see ‘Sensitivity Analyses’).

Analysis and Estimation

All analyses should be pre-defined in a detailed statistical analysis plan before the SAT starts, i.e. before inclusion of the first patient. For the statistical analysis of a SAT, applicable non-parametric or parametric statistical methods may be applied.

The statistical analysis model 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 4.2), 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 factor levels are dealt with in the statistical analysis model for estimation. Due to the comparison against a randomised control this is not a problem in RCTs when using linear models (while 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’). While this issue is also related to selection bias, the operational handling of data cannot 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 different results, this may question the overall reliability of the study result.

Interpretation of results

For endpoints that unambiguously isolate the drug effect, the statistical analysis can be fully based on the data resulting from the SAT. In some cases, a threshold can be pre-specified which the summary measure at the trial population level must exceed to ensure that the summary measure observed in a SAT on such an endpoint reflects clinical benefit at the target population level. Such thresholds can be based on external clinical information, which however bears the inherent risk of erroneous conclusions due to comparing results across different databases. In any case, the basis of the threshold needs to be given upfront and its clinical validity in the therapeutic context needs to be carefully justified.

In the case of endpoints for which the desired individual outcomes are expected to occur to a negligible extent in the absence of treatment, variability may be observed such that the drug effect 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 versus the treatment effect (see Section 3). Often a pre-defined threshold to surpass is set in order to convincingly demonstrate efficacy. Conceptually, if that threshold was the corresponding summary measure under the hypothetical scenario of no treatment, then the size of the effect attributable to the treatment would only be the difference between this threshold and the summary measure observed in the study. However, such 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 'Multiplicity'). In addition, it is noted that the defined threshold will usually not truly be known as a constant but 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 in its definition and a sufficiently conservative threshold should be chosen. For example, in some settings 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 data. Moreover, results depend on the selection mechanism in the trial. Overall, this makes the choice of a threshold difficult to justify. Therefore, such a scenario represents a critical risk for the interpretation of a SAT.

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 derived 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.

Sample size

As for any other study 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 experimental treatment, uncertainty with respect to bias may outweigh any gains in precision compared to a randomised controlled design.

4.5. 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 derived 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 derived from a SAT.

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

Type of bias	Description	Potential bias reduction measures
Ascertainment bias	External information will likely differ in certain aspects of trial conduct and data collection, e.g. different frequencies and standards for outcome assessment than for the data collected within the SAT.	When using external information, the ascertainment and data collection practices need to be comparable to the practices from the SAT.
Assessment bias	Knowledge of the therapy can influence the outcome assessment.	Endpoints in SATs should be sufficiently objective and, if possible, assessments should be made independently and preferably unaware of timing in relation to treatment.
Attrition bias	Attrition of patients and missing data in general constitute an additional source of confounding that is difficult to resolve.	Avoid missing data through study design and conduct. Pre-specify methods for handling of missing data that do not overestimate the response to treatment and conduct suitable sensitivity analyses. When using external data, the data needs to be of high quality with follow-up including all patients, to avoid bias due to missing data.
Bias due to lack of pre-planning	Any post trial-initiation changes in design, conduct and planned reporting (e.g. in statistical analysis plan, adaptations in treatment, follow-up, protocol amendments on inclusion or exclusion criteria, 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	Patients selected based on their outcome values during the monitoring period are expected to show improved outcomes due to regression to the mean.	Define target population independently of disease severity during pre-treatment monitoring period. 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 drug 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.
Calendar time bias	Standard therapy and trends in the overall management of the disease may change the disease course and individual outcomes over time. In SATs the impact of these trends cannot be disentangled from the treatment effect.	Use of contemporaneous external information.
Immortal time bias	Study or treatment start relative to previous studies or external data is difficult to determine as an anchor for patient specific time scale.	The start time of being at risk (time 0) needs to be clearly defined, should time-to-event endpoints including a comparison to external data be required. Moreover, sensitivity analyses are needed.
Intercurrent event bias after study entry	Failure to clearly define the main estimand(s) and intercurrent events of interest at the trial planning	Follow ICH E9 (R1), anticipating the intercurrent events at trial planning stage and ensuring the definition of estimand(s)

	stage carries the risk that the clinical question of interest cannot be addressed.	as well as detailed collection of information on intercurrent events.
Retrospective selection bias	Retrospective selection of external information to use as reference and the specification of key analysis features post trial-initiation carry risk of introducing bias.	Pre-specification of the use of external information, including details of the statistical analysis, prior to trial start. The statistical analysis plan needs to be finalised before trial initiation.
Selection bias in relation to the hypothetical control group	Patients enrolled in a SAT may systematically differ from the hypothetical control group in ways that impact their prognosis.	Precisely pre-specify inclusion and exclusion criteria such that the enrolled trial population matches well the external information that assumptions are based on.
Selection bias in relation to the target population	Patients 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 pre-specify 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	Patients selected based on a pre-defined biomarker for targeted treatment may differ in prognosis compared to the full population.	Ensure prognosis of the biomarker targeted patient subgroup is known sufficiently accurate prior to study start.
Stage migration bias	The improvement of assessment methods leads to improvement in prognosis of both earlier and later stages.	Ensure that the same assessment methods are used in the SAT as in the sources of external information; demonstrate a magnitude of effect that exceeds the maximum possible effect of stage migration effect.
Study bias	Patients in a SAT may have systematically different outcomes (independent of the experimental treatment) than in the targeted clinical practice, e.g. due to different care in the trial setting.	Assure and show that the auxiliary care reflects the current standard.