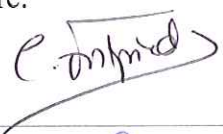



Phase III / Confirmatory Classical Designs

ST-001-WIN-03

Version: 2

ALWAYS REFER TO INTRANET TO CHECK THE VALIDITY OF THIS DOCUMENT

Author <i>Senior Biostatistician</i> Catherine Fortpied	Signature: 	Date: 7 Dec 2015
Approved and authorized by <i>Methodology Director</i> Jan Bogaerts	Signature: 	Date: 7 DECEMBER 2015

This document is the property of EORTC.
No release of this document is granted for any use without the written agreement of the EORTC.

Contents

1	PURPOSE.....	5
2	Definition of Phase III studies	5
3	Objectives	5
4	Patient selection criteria.....	5
5	Randomization and Stratification	6
5.1	Basic principles	6
5.2	Studies with multiple randomizations.....	6
6	Blinding	7
7	Endpoints	7
7.1	Recommended primary endpoint, by disease setting.....	8
7.1.1	Early stage disease studies.....	8
7.1.2	Adjuvant studies	8
7.1.3	Advanced disease studies	8
7.1.4	Palliative setting.....	8
7.2	Definition of Endpoints.....	8
7.2.1	Time to event endpoints.....	8
7.2.2	Response and related endpoints.....	9
7.2.3	Safety endpoints.....	10
7.2.4	Quality of Life endpoints.....	10
7.2.5	Other endpoints.....	10
8	Superiority, equivalence and non-inferiority studies.....	11
8.1	Targeted treatment effect in superiority/difference studies	11
8.2	Margin in equivalence and non-inferiority studies	12
9	Sample size calculation.....	12
9.1	Number of events (time to event endpoint) / number of patients (binary endpoint)	12
9.2	Number of patients (time to event endpoint)	13
9.3	Assumptions underlying the sample size calculation for time-to-event endpoints.....	14
9.4	Sample size adjustment during the course of the study	14
9.4.1	Sample size adjustments not changing the total information	14
9.4.2	Sample size adjustments changing the total information without access to interim efficacy treatment comparisons	15
10	General considerations on early stopping rules and interim analyses	15

11	Stopping rules for efficacy and/or futility	16
11.1	Stopping rules for rejection of H ₀ only, of both H ₁ or H ₀ and of H ₁ only.....	16
11.2	Overview of methods used at the EORTC	16
11.3	Repeated significance testing.....	17
11.4	Spending function boundaries.....	18
11.4.1	Basic principles.....	18
11.4.2	Early stopping for superiority in a superiority study (early rejection of H ₀)	20
11.4.3	Early stopping for futility in a superiority study (early rejection of H ₁).....	20
11.4.4	Early stopping rules in non-inferiority or equivalence trials	20
11.5	Stochastic curtailment based on conditional power	21
11.6	Number and timing of interim analyses	22
11.7	Planned versus actual information time at interim analysis.....	22
11.8	Binding versus non-binding stopping futility boundaries.....	23
12	Stopping rules for safety/toxicity.....	23
12.1	Basic principles	23
12.2	Comparative stopping rules	24
12.3	Non-comparative stopping rule.....	24
13	Designs involving multiple arms, multiple hypotheses, subgroups	24
13.1	More than two treatment arms	25
13.2	Factorial designs	26
13.3	Multiple primary endpoints, multiple hypotheses or comparisons	26
13.4	Studies with targeted therapies	27
13.4.1	Study population: “all-comers” versus “marker positive patients”	27
13.4.2	Targeted design.....	28
13.4.3	Biomarker stratified design.....	28
13.4.4	Randomized strategy designs	29
14	References.....	31
14.1	On sample size calculation.....	31
14.2	With competing risks	31
14.3	On non-inferiority designs	31
14.4	On stopping rules and interim analyses	32
14.5	On multi-arm studies.....	33
14.6	On targeted designs	34
14.7	On Bayesian designs	35

14.8 On trial designs for rare cancers 35

15 ASSOCIATED DOCUMENTS..... 35

16 DOCUMENT HISTORY 35

1 PURPOSE

The current work instruction details various aspects of the statistical design of EORTC confirmatory studies as a function of the study objectives

This WIN covers classical phase III designs, including design adaptations considered as “well-understood” according to FDA Guidance on Adaptive Design Clinical Trials for Drugs and Biologics (2010). For designs with adaptive methods considered as “less well-understood”, according to the same FDA Guidance please refer to the published literature.

2 Definition of Phase III studies

Phase III studies are randomized comparative studies using a low type I error (5% two-sided or less), and proper power (80% or more) against H_0 , at an alternative hypothesis that is both realistic (i.e. possibly attainable) and clinically meaningful. Because of these properties, such studies are potentially practice-changing but generally of large sample size.

These studies must be randomized in order to reduce the possibility of systematic bias and to reduce the risk of random errors (Type I and II). One of the randomized arms must unambiguously be considered a control arm; if possible this should be a treatment that is considered as standard of care for the study population.

For rare cancers, less strict trial designs that may however be considered to be potentially practice-changing in these circumstances, and that have no plan for a further follow-through study if successful, can be designated Phase III as well. Bogaerts et al (2015) describe possible approaches to design trials in rare cancers.

3 Objectives

The objective(s) of a phase III study is (are) comparative, between randomized arms. Primary objectives are generally one of the following:

- To determine the effectiveness of a treatment relative to the natural history of the disease. In this case the study has a no treatment or a placebo control arm (difference or superiority study)
- To determine the effectiveness of the new treatment as compared to the best current standard therapy (difference or superiority study)
- To determine if a new treatment is as effective as the standard therapy but is associated with less severe toxicity or a better quality of life (equivalence or non inferiority study)

4 Patient selection criteria

The eligibility criteria are selected keeping the following principles in mind:

- The criteria should ensure that the primary endpoint will be observable in eligible patients
- Restrictions to ensure observability of endpoints other than the primary are extraneous. For example, if overall survival is the primary endpoint, there is no reason to restrict accrual to patients with measurable disease; in this case the analysis of progression-free survival can be stratified for patients with/without measurable disease or alternatively be restricted to patients with measurable disease.

An appropriate balance needs to be found between narrow and broad selective criteria:

- Narrow eligibility criteria make the patient sample as homogeneous as possible thus resulting in more precise estimations of treatment effects; they are preferable when there are a priori reasons to believe that the treatment is beneficial only to a subgroup of patients.
- Broad eligibility criteria are preferable to narrow ones when little is known at the time of designing the study concerning the possible benefit expected from the treatment under investigation in different subgroups of patients (e.g. those with a specific genetic / biomarker profile). In this (common) situation, the most plausible hypothesis is that the treatment effect will be similar in different subgroups of patients. Broad eligibility criteria allow the broadest generalizability of results to the population of interest.

Please refer to section 13.4 for a more specific discussion on targeted designs.

Adaptation of study eligibility criteria does not affect the integrity of the study if it is based on analyses of pretreatment (baseline) data or on information from an external source.

5 Randomization and Stratification

5.1 Basic principles

Treatment allocation by a random process is the method of choice in phase III studies. In case a phase III study does not include randomization for the primary objective, the Statistician writes an explicit justification for this deviation. The randomization should be stratified for a small number of reliable factors of known prognostic value. For all information and details on randomization in EORTC studies, refer to ST-002-SOP "Randomization/Registration", and for technicalities of ORTA implementation to ST-002-WIN-01 "Registration/Randomization Procedure in ORTA" and ST-002-WIN-02 "Randomization and stratification of clinical studies".

The Statistician carefully considers the following elements, and describes the first four in the outline and protocol. The unpredictability settings are part of the ORTA implementation and will not be described in the protocol:

1. Timing of the randomization: in order to minimize non-compliance and differential effects between randomized arms, the randomization has to be performed as close as possible prior to the start of the randomized treatment. In some studies, the setup is such that the assigned treatment arm will not take effect in all randomized patients. In such cases, a temporary blinding (until start of treatment) can be considered.
2. Method of randomization: dynamic minimization algorithm or permuted block method.
3. Randomization ratio.
4. Stratification factors: by default, EORTC Phase III studies are stratified by institution. All studies should also be stratified by at least one factor in addition to institution, in order to avoid predictability. Stratification factors should be taken among key patient, disease and prior treatment characteristics that need to be balanced to avoid confounding (as a statistical term).
5. Elements of balance and unpredictability: the Statistician determines the parameters of the dynamic minimization algorithm or permuted block method which provide adequate unpredictability of the allocations and balance over the treatment arms (ST-002-WIN-002).

5.2 Studies with multiple randomizations

If more than one randomization are planned for a Phase III study, the following must be considered:

- Later randomizations have to be stratified for the treatment allocation of the earlier one(s).
- In studies where the second randomization is conditional on the outcome of the treatment given per the first randomization (e.g. only responders have the second randomization) the primary endpoint for the first randomization is recommended to be an endpoint which is assessed prior to the second randomization (such as response to treatment). Comparisons of endpoints for the first randomization that are assessed after the second randomization may be biased if there is a difference in treatment efficacy between the treatments assigned by the second randomization when they are given only to the responders.

6 Blinding

Blinding study team (including the Statistician), investigator and patient to treatment allocation is becoming more common in oncology. It is a gold standard for randomized clinical studies. It must be considered as a possibility, and when done, will improve the scientific value of the study results. It should specially be considered if there is concern for operational bias, such as:

- the primary endpoint is subjective (e.g. reduction of pain or in the consumption of analgesics, etc.)
- oral drug therapy in which one treatment arm is compared to a no treatment control arm (where blinding means the use of an oral placebo)
- in studies in which two different dose schedules of the same drug are tested.

Factors rendering blinding unfeasible can include:

- harm or undue risk to a patient
- treatments of very different nature or toxicity. E.g. studies of cytotoxic drugs are not usually blinded because complicated dose schedules, the likelihood of serious side-effects and the need for dose modifications make blinding impossible.

The protocol will describe how the blinding is implemented in the trial:

- Who is blinded to treatment allocation,
- Measures to preserve blinding e.g. in connection with randomization procedure and drug supply,
- Planned unblinding during the course of the study (e.g. at interim analysis)
- Procedure for unblinding in case of emergency.

For operational implementation of a blind study, please refer to CM-011-SOP.

7 Endpoints

The primary endpoint of a phase III study should measure the efficacy of the treatment in terms of clinical benefit relevant for the patient.

The sections below provide general recommendations on the selection and definition of endpoints. Specific studies may justify using different endpoints.

7.1 Recommended primary endpoint, by disease setting

7.1.1 Early stage disease studies

Early stage disease studies evaluate potentially curative primary therapies, which are generally local therapies (radiotherapy, surgery) with or without systemic therapy.

Generally, the primary endpoint is a locally defined endpoint (such as time to local or loco-regional recurrence/progression).

7.1.2 Adjuvant studies

Adjuvant therapy is given to patients treated by a potentially curative primary therapy, but for whom there is a substantial risk of recurrence. Adjuvant studies study potentially curative treatments (surgery, radiotherapy, chemotherapy) given to patients who are clinically disease free after a primary treatment (usually surgery).

Generally, in adjuvant studies, long term endpoints such as overall survival and relapse/recurrence free survival are recommended. In studies with a particular question on local therapy (radiotherapy, surgery), possibly a locally defined endpoint (such time to local or loco-regional recurrence/progression) may be selected.

7.1.3 Advanced disease studies

Advanced disease includes all patients for whom local treatment is no longer curative. There are two types of advanced disease: recurrent or locally advanced disease in which the disease is still confined to the region of the primary tumor as opposed to metastatic disease where the disease has spread to distant sites.

For Phase III studies in this setting, overall survival is the preferred endpoint.

Progression free survival is acceptable if it has been proven to be a surrogate endpoint for overall survival. Progression-free survival may also be used when a design based on overall survival would lead to an unfeasible study in terms of number of patients and study duration; or when subsequent treatment post progression has a significant impact on patient overall survival.

7.1.4 Palliative setting

In palliative care studies, the objective is to improve the quality of life of the patients.

Primary endpoints for these phase III studies are generally symptom control together with quality of life evaluations.

These non-clinical endpoints may also be used as secondary endpoints in phase III studies in other settings.

7.2 Definition of Endpoints

7.2.1 Time to event endpoints

In phase III studies the primary endpoint is most often the time to an event.

The Statistician ensures that a proper and unambiguous definition is provided of all planned time to event endpoints in the protocol, notably:

- Starting time point of counting the time to event: in most cases this is the date of randomization. Use of the date of surgery or start of treatment as the starting point may be biased if patients did not receive their randomized treatment or if delays between randomization and start of treatment can be expected.
- Which types of observations will constitute events of interest: depending on the tumor type, clear description will need to be provided of which progressions, relapses or recurrences will be counted.
- Which types of observations will constitute competing risks, i.e. where observing one type of event may preclude observing the event of interest. This may be the case, for example, if the endpoint is death due to malignant disease or local-regional recurrence. The use of local-regional recurrence as an endpoint may prove difficult if a large majority of patients develop distant metastases as the first evidence of relapse. In this case the disease free interval or disease free survival is to be preferred.
- Censoring date: If the event of interest or competing risk event has not yet been observed at the time of the last available information, then the patient is censored at this date.

Examples of definition of time-to-event endpoints:

- Time to loco-regional recurrence (early stage or local therapy): the time interval between the date of randomization and the date of first local or regional progression. Distant recurrence/progression and second cancers diagnosed before locoregional recurrence and death in absence of locoregional recurrence are considered as competing risk events. Patients without any of the listed events (i.e. events of interest or competing risks events) are censored at the date of the last follow-up examination.
- Disease free survival (adjuvant studies): the time interval between the date of randomization and the date of disease recurrence or death, whichever comes first (the period during which there is no evidence of disease activity). If neither event has been observed, then the patient is censored at the date of the last follow up examination. Again, the description of what constitutes "recurrence" needs to be further clarified, depending on the tumor type.
- Disease free interval (adjuvant studies): the time interval between the date of randomization and the date of first disease recurrence (the period during which there is no evidence of disease activity). For patients who die prior to progression, death is analyzed as a competing risk. If no disease recurrence or death has been observed, the patient is censored at the date of the last follow up examination. Progression free survival (advanced disease studies): the time interval between the date of randomization and the date of disease progression or death, whichever comes first. If neither event has been observed, then the patient is censored at the date of the last follow up examination. The description of what constitutes "disease progression" needs to be further clarified, depending on the tumor type.
- Time to progression (advanced disease studies): the time interval between the date of randomization and the date of disease progression. Deaths in the absence of progression will be interpreted as competing risks. If no progression or death has been observed, the patient is censored at the date of the last follow up examination.
- Duration of survival (synonym: overall survival): the time interval between the date of randomization and the date of death for any cause. Patients who were still alive when last traced are censored at the date of last follow up. A separate analysis of death due to malignant disease may be carried out where deaths due to other causes are censored or treated as competing risks), but the main analysis should include deaths due to any cause.

7.2.2 Response and related endpoints

Generally, response to treatment should not be taken as a primary endpoint in Phase III studies.

The following endpoints related to response to treatment may be selected as secondary endpoints:

- Objective response to treatment: this is an indicator of anticancer activity, but is generally not considered as a valid surrogate for end-points quantifying therapeutic benefit (i.e. Phase III). Standard methods for assessing objective response are defined by the RECIST criteria, in case of solid tumors; by Cheson criteria for Lymphoma; etc.... As a general rule, when computing response rates, the denominator of the response rate must include all patients and not only those evaluable for the response.
- Duration of complete response (complete responders only): the time interval between the date that the criteria of complete response are met for the first time and the date of first disease progression after complete response.
- Duration of overall response (complete and partial responders taken together): the time interval between the date that the criteria of complete or partial response are met for the first time and the date of first disease progression after complete/partial response.

7.2.3 Safety endpoints

Safety endpoints are generally not selected as primary endpoints in a phase III study. They are analyzed as part of the overall safety evaluation of the studied treatments. Safety endpoints are generally not formally compared using a statistical test of significance; if a formal comparison between arms is to be made, the protocol will specify the significance level to be used.

Some safety endpoints of special interest may be selected for in-depth evaluation, depending on the disease and the treatment, the availability of earlier clinical trial data, the studied population, the mode, dose and schedule of administration. In this case, the protocol will specify which toxicities will be analyzed in depth, preferably using published guidelines such as CTCAE. In general, all adverse events, whether related or not related to study drug, are included in the definition. This is especially important in open-label studies.

If the onset of adverse events is assessed as a function of time or of the cumulative dose of a drug, the “time to onset” or “dose to onset” of the adverse event may be defined as an endpoint. In such cases, patients starting a new therapy are generally censored at the date of treatment change.

For chronic toxicity, none of the existing scales may be considered as a validated standard. For studies evaluating chronic toxicity, appropriate scales must be included in the protocol. See the EORTC Investigator’s Handbook for a description of the different scales. For chronic toxicity, the “time to onset” or “dose to onset” is generally used as an endpoint. After progression of the disease, chronic toxicity is often difficult to assess (at least if they are local) and progressing patients are generally censored in this type of analysis.

The “time to event” approach is preferred for the analysis of late toxicity.

7.2.4 Quality of Life endpoints

Quality of Life endpoints are generally not selected as primary endpoints in a phase III study, but as secondary endpoints. The protocol will specify the Quality of Life instrument used and the scales, items and derived variables that are used as endpoints.

7.2.5 Other endpoints

Whatever other endpoint may be used, the Statistician takes care that they are properly and unambiguously defined in the protocol without introducing bias between randomized arms.

8 Superiority, equivalence and non-inferiority studies

A distinction is made between studies attempting to show a difference in therapeutic effect (superiority studies, one sided, or difference studies, two-sided) and studies designed to show non inferiority (one sided) or equivalence (two sided).

In studies attempting to show a difference (superiority study), one attempts to reject the null hypothesis of no difference ($d = 0$) in treatment efficacy versus the alternative $d > 0$, generally based on a two sided test.

If it is desired to show that a new more conservative treatment is equivalent (or not inferior) in efficacy to a standard more intensive therapy, one will select the more conservative treatment as being equivalent (or not inferior) if it is not different from (worse than) the standard treatment by more than some small amount Δ judged to be acceptable by the investigator. One tests the null hypothesis that the standard therapy is actually different from (more effective) than the new therapy by at least some margin $\Delta > 0$. The alternative hypothesis is typically that there is no difference; in some circumstances the alternative hypothesis is that a small benefit in favor of the experimental arm is present. For a discussion of this approach, please refer to Freidling B, Korn EL, George S, Gray R.(2007).

This is summarized in the following table:

Testing to show	Null Hypothesis	Alternative Hypothesis
Difference	No difference	$ \text{Difference} > d > 0$
Equivalence	$ \text{Difference} > \Delta$	$ \text{Difference} < \Delta$
Non inferiority	$\text{Difference} > \Delta$	$\text{Difference} < \Delta$ ($\Delta = 0$)

When interpreting the results of a study attempting to show a difference in efficacy between two treatments, a non significant result does not imply that the two treatments are equivalent. Confidence intervals around the estimated treatment effect can provide information about the possible size of the treatment effect; however they can only be used to assess non-inferiority if the margin Δ has been prospectively defined.

Similarly, when a study is designed to demonstrate non-inferiority and fails to demonstrate it, it cannot be interpreted that the new more conservative treatment is worse than the standard treatment. This would require a difference test (in this case: inferiority) looking at one bound of the confidence interval, while the non-inferiority test is a one-sided test looking at the other bound of that confidence interval. Such a test cannot be formally interpreted, unless it was prospectively planned and proper adjustment for multiple testing was done.

8.1 Targeted treatment effect in superiority/difference studies

The targeted treatment effect d must satisfy the following criteria:

- The effect must be one that could occur in a realistic scenario. A treatment effect that would represent a major medical breakthrough is generally too optimistic to use as a realistic basis for power calculation.
- The effect has to be clinically worthwhile. What is clinically worthwhile depends on the frequency and severity of the disease and on the balance between efficacy and undesirable adverse reactions of the compared treatments.

Examples:

A small improvement gained at the expense of much toxicity may not be worthwhile.

More than a small survival improvement with an adjuvant therapy in good prognosis patients is probably not realistic. Still a small improvement in the adjuvant setting can be clinically relevant.

8.2 Margin in equivalence and non-inferiority studies

The margin Δ must satisfy the following criteria:

- The margin should be small, reflecting a difference clinically non-relevant
- The margin should be small, i.e. no more than one third or one half the magnitude of effect size that was shown when the standard treatment was established. The effect size in the standard treatment must be well documented.
- The choice of the margin should be balanced with the advantages of the new treatment in terms of safety or other secondary endpoints

9 Sample size calculation

The EAST software is used for the calculation of sample size for Phase III studies. This section deals with sample size calculation for designs with no stopping rules. The design of studies with early stopping rules is discussed in sections 10 and 1112.

9.1 Number of events (time to event endpoint) / number of patients (binary endpoint)

The sample size calculation depends on the primary endpoint of the study, the objective of the study (superiority, non-inferiority, equivalence) and on the statistical analysis technique that will be used to analyze this endpoint. The number of patients entered into a phase III study (the sample size) is calculated to ensure a reasonable power (≥ 0.80) to reject the null hypothesis at a pre-specified significance level (α) which should be sufficiently small (usually 5% 2-sided for superiority test, 2.5% 1-sided for a non-inferiority test).

For continuous or binary endpoints, the power depends on the number of patients. For time to event endpoints the power depends on the number of events observed, not on the number of patients entered. The number of events for time to event endpoints (duration of survival) and the number of patients for binary endpoints (response rate, percent with unacceptable toxicity) depends on the following factors:

- A prior estimate for the primary endpoint in the control arm.
- Estimates of the targeted treatment effect or the equivalence/non-inferiority margin (see section 8).
- The test statistic used to compare the treatment arms.
- The size of the type I error α (false positive rate; ≤ 0.05 ; two-sided except for non inferiority studies where it is recommended to take a one sided $\alpha = 0.025$) and type II error β (false negative rate; ≤ 0.20). $1 - \beta$ is called the power. If one plans to have asymmetric futility testing at interim, it is more straightforward to describe the sample size from a one-sided $\alpha = 0.025$ perspective. The analysis can then specify that two-sided 95% confidence intervals will be provided. This process is equivalent to the two-sided 5% version, allowing a "natural" futility test.

For example, to detect a difference in response rate, assuming a response rate of 50% in the control arm, the number of patients required on each arm for a two sided test, with type I and type II errors of 5% and 20% respectively, is approximately:

Number of Patients Per Arm	Difference
1565	5%
385	10%
170	15%
90	20%
60	25%

For a time to event studies, the targeted treatment effect or the equivalence/non-inferiority margin is expressed as a ratio of medians or as a hazard ratio. The total number of events required to detect such ratio, for a two sided test, with type I and type II errors of 5% and 20% respectively, is approximately:

Number of Events	Median Ratio	Hazard Ratio
3460	1.1	0.91
945	1.2	0.83
630	1.25	0.8
460	1.3	0.77
280	1.4	0.71
192	1.5	0.67

As the same absolute difference will yield a different median/hazard ratio depending on the baseline event rate, care must be taken when specifying absolute event rates to determine the plausibility of the resulting hazard ratio.

Non inferiority studies and equivalence studies require considerably more events than superiority studies since the margin in such studies should be substantially lower than the effect size of the active comparator from previous superiority studies (see section 8.2). The two confidence interval procedure proposed by Rothmann is not felt to be feasible due to the large sample size that is required.

9.2 Number of patients (time to event endpoint)

For time to event studies, a sufficiently large number of patients must be entered and followed for a sufficiently long time in order to observe the required number of events. Once the required number of events is determined, a compromise is made between the number of patients and the duration of accrual/study, the two extremes being of entering only the same number of patients as the number of events that are required (minimizes the number of patients, maximizes the duration of the study) or continuing to enter patients until the required number of events is observed (minimizes the duration of the study, maximizes the number of patients). The link between number of patients and study duration depends on the expected accrual rate.

A piecewise exponential accrual rate over the first year will be used to account for progressive opening of trial sites.

To account for possible loss to follow up in studies requiring long term follow up, the required number of patients should either be increased by 5% in metastatic studies and 10% in adjuvant studies or the duration of follow up extended in order to ensure that the required number of events will be observed.

9.3 Assumptions underlying the sample size calculation for time-to-event endpoints

When calculating the sample size and number of events, the simplest assumption is an exponential distribution (i.e. constant event rate) and proportional hazards. The following different scenarios are not uncommon:

- When the event rate decreases over time: in this scenario, additional follow up will not necessarily yield the required number of events. For example, if nearly all events are expected to occur within the first two years, then long term follow up beyond two years will not produce many additional events. In this case it is important that enough patients are entered to yield the required number of events assuming a follow up of just two years for each patient.
- Delayed treatment difference, resulting in non-proportional hazards: in this scenario, in case of a superiority study, additional events are required as the first events observed do not contribute to the treatment difference; in case of a non-inferiority study, caution should be taken that the interim/final analyses are not planned too early since they may be biased towards no difference between the two arms.

EAST provides the possibility of specifying variable accrual patterns, drop outs, piecewise exponential survival distributions and fixed follow up per patient. In addition, simulations can be done taking these different factors into account as well as non-proportional hazards.

Specific methods have been developed for sample size calculations in the presence of competing risks, however the results of these methods differ and no one specific method is currently recommended over another.

9.4 Sample size adjustment during the course of the study

In this WIN, we only refer to sample size adjustment not based on interim treatment comparisons. These sample size adjustments do not lead to an increase of Type I error and as such are considered to be well-understood adaptive design.

We will distinguish between sample size adjustments not changing the total information (i.e. number of events for time-to-event studies) and sample size adjustments changing the total information but without access to interim treatment comparisons.

Documentation and justification of these sample size adjustments should be documented in ST-001-AF-01 and the procedure for review and approval by PRC and/or IDMC is described in ST-001-SOP.

9.4.1 Sample size adjustments not changing the total information

Sample size adjustments based on interim data from the control arm only or on blinded aggregate data are considered as “well-understood” adaptations of the study design as they are not based on interim treatment comparison and do not modify the total information (number of events, for time-to-event studies).

It is recommended that some time before Phase III trial reaches full accrual, the statistician checks that the original assumptions for the control arm (or pooled arms) hold. If they do not, the statistician may come to the conclusion that the sample size should be adjusted to guarantee that the originally planned total information. As stated in ST-001-SOP, these sample size adjustments need not to be submitted to the EORTC IDMC.

9.4.2 Sample size adjustments changing the total information without access to interim efficacy treatment comparisons

These sample size adjustments can be motivated by:

- New data external to the trial necessitate a change in the alternative hypothesis (either more or less strict)
- Enrollment is less than projected, but there is still the possibility of powering the study for a larger (but still realistic) effect size.
- A change in eligibility criteria leads to a different effect size being hypothesized. For example a subgroup has been detected in another trial and it is being aimed for. In this subgroup the effect may be (believed to be) larger than overall.
- Dropping one treatment arm because of reasons unrelated to efficacy (e.g. based on interim safety data).

These sample size adjustments may require an amendment to the protocol and/or submission to the IDMC.

10 General considerations on early stopping rules and interim analyses

Phase III studies of long duration, or of large size, should preferably have interim analyses built in at the design stage. These aim to stop the study (and if possible the recruitment) but before complete data maturity as defined by the primary trial objective where there is sufficient statistical evidence at the interim stage to allow rejection of the null hypothesis and/or the alternative hypothesis.

Early stopping rules provide the statistical rules that guide the decisions to stop a trial early during its recruitment or to reporting its results before the trial is mature for the primary endpoint. Early stopping rules are meant to stop trial as early as possible if:

- Treatment effect larger than anticipated: stopping for efficacy using (group) sequential designs
- No treatment effect: stopping for futility using (group) sequential designs or stochastic curtailment
- Effect in wrong direction: stopping for harm using (group) sequential designs
- Unacceptable adverse events/ toxicity

The process of correctly gathering, analyzing and interpreting accumulating information during a clinical trial is often referred to as “interim analyses”.

The fundamental problem with regularly and continually analyzing the emerging data from a trial is that more than one statistical test is performed during the course of the study which induces multiplicity issues among tests that are correlated because the data used in the i^{th} analysis include the data used in the former analyses.

Whatever the approach used, investigators pay a price for terminating trials early. Studies that permit or require formal data-dependent stopping methods as part of their design are more complicated to plan and carry out and, when such trials are stopped early, the estimates of treatment differences can be biased and imprecisely estimated. Furthermore, one would only want to take a decision to stop the study or not if the data are convincing and precise enough to impact further research and/or clinical practice.

Formal interim analyses are an integral part of the statistical design of a trial and as such must be precisely defined in the study protocol.

The Statistician clearly defines in the protocol:

- The number of stages and timing of the interim analyses must be specified either in terms of information (number of events or patients) or by a calendar date.
- Who will perform the interim analysis and who will have access to the interim data

- The stopping rules and how they are calculated. In some cases this information may be left out of the protocol to avoid operational bias..

IDMC Analysis Reports will be prepared in accordance with ST-006-SOP. The policy and procedure regarding the role of the IDMC regarding interim analyses and stopping rules can be found in EORTC Policy POL004 and ST-004-SOP.

Section 11 provides technical guidance on group sequential methods for implementing early stopping rules for efficacy and/or for futility. These are part of the “well-understood” adaptive designs, as unblinded interim analyses are used in a planned and confidential manner that controls Type I error and maintain study integrity.

Section 12 provides technical guidance on methods for implementing early stopping rules for safety/toxicity. Designs that include such stopping rules are also considered as “well-understood” adaptive designs, provided safety/toxicity outcomes are independent of, and uninformative about, the treatment-related efficacy effect.

11 Stopping rules for efficacy and/or futility

The terms used in the section are those defined in the Glossary of the EAST manual.

11.1 Stopping rules for rejection of H_0 only, of both H_1 or H_0 and of H_1 only

For a trial that tests a null hypothesis H_0 and for which the power is computed under a specified alternative H_1 , one will generally distinguish stopping rules that attempt either

- To stop the trial for early evidence in favour of H_1 (early rejection of H_0)
 - If H_1 is a two-sided, this may be either in the direction of H_{1+} (early evidence of efficacy in a superiority study)
 - Or in the direction of H_{1-} (early evidence of harm)
- To stop the trial for early rejection of H_1 (futility)
- To stop the trial for either early rejection of H_0 or early rejection of H_1 .

It is important to note that depending on whether a one-sided or two-sided test is used at the end of a trial; and whether the trial aims to show a difference in efficacy or to demonstrate the absence of any meaningful difference, the early rejection of H_0 and of H_1 take on different meanings.

In designing early stopping boundaries, one should always consider the compatibility of the stopping rules with the final trial objectives, in particular when this one is two-sided. For example: an early stopping rule for harm or futility in a superiority study is not compatible with a 2-sided final test for a difference; it is advised to use a 1-sided hypothesis test if one intends to stop a difference trial for early evidence of harm or of futility. Conversely, the results of a non-inferiority trial are difficult to interpret if it is stopped early for futility, as this is not a proof of harm; it is advised to design the study as an 2-sided equivalence study if one wishes to early stop early for harm.

11.2 Overview of methods used at the EORTC

Statistical methods for early stopping used at the EORTC can be classified as follows:

- Methods based on just the currently available evidence:

- Repeated significance testing
- Spending function boundaries
- Methods based on that evidence and specified scenarios for the future observations: stochastic curtailment based on conditional power

At the EORTC, mostly for logistic reasons, strictly sequential methods are not used but only group-sequential approaches.

Bayesian methods, besides stochastic curtailment, are not used within the EORTC and are therefore not described here. A general discussion and further references about Bayesian methods can be found in Berry [1995], Freedman and Spiegelhalter [1989] and Freedman, Spiegelhalter and Parmar [1994].

One must always verify by simulations of the envisaged method has a sufficiently high chance of leading to a decision at each interim look, in particular at the earlier ones. Indeed, there is little justification for conducting a formal interim analysis when the stopping boundary is so stringent that the chances of effectively modifying the study conduct on its basis are virtually nil. One must also avoid performing too early interim looks (<25% information) especially if non proportionality of effects is expected.

11.3 Repeated significance testing

Repeated significance testing is a group-sequential procedure constructed on the basis of repeatedly performing statistical tests on accumulating data.

If the plan is to perform K analyses ($K-1$ during the course of the trial and one at the planned end of the trial), the method consists of adopting more stringent nominal significant levels α'_k , $k=1, \dots, K$, for the tests at each of the K interim looks, to ensure that the overall significance level over the K analyses is α . Clearly as more α'_k is used at each interim analysis, less is available for use at the final analysis.

Several procedures were developed by Pocock [1977], Haybittle-Peto [1971] and O'Brien and Fleming [1979]:

- The procedure developed by Pocock [1977] consists in considering the same constant nominal level α' at each look (including the last one).
- Haybittle [1971] proposed a very simple procedure, known as the Haybittle-Peto boundaries, which consists in specifying a fairly small α' , say 0.001, for early stopping at the first $K-1$ looks, and to compute the last look α'_K needed to achieve an overall α over the K looks.
- O'Brien and Fleming [1979] proposes a procedure with α'_k increasing from one look to the next.

These different methods are illustrated in the table below considering two-tailed testing with an overall α equal to 0.05 and $K = 2$ or 3; the table gives values of α'_k for the Pocock, Haybittle-Peto and O'Brien-Fleming procedures at equally spaced analyses:

Look number	Pocock	Peto-Haybittle	O'Brien and Fleming
(k)	α'_k	α'_k	α'_k
1	0.029	0.001	0.005

2	0.029	0.05	0.048
1	0.022	0.001	0.0005
2	0.022	0.001	0.014
3	0.022	0.05	0.045

The most widely used approaches are the O'Brien Fleming and the Peto-Haybittle procedures, as they retain a large proportion of α for the final analysis. However these approaches are very conservative when interim tests are done; instead of stopping the trial whenever a p-value is 0.05, the trial stops only when the p-value is considerably less than 0.05 at pre-specified interim looks. In this case, the power to detect a significant difference during an interim analysis is likely to be low.

These methods require that the number of interim analyses is pre-specified and the interim analyses are performed at equal increments of information.

11.4 Spending function boundaries

11.4.1 Basic principles

This method generalizes the “Repeated significance testing approach”, escaping the restrictions of a pre-specified number of interim analyses and interim analyses performed at equal increments of information.

The method was developed by Lan and DeMets [1983]. It defines boundaries by relying upon the α spending function (see also DeMets and Lan [1994]). Assume that the trial is completed at information time T , scaled arbitrarily such that $T=1$. The α spending function $\alpha(t)$ represents the cumulative amount of the type I error that has been “spent” at each analysis carried out at information time t ($0 \leq t \leq 1$). It can be any monotone function defined on the unit interval such that $\alpha(0) = 0$ and $\alpha(1) = \alpha$. For the whole trial the risk of the type I error is α regardless of the number of interim analyses.

Several pre-defined α spending functions are available in EAST:

- The Lan-DeMets spending function with O'Brien-Fleming flavor (Lan-DeMets [1983]) which generates stopping boundaries that closely resemble the O'Brien-Fleming stopping boundaries.
- The Lan-DeMets spending function with Pocock flavor (Lan-DeMets [1983]) which generates stopping boundaries that closely resemble the Pocock stopping boundaries.
- The Gamma spending function (Hwang, Shih and Decani [1990]) whose functional form depends on a parameter γ . Negative values of γ yield convex spending functions that increase in conservatism as γ decreases, while positive values of γ yield concave spending functions that increase in aggressiveness as γ increases. The choice $\gamma=0$ spends the error linearly. Interestingly, the choice $\gamma=-4$ produces stopping boundaries that resemble the O'Brien-Fleming boundaries while the choice $\gamma=1$ produces stopping boundaries that resemble the Pocock boundaries.
- The Rho spending function (Kim and DeMets [1987], Jennison and Turnbull [2000]) whose functional form depends on a parameter ρ . Larger values of ρ yield increasingly conservative boundaries. When $\rho=1$, the corresponding stopping boundaries resemble the Pocock boundaries while when $\rho=3$, the boundaries resemble the O'Brien-Fleming boundaries.

Boundary	Gamma	Rho
Very Aggressive	> 1	$0 < \rho < 1$
Pocock	1	1
O'Brien-Fleming	-4	3
Very Conservative	< -4	> 3

It is also possible in EAST to make your own spending function by using the interpolated spending function approach, where the user defined the cumulative error to be spent by each time point.

Pampallona, Tsiatis and Kim [1995, 2001] developed the notion of a β spending function to derive stopping boundaries for the early rejection of H_1 for futility. Similar to the α spending functions, the β spending function is any monotone function defined on the unit interval such that $\beta(0) = 0$ and $\beta(1) = \beta$.

One may either use α or β spending functions on their own, or combine both α and β spending functions in a single trial, with one-sided or two-sided, symmetric or asymmetric boundaries. In two-sided alternative hypotheses, one may in fact decide on 4 stopping boundaries (two for the alpha-spending functions, for rejecting H_0 in favor of respectively H_{1+} and H_{1-} and 2 for the beta-spending function, for concluding to H_0 for futility of H_{1+} or of H_{1-} . In general, the boundaries are chosen to be symmetric but they need not to be. (see Kittelson JM and Emerson SS (1999) for more details).

In order to help choose between different error spending functions, the power to detect a difference at the interim analysis should be calculated via simulation for each spending function under consideration.

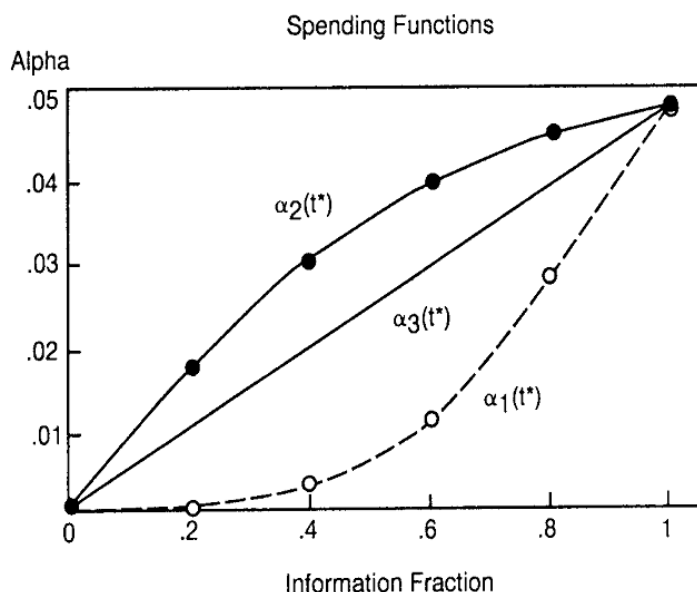


Fig. 15-6 Alpha-spending functions for $K = 5$, two-sided $\alpha = 0.05$ at information fractions 0.2, 0.4, 0.6, 0.8, and 1.0. $\alpha_1(t^*) \sim$ O'Brien-Fleming; $\alpha_2(t^*) \sim$ Pocock; $\alpha_3(t^*) \sim$ uniform.

For a general discussion on group sequential stopping boundaries, see Kittelson JM and Emerson SS (Biometrics 1999)

The next sections provide recommendations on the error-spending functions depending on the objective of the stopping rule.

11.4.2 Early stopping for superiority in a superiority study (early rejection of H_0)

A formal alpha spending function is used to calculate efficacy boundaries for rejecting H_0 and thus preserve the overall type I error.

It is recommended to choose an alpha spending function and interim analysis times that leave a significant fraction (at least 80%, i.e. 0.04) of the type I error to be spent at the final analysis while if possible maintaining a realistic power to detect differences at the interim analyses. An alpha spending function falling between a Pocock design and an O'Brien-Fleming design should be considered (for example gamma between 1 and -4 or rho between 1 and 3).

11.4.3 Early stopping for futility in a superiority study (early rejection of H_1)

A formal beta spending function is used to calculate futility boundaries for rejecting H_1 and thus preserve the overall type II error and power.

The use of an O'Brien-Fleming beta spending function may lead to early stopping for futility (rejecting H_1) even when the p value may appear to be headed in the direction of rejecting the null hypothesis. It is thus recommended to use very conservative stopping boundaries, for example a beta spending function from the Gamma family with a parameter value more extreme than -4 (O-F), for example at least -6 or -6.5.

For reference, an O'Brien-Fleming beta-spending function (as originally formulated, not the O'Brien-Fleming-like functions in East) has the characteristic that it has 50% conditional power (under the alternative) at threshold interim values. For the East O'Brien-Fleming-like beta spending functions, this is somewhat lower (30 to 40%).

In case one wishes to define a stopping rule for inferiority in a superiority study, the design has to be specified with a one-sided Type I error.

For further guidance on the implementation of stopping rules for lack of benefit, see Freidlin and Korn (2009)

11.4.4 Early stopping rules in non-inferiority or equivalence trials

Interim analyses in non-inferiority trials can be designed to:

- (1) stop early if the new treatment is as good as or better than the standard: Reject H_0 (reject null hypothesis of inferiority and conclude that the new treatment is not inferior). This is based on alpha-spending function.
- (2) stop early for futility, i.e. if the new treatment is not non inferior to the standard: Reject H_1 . Techniques used for early stopping due to futility may be applied either via beta-spending or stochastic curtailment.

Early stopping for futility in a non inferiority study reduces to a one sided test for superiority where the significance level is the type II error (beta) of the non inferiority test. If a very conservative beta spending function is used (gamma = - 6 or more extreme), a sufficiently powered interim analysis may only be able to be carried out after an unacceptably large number of patients have been entered and events observed. A less conservative beta stopping boundary is used (for example O-F or even Pocock) has more power to stop earlier if the null hypothesis of inferiority is true, but has the potential disadvantage of stopping when the conditional power of rejecting H_0 is still relatively high.

The consequences of early conclusions from a non inferiority or an equivalence study on medical practice must be carefully weighted. These may depend whether one or both treatments are used in practice and whether the trial endpoint is overall survival or an earlier endpoint. In some instances, early release of results from non-inferiority studies may be allowed independently of the observed data, when the information is likely to help patients facing a treatment decision but would not compromise the integrity of the trial or interfere with the completing of the trial to its definitive analysis (Korn, Hunsberger, Freidlin et al 2005).

Special caution must be taken in the case of early stopping in a non-inferiority study that uses a composite primary endpoint, since one component of the endpoint may drive the picture or opposite effects on various components may mask the overall picture. For a discussion of these aspects please see Cairns et al (2008).

11.5 Stochastic curtailment based on conditional power

The stochastic curtailment approach, proposed by Lan et al. [1982], attempts, by taking into account the information available at a given interim analysis, to predict the final results that would be obtained if the trial was allowed to continue its planned course until its normal completion under a specified behavior for the future observations.

This approach is most commonly used to stop early for futility. For example, one would stop the study for futility if there is a low enough probability of rejecting H_0 conditional on the observations already available and the hypothesis that future observations follow H_1 .

The conditional power is defined as the probability that the study will demonstrate statistical significance in its primary efficacy endpoint at the end of the study conditional on the data observed in the trial thus far and assuming a given treatment effect θ for the future observations:

$$CP = \text{Prob (Reject } H_0 \text{ at the end of the study} \mid \text{current data, } \theta)$$

Conditional power relies on the choice for the parameter value of θ [van der Tweel, 2003].. θ is usually the value specified at the design stage under the alternative hypothesis H_1 . θ can also be the estimated treatment effect based on the data accumulated at the interim analysis. As the IDMC may wish to know the conditional power under various assumptions for θ , a graphical representation of the conditional power as a function of θ may be helpful.

Using this approach, futility will be concluded if the conditional power for rejecting H_0 for a given value of θ becomes too low, say below γ . The critical value of γ can either be arbitrarily fixed or can be determined by expressing other stopping rules (such as spending function boundaries) in terms of conditional power boundaries. Usually γ is fixed to about 5%, corresponding in fact to stopping rules much more conservative than the typical spending function boundaries used (even more conservative than O'Brien Fleming boundaries, which corresponds to $\gamma=0.5$). However, use of larger values of γ will usually make people feel uncomfortable in case the trial should be stopped early. Of note there is a one-to-one correspondence between conditional power based rules and beta-spending futility boundary rules. It suffices to calculate the conditional power at the threshold value for any beta-spending function, or conversely to calculate the probability of stopping under the alternative hypothesis for any stochastic curtailment approach, to show this correspondence in operational characteristics. [see Emerson SS, Kittelson JM and Gillen GL (Statistics in medicine 2007).

Early stopping rules based on the conditional power potentially affect the type I and II error rates but the necessary increase in sample size to maintain these is usually marginal for small values of γ .

For a review of methods for futility stopping based on conditional power, see Lachin (Stats in Medicine 2005).

Less commonly, stochastic curtailment can also be used to stop early in favor of H1.

Computation of conditional power can also be used as an interesting additional tool as it provides additional information to decide if the trial should be continued further.

11.6 Number and timing of interim analyses

Statistical techniques allow a high flexibility with respect to the number and timing of interim analyses and the spending of the overall type I error α and/or Type II error β at each interim analysis. If adequate statistical methods are applied and strictly adhered to, there is no statistical argument against performing a large number of interim analyses. However, for several reasons, only a limited number of well justified interim analyses can generally be recommended [Koch, 2005].

To maintain the overall significance alpha level, the nominal significance level to be used at each analysis becomes smaller as the number of interim analyses increases. The more interim analyses are performed, the larger will be the maximum sample size [McPherson, 1982]. Furthermore, increasing the number of interim analyses may lead to practical and logistics issues, for example obtaining sufficient additional validated data at each interim analysis. For EORTC randomized phase III clinical trials, a maximum of 3 interim analyses is recommended.

In some situations, it might not be practical to carry out any interim analyses, for example if the trial's duration is expected to be very short and/or when considering a late endpoint as the primary endpoint.

The number and timing of interim analyses should be fixed in advance based on the trial's objectives and endpoints, sample size, required number of events, event rate, duration and speed of accrual, and size of expected treatment differences and be specified in the study protocol.

Together with the choice of the method and the error spending function, one must always verify by simulations (in EAST) if the envisaged number and timing of interim analyses has a sufficiently high chance of leading to a decision at each interim look, in particular at the earlier ones. Indeed, there is little justification for conducting a formal interim analysis when the information fraction is very small. In addition, the statistician should pay attention to the fact that early interim analyses might be misleading especially in settings when the results obtained at a very early stage might not be representative of the final results (e.g. in case of non-proportional hazards (PH)). The statistician will also consider the timing of the interim analyses with respect to the patient recruitment, since stopping rules applied after completion of accrual can only affect the timing of release of results but will not spare patients.

Finally, the statistician will take due consideration for the time needed for data processing, database lock, data analysis and reporting to the IDMC. This process must be planned ahead of time with the Trial team. It is recommended to count around 6 months between the first pre-analysis meeting and the IDMC meeting.

11.7 Planned versus actual information time at interim analysis

It should be noted that interim analyses will usually be conducted at an observed information time that can be somewhat different from the planned interim analysis time. This is because it is logistically impossible to force the actual number of events observed by the time the interim database lock to be exactly equal to the planned number. The interim analysis stopping boundaries should be recalculated using the actual observed

information fraction, not the planned information fraction. This must be documented in the interim analysis report.

11.8 Binding versus non-binding stopping futility boundaries

When using non-binding stopping futility boundaries, the trial may continue even if the boundaries are crossed. EAST computes non-binding stopping futility boundaries that produce the desired power and guarantees the Type I error is at most the design alpha. In fact, the Type I error is exactly equal to alpha if the futility boundary is overruled and is lower than alpha if the futility boundary is respected.

When using binding futility boundaries, the study must be stopped if the boundaries are crossed, otherwise Type I error is no longer preserved.

12 Stopping rules for safety/toxicity

12.1 Basic principles

Generally, in phase III trials, toxicity is not the primary endpoint. However there may be a need to install some type of stopping rule for the possibility of too high a level of toxicity.

To implement early stopping rules based on toxicity, one will essentially implement a phase II type of design with the toxicity endpoint, early on in the study.

Depending if on is concerned:

- a) by an increased toxicity of the experimental arm as compared to the reference arm (comparative situation)
- b) or by the rate of a specific toxicity of the experimental arm which would become unacceptable if it rises above a certain incidence rate (non-comparative situation),

one will opt for a comparative or a single-arm design.

Usually toxicity stopping rules will be one-sided since it is assumed that the toxicity of the standard treatment arm is sufficiently known before the trial so that no specific stopping rules are needed for that arm.

A challenge with such stopping rules is to clearly define which adverse events and of which grade will qualify as a "toxic event" in the early stopping rule and to obtain reference values for informing the design. The definition can be of a general nature (e.g. toxic deaths, toxic discontinuations, related grade 3 or 4 adverse events) or very specific depending on the case and provided primarily by medical experts (. It is important to keep in mind that the incidence can be affected by the mode of collection. For example, a specific question on the CRF may result in higher incidence than if one relies on spontaneous reporting of the event. An additional point of attention should be the time at which the toxicity can reasonably be expected to occur. For example neuropathy with chemotherapy, as well as radiotherapy toxicity, can be long term rather than immediate.

In all of the above situations, the primary interest of the statistical testing of toxicity endpoint is not on the final analysis at the end of the trial, but on the need to prematurely stop the trial (or the specific treatment arm) if the toxicity rate becomes unacceptably high. Therefore it is strongly recommended to:

- Plan a first analysis early on to exclude a very high and unacceptable rate of toxicity with sufficient power (e.g. 90% power). An alpha spending function of the Pocock type (or similar) can be used.

- Alternatively, use a phase II type of design for the first part of the trial (e.g. A'Hern if non comparative, or Exact chisquare test if comparative) where the endpoint of “clinical response” is replaced by “toxicity”. In this case, the entire alpha can be spent at the interim analysis.

Since the trial total sample size is already determined by the primary endpoint, one may compute the statistical power of the comparison at each interim look. It is useful to document the power for the toxicity stopping rule in the protocol.

P-values for toxicity other than the toxicity stopping rule should not be presented, unless specified otherwise in the protocol.

The timing of the interim looks should be specified on the information scale of the toxicity endpoint (i.e. based on number of patients). These may be chosen unequally spaced. It is important to keep in mind that the information time for the toxicity endpoint may be different from the information time for efficacy endpoints. One should carefully consider for how many patients in the database the toxicity endpoint can be evaluated.

12.2 Comparative stopping rules

This is a stopping rule to reject the hypothesis $H_0: \pi_0 \geq \pi_1$ in the direction $H_1: \pi_0 < \pi_1$, where π_0 is the percent of patients with toxicity on the standard arm, and π_1 is the percent of patients with toxicity on the experimental arm.

Stopping rules can be designed using the EAST software, for binomially distributed data. In the calculation it is recommended to use the unpooled estimates of the variance.

An appropriate Type I error (alpha) should be first determined, which corresponds to the risk of erroneously concluding that the treatment is not toxic. The whole alpha will be spent at the interim looks. The alpha-spending function is recommended to be closer to the Pocock type, rather than O'Brien-Fleming which is too stringent early on and would unlikely stop the trial if the toxicity was excessive.

12.3 Non-comparative stopping rule

This is a stopping rule to reject the hypothesis $H_0: \pi_1 \leq \delta$, where δ is a fixed percentage giving the maximum acceptable toxicity rate on the experimental arm.

For a single interim look or two interim looks, one may design the stopping rule using the classical phase II methods used for single-arm phase II designs (see ST-001-WIN-02), using A'Hern design or a Simon-two stage design using toxicity instead of response as the endpoint (i.e. a "response" in this case is a patient with "no toxicity" and the null and alternative hypotheses are specified accordingly).

Alternatively one may elaborate a stopping rule using EAST for a Binomial Superiority One-Sample test, similarly to the way it was done above for the comparative setting above, specifying a spending function close to Pocock and the information times at which the looks take place.

13 Designs involving multiple arms, multiple hypotheses, subgroups

In order to ensure the feasibility of a phase III study, the design should be kept as simple as possible in order to meet the objectives of the study. In most cases a simple randomization between just two treatments is recommended.

More complex but still classic designs are described in the subsection below.

Crossover studies are generally to be avoided since the underlying assumptions for carrying out such studies are almost never valid in cancer studies are not considered here.

13.1 More than two treatment arms

When comparing more than 2 treatment arms, the control arm and experimental arms should be clearly identified along with the treatment comparisons that will actually be carried out. It must be clearly stated which of these comparisons are considered primary comparisons, because:

- Any significant result of a primary comparison is a valid stand-alone study conclusion
- Multiple primary comparisons with or without an appropriate adjustment of the type I error. Possible techniques for adjustment include, but are not restricted to:
 - Bonferroni correction (equal or unequal splitting of the Type I error across the co-primary comparisons and/or endpoints)
 - Establishing an order of testing, and conditioning the performance of lower-ranked tests on the higher-ranked ones being significant
 - More refined testing procedures, such as Holm, Hochberg.
 - Other closed testing procedures, for example comparing the results of the pooled experimental arms versus the control arm at the interim looks and final analysis, and performing pairwise testing at the last look only if significant results are obtained for this pooled analysis.
 - Other options are discussed by Proschan et al. [1994].

Multiple primary comparisons may be performed without adjustment of Type I error, using the argument that if the experimental arms were compared with controls in two-armed separate trials, no adjustment would be made. However, when the research questions corresponding to the experimental arms are related, for example when several treatments are evaluated alone and in combination, or when several different schedules or doses of an agent are investigated, multiplicity adjustment is appropriate.

Whatever approach is chosen, the design and sample size calculation need to reflect this approach. Except for binomial and continuous endpoints, EAST cannot design or analyze trials where there is an overall comparison of more than 2 treatment arms.

The considerations given in the chapter 13.1 dealing with multiple primary endpoints or hypotheses may also apply.

Stopping rules may be applied to multi-arm studies using group-sequential methods applied independently to individual treatment-control comparisons. These are called group-sequential multi-arm multi-stage (MAMS) designs and are considered as “well-understood” adaptive methods, according to FDA Guidance on Adaptive Design Clinical Trials for Drugs and Biologics (2010). In particular, the Royston, Parmar, Qian (2003) design is a design where all treatments are continued at each stage, provided the observed interim effect measured on an intermediate endpoint surpasses a predetermined threshold. These are attractive designs but complex and difficult to conduct in an international setting.

For other approaches for treatment selection in multi-arm studies please refer to the published literature.

13.2 Factorial designs

When two or more questions are to be studied the use of factorial designs should be considered. It may be possible to use a 2 x 2 factorial design to answer two questions for the price of one. An example of such a study is a study with the following treatment arms:

1. surgery
2. surgery + radiotherapy
3. surgery + chemotherapy
4. surgery + radiotherapy + chemotherapy.

Through the proper use of retrospective stratification, each treatment arm can be used twice, once to study the benefit of radiotherapy and once to study the effect of chemotherapy. For example, the effect of radiotherapy is assessed by comparing radiotherapy to no radiotherapy (2 and 4 versus 1 and 3) with a retrospective stratification for whether or not patients received chemotherapy.

The number of patients required for this study will be similar to that needed for a two arm study under the assumption of no radiotherapy/chemotherapy interaction and a treatment difference equal to the smallest of the powered treatment differences in the factorial design. No adjustment to the size of the type I error is made for the two comparisons.

It is to be noted that the assumption of no interaction is a strong one, rarely met and can only be tested at reduced power within the study.

In the presence of an interaction, the hypotheses underlying the 2 x 2 factorial design are not satisfied and such a pooling may yield meaningless results. In order to take into account the fact that the assumptions underlying the 2 x 2 factorial design may not be completely satisfied, one could increase the power to 85% or 90%.

2 x 2 factorial design studies where one question involves a superiority comparison and the other a non inferiority comparison are not recommended.

When planning stopping rules in a factorial design, one should envisage the overall impact on the study of a stopping boundary applied to one particular comparison, i.e. how an early stopping of one arm of one randomization in a 2x2 factorial design may affect the other randomized comparison.

13.3 Multiple primary endpoints, multiple hypotheses or comparisons

In some studies, the research logic dictates co-primary endpoints (e.g. an efficacy endpoint and a specific safety endpoint. As in the previous section, the method of protecting for multiplicity is specified and applied. In case the study is defined to be successful if both co-primary endpoints show to be positive, the sample size should be adjusted for multiplicity to ensure adequate power.

In case the study is defined to be successful if at least one of the co-primary endpoints show to be positive, the sample size should be adjusted for multiplicity to ensure adequate Type I error.

The design will have to be considered on a case by case basis to determine:

- The need for adjustment of the overall significance level to be used for any or all of the endpoints or hypotheses to be tested (due to the comparison of multiple endpoints or hypotheses).
- The hypotheses, endpoints or comparisons for which interim analyses will be done

- The effect of an interim analysis for one endpoint on the significance levels to be used for that and all other primary endpoints, hypotheses or comparisons at the time of additional interim analyses and for the final analysis.

See section 13.1 concerning multi-arm multi-stage designs such as the Royston, Parmar, and Qian design where the decision to drop a treatment arm at an interim analysis may be based on an intermediate endpoint.

13.4 Studies with targeted therapies

In studies with targeted therapies, a larger treatment benefit is generally expected in patients expressing the target than in patients for whom the target is not expressed. This leads to questions concerning the optimal study design and analysis strategy.

13.4.1 Study population: “all-comers” versus “marker positive patients”

The first consideration is whether to include "all-comers" (i.e. both marker positive and marker negative patients) or only the targeted subset of patients (called "targeted " design):

- The main advantage of the "all-comers" approach is that it enables formal assessment of the predictive effect of the marker and allows testing other markers.
- However there may be circumstances where it may be unethical to enter patients not presenting the marker of interest (either because of lack of biological rationale of the new drug in that subset, or because a negative effect is anticipated). The targeted design is most efficient if a large treatment effect is expected in the marker positive patients, no effect is anticipated in marker negatives and if the prevalence of marker positives is relatively low (Maitournam A and Simon R, 2005)

In the absence of strong data about the effect being restricted to an adequately detectable sensitive subgroup with a specific genetic / biomarker profile, the eligible population must be kept broad.

Other factors to be taken into account for studies with targeted therapies are:

- Prevalence of the biomarker of interest. A low prevalence can lead to a high cost per randomized patient in a design that puts most emphasis on that subgroup, because many other (negative) screened cases are needed.
- The test reliability of the assay to assess sensitivity and specificity.
- The danger of over-treating non-sensitive patients (marker negative): this needs to be controlled by means of early stopping rules in the putative non-sensitive patients
- The feasibility of obtaining marker results before randomization, which is a prerequisite for using an enrichment approach and in the all-comers design for stratification of the randomization by marker status

Three designs are discussed below:

- Targeted design
- "All-comers" design:
 - Biomarker stratified
 - Randomized strategy

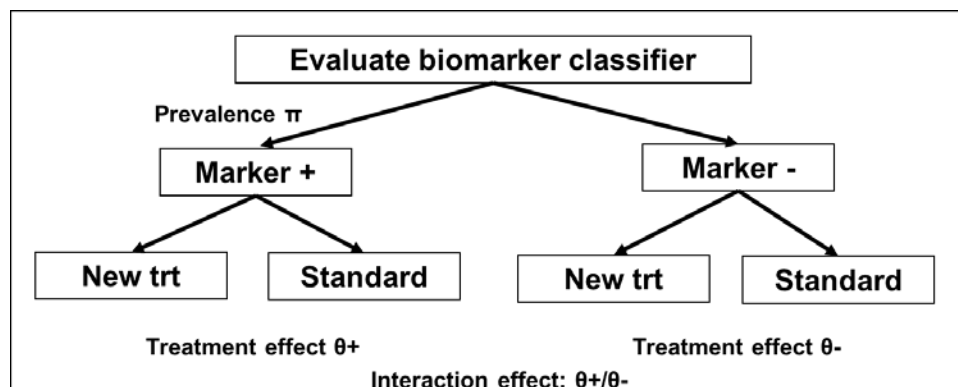
Designs where a subpopulation is selected during the course of the study are called enrichment designs. For the design of this type of study, please refer to the published literature.

13.4.2 Targeted design

If an "enrichment" design is chosen, the classical phase III designs can be used with the additional consideration that the number of patients to screen for the marker will be back calculated using the trial sample size and the marker prevalence, and that the marker prevalence and marker assessment failure rate will need to be monitored during the trial.

13.4.3 Biomarker stratified design

A "biomarker-stratified" design generally presents as follows:



Stratification for marker status may not be possible if the marker is tested retrospectively; even in that case, we will remain with a general framework of having two (or more) populations of interest in the study:

- all patients
- the marker positive subgroup (putative sensitive group)
- the marker negative subgroup (or alternatively, if the marker is tested retrospectively and entails some test failures, the "non-positive" group)

The statistical design for a biomarker stratified design may be built along 3 routes:

- Interaction-based approach: whereby the statistical power is based on the test for treatment by marker interaction. A closed testing may follow whereby a test is conducted in each marker subset, if the interaction test is statistically significant, and is conducted in the overall population otherwise. This approach generally requires huge sample size given the low power of interaction tests, unless a very sensitive endpoint is used or a qualitative interaction may be envisaged.
- Designs aiming to test the treatment effect in the marker group of interest and overall (or alternatively in the marker negative subgroups). This approach requires controlling for the type I error rate at the study level (family-wise error, rate FWER) since two co-primary analyses are envisaged. This may be done either using
 - Hierarchical testing approaches: by pre-specifying a sequence of tests, each one being performed at the nominal type I error rate, conditional on the former one being significant at that level, or use sequential testing procedures such as Hochberg step-up procedure

- Split- α approaches where two sets of interest are defined, say the whole group and the marker positive group and claim for efficacy will be made in either of the two sets. The FWER alpha is split in two parts, $\alpha = \alpha_{\text{whole}} + \alpha_{\text{marker+}}$. Efficacy in the whole group is claimed if the test in the whole group is significant at the α_{whole} level. Otherwise, efficacy is claimed in the marker positive subgroup only if the test in the marker positive subgroup is significant at the $\alpha_{\text{marker+}}$ level. The significance level allocated to the whole group is generally at least half of the family-wise error rate α . One can take into account the correlation that exists between the test statistic for the subgroup analysis and the test statistic for the overall analysis, since patients that are part of the subgroup analysis also contribute to the overall analysis. The significance levels that incorporate correlation are higher than the significance levels that do not incorporate this correlation, but the FWER will be controlled at the prespecified level. To calculate the alpha level, Spiessen and Debois (2010) suggest using existing softwares aimed at planning for interim analyses, since there as well, the data at the interim is part of the data available at the final analysis

You may refer to Wang (2007) and to Hoering and Leblanc (2008) for a discussion of the relative efficiency of various approaches for controlling the overall type I error and to Song and Chi (2007) for a general framework of this approach.

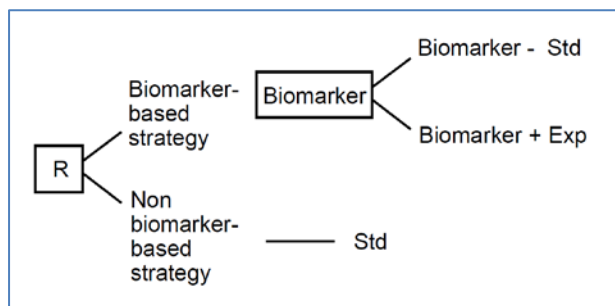
When planning the trial, one must ensure that the sample sizes in the two groups of interest is compatible and specify upfront rules for stopping the trial. One may wish to guarantee a minimum number of patients in the marker+ subgroup. For time to event endpoints, events may not accumulate equally fast in the two marker groups (depending on patient numbers, hazard rates etc...) if the marker is also prognostic, this may impact timing of analyses. It is also recommended to monitor the marker prevalence during the study, so that adjustment of the sample size is applied whenever the assumed prevalence was incorrect.

Wang SJ et al. (2007) suggests implementing an interim stopping rule for futility in the marker negative subgroup. Recruitment of marker negative patients may be stopped and the trial continues with an enriched population to the initially planned total sample size. The final test(s) are conducted at the initially planned α -level(s).

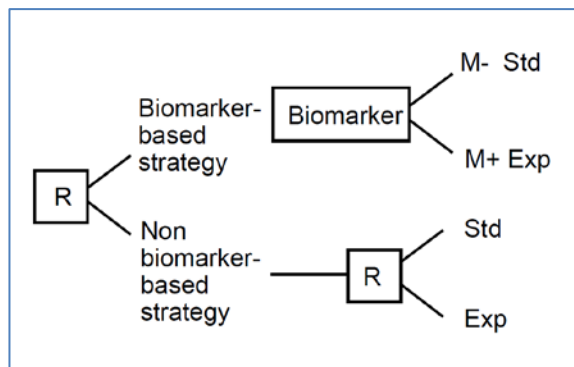
13.4.4 Randomized strategy designs

The "randomized strategy designs" come in essentially two variants:

Variant1:



Variant 2:



- Variant 1 is rarely efficient. It is particularly inefficient when the marker prevalence is low and in that design, predictive effect of the marker and treatment effect are confounded.
- Variant 2 offers a design that mimics the practical use of the assay, but is generally inefficient because the random allocation will result in the correct assignment of patients to treatment in a proportion of the patients. The design is equivalent to a design that randomizes only the patients in who marker-based and random allocations differ. The randomization ratio should preferably be adjusted to reflect the marker prevalence.

For a complete discussion of the efficiency of the marker-based design, refer to Sargent et al (2005) or to Eng (2014) who proposed a "reverse marker based strategy design" to try to alleviate the dilution problem by having a reference arm where patients are allocated to the arm they are not expected to be most sensitive to. Although statistically appealing, that design is unlikely ethically acceptable.

Finally, when no pre-specified marker of interest is available, the adaptive signature designs proposed by Freidlin and Simon may be appropriate. That design works as follows:

- Conduct a two stage study where sensitive patients are defined in the first stage, a test in sensitive cases is carried out in the stage II population and an overall test is carried out in the combined stage I + II population.
- Use 80% (0.04) of the alpha for testing the whole population and 20% (0.01) of the alpha to test the sensitive population. The study is considered to be positive if either test is significant. In such design, , the statistical method used for developing the classification should be prospectively defined in the protocol and the classification system must be available prior to starting the final analysis. A different split of the alpha may also be considered and the sample size may be adjusted to have sufficient power for the overall test based on an $\alpha < 0.05$.

14 References

14.1 On sample size calculation

- ◆ Collett. Modelling Survival Data in Medical Research, Third Edition. Chapman and Hall/CRC.
- ◆ Curran D., Sylvester R. and Hocht Boes G. (1999). Sample size estimation in phase III cancer clinical trials. *European Journal of Surgical Oncology*, 25: 244 – 250.
- ◆ George S. (1984). The required size and length of a phase III clinical trial. In: Buyse, Staquet, Sylvester, *Cancer clinical trials, methods and practice*, pages 287-310. Oxford Medical Publications.
- ◆ George S. and Desu M. (1974). Planning the size and duration of a clinical trial studying the time to some critical event. *J Chron Dis*, 27, 15-24.
- ◆ Roebruck P and Kuhn A. (1995). Comparison of tests and sample size formulae for proving therapeutic equivalence based on the difference of binomial probabilities. *Statistics in Medicine*, 14, 1583 - 1594.
- ◆ Sahai H and Khurshid A. (1996). Formulae and tables for the determination of sample sizes and power in clinical trials for testing differences in proportions for the two sample design: a review. *Statistics in Medicine*, 15, 1 - 21.
- ◆ East® Version 6.2 © Cytel Inc. Copyright 2013 Manual Version 6.2:01, July 29, 2013.

14.2 With competing risks

- ◆ Latouche A, Porcher R., Chevret S. (2004) Sample size formula for PH modelling for competing risk. *Statistics in Medicine*, 23: 3263-3274.
- ◆ Pintilie M (2002). Dealing with competing risks: testing covariates and calculating sample size. *Statistics in Medicine*, 21: 3317-3324.
- ◆ Schulgen G, Olschewski M, Krane V, Wanner C, Ruf C, Schumacher M. (2005). Sample sizes for clinical trials with time-to-event endpoints and competing risk. *Contemporary Clinical Trials*, 26: 386-396.

14.3 On non-inferiority designs

- ◆ Cairns JA, Wittes J, Wyse G et al. (2008). Monitoring the ACTIVE-W trial: Some issues in monitoring a noninferiority trial. *American Heart Journal*, 155(1): 33-41.
- ◆ Carroll. Active-controlled, non-inferiority trials in oncology: arbitrary limits, infeasible sample sizes and uninformative data analysis. Is there another way? *Pharmaceut. Statist.* 2006; 5: 283–293
- ◆ European Medicines Agency (2006): Guideline On The Choice Of The Non-inferiority Margin.

http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003636.pdf

- ◆ European Medicines Agency (2000): Points to consider on switching between superiority and non-inferiority.

http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003658.pdf

- ◆ Freidlin B, Korn EL, George S, Gray R.(2007). Randomized Clinical Trial Design for Assessing Noninferiority When Superiority Is Expected. *J Clin Oncol*. 2007 Nov 1;25(31):5019-23.
- ◆ Pocock. The pros and cons of noninferiority trials. *Fundam Clin Pharmacol*. 2003 Aug;17(4):483-90.
- ◆ Rothmann. Design and analysis of non-inferiority mortality trials in oncology. *Statist. Med*. 2003; 22:239–264.

14.4 On stopping rules and interim analyses

- ◆ DeMets, D.L. and Lan, K.K.G. (1994). Interim analysis: The alpha spending function approach. *Statistics in Medicine*, 13: 1341-1352.
- ◆ EMA Reflection Paper On Methodological Issues In Confirmatory Clinical Trials Planned With An Adaptive Design (2007)
http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003616.pdf
- ◆ Emerson SS, Kittelson JM, Gillen DL (2007) Frequentist evaluation of group sequential clinical trial designs; *Statistics in medicine* 26: 5047-5080.
- ◆ FDA Guidance on Adaptive Design Clinical Trials for Drugs and Biologics (2010)
<http://www.fda.gov/downloads/Drugs/Guidances/ucm201790.pdf>
- ◆ Freidlin B.; Korn EL (2009). Monitoring for Lack of Benefit: A Critical Component of a Randomized Clinical Trial, *Journal of Clinical Oncology*, 27:629-633
- ◆ Freidlin B.; Korn EL (2002). A comment on futility monitoring, *Controlled Clinical Trials*, 23:355-366.
- ◆ Haybittle J.L. (1971). Repeated assessment of results in clinical trials of cancer treatment. *British Journal of Radiology*, 44: 793-797.
- ◆ Hwang I.K., Shih W.J. and DeCani J.S. (1990). Group sequential designs using a family of type I error probability spending functions. *Statistics in Medicine*, 9: 1439-1445.
- ◆ Jennison C. and Turnbull B.W. (2000). *Group sequential methods with applications to clinical trials*. Chapman and Hall/CRC, London.
- ◆ Kim K. and DeMets D.L. (1987). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika*, 74; 149-154.
- ◆ Kittelson JM and Emerson SS (1999). A unifying family of group sequential test designs. *Biometrics* 55; 874-882
- ◆ Koch A. (2005). How much do we have to pay for how many interim analyses. *Contemporary Clinical Trials*, 26; 113-116.

- ◆ Korn EL, Hunsberger S, Freidlin B et al (2005). Preliminary data release for randomized clinical trials of non inferiority: a new proposal. *Journal of Clinical Oncology*, 23:5831-36.
- ◆ Lachin JM (2005) A review of methods for futility stopping based on conditional power. *Statistics in Medicine* 24: 2747-2764.
- ◆ Lan K.K.G. and DeMets D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70: 659-663.
- ◆ Lan K. and Zucker D. (1993). Sequential monitoring of clinical trials: the role of information and brownian motion. *Statistics in Medicine*, 12
- ◆ McPherson K (1982). On choosing the number of interim analyses in clinical trials. *Statistics in Medicine*, 1: 25-36.
- ◆ O'Brien P.C. and Fleming T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, 35: 549-556.
- ◆ Pampallona S., Tsiatis A.A. and Kim K. (1995). Spending functions for type I and type II error probabilities of group sequential trials. Technical report, Dept. of Biostatistics, Harvard School of Public Health, Boston.
- ◆ Pampallona S., Tsiatis A.A. and Kim K. (2001). Interim monitoring of group sequential trials using spending functions for the type I and type II error probabilities. *Drug Information Journal*, 35, 1113-1121.
- ◆ Pocock S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64: 191-199.
- ◆ Proschan MA, Follmann DA, Geller NL (1994). Monitoring multi-armed trials. *Statistics in Medicine*, 1, 25-36.
- ◆ Snappin S., Chen M., Jiang Q. and Koutsoukos T. (2006). Assessment of futility in clinical trials. *Pharmaceutical Statistics*, 5: 273-281
- ◆ Sydes M.R., Spiegelhalter D.J., Altman D.G., Babiker A.B., Parmar M.K.B., and the DAMOCLES Group (2004). Systematic qualitative review of the literature on data monitoring committees for randomized controlled trials. *Clinical Trials*, 1: 60-79.
- ◆ Stallard N and Friede T (2008). A group sequential design for clinical trials with treatment selection. *Statistics in Medicine* 27: 6209-6227.
- ◆ van der Tweel I., van Noord P.A.H. (2003). Early stopping in clinical trials and epidemiologic studies for "futility": Conditional power versus sequential analysis. *Journal of Clinical Epidemiology*, 56; 610-617.
- ◆ Whitehead J. (1997). *The Design and Analysis of Sequential Clinical Trials*. John Wiley & Sons Ltd, England.

14.5 On multi-arm studies

- ◆ Barthel FM, Parmar MK, Royston P. How do multi-stage, multi-arm trials compare to the traditional two-arm parallel group design--a reanalysis of 4 trials. *Trials*. 2009

- ◆ Choodari-Oskooei B, Parmar MK, Royston P, Bowden J. Impact of lack-of-benefit stopping rules on treatment effect estimates of two-arm multi-stage (TAMS) trials with time to event outcome. *Trials*. 2013
- ◆ Freidlin B, Korn EL, Gray R, Martin A. Multi-arm clinical trials of new agents: some design considerations. *Clin Cancer Res*. 2008.
- ◆ Jaki. Multi-arm clinical trials with treatment selection: what can be gained and at what price? *Clin. Invest. (Lond.)* 2015
- ◆ Parmar et al. Speeding up the evaluation of new agents in cancer. *J Natl Cancer Inst*. 2008
- ◆ Parmar et al. More multiarm randomised trials of superiority are needed. *Lancet*, 2014.
 - ◆ Royston P., Parmar M. and Qian W. (2003). Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Statistics in Medicine*, 22, 2239 – 2256.
- ◆ Royston P, Barthel FM, Parmar MK, Choodari-Oskooei B, Isham V. Designs for clinical trials with time-to-event outcomes based on stopping guidelines for lack of benefit. *Trials*. 2011
- ◆ Stampede website : <http://www.stampedetrial.org/>
- ◆ Sydes et al. Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial. *Trials*, 2009.
- ◆ Sydes et al. Flexible trial design in practice - stopping arms for lack-of-benefit and adding research arms mid-trial in STAMPEDE: a multi-arm multi-stage randomized controlled trial. *Trials*, 2012.
- ◆ Wason et al. Some recommendations for multi-arm multi-stage trials. *Stat Methods Med Res*. 2013.
- ◆ Wason J, Stallard N, Bowden J, Jennison C. A multi-stage drop-the-losers design for multi-arm clinical trials. *Stat Methods Med Res*. 2014

14.6 On targeted designs

- ◆ Eng KH randomized reverse marker strategy design for prospective biomarker validation. *Statistics in Medicine* 2014 33(18): 3089–3099
- ◆ Freidlin and Simon. Adaptive Signature Design: An Adaptive Clinical Trial Design for Generating and Prospectively Testing A Gene Expression Signature for Sensitive Patients. *Clin Cancer Res* 2005;11(21): 7872- 78
- ◆ Hoering A, Le Blanc M, Crowley JJ, Randomized Phase III Clinical Trial Designs for Targeted Agents *Clin Cancer Res* 2008;14(14): 4358-67
- ◆ Maitournam A., Simon R. On the efficiency of targeted clinical trials. *Statistics in Medicine* 24:329-339, 2005
- ◆ Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *J Clin Oncol* 2005; 23: 2020–27
- ◆ Song, Y. and Chi, G. Y. H. (2007), A method for testing a prespecified subgroup in clinical trials. *Statist. Med.*, 26: 3535–3549.
- ◆ Spiessens B, Debois M. Adjusted significance levels for subgroup analyses in clinical trials. *Contemporary Clinical Trials* 31 (2010) 647–656

- ♦ Wang SJ, O'Neil RT, Hung HMJ Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceut. Statist.* (2007)

14.7 On Bayesian designs

- ♦ Berry D.A. (1985). Interim analyses in clinical trials: Classical vs. Bayesian approaches. *Statistics in Medicine*, 4: 521-526.
- ♦ Freedman L.S., Spiegelhalter D.J., and Parmar M.K.B. (1989) Comparison of Bayesian with group sequential methods for monitoring clinical trials. *Controlled Clinical Trials*, 10: 357-367.
- ♦ Freedman L.S., Spiegelhalter D.J., and Parmar M.K.B. (1994) The What; Why and How of Bayesian Clinical Trials Monitoring. *Statistics in Medicine*, 13: 1371-1384.

14.8 On trial designs for rare cancers

- ♦ Bogaerts, Sydes et al. Clinical trial designs for rare diseases: studies developed and discussed by the International Rare Cancers Initiative. *Eur J Cancer*. 2015.

15 ASSOCIATED DOCUMENTS

None

16 DOCUMENT HISTORY

Version N°	Brief description of change	Author	Effective date
1.00	Initial release, supersedes WP1102 version 1.5	Jan Bogaerts	25 Feb 2011
1.00	No change	Jan Bogaerts	25 Feb 2014
2	Merge ST-001-WIN-03 and former WP1305 + update according to current state-of-the-art.	Catherine Fortpied	11 Dec 2015