



Trial Design – Phase II / Exploratory studies

ST-001-WIN-02

Version 3

ALWAYS REFER TO THE INTRANET TO CHECK THE VALIDITY OF THIS DOCUMENT

<p>Author:</p> <p><i>Associate Head of Statistics Department</i></p> <p>Saskia Litière</p>	<p>Signature:</p> 	<p>Date: (ex:10-Feb-2017)</p> <p>23-APR-2018</p>
<p>Approved and authorized by</p> <p><i>Head of Statistics Department</i></p> <p>Laurence Collette</p>	<p>Signature:</p> 	<p>Date: (ex:10-Feb-2017)</p> <p>24 Apr 2018</p>

This document is the property of EORTC.

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

1 PURPOSE

The current work instruction details various aspects of the statistical design of EORTC phase II / exploratory trials as a function of the objectives of the study.

2 DEFINITIONS

Phase II trials or exploratory trials are carried out after the phase I assessment of a new agent or combination of agents, but before large scale confirmatory comparative studies are launched in the framework of randomized phase III trials. Exploratory trials are performed to get a minimal assurance of activity of the treatment being studied, as well as getting a better understanding of the safety profile (common short-term side effects and risks associated with the treatment) in a well-defined and relatively homogeneous, relatively small number of patients. They may sometimes be integrated as expansion phase of a phase I trial (phase I/II trial, generally non-randomized) or be integrated as a first step of a phase III trial (phase II/III randomized trials).

3 Design considerations for phase II / exploratory trials

3.1 Objectives

The general objectives of a phase II / exploratory trial are:

- ♦ to identify possible biological anti-tumor activity or early signs of efficacy
- ♦ to obtain a more detailed description of the drug's toxicity and mode of action

Additional objectives specific to certain designs are discussed in the following sections.

3.1.1 Early versus late phase II trials

Early phase II trials (also called phase IIa trials) are designed to identify anti-tumor activity in a representative sample of tumor types, usually selected on the basis of the targets used for drug development and/or the results of both pre-clinical studies and phase I trials. A further aim of early phase II trials is to obtain a more detailed description of the drug's toxicity, particularly cumulative toxicity (which is easier to study in a phase II patient population than in phase I) and to study ways to manage toxicity (e.g. by preventive measures, concomitant medication, etc.). Early phase II trials may also study the pharmacokinetic / pharmacodynamics relationship. At the conclusion of an early phase II trial a decision is made whether to continue the research and clinical development of a new treatment strategy.

Once early phase II trials have been successfully conducted in an initial sample of tumor types, additional phase II trials may be conducted to document the drug's anti-tumor activity or look for early signs of efficacy in other types of tumors, or in combination with other treatments in *late phase II trials* (also called phase IIb trials). Further documentation of the toxicity profile is also a goal, but at this stage the principal toxicities have generally already been identified. At the conclusion of a late phase II trial a decision is made whether to further evaluate the drug in phase III trials in a specific tumor type.

3.1.2 Feasibility studies

In many situations the information provided by phase I and early phase II trials may not be sufficient to justify the treatment of a large number of patients with an experimental treatment in a randomized phase III trial. Particularly

when a new agent is incorporated in a combined therapy, the feasibility of the new therapeutic approach is unknown. This happens namely in the following situations:

- ◆ it is unknown if the drug may be safely combined with other agents
- ◆ the compatibility of the new agent with surgery or radiotherapy is unknown
- ◆ new schedules of established agents or fractionation schemes of radiotherapy are required
- ◆ the feasibility or morbidity of a new surgical procedure is unknown
- ◆ new strategies such as high dose therapy with stem cell support, hyper-fractionation of radiation, synchronous drug and radiation treatment, use of chemo-protectors, etc, are under investigation.

This may be addressed by a dedicated study. These trials are often labeled as “feasibility studies” and must be specifically designed according to their objectives: to provide the justification for a subsequent large randomized phase III trial that will assess the potential therapeutic value of a new treatment. Feasibility studies alone never provide evidence of a possible therapeutic benefit.

3.2 Treatment mechanism of action

Before thinking about the statistical design to address a research question, it is important to have insight into the mechanism of action of the treatment, i.e. how it works. Investigations into local therapy require a different strategy (outcome measure and design) from investigations into systemic therapy.

Investigations with *cytotoxic agents* may require a different approach from investigations with *molecular targeted agents*. For instance, research involving so-called cytotoxic agents will often use tumor shrinkage as a measure of anti-tumor activity. However, many of the new therapies currently being developed are targeting specific molecular pathways which impact tumor growth, programmed cell death (apoptosis) and/or new blood vessel formation (angiogenesis). For such agents, tumor shrinkage may not be the most appropriate endpoint to capture activity of the treatment(s) since these drugs aim at halting or slowing down disease progression without necessarily inducing tumor shrinkage.

Furthermore, for targeted agents, a larger treatment benefit can be expected in a sub-group of patients expressing the target than in those for whom the target is not expressed. This has implications for the optimal study design and analysis strategy.

3.3 Patient selection criteria

Phase II / exploratory studies generally focus on a more *advanced patient population* than confirmatory studies. Indeed, it would be unethical to include in an early phase II trial any patients who could not contribute to the evaluation of the treatment’s activity or for whom established active treatment options are available. In late phase II trials selection criteria may be more restrictive, for example in terms of histology and/or prior therapies, if the aim is to test a drug in a very specific patient population.

A treatment may also be tested simultaneously in different populations within the setting of *one large stratified phase II trial* (also called “*basket trials*”). In this case, the sample size calculation applies to each stratum separately and the recruitment is closed for each stratum independently of the others as soon as the required sample has been recruited. The number of cases required for each stratum is not necessarily equal in all strata since the expected level of activity or toxicity may differ.

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 population. To enrich the study population upfront (i.e. to limit the population to patients believed more likely to benefit from the

experimental therapy, e.g. based on biomarker expression) may be tempting to try to improve the chance that the drug will show benefit in the tested subgroup.

In general, enrichment in the early drug development will be efficient if

- there is a very solid biological rationale (preclinical and confirmed in early clinical testing),
- if the biomarker could be identified in very early studies,
- if an assay is available that is both technically valid (i.e. it measures what it is intended to measure) and reliable (its results are reproducible, both within a patient and between observers), and clinically validated with at least a putative cut-off to classify patients as positive or negative.
- the prevalence of the sensitive patients is sufficiently high to justify the development and low enough to justify selection.

One must also consider that all potential patients will need to be screened for marker expression upfront, which induces costs. The marker prevalence will also directly impact the speed of recruitment into the study. Therefore, there should also be already a very strong signal in the sensitive subgroup and only in the sensitive subgroup before considering this approach.

Further considerations relevant to trial feasibility is the availability of tissue material for testing in all patients, the need for central laboratory testing, and the processing time and failure rate of the diagnostic test. Diagnostic test performance (i.e. sensitivity, specificity, positive predictive value, negative predictive value) will directly impact the study performance because of miss-classification of sensitive patients as ineligible and inclusion of insensitive patients erroneously diagnosed positive that tend to dilute the treatment effect.

When limited knowledge concerning the drug's mechanism of action is available at the time of starting the trial, it is recommended not to enrich the population upfront but to use the trial to explore the biomarker further. Indeed enrichment based on the wrong target will result in false negative results and does not leave the option to study an alternative biomarker in the same study since only a subset is represented.

3.4 Endpoints of phase II/ exploratory trials

Endpoints in phase II trials are most often binary outcome measures of activity and/or toxicity, although time to event endpoints may also be used, especially in randomized trials.

Response to treatment is assessed on the basis of objective criteria which measure the decrease in size of prospectively selected "target lesions". International standards are available for measuring response in phase II trials, the most common being the RECIST criteria (<http://www.eortc.org/investigators-area/recist>). The protocol should state whether a complete response (CR), response (CR + PR), or clinical benefit (CR + PR + SD) will be used as the primary endpoint.

It should be underlined that response to therapy is not an appropriate surrogate for therapeutic benefit but only an indicator of anti-tumor activity. Furthermore, tumor response classification is often criticized for the loss of information which is inherent to categorizing a continuous measure of tumor shrinkage. More and more often waterfall plots are used to show the individual changes in tumor size for all patients in a study, and to graphically show the benefit of a new treatment. Designs based on the continuous assessments have also recently been re-discussed (e.g. Karrison et al. 2007).

Progression free survival (PFS) rate or a **similar time to event endpoint** (TTE) evaluated **at a fixed point in time** after randomization / start of treatment (usually at 3 or 6 months) may also be the primary endpoint in phase II trials with

non-cytotoxic agents and in trials where the response to treatment is difficult to assess. The protocol should ensure to obtain an assessment at that time for all patients within a pre-specified time window. The analysis can follow a binary logic, declaring all patients without appropriate follow-up to be failures. However, depending on the setting, this may not be beneficial to the interpretation of the trial. If any drop-out or failure of follow-up (which is not due to or related to the event of interest) is to be feared, it is advisable to design the trial using a Kaplan-Meier or interval-censored estimate at the time point of interest, and carefully calculate (or simulate) the operating characteristics of the proposed design. Furthermore, if historical information on the control treatment is lacking or limited, a randomized approach using the logrank test for comparison might be more appropriate (e.g. Korn's design in Section **Error! Reference source not found.**).

Toxicity is graded according the “Common Terminology Criteria for Adverse Events” (CTCAE) (<http://ctep.cancer.gov/reporting/ctc.html>).

In feasibility studies it may be desirable to document the proportion of patients completing therapy, rate of patients without severe toxicity (to be well defined in the protocol), success rate of a post-treatment surgery, completion rate of treatment at a minimum specified dose in a maximum specified time, or similar endpoints.

For a more extensive overview of endpoints for screening phase II / exploratory trials, please refer to Dhani et al. (2009).

3.5 Type I and type II error

Phase II/exploratory trials are generally designed to reject the null hypothesis reflecting a lack of or an insufficient level of activity. They are not meant to demonstrate that the H1 hypothesis stated for the power calculations is true. This must be taken into account when defining the null hypothesis of the trial.

When deciding on the future of a new agent, two types of errors can be made: a false positive or **type I error** (recommending an inactive agent for further study,) and a false negative or **type II error** (declaring an active agent to be inactive). The false negative is generally considered to be the more serious error of the two when testing a new drug for activity in a disease where few treatment alternatives (established or in development) are available. Indeed, once rejected, the drug will generally have no further chance to show its activity, whereas an inactive drug can still be rejected if its activity is not confirmed at a later stage. A type I error may be seen as more important if the objective is to screen for an active agent in an area where many new drugs are to be tested.

The probability of rejecting an effective drug (*beta*, the size of the type II error) should not be greater than 0.20; 0.05 or 0.10 are the recommended levels. *Alpha*, the size of the type I error, should not be greater than 0.20; 0.10 is recommended in a phase II trial. Preferably the sum of the two error rates should not exceed 0.20 in non-comparative phase II trials. The balance between alpha and beta needs to be considered according to the setting of the trial: number of alternative candidates available, relative cost of losing a candidate drug or of incorrectly concluding activity, etc.

When *toxicity is the primary endpoint*, the consequences of type I and type II errors must be clearly evaluated: a type I error results in recommending a toxic agent for further study, whereas a type II error would stop the development of a non-toxic agent for reasons of toxicity. In this case, the false positive type I error would be considered to be the more serious of the two. Therefore, the levels for each of these error rates need to be selected accordingly.

3.6 Randomization

Randomization protects against selection bias, balances treatment groups for prognostic factors and contributes towards ensuring a valid comparison of the treatments under investigation, such that any treatment effect observed can be reasonably attributed to the treatment under investigation.

Early phase II trials are generally non-randomized, however randomized early phase II trials may be carried out and are of potential interest in the following situations:

- if different schedules of the same drug are to be tested
- if two or more drugs are ready for testing at the same time
- no solid reference data for the primary trial endpoint in the trial population of interest is available to inform the design of a non-randomized or non-comparative phase II design

It is generally recommended that late phase II trials be randomized, especially when

- an *analogue* of an active compound is to be screened,
- multiple experimental arms are to be tested simultaneously,
- a combination of two or more drugs is being studied (Van Glabbeke et al, 2002),
- no solid reference data for the primary trial endpoint in the trial population of interest is available to inform the design of a non-randomized or non-comparative phase II design
 - o e.g. a targeted drug is tested in a molecularly defined subset for which little reference data on outcome is available,
- PFS is used as endpoint.

When screening *analogues* it is recommended to carry out a randomized phase II trial including the parent compound as a control arm in order to reduce the incidence of false negatives. That is, if the analogue is found to be inactive but the parent compound is also found to be inactive in a patient population for which it is normally active, then the negative results with the analogue may be due to the patient population studied and does not necessarily mean that the analogue is inactive.

The main purpose of randomization in phase II trials is to provide an assessment free of selection bias by ensuring reference data (either control or other experimental groups) on a randomized basis, however it also substantially increases the required sample size. Randomization ratios other than 1:1 may be envisaged in this case to reduce the total number of patients and number randomized to the reference arm.

For recommendations on randomization in EORTC trials, please refer to the corresponding section in ST-001-WIN-03. The same rules apply as for phase III / confirmatory trials.

3.7 One stage versus two/multi stage designs

One stage designs are relatively straightforward, require a fixed number of patients and avoid any complexities associated with interrupting accrual at the time of interim analyses. They are considered appropriate when the safety of a treatment is well known or data are already available confirming some level of activity.

If not, early stopping rules (two/multi stage designs) are often recommended to allow for early termination due to drug ineffectiveness or underestimated toxicity. They generally allow the drug to be rejected at the end of each interim stage if there is insufficient activity to warrant further testing. Continuing into the final stage allows a more precise evaluation of the response rate (e.g. Gehan two stage designs, Section **Error! Reference source not found.**) or a decision rule for further investigation of the agent (e.g. Simon, Fleming two stage designs, Sections **Error! Reference source not found.** and **Error! Reference source not found.**).

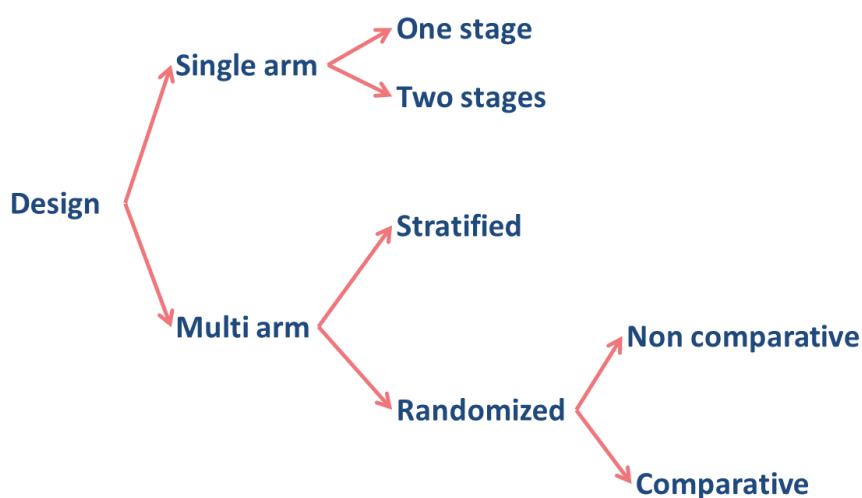
However two/multi stage designs have the practical disadvantage that the trial must be temporarily closed to patient entry between the consecutive stages in order to determine if the criteria for continuation are met, especially if the recruitment is fast and/or if the time to observing the event of interest is long.

4 Statistical designs of phase II / exploratory trials

Options for phase II designs are manifold, even for the simple case when no biomarker is involved.

Figure 1 provides a general summary based on the design considerations discussed in the previous section. In what follows, these categories will be supplemented with a review of examples of theoretical designs used in EORTC trials. For a more detailed compendium on designs for phase II trials in cancer studies, please refer to Brown et al. (2010).

FIGURE 1 GENERAL FLOWCHART FOR PHASE II / EXPLORATORY DESIGNS



4.1 Single arm designs

4.1.1 Fleming or A'Hern design

The Fleming (1982) and A'Hern (2001) design are appropriate designs for late phase II trials with a binary outcome of anti-tumor activity because they provide a decision rule for accepting or rejecting a drug from further study. The A'Hern design uses the exact binomial distribution, whereas the Fleming design is based on a normal approximation of the binomial distribution.

The Fleming design can also be extended to a two-stage design with an early stopping rule for inactivity. A disadvantage of this two-stage design is that although it satisfies the constraints on the size of the type I and type II errors, it does not satisfy any optimality criteria as far as the sample size is concerned.

Recommended software: Sample Size Tables for Clinical Studies Software (based on Machin et al. 2009)

4.1.2 Simon two-stage designs

The Simon two stage designs (1989) provide an extension to the Fleming/A'Hern two-stage design, optimizing the expected sample size while satisfying the constraints on the size of the type I and type II error.

There are two possible strategies for optimization:

- the optimum design: a two-stage design which minimizes the expected sample size if the drug has low activity.
- the minimax design: a two-stage design that minimizes the maximum sample size.

Note that the decision rule of the Simon two-stage design is valid for the planned sample size and adjusted for the interim look. However, the final estimate of the response rate is often reported with an unadjusted p-value and confidence interval, based on the actual sample size. This may in some cases lead to situations where the conclusion from the decision rule differs from the one based on the p-value / confidence interval. It is therefore recommended to report the conditional p-value and confidence interval proposed by Koyama and Chen (2008).

Recommended software:

- Sample Size Tables for Clinical Studies Software (based on Machin et al. 2009)
- For all admissible Simon 2-stage type of phase II designs with other optimality criteria: CTD Systems – Clinical Trial Design Systems.
- For the conditional p-value and confidence interval: SAS macro K:\SAS\Extra EORTC Macros\Simon2st_inf.sas

4.1.3 Gehan two stage design for very early exploratory trials

The Gehan design (1961) is generally preferred for early phase II trials because it minimizes the number of patients treated during the first stage if the drug is totally inactive. If one or more patients respond, the sample size in the second stage is determined so that the true effectiveness of the drug is estimated with approximately a pre-specified precision. If responses were already observed in the phase I stage, the logic of the Gehan approach (to look for at least one response) no longer applies.

Another major drawback is that, unlike the Simon two-stage designs, it does not provide a clear statistical decision rule concerning the further development of the drug. Therefore, it is mentioned here only for completeness, it is rarely applicable for use in contemporary EORTC trials.

4.1.4 Bryant and Day design

The two stage Bryant and Day design (1995) is of interest when response and toxicity are considered to be co-primary endpoints. This design incorporates toxicity in addition to response in the design, providing a decision rule for accepting or rejecting a drug from further study based on both response and toxicity. In this design both α and β should generally be set to 0.10 or less.

In addition to having to temporarily close the trial to patient entry between the two stages, the two stage Bryant and Day design has the additional problem that the denominators for the analysis of response and the analysis of toxicity are potentially different. A one stage Bryant and Day design, which has not been published, is not generally employed. Its merits should be considered on a case by case basis. In some cases a composite endpoint encompassing both treatment activity and toxicity might be preferable to the Bryant and Day design.

Recommended software: SAS Macro K:\SAS\EORTC macros\BRYADAY.SAS.

This macro is preferred over Sample Size Tables for Clinical Studies Software (based on Machin et al. 2009) as it provides summary statistics on the design characteristics, e.g. the probability of early termination, and it provides one-stage as well as two-stage solutions (Statsclub 01/07/2010).

4.1.5 Designs for co-primary activity endpoints

In some circumstances it may be of interest to consider two co-primary endpoints of activity, such as response and progression free survival at a given time (as in the Rinehart “modified Simon design”, 2004) or overall response (CR+PR) and complete response (CR; as in the Lu design, 2005). The treatment is then declared to be active if it meets the criteria for activity for at least one of the two endpoints. The Rinehart design has type I error $< 2 \times \alpha$, whereas the Lu design is optimized to fulfill error conditions on α (for the trial) and the marginal powers for the two endpoints. The Lu approach is to be preferred to that of Rinehart since it gives the possibility to weight the two endpoints differently and doesn't inflate the size of the type I error.

Recommended software: K:\STAT\crr2stage\start.bat

4.1.6 Stratified single or two stage designs

In case it is known in advance that the selected patient response distribution will be heterogeneous, ignoring this heterogeneity may result in a considerable loss of power of the single arm study. In this case, it may be advisable to stratify patients into subgroups, each with a different prognosis and use a weighted combination of response rates and variances in the different strata to come to an overall conclusion (London and Chang, 2005; Chang et al. 2012). Both one-stage and two-stage versions of this design are available. This type of design is generally *not recommended* due to the high risk of rejecting a drug even if it works in one of the subgroups. Furthermore, if the assumed distributions of the strata differ from the observed distribution, this may require a sample size re-estimation prior to completing the study accrual.

Software: SAS code K:\SAS\Extra EORTC Macros\Improved two-stage tests.sas

4.1.7 Sargent's three-outcome design

This design plays on the interpretation of a grey zone of outcomes as "inconclusive" to reach lower sample sizes (Sargent et al., 2001). Rather than limiting the outcome of the hypothesis testing framework to either rejecting the null hypothesis or rejecting the alternative hypothesis based on the observed level of activity, it allows for a third outcome, to reject neither, and use of information from secondary endpoint to formulate an overall conclusion. Note that in the context of a binary primary endpoint in a non-randomized setting it is considered to be less informative than the classical Simon two stage design, and is not recommended for use at EORTC in this particular context.

Recommended software: R macro J:\UNIT\Stat\R code\Sargent-single arm-three outcome\threeoutcome.R

4.1.8 Mick design for the growth modulation index

The Mick design is an example of an alternative single arm phase II design, where the patient serves as its own control. Mick et al. (2000) propose to compare the time to progression for the new treatment to the time to progression for the previous treatment that the patient received. This endpoint is called the growth modulation index (GMI) and the design is based on the proposition that a treatment can be considered effective if the index is greater than 1.33. This design may be acceptable within the EORTC if it is restricted to trials in rapidly progressive disease where the time to progression is short and the eligibility criteria specify: (1) that the initial time to progression is known, (2) that the patient was previously treated in the same institution, (3) the previous treatments are acceptable, (4) that the definition of progression and the follow up schedules are the same in order to minimize potential bias.

4.2 Multiple arm randomized designs

4.2.1 Non-comparative designs

4.2.1.1 Experimental arms only

When appropriate historical control data are available, a preferred approach could be a design where each arm is considered as a separate one arm, one / two stage study (Fleming, A'Hern, Simon, see Section **Error! Reference source not found.**). One or more of several experimental arms can be selected to take forward into a phase III trial.

Alternatively, a drift adjusted A'Hern design can be considered to account for available information from the control arm. It's an extension of the classical A'Hern design (Section 4.1) that introduces an additional step to the final decision rule by comparing the observed response rate in the experimental arm to the response rate observed in the control arm. In situations where the historical control rate might be higher than anticipated, this approach results in a reduced risk of a false positive result.

More information: J:\UNIT\Stat\1. Methodological references\Phase II\Neven et al - flexible screening design allowing for comparison with historical control.pdf

Recommended software: R macro J:\UNIT\Stat\12. Statistical designs\Screening design\screening_1exparm_AnoukV2.R

4.2.1.2 With a reference control arm

In non-comparative randomized phase II trials with an internal (randomized) control, the sample size calculations are carried out as described above using one of the non-randomized designs and applied in the same fashion to each experimental arm. Therefore, while randomized, there is no formal comparative intent in such trials, and the internal control is used to evaluate the veracity of the prior assumptions in the enrolled population.

4.2.2 Comparative designs

4.2.2.1 Randomized discontinuation design

The Randomized Discontinuation Design (Rosner et al., 2002), randomizes a subgroup of patients based on previous response to treatment to continuing or stopping therapy, and is not recommended for use at the EORTC. There are a number of potential problems related to it: (1) the ethics of stopping treatment (2) the randomized subset is a potentially biased selection of patients, (3) randomized comparisons may be misleading in the phase II setting if they are interpreted as definitive evidence, (4) the difficulty of knowing the number of patients to be registered at the first step.

4.2.2.2 Selection designs (no reference arm)

Selection designs are used when several treatment regimens are available for testing. The objective of these multiple arm randomized phase II / exploratory trials with experimental arms only, is to select one of several arms to take forward to a phase III study. The goal of these designs is not to conclude superiority of one arm over another, rather to ensure that if one treatment is clearly inferior to the other, there is a small probability that the inferior treatment will be carried forward to a phase III trial. An indifference region may be set a priori to allow flexibility when the choice based on the primary endpoint is not clear.

Of note, this approach requires that there is a standard of care available against which the selected arm can be tested in the consecutive phase III.

Several selection procedures are available.

4.2.2.2.1 The Simon randomized selection design (Simon, Wittes and Ellenberg)

The Simon randomized selection design (Simon et al. 1985) selects the regimen that results in the highest observed response rate. Sample sizes are given that assure 90% probability to select the best study arm, so long as the true expected response rate exceeds that of any other arm by at least 15% (in absolute terms; e.g. 35% v 20%).

Note that this approach does not specify a minimum level of activity, and therefore does not guarantee that the selected treatment has a clinically relevant effect.

Recommended software: Sample Size Tables for Clinical Studies Software (by Machin et al., 2009)

4.2.2.2.2 The Sargent and Goldberg multiple arm screening design

The Sargent and Goldberg flexible design for multiple arm screening trials (2001) requires the observed difference in success rates of the treatment arms to be larger than a pre-specified quantity. If not, other factors can be considered in the process. The authors argue that this type of design does not have type I error control as primary consideration, therefore any conclusion regarding comparative efficacy with an established control is inappropriate. Nevertheless, it is recommended to perform simulations to document the characteristics of the design in terms of power and type I error under several possible scenarios. An extension of this design including provisions for an interim analysis and/or a comparison to a historical control is available from Wu et al. (2013).

Note that this approach does not specify a minimum level of activity, and therefore does not guarantee that the selected treatment has a clinically relevant effect.

Recommended software: SAS macro K:\SAS\EORTC macros\sargent_goldberg.sas

4.2.2.2.3 Hybrid screening design

The hybrid screening design (Neven et al XXX) combines the principles of a screening design with those of classical phase II designs looking for a minimal level of activity. In a first step, activity is evaluated in each experimental arm using an A'Hern design (Section 4.1). Active experimental arms (if more than one) are then compared using the Sargent and Goldberg design. When using this approach, it is important to consider sufficient power in both stages of the design to ensure the overall power of the design.

More information: J:\UNIT\Stat\1. Methodological references\Phase II\Neven et al - flexible screening design allowing for comparison with historical control.pdf

Recommended software: J:\UNIT\Stat\12. Statistical designs\Screening design\screening_2exparm_AnoukV2.R

4.2.2.3 Screening designs (including a reference arm)

Screening designs typically cover the setting of multiple arm randomized trials, including a standard treatment arm against which all experimental arms are to be compared.

Hybrid designs can be used as well, such as screening based on a comparison with the control arm, among arms that successfully pass a selection procedure (e.g. Fleming or Simon type of condition).

4.2.2.3.1 Korn / Rubinstein randomized screening design

If a randomized trial is used in screening mode, for a treatment regimen that has not previously shown clinical activity, and a larger definitive trial is planned if the agent would show activity in the present trial, Korn et al. (2001) and Rubinstein et al. (2005) propose to use an inflated α (for example 0.20, one-sided test) for a comparative test against a control. Therefore, it uses the sample size calculation of a Phase III trial, with a comparative purpose, to obtain minimal comparative data. This design can be needed in cases where historical controls are of little use, and it is hard to come up with a null hypothesis estimate.

The danger of over-interpretation of the data (as if it were a Phase III trial) needs to be taken into account. Due to the elevated type I error, an inactive agent may lead one in five times to a p-value smaller than 0.20. The results conducted in a smaller subset of patients in the context of a phase II trial may not reflect those obtained in a more representative population enrolled in a large phase III. The estimate of effect size is not as accurate (wide confidence intervals). Rubinstein et al. (2005) suggested that a p value must be smaller than or equal to 0.005 for the phase II to be considered definitive. This cutoff is in line with cutoffs used for phase III interim looks for efficacy. Finally, due to the relatively small number of patients, signals related to toxicity and regimen tolerance may differ from those expected in a phase III setting (Cannistra, 2009).

Note that this approach could be extended to the case of testing K experimental arms versus 1 control. Each arm is formally compared to the control with alpha and beta errors rates of 10% or 20%, based on large differences. This approach increases the number of patients compared to the Fleming design and there are no multiplicity adjustments. It is thus generally not to be recommended.

Recommended software: EAST for phase III with extreme HR and large Alpha

4.2.2.4 Comparative three outcome design

Hong and Wang (2007) proposed an extension of the single arm three-outcome design of Sargent et al. (2001, see Section **Error! Reference source not found.**) into a randomized comparative setting for a binary endpoint. Again, in addition to the usual two outcomes of hypothesis testing, this design allows a grey zone where the final clinical decision is based on the overall evaluation of trial outcomes and other relevant factors.

It's possible to extend this design also for the setting of a time to event endpoint. An Excel sheet is available that can help with the determination of the required number of events, and which provides the parameters to be used in EAST to calculate the required sample size.

Recommended software: Excel sheet available in J:\UNIT\Stat\6. Statsclub\1. Presentations\Phase II - three outcome design _ SLitiere_2014_03_04

4.3 Designs exploring targeted subsets

Options for designs that recruit all-comers and investigate a biomarker are manifold.

4.3.1 Unselected designs

Unselected (possibly stratified parallel) phase II /exploratory trials with strong TR component can be used to investigate the marker effect. Generally referred to as "basket studies", this kind of trial may be stratified for various disease sites, histologies or else in one single protocol.

4.3.2 Biomarker-adaptive designs for one biomarker and one drug

Biomarker-adaptive designs allow switching to an enriched population during the study.

Note that this type of design needs a predefined diagnostic tests (with threshold) and needs to specify the expected effect overall and in the subset of interest (or alternatively in the subset and its complement)

Several strategies can be adopted. In this section we focus on one drug and one biomarker.

4.3.2.1 Biomarker-adaptive parallel Simon two stage design

The proposed phase II design allows for preliminary determination of efficacy that may be restricted to a particular sub-population defined by biomarker status (M+ or M-). It is an extension of the Simon two stage design, starting by enrolling unselected pts during stage I; if very few success in M- subgroup, continue only with M +, else continue with entering all pts. At the end, a conclusion is reached for either the whole group or M+ group only (Jones and Holmgren, 2007).

4.3.2.2 Randomized biomarker-adaptive phase II / exploratory design

This approach is an extension of the biomarker-adaptive parallel Simon two stage design, including randomization and allowing the treatment effect to be specified as a HR (Freidlin et al., 2012).

In a first step, the treatment comparison is done in the M+ subgroup. If a significant effect is observed, a CI is estimated for the HR in the M- subgroup.

If the therapy is only marginally helpful in the M- group, the recommendation will be to perform a biomarker-enriched design, i.e. in the M+ group; if the treatment effect is inconclusive in the M- group, a biomarker-stratified design will be recommended; if the targeted therapy is better than the standard in the M- group, a standard phase III design (i.e. dropping the biomarker) will be recommended.

If on the other hand the targeted therapy is not better than the standard in the M+ group, then a test in the overall population is performed, which will result in either a recommendation to drop the biomarker or stop further testing of the therapy.

4.3.3 Biomarker-adaptive designs for multiple biomarkers and one drug

Several extensions of the biomarker-adaptive designs are available that allow considering multiple biomarkers at the same time.

4.3.3.1 Tandem two-steps phase II

In this approach proposed by Puzstai et al (2007), the goal is to determine whether a new drug has enough clinical activity in an unselected patient population. If it is below the level of interest, a patient selection method is applied to enrich the responding population to meet the targeted level of activity in the molecularly selected group. This can be done simultaneously in several subgroups, and thus extends the design strategies in section 4.3.2 allowing to test several markers at the same time.

First, unselected patients are enrolled during stage I; if sufficient responses are observed, the trial continues unselected, otherwise, a new 2-stage study is started in only M+ patients. At this stage, several biomarkers can be tested simultaneously.

4.3.3.2 Adaptive signature designs

An adaptive signature design (Freidlin and Simon, 2005) is a strategy for generating and prospectively testing an assay or gene expression-signature for sensitive patients, when the signature is not available at the onset of the trial. It can also be used for developing and testing a single marker assay. The design includes a screening stage and a confirmation stage once the biomarker is selected. Such a design was used in the ToPARP trial

(<http://www.clinicaltrials.gov/show/NCT01682772>)

4.3.3.3 Other

When no defined biomarkers are available but enrichment is needed, the randomized discontinuation design (see Section **Error! Reference source not found.**) is an option that allows searching for markers associated with response or long benefit from the treatment and may form the basis of subsequent phase II or phase III development.

4.3.4 Biomarker-adaptive designs for multiple biomarkers and drugs

4.3.4.1 Bayesian adaptive randomization

Bayesian Outcome-adaptive Randomization (BAR) uses patients' response to treatment to increase the probability of randomizing subsequent patients to the most promising arm, whereby the allocation ratio is allowed to change over the course of the trial. Both binary (e.g. response rate) and time to event (e.g. PFS) endpoints can be used.

In general, BAR consists of an allocation scheme based on the posterior probability of one treatment arm being more effective than another. As the posterior probability can be highly variable in the beginning of a trial, Thall and Wathen (2007) proposed to stabilize this quantity by adding a positive tuning parameter. Although developed initially in a two-arm setting, their approach can be extended to a setting with multiple arms. Various variations on constructing the allocation probability in a K-arm ($K > 3$) trial (e.g. computing the maximum or the average samples posteriors) have been suggested in the literature. Due to the allocation imbalance, an increased sample size is generally needed compared to an equally randomized design (ER) to satisfy the type I and II error constraints. Therefore, Trippa and al. (2012) proposed a BAR procedure in which the number of patients in the control arm approximates the number of patients in the best-performing arm. This approach can only be used in a multi-arm setting and is more powerful than ER whenever only one experimental treatment arm is effective.

The use of BAR can result in a flexible trial with a higher overall response rate (compared to ER). Bayesian early stopping rules can be included in all types of BAR designs, but this reduces the difference between BAR and ER. Furthermore, there are multiple drawbacks associated with the use of BAR: changes in randomization ratio are based on efficacy but neglect safety, more resources are required to plan and implement such a trial, and a robust infrastructure is needed to overcome some of the operational and logistical challenges. Both risk of unblinding and population drift have to be taken into account. Additional problems can arise as a result of using a Bayesian framework (e.g. subjective priors) and adaptive analysis (e.g. biased treatment effect).

Due to its challenges, BAR has not been widely adopted in the design of clinical trials. Most well known examples include the BATTLE and the I-SPY 2 trials. In the BATTLE trial (Zhou et al., 2008), NSCLC cancer patients were randomized to one of 4 molecular targeted therapies, according to 5 different biomarker profile groups assessed prior to randomization. Following an initial ER period, the allocation rates were updated proportionally to the marginal posterior of the disease control rate. Difficulties result from the absence of a control group, and the

challenge to make reliable assumptions about the biomarker prevalence and related inferential issues. The software and web-based database application are available on the MT Anderson Cancer website. The I-SPY 2 trial (Barker et al., 2009) is being conducted in the adjuvant breast cancer setting. In contrast to the BATTLE trial, it contains a control arm and new drugs may be added throughout the trial. Currently, insufficient details are available to evaluate the design (no public software, priors require patient-level data from I-SPY 1, results not disseminated).

Overall, BAR is most useful in a rare cancer setting where a short-term endpoint for treatment efficacy is available and accrual is not expected to be too fast relative to the outcome observation time. The accrual period and trial duration have to be long enough for BAR to be implemented. Advantages of this design are most pronounced when a large treatment effect is anticipated (e.g. biomarker-stratified designs) and more than 2 treatment arms are envisaged. Nevertheless, the use of BAR has to be considered carefully and possible advantages have to be weighed against its disadvantages and complexities. Extensive simulations have to be performed to obtain desirable trial characteristics and test the robustness of the design. No software is currently available at EORTC.

More detailed information and references can be found in the relevant Statsclub presentation. (J:\UNIT\Stat\6. Statsclub\1. Presentations\Bayesian adaptive randomization_WG1_2018_01_16.pptx).

4.4 Bayesian designs

Please refer to the published literature.

5 References

- ◆ A'Hern. Sample size tables for exact single-stage phase II designs. *Statistics in Medicine* 2001, 20, 859-866
- Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* 1989, 10, 1-10.
- ◆ Barker AD, Sigman CC, Kelloff GJ, Hylton NM, Berry DA and Esserman LJ. I-SPY 2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy. *Clinical pharmacology & Therapeutics*, 2009;86(1):97-100
- ◆ Brown SR, Brown J, Buyse M, Twelves C, Parmar M, Seymour M and Gregory W. 2010 Choosing your phase II trial design: a practical guide for cancer studies. CTRU, University of Leeds manual.
- ◆ Bryant J and Day R (1995) Incorporating Toxicity Considerations into the Design of Two-Stage Phase II Clinical Trials. *Biometrics*, 51: 1372-1383.
- ◆ Cannistra SA (2009) Phase II trials in *Journal of Clinical Oncology*. JCO, 27(19), 3073-3076.
- ◆ Chang NM, Shuster JJ and Hou W (2012) Improved two-stage tests for stratified phase II cancer clinical trials. *Stat Med* 31, 1688-1698
- ◆ Dhani N et al. Alternate endpoints for screening phase II studies. *Clin Cancer Res* 2009; 15(6): 1873-82.
- ◆ Fleming T (1982). One Sample Multiple Testing Procedure for Phase II Clinical Trials. *Biometrics*, 38: 143-151.
- ◆ Freidlin B, McShane LM, Polley M-YC, Korn EL. Randomized Phase II Trial Designs With Biomarkers. *J Clin Oncol* 2012, 30:3304-3309.
- ◆ Freidlin B and Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clin Cancer Res* 2005;11:7872-8
- ◆ Gehan E (1961) The Determination of the Number of Patients Required in a Preliminary and a Follow-up Trial of a New Chemotherapeutic Agent. *Journal of Chronic Diseases*, 13: 346-353.
- ◆ Hong S, Wang Y. A three-outcome design for randomized comparative phase II clinical trials. *Statistics in Medicine* 2007, 26, 3525-3534

- ◆ Jones CL, Holmgren E. An adaptive Simon Two-Stage Design for Phase 2 studies of targeted therapies. *Contemporary Clinical Trials* 28 (2007) 654–661
- ◆ Kaplan R, Maughan T, Crook A, Fisher D, Wilson R, Brown L, and Parmar M. Evaluating Many Treatments and Biomarkers in Oncology: A New Design. *J Clin Oncol* 2013; 31:4562-4568.
- ◆ Karrison TG, Maitland ML, Stadler WM, Ratain MJ (2007) Design of phase II cancer trials using a continuous endpoint of change in tumor size: application to a study of sorafenib and erlotinib in non small-cell lung cancer. *J Natl Cancer Inst*, 99, 1455-61.
- ◆ Korn EL, Arbuck SG, Pluda JM, Simon R, Kaplan RS, Christian MC. Clinical trial designs for cytostatic agents: are new approaches needed? *Journal of Clinical Oncology* 2001, 19, 265-272.
- ◆ Korn EL and Freidlin B. Outcome-adaptive randomization: is it useful? *J Clin Oncol* 2011; 29: 771-776.
- ◆ Korn EL and Freidlin B. Reply to Y. Yuan et al. *J Clin Oncol* 2011; 29(13): e393.
- ◆ Koyama T and Chen H (2008). Proper inference from Simon's two-stage designs. *Statistics in Medicine*, 27(16): 3145-54.
- ◆ London WB and Chang MN (2005) One- and two-stage designs for stratified phase II clinical trials. *Stat Med* 24, 2597-2611
- ◆ Lu Y., Jin H., and Lamborn KR (2005). A design of phase II cancer trials using total and complete response. *Statistics in Medicine*, 24(20): 3155-3170.
- ◆ Machin D., Campbell M.J., Tan S.B., Tan S.H. (2009). *Sample Size Tables for Clinical Studies*, Third Edition. Wiley-Blackwell, Hoboken, NJ.
- ◆ Mick R., Crowley JJ, Carroll RJ (2000). Phase II clinical trial design for non-cytotoxic anticancer agents for which time to disease progression is the primary endpoint. *Control Clin Trials* 21, 343 – 359.
- ◆ Puzstai L, Anderson K, and Hess KR. Pharmacogenomic Predictor Discovery in Phase II Clinical Trials for Breast Cancer. *Clinical Cancer Research* 2007;13(20): 6080-86
- ◆ Rinehart J, Adjei AA, LoRusso PM, Waterhouse R, Hecht JR, Natale RB, et al. (2004). Multicenter Phase II Study of the Oral MEK Inhibitor, CI-1040, in Patients With Advanced Non-Small-Cell Lung, Breast, Colon, and Pancreatic Cancer. *J Clin Oncol*, 22:4456-4462.
- ◆ Rosner G, Stadler W and Ratain M (2002). Randomized discontinuation design: application to cytostatic anti-neoplastic agents. *J Clin Oncol*, 20: 4478 – 4484.
- ◆ Rubinstein LV, Korn EL, Freidlin B, Hunsberger S, Ivy SP and Smith MA. (2005) Design issues of randomized phase II trials and a proposal for phase II screening trials. *J Clin Oncol*, 23(28): 7199-7206.
- ◆ Sargent DJ, Chan V, Goldberg RM. A three-outcome design for phase II clinical trials. *Controlled Clinical Trials* 2001, 22, 117-125.
- ◆ Sargent D. and Goldberg R. (2001). A flexible design for multiple armed screening trials. *Statistics in Medicine*, 20: 1051 – 1060.
- ◆ Seymour L. et al. The design of phase II cancer trials testing cancer therapeutics. *Clin Cancer Res* 2010; 16(6): 1764-69
- ◆ Simon R., Wittes, R. and Ellenberg, S. (1985). Randomized phase II clinical trials. *Cancer Treatment Reports*, 69: 1375 – 1381.
- ◆ Simon R (1989) Optimal Two-stage Designs for Phase II Clinical Trials. *Controlled Clinical Trials*, 10: 1-10.
- ◆ Thall PF, Wathen JK (2007). Practical Bayesian adaptive randomization in clinical trials. *Eur J Cancer*; 43: 860–867.
- ◆ Trippa et al. (2012). Bayesian Adaptive Randomized Trial Design for Patients with Recurrent Glioblastoma, *J Clin Oncol*; 30: 3258-3263

- ◆ Van Glabbeke M., Stewart W and Armand J.P. (2002). Non randomized phase II trials of drug combinations: often meaningless, sometimes misleading. Are there alternative strategies? *Eur J Cancer*, 38, 635 – 638.
- ◆ Wason, J. M. S. and Trippa, L. (2014), A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. *Statistics in Medicine* 33 (13), 2206–2221.
- ◆ Wu W. Bot B., Hu Y., Geyer SM, Sargent DJ. (2013). A phase II flexible screening design allowing for interim analysis and comparison with historical control. *Contemporary Clinical Trials*, 35, 128-137.
- ◆ Yuan Y. and Yin G. (2011) On the usefulness of outcome-adaptive randomization. *J Clin Oncol* 29(13) e390-e392.
- ◆ Zhou X., Liu S., Kim ES, Herbst RS and Lee JJ. Bayesian adaptive design for targeted therapy development in lung cancer - a step toward personalized medicine. *Clin Trials* 2008; 5; 181-193
- ◆ Phase II Trials in the EORTC (1997). *Eur. J. Cancer*, 33: 1361 – 1363.

6 ASSOCIATED DOCUMENTS

None

7 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	<p>Major changes involve</p> <p>The content reorganized to start with the definition of a phase II/exploratory study and to provide some general design considerations before going to specific examples of designs (Section 2 & 3).</p> <p>The designs reorganized according to the structure illustrated in Figure 1. For each design, the recommended software is updated/provided, if available.</p> <p>Significant changes made in some of the design sections to reflect current practice (Fleming/A'Hern, Simon two stage, stratified designs, Sargent's three outcome design, randomized selection/screening designs)</p> <p>Addition of a section dedicated to designs exploring targeted subsets (4.3)</p>	Saskia Litière	11 Dec 2015
3	<p>Update of non-comparative designs (section 4.2.1.1)</p> <p>Clarifications on:</p> <ul style="list-style-type: none"> - Section 4.2.2.2.1 (Simon, Wittes, Ellenberg design) - Section 4.2.2.2.2 (Simon & Goldberg design) <p>Addition of hybrid screening designs (section 4.2.2.3)</p> <p>Update of "Bayesian adaptive randomization" (section 4.3.2)</p>	Saskia Litière	25 Apr 2018