WHEN SHOULD A PHASE II TRIAL BE RANDOMIZED?

Author: Martin Bigler
Review: Participants of PRC meetings, selected board members

Abbreviations

CC	Coordination Center
GIST	Gastrointestinaler Stromatumor
HR	Hazard ratio
IDMC	Independent data monitoring committee
JCO	Journal of Clinical Oncology
MAMS	Multiple arms and multiple stages
PFS	Progression-free survival
SAKK	Schweizerische Arbeitsgemeinschaft für klinische Krebsforschung

Main goal of phase II trials

The main goal of phase II trials is to screen promising agents for further investigations in phase III trials. Carefully planned phase II trial designs will reduce the chance of observing negative results in the subsequent phase III trials.

Randomized comparative phase II designs are generally acceptable and preferred, if feasible. Non-randomized phase II trials should be the exception (Ratain and Sargent 2009).

Design alternatives

We distinguish between randomized comparative designs, randomized designs with a “calibration” or “reference” arm, and single-arm designs.

Randomized comparative designs

Randomized comparative phase II designs have several advantages.

- They reduce selection bias (Van Glabbeke et Al. 2002, Wieand 2005).
- They provide more reliable data for phase III trial planning both for cytotoxic and non-cytotoxic agents, as they have a better control on uncertainty and potential confounding factors (SAKK consensus 2009).
- The primary analysis might identify a statistically significant difference between the regimens.
- They allow for unbiased comparisons in secondary analyses. Secondary analyses should, however, be reported with extreme caution and clearly identified as exploratory, as the potential for false positives increases with the number of secondary analyses (Wieand 2005).
- For combination therapies, data from single-arm phase II trials without control are difficult to interpret.

- Values such as response rates in the historical control may not be known exactly, i.e. there is uncertainty (Rubinstein et Al. 2005). Also, the historical control patients may differ from the patient population in a given institution for the new trial. Misspecification of the assumed rates may have a serious impact on the validity of the results (Baey and Le Deley 2011).
- Randomized designs are preferred by editors, e.g. JCO (Cannistra 2009).
- Non-randomized trials should never be used if there is an uncertainty regarding the optimal dose (Ratain and Sargent 2009).

There are also some disadvantages:

- Randomized comparative designs substantially increase the sample size requirements (Rubinstein et Al 2009). In addition, accrual is becoming more difficult.
- They have a risk of being misleading and unreliable unless it is known what tumor characteristics are required in patients for a drug to be effective and it is known which patients have those tumor characteristics (Stewart 2010).
- A positive phase II result can never guarantee a positive phase III result, no matter if randomized or not.

Randomized designs with a calibration arm

These designs were used at SAKK in the past. We do not consider these designs the preferred option anymore. The only advantage over single-arm designs they share with randomized comparative designs is that they can reduce selection bias.

They require fewer patients that randomized comparative trials, but the information from the control arm is small. Therefore it can be unethical to use such designs.

Single-arm designs

As stated before, a phase II trial should generally be randomized. There are situations, however, where single-arm designs are an option. Only "not enough patients" is not valid as an argument.

If investigators want to perform a single-arm phase II trial they are asked to complete the *checklist for single-arm phase II designs* (attachment 1). On this list they can provide reasons, why a single-arm design is appropriate in their case. With this list the arguments for a single-arm trial are documented in a clear way.

If a single-arm design is chosen, an effort has to be made to avoid bias. In the *checklist for single-arm phase II designs* the investigators are also asked to provide the points considered for avoiding bias.

Sample size calculation for randomized phase II trials

Consider individual arms separately or jointly?

- Sample size calculations based on between-arm comparison (e.g. with α=0.10 or 0.20) are the preferred option and have become more important (Cannistra 2009, Ratain and Sargent 2009).
- A sample size calculation based on between-arm comparisons with an unequal randomization ratio will always require a higher total sample size than one with an equal ratio. Despite

that, the unequal ratio might be attractive for patients in some situations, e.g. if the patients in the control arm receive no active treatment.

- As mentioned before, sample size calculations separately for each arm were performed at SAKK in the past, but should be the exception and carefully explained.

Hypothesis test, confidence interval or Bayesian approach?

- Most cases: hypothesis test.
  - For between-arm comparison designs as for phase III trials, but with a one-sided α of 0.10–0.20 and power 80–90%, can be used. The experimental treatment will be considered promising for further treatment if the one-sided p-value is smaller than prespecified α.
  - For the exceptional case of randomized design with calibration arm and single-arm designs: Simon two-stage design, A'Hern single-arm designs, etc., for the arms individually. No formal comparison between the arms is planned.
- Other designs are possible, e.g. pick-the-winner (Simon et Al., 1985), randomized discontinuation (Rosner et Al., 2002). See also Seymour et Al. (2010) and the list provided below.
- Bayesian approach: *To be explored*.

Discussion about primary endpoint

- Ideally, an agreement on the choice of the primary endpoint should be reached before sample size calculation.
- The choice depends on the disease, treatment, etc. and should be based on references. These references of related trials should be provided. Sample size should not be considered for the choice of the primary endpoint.
- Clear definition of "what", "when" and "how" for the primary endpoint should be provided.
- Usually, phase II trials use rather a short-term endpoint.
- Often the endpoints in phase II trials are binary. This might be ineffective (Senn 2013). Other authors favor comparisons of a primary outcome measure at a single time point, e.g. PFS rate @ 12 weeks (Ratain and Sargent 2009).
- If the endpoint is a time-to-event endpoint at a specific time point (e.g. PFS @ 12 weeks) the sample size should generally be based on the survival endpoint (Kaplan Meier/Nelson Aalen estimator).

Randomized phase II designs

The following list provides some example of randomized phase II designs.

- MAMS design (Parmar et Al. 2008, Royston et Al. 2003). A design with multiple (≥2) arms – one of them a control arm – and multiple (≥2) stages in a phase II/III setting. Might be feasible for SAKK (SAKK consensus, 2009).

  Definitive outcomes *D*, usually overall survival, and intermediate outcomes *I*, such as progression-free survival, are considered. n_1 patients are randomized to the treatment arms. Based on the observed hazard ratios for *I*, it is decided whether to go to the second stage or not. There, n_2 patients are randomized, and the final test is based on all n_1+n_2 patients and the hazard ratio for *D*.

The STAMPEDE trial is an example of the MAMS design with a pilot stage (outcome safety), three “activity” stages (outcome failure-free survival) and a final “efficacy” stage (outcome overall survival).

- Three-outcome design (Hong and Wang 2007). The standard outcomes of a clinical trial are *reject H_0* and *reject H_1*. Here, a third outcome *statistically inconclusive* is added.
  With p_E and p_C being the observed response rates and π_E and π_C being the true response rates for the experimental and the control treatment, respectively, the outcomes are:
  - If $p_E - p_C \geq d_2$, reject H0: $\pi_E \leq \pi_C$
  - If $p_E - p_C \leq d_1$, reject H1: $\pi_E \geq \pi_C + \delta$
  - If $d_1 < p_E - p_C < d_2$, the result is statistically inconclusive and the clinical decision would be based on the overall results along with many other factors, e.g. cost etc.

  The sample size for the three-outcome design is always less than the standard two-outcome design with the same error rates α and β.
- A randomized screening trial design is proposed by Rubinstein et Al. (2005), a design for the primary endpoint PFS in a randomized phase II trial with modest sample sizes. The idea is to allow α and β to be up to 20% and to assume a hazard ratio between 1.3 and 1.75.
  Note that the power is the probability of rejecting the null hypothesis, if the alternative hypothesis is true. I.e. assuming a too optimistic hazard ratio is equivalent to use a lower power, as the true probability of rejecting the null hypothesis is lower than under the too optimistic assumptions.
- A randomized discontinuation design, proposed by Rosner et Al. (2002). Here, all patients are treated with the study agent for a defined time period and then randomizes patients with SD to either continuation or discontinuation of the experimental treatment.
  The problem is that, depending on the proportion of patients with SD, many patients have to be included in the first, non-randomized stage.
- Case and Morgan (2003) suggest a design for trials evaluating survival probabilities, minimizing either the expected duration of accrual or the expected total study length under H0.
- Pick-the-winner design of Simon et Al. (1985). Here several treatment arms are compared, and the sample size is calculated in a way that the correct arm is selected for further testing in a phase III trial, if there is at least one arm which is superior to the others by some margin.
- Brown et Al. 2011 provide several examples of phase II trial designs together with a guidance for designing such a trial.
- *To be extended*.

Between-arm comparisons in manuscripts

- If the sample size was calculated for a between-arm comparison, then the result of this comparison should be included in the manuscript.
- A confidence interval for the between-arm difference should always be included.
- The p-value of a comparison between arms should be considered as a flag for “promising” or not. The chance of false positive is high due to the high α level. If the p-value is included in the manuscript, caution about its interpretation should be mentioned.
- If the sample size was not calculated for a between-arm comparison then no formal comparison (e.g. tests, HR) should be done.

Other issues

- The stopping rule and decision rule given in the statistical considerations of the trial design form a useful and objective tool for decision making. However, the real decision making is not solely driven by statistical considerations. Information about other endpoints, results from other trials, etc., are also relevant. To avoid biased decision making, the use of an independent data monitoring committee (IDMC) is recommended, mainly in randomized trials.
- Underpowered studies should be avoided. They will most likely result in a "negative trial". The studied agent then will not be explored further, even if it is potentially active (Coffey 2014).
- Investigators may fail to acknowledge the large uncertainty in the estimated response rates from historical control and interpret the results as if they came from a phase III trial.

References

A'Hern (2001). *Sample size tables for exact single-stage phase II designs.* Stat Med **20**(6): 859–66.

Amiri-Kordestani and Fojo (2012). *Why do phase III clinical trials in oncology fail so often?* J Natl Cancer Inst **104**(8): 568–9.

Baey and Le Deley (2011). *Effect of a misspecification of response rates on type I and type II errors, in a phase II Simon design.* Eur J Cancer **47**(11): 1647–52.

Booth et Al. (2008). *Design and conduct of phase II studies of targeted anticancer therapy: Recommendations from the task force on methodology for the development of innovative cancer therapies (MDICT).* Eur J Cancer **44**: 25–9.

Brown et Al. (2011). *Designing phase II trials in cancer: a systematic review and guidance.* Br J Canc **105**: 194–9.

Cannistra (2009). *Phase II trials in journal of clinical oncology.* J Clin Oncol **27**(19): 3073–6.

Case and Morgan (2003). *Design of phase II cancer trials evaluating survival probabilities.* MBC Med Res Methodol **3**(6).

Coffey (2014). *Webinar: Challenges in designing small clinical trials.* Society for Clinical Trials.

Estey and Thall (2003). *New designs for phase 2 clinical trials.* Blood **102**(2): 442–8.

European Science Foundation (2009). *Investigator-Driven Clinical Trials.*
http://www.esf.org/publications/medical-sciences.html, accessed 09.04.2014.

Herson and Carter (1986). *Calibrated phase II clinical trials in oncology.* Stat Med **5**(5): 441–7.

Holmgren (2008). *Are phase 2 screening trials in oncology obsolete?* Stat Med **27**: 556–67.

Hong and Wang (2007). *A three-outcome design for randomized comparative phase II clinical trials.* Stat Med **26**: 3523–34.

Lee and Feng (2005). *Randomized phase II designs in cancer clinical trials: current status and future directions.* J Clin Oncol **23**(19): 4450–4457.

Parmar et Al. (2008). *Speeding up the evaluation of new agents in cancer.* J Natl Cancer Inst **100**(17): 1204–1214.

Ratain and Sargent (2009). *Optimising the design of phase II oncology trials: the importance of randomization.* Eur J Cancer **45**: 275–80.

Rosner et Al. (2002*). Randomized discontinuation design: application to cytostatic antineoplastic agents.* J Clin Oncol **20**(22): 4478–84.

Royston et Al. (2003). *Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer.* Stat Med **22**: 2239–56.

Rubinstein et Al. (2005). *Design issues of randomized phase II trials and a proposal for phase II screening trials.* J Clin Oncol **23**(28): 7199–206.

Rubinstein at Al. (2009). *Randomized phase II designs.* Clin Cancer Res **15**(6): 1883–90.

Rubinstein et Al. (2011). *More randomization in phase II trials: necessary but not sufficient.* J Natl Cancer Inst **103**(14): 1075–7.

SAKK consensus on phase II trial designs, March 31, 2009.

Senn (2013). *Being efficient about efficacy estimation.* Statistics in Biopharmaceutical Research **5**(3): 204–10.

Seymour et Al. (2010). *The Design of Phase II Clinical Trials Testing Cancer Therapeutics: Consensus Recommendations from the Clinical Trial Design Task Force of the National Cancer Institute Investigational Drug Steering Committee.* Clin Cancer Research **16**(6): 1764–9.

Sharma et Al. (2011*). Randomized Phase II Trials: A Long-term Investment With Promising Returns.* J Natl Cancer Inst **103**(14): 1093–100.

Simon et Al. (1985). *Randomized phase II clinical trials.* Cancer Treat Rep **69**(12): 1375–81.

Simon (1989). *Optimal two-stage designs for phase II clinical trials.* Control Clin Trials **10**(1): 1–10.

Stewart (2010). *Randomized phase II trials: misleading and unreliable.* J Clin Oncol **28**(31): e649–e650.

Tang et Al. (2010). *Comparison of error rates in single-arm versus randomized phase II cancer clinical trials.* J Clin Oncol **28**(11): 1936–41.

Taylor et Al. (2006). *Comparing an experimental agent to a standard agent: relative merits of a one-arm or randomized two-arm phase II design.* Clinical Trials **3**: 335–48.

Van Glabbeke et Al. (2002*). Non-randomised phase II trials of drug combinations: often meaningless, sometimes misleading. Are there alternative strategies?* Eur J Cancer **38**: 635–8.

Vickers et Al. (2007). *Setting the bar in phase II trials: The use of historical data for determining "Go/No Go" decision for definitive phase III testing.* Clin Cancer Research **13**: 972–6.

Wieand (2005). *Randomized phase II trials: what does randomization gain?* J Clin Oncol **23**(9): 1794–5.