Published in final edited form as:
Int J Radiat Oncol Biol Phys. 2022 January 01; 112(1): 22–29. doi:10.1016/j.ijrobp.2021.07.1700.

An Overview of Phase II Clinical Trial Designs

Pedro A. Torres-Saavedra, Ph.D.[*], **Kathryn A. Winter, M.S.**[*]
[*]NRG Oncology Statistics and Data Management Center, American College of Radiology, Philadelphia, PA

Abstract

Clinical trials are studies to test new treatments in humans. Typically, these treatments are evaluated over several phases to assess their safety and efficacy. Phase I trials are designed to evaluate the safety and tolerability of a new treatment, typically with a small number of patients (e.g., 20–80), generally spread across several dose levels. Phase II trials are designed to determine if the new treatment has sufficiently promising efficacy to warrant further investigation in a large-scale randomized phase III trial, as well as to further assess safety. These studies usually involve a few hundred patients. This article provides an overview of some of the most commonly used phase II designs for clinical trials and emphasize their critical elements and considerations. Key references to some of the most commonly used phase II designs are given to allow the reader to explore at more detail the critical aspects when planning a phase II trial. A comparison of three potential designs in the context of the NRG-HN002 trial is presented to complement the discussion about phase II trials.

Introduction

Clinical trials are studies to test new treatments in humans. Typically, these treatments are evaluated over several phases to assess their safety and efficacy. Phase I trials are designed to evaluate the safety and tolerability of a new treatment, typically with a small number of patients (e.g., 20–80), generally spread across several dose levels. Phase II trials are designed to determine if the new treatment has sufficiently promising efficacy to warrant further investigation in a large-scale randomized phase III trial, as well as to further assess safety. These studies usually involve a few hundred patients. According to the FDA, approximately 33% of drugs in phase II move to the next phase.[1] Phase II trials also generate insights on adverse events and their management, the types of cancer in which the treatment is effective, and the best regimen for future use in a later phase, depending on the trial design.[2,3] Phase III trials are conducted to obtain definitive evidence on the comparative

Corresponding author: Pedro A. Torres-Saavedra, torresp@nrgoncology.org.
Author responsible for statistical analysis: N/A

Data Availability Statement for this Work: N/A

Disclosures: None

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

efficacy (and safety) of a new therapy in comparison to the current standard-of-care. There are several examples in the literature of promising efficacy or safety results from phase II trials that were not confirmed in subsequent phase III trials.[4]

The development of new therapeutic options in cancer and other diseases has led to innovative designs that combine the phases of a clinical trial to maximize the efficiency (i.e., cost and time) of discovery. The traditional phase I-II-III paradigm in clinical trials is no longer clear-cut. Many contemporary phase I trials incorporate dose-expansion cohorts with a few dozens of patients once the maximum tolerated dose or the optimal biological dose is established to assess toxicity further and obtain preliminary efficacy evidence.[5] These early phase trials with dose-expansion cohorts resemble single-arm phase II trials discussed later. Phase I/II trials incorporating toxicity and efficacy endpoints, such as the EffTox, BOIN12, and TITE-BOIN12 designs, have been commonly used across several therapeutic areas.[6,7] Likewise, seamless phase II/III designs that reduce the time to transition to confirmatory phase III trials have also been extensively used. Many trials, especially those involving radiation therapy, evaluate regimens combining drugs with radiation and possibly surgery, looking for superiority in outcomes, while others assess de-intensification therapies with non-inferiority designs in diseases with a good prognosis.

The main goal of this manuscript is to provide an overview of some of the most commonly used phase II designs and emphasize their critical elements and considerations. Some authors have listed the key factors to consider in a phase II design[8]: therapeutic considerations (e.g., cytotoxic, cytostatic, immunotherapy, combination therapy, biomarker dependent), trial aim (e.g., treatment selection or “go/no-go” decision for phase III), the outcome of interest (e.g., binary or time-to-event endpoint), characteristics of the design (e.g., randomization, stages, etc.), and other practical considerations such as early stopping for futility, availability of previous data, among others. Instead of providing a list of key design considerations, which are intertwined, they are addressed in the discussions of the designs. A comparison of three potential designs in the context of the NRG-HN002 trial is presented to complement the discussion about phase II trials.

I. Endpoints of Phase II Trials

Phase II trials are designed to assess if a treatment has sufficient activity or promise of efficacy or if it provides another benefit to warrant further investigation in a definitive phase III trial. Efficacy in phase II trials could be assessed using the same phase III endpoint when feasible or a validated surrogate endpoint such as tumor response, a time-to-event endpoint, or a biomarker.[9] A common feature of phase II primary endpoints is their ability to be assessed quickly. Traditionally, single-arm phase II oncology trials with single novel cytotoxic agents have used tumor response assessed using the Response Evaluation Criteria in Solid Tumors (RECIST) as the primary endpoint.[10] The proportion of best overall complete and partial responses, called objective response rate (ORR), is often the primary measure of efficacy. However, response as the primary endpoint in phase II trials may not appropriate for combination therapies and molecularly targeted agents, which can have significant disease control manifested through mechanisms not captured by disease response.

Due to the questionable utility of tumor response assessment in contemporary oncology clinical trials, progression-free survival (PFS) has become a commonly preferred primary endpoint in phase II trials.[11,12] In randomized trials, PFS is defined as the time from randomization to disease progression or death of any cause. The definition of disease progression, including what constitutes a failure and how to assess progression, needs to be included in the trial protocol. The use of PFS is beneficial for several reasons, including the need for a shorter follow-up than with an OS endpoint and the fact that salvage therapies or supportive care measures post progression do not dilute the PFS treatment effect. Because there are naturally more PFS than OS events, PFS-based studies typically require smaller sample sizes and can be completed faster. Of course, OS could also be used in diseases with poor prognosis, diseases lacking salvage therapies, or if the PFS cannot reliably be measured. For instance, OS was the primary endpoint in NRG/RTOG 0912 (NCT01236547). This trial tested the addition of pazopanib to intensity-modulated radiation therapy plus paclitaxel in anaplastic thyroid cancer, a disease with a median OS of about seven months (2-year OS 12.9%) for patients treated with the standard therapy.[13] There are other time-to-event endpoints closely related to PFS, such as disease-free survival (DFS) and event-free survival (EFS).[14] DFS is typically used in post-operative phase II head and neck trials with adjuvant therapies combined with radiation. It captures the length of time after a patient survives without any signs or symptoms of the disease after surgery. For example, DFS is the phase II primary endpoint of NRG/RTOG 1216 (NCT01810913), a phase II/III trial comparing two experimental arms combining radiation and cetuximab plus docetaxel or cisplatin plus atezolizumab against the standard chemoradiotherapy. DFS is also used in non-post-operative trials. Finally, when the primary benefit of a new therapy is local disease control, such as in some radiation trials, locoregional control (LRC) could be an appropriate phase II primary endpoint.[15]

Alternative primary endpoints such as patient-reported outcome quality of life (PRO/QOL) or toxicity are becoming more common, particularly in seamless phase II/III clinical trials.[9,16] These endpoints are often used in non-inferiority trials where improved clinical outcomes from the new intervention are not typically expected. Instead, the focus is on benefits to patients that may include better PRO/QOL, fewer toxicities, symptoms, or costs associated with the new therapy. In general, other secondary endpoints in phase II trials, such as toxicity, biomarkers, PRO/QOL, tolerability, treatment compliance, efficacy measures (e.g., distant metastasis, LRC), are encouraged to help with the interpretation of the trial results.

II. Non-randomized Trials

A. Single-arm Trials—In single-arm trials with one stage, eligible patients receive the new intervention, and the primary endpoint analysis is only done at the end of the trial, i.e., no protocol-specified interim futility or efficacy analyses are incorporated into the design. However, a critical element of a phase II trial is that it should minimize the number of patients exposed to ineffective or overly toxic treatments. Therefore, often phase II designs include a planned interim futility analysis to determine whether the trial is unlikely to recommend the new treatment for further testing.[17] In general, single-arm trials that include an interim futility (or efficacy) analysis use a two-stage design. Investigators need to examine the inclusion of a futility rule as it could have a non-trivial impact on the power of the trial. Unlike phase III trials, early stopping due efficacy is discouraged in phase II trials due to the small sample size and high uncertainty around the trial results. Other early stopping rules based on key secondary endpoints such as toxicity, tolerability, or treatment compliance can also be incorporated into a phase II trial design.

Traditionally, two-stage single-arm phase II trials have been used to assess the activity and toxicity of new single agents. There are several two-stage designs for phase II trials[18,19]. Simon's design is a popular option that minimizes exposure of too many patients to ineffective and perhaps too toxic agents[20]. A relatively small number of patients (e.g., <30) are enrolled in the trial and receive the new agent in the first stage. At the end of the first stage, the analysis is an interim futility look to rule out ineffective agents earlier in the trial. After assessing the primary endpoint, typically tumor response using the RECIST in oncology[10], a decision based on the number of responses at the end of this stage is made either to stop the trial and declare the new treatment ineffective or to continue enrolling additional patients in the second stage. For the later, the conclusion on the activity of the new agent is made based on the number of responses at the end of the second stage.

While historical control data are often used to design phase II trials, it is essential to recognize several key limitations of this approach. The interpretation of the trial results against historical data can be quite challenging due to potential differences in populations, the current standard-of-care, or the frequency or method of disease assessments.[18] In the case of uncertain applicability of historical outcomes to the population under study, investigators should avoid using time-to-event primary endpoints, such as PFS or OS. Using the PFS rate at a pre-specified time (e.g., 2-year PFS) could help to reduce bias due to differential assessments between the trial and historical controls. Still, this approach is statistically inefficient compared to time-to-event PFS calculations (i.e., it usually requires a larger sample size).[18] The optimal use of historical controls, synthetic historical controls, and real-world data and evidence is an active research area.[21,22] There has been an explosion of methods and designs using various strategies, such as combining data from different sources that seek to minimize bias from historical controls.[23,24]

Despite the caveats of single-arm phase II trials, these designs are broadly used and sometimes are the most feasible alternative. A controlled randomized phase II trial, ideal in many scenarios, is not always possible, particularly in rare diseases or single-institution trials.[25] In some rare diseases, where there is no standard of care, or clinical outcomes are poor, randomizing patients in a controlled phase II trial could be problematic. Some strategies such as unequal randomization could help in these situations, but they could be insufficient to enhance trial participation—these following references discuss additional considerations when designing single-arm phase II trials.[26–28] Finally, other types of trials use single-arm designs, such as window-of-opportunity trials that aim to assess a treatment before definitive therapy and basket trials that assess a new drug or drug combination in multiple disease populations or tumor types ("the basket").[29–31] Basket trials can be conducted using a randomized design.

B. Non-comparative Randomized Trials—In non-comparative randomized trials (NCRT), patients are randomized to two or more experimental arms. A concurrent control arm is not included in this design. These designs are closely related to "selection designs," in which the arm with the highest observed response rate is selected for further study.[18,32] In an NCRT, each experimental arm is strictly compared against historical controls, either using patient-level data or a benchmark. In this respect, NCRT's resemble multiple single-arm trials. As a consequence, they inherit the same shortcomings from single-arm trials discussed in the previous section. In addition, NCRT's are powered to compare each experimental arm against historical controls, so they are not structured to statistically compare the experimental arms with each other.

The NRG-HN002 trial is an example of a phase II study that used this design.[33] This non-inferiority phase II trial aimed to select a de-intensification arm for further testing in a definitive trial. The primary endpoint was 2-year PFS. The source of the historical control data was a randomized phase III multicenter trial, NRG/RTOG 0522 (NCT00265941), conducted by the same NCI's National Clinical Trials Network (NCTN) group (formerly referred to as a cooperative group). The findings in NRG-HN002 informed the design of its successor trial, NRG-HN005 (NCT03952585)[33]. As a result, a de-intensification regimen with a lower radiation dose plus cisplatin is now being compared against the standard chemoradiation in NRG-HN005. This is a randomized phase II/III non-inferiority trial that was proposed to determine if either or both of two reduced-dose radiation regimens (60 Gy radiation plus cisplatin or Nivolumab) are non-inferior to the standard chemoradiation therapy (70 Gy radiation plus cisplatin) in the same target population as in NRG-HN002. PFS (time-to-event) is the primary endpoint in phase 2. For the phase 3 portion, PFS and MD Anderson Dysphagia Inventory (MDADI) score at 1-year post-IMRT are coprimary endpoints. It is worth mentioning that the randomized phase II in NRG-HN005 is being used to determine if either or both of the two de-escalation regimens warrant further evaluation in the definitive randomized phase III portion based based on preliminary efficacy and quality of life outcomes,.

III. Comparative Randomized Phase II Trials

Randomization aims at balancing the prognostic factors (both known and unknown) between treatment arms. It also provides the proper framework to draw causal inferences. There is a consensus that controlled randomized trials are the "gold standard" to establish a signal of treatment benefit in phase II trials prior to proceeding with a definitive phase III trial[27]. Expert consensus is particularly united on the utility of this randomized phase II to phase III pathway for trials with time-to-event endpoints, disease processes with unclear natural histories, and those with biomarker-guided designs. This idea of conducting small, randomized phase II trials to obtain non-definitive evidence of an experimental regimen against a standard therapy was initially called "randomized phase II screening trials."[32] Some authors have criticized the use of randomization in phase II trials, arguing that the selection of patients more likely to benefit from therapy should be the aim instead of randomizing a heterogeneous cohort of patients to treatment groups.[28] However, these concerns can be addressed using better designs, such as biomarker-enrichment approaches, which include randomization.[34–38]A thorough discussion on the use of randomization in phase II trials can be found in Grayling et al.[27]

Randomized phase II trials usually incorporate an interim futility analysis. As an example of a simple futility rule, NRG-HN004, a randomized phase II/III trial comparing IMRT plus durvalumab against IMRT plus cetuximab (control arm) in cisplatin-ineligible head and neck cancer patients (NCT03258554), included an interim futility analysis in phase II after 50% of the required PFS events transpired[39]: If the observed hazard ratio is ≥ 1, favoring the control arm, then early stopping is considered, with the conclusion being that the new regimen would not be a candidate for further evaluation in phase III. Futility rules in randomized phase II trials can be beneficial for early identification of treatment regimens that should not move forward, saving time and resources; however, careful consideration needs to be given to the futility rule and its timing to minimize the risk of an erroneous conclusion.

IV. Sample size and characteristics of a non-comparative, randomized trial vs. controlled, randomized trials: the NRG-HN002 case

The NRG-HN002 trial is an ideal framework to discuss potential alternative phase II designs that could have achieved the same goal of selecting a de-intensification therapy for a subsequent definitive non-inferiority trial. Three options will be discussed: the original NRG-HN002 design[33] with a minor modification and two alternative controlled randomized designs. In NRG-HN002, an estimate of the 2-year PFS rate for the target population with the standard chemoradiation was 91%, based on historical data from NRG/RTOG 0522. The protocol specifies that a 2-year PFS less than 85% for a de-intensification regimen would be considered unacceptable.

- ***Scenario 1* - A two-arm non-comparative randomized trial (NCRT) with 2-year PFS endpoint (NRG-HN002 design with a chi-square test for a single proportion instead of an exact binomial test):** Following the non-inferiority criteria established in this trial, the null hypothesis H_0 with a de-intensification regimen states that 2-year PFS ≤ 85% and the alternative hypothesis H_1 that 2-year PFS > 85%, with a targeted rate of 91% based on historical controls. This design involves the comparison of the two de-escalation regimens against the historical controls using a performance criterion approach.[40] A significant caveat of using a point estimate based on historical controls is that the uncertainty around this number is effectively ignored[26]. If H_0 is rejected, then the 2-year PFS for a de-intensified IMRT arm is deemed acceptable for further evaluation. With one-sided α=0.10 and 80% power for each comparison, an NCRT would need 272 patients overall (136/arm) using a chi-square test for one proportion. Assuming a projected accrual of 15 patients/month, the trial duration, including accrual and follow-up for primary endpoint assessment, with this design is projected to be 3.5 years.

- ***Scenario 2* - A three-arm controlled randomized trial (CRT), using a non-inferiority (NI) design with 2-year PFS endpoint:** This design dictates that a de-intensification regimen arm would be non-inferior to the concurrent control arm if the difference in 2-year PFS rates between the control and a de-intensification arm is less than

6%. This design involves the comparison of the two de-intensification regimens against the concurrent control arm using 2-year PFS endpoint. That is, the null hypothesis is that the 2-year PFS probability for the de-intensification is at least 6% worse (i.e., inferior) than the control regimen, and the alternative hypothesis is that the 2-year PFS probability for the deintensification arm is no more than 6% worse than the control regimen (i.e., non-inferior) If the upper limit of a two-sided 80% confidence interval (CI) for the difference in 2-year PFS rates between de-intensification and control arms, based on the normal distribution approximation, is less than 6%, then a de-intensified IMRT arm is declared non-inferior and will move to further testing.[41] This design would have required 618 patients overall (206/arm and 412/comparison of each experimental arm against the concurrent control arm). The trial duration is projected to be 5.4 years.

- ***Scenario 3* - A three-arm controlled randomized trial (CRT), NI design with PFS endpoint:** Instead of using 2-year PFS (i.e., binary outcome), this design uses PFS, a time-to-event endpoint. We use the same 2-year PFS rates for the control arm (91%) to derive the NI margin based on the hazard ratio (HR) for PFS. A de-intensified IMRT arm would still be considered non-inferior to the control arm if 2-year PFS absolute difference is less than 6% (i.e., <85%). Under the exponential assumption distribution on PFS rates, the previous absolute difference in 2-year PFS rates translates into an HR(experimental/control)=1.72, the NI margin to test the NI hypothesis on PFS between a de-intensification therapy and the control arm. The hypothesis testing for each comparison is done using a two-sided 80% CI for the HR(experimental/control) based on a Cox proportional hazards model with the treatment arm as a covariate.[41] This design would have required 540 patients overall (180/arm and 360/comparison). The trial duration is projected to be 5.4 years. Note that this design requires fewer patients for the same trial duration as for the fixed time point (2-year PFS) design. Conversely, if 618 patients were used in the design with PFS, the trial duration would be shorter than the fixed time point (2-year PFS) design (61.7 months).

Table 1 displays some of the main elements of the three designs considered above. These results suggest that the required number of patients and duration of the randomized controlled trials compared to the NCRT is almost double. Nonetheless, randomized controlled trials are more likely to provide reliable information on the treatment effect, as they are not hampered by the potential biases of single-arm phase II trials. These biases could arise from changes in the standard-of-care, differences in disease assessments (follow-up schedule and techniques), or changing patients' characteristics in the trial versus historical controls. The sample size alone should not be used to make a final decision on the design of a trial. Other considerations, particularly the appropriateness of historical control data, are crucial, as previously discussed. It is essential to highlight that this exercise was done in the context of a NI trial with specific PFS rates for the target population and a set of assumptions. Therefore, these differences in sample sizes between the NCRT and the CRT alternatives are likely to differ in other scenarios; for instance, they might be less remarkable in the context of a phase II superiority trial.

When NRG-HN002 was designed approximately eight years ago, an NCRT was deemed appropriate and the most feasible option to select a de-intensification arm for further evaluation in a subsequent phase III trial. The available historical data from a multicenter phase III trial was considered appropriate to establish a performance criterion for the 2-year PFS for the de-intensification arms. However, the importance of critically assessing the appropriateness of using historical control data in an NCRT design cannot be emphasized enough. Moreover, if this trial were designed today, it most likely would have been a randomized phase II/III trial with a concurrent control arm.

V. Beyond the conventional randomized phase II trial

Phase II clinical trials now play a more diverse role in the search for new therapies. Efficient strategies such as seamless phase II/III designs have been adopted across different therapeutic areas[42]. For example, NRG/RTOG 1216 (NCT01810913), NRG-HN005 (NCT03952585), and NRG-HN006 (NCT04333537) are NRG Oncology trials conducted through the NCI's NCTN that use such designs. In seamless phase II/III trials, a phase II portion is conducted with one or several experimental arms against a common control arm. Phase II serves as a "go/no-go" decision, sometimes involving arm selection. Endpoints such as PRO/QOL or toxicity are becoming more common in seamless phase II/III clinical trials to further inform this selection[9,16]. Critically important is the fact that the phase III analysis set *includes* patients enrolled during the phase II portion. This strategy is more efficient than conducting separate phase II and III trials.[43,44] However, careful consideration should be given to the phase II endpoint selection as it plays a crucial role in the operating characteristics of the design.

Umbrella designs are relatively new players and are used to evaluate multiple investigational drugs administered as single drugs or combination in a single disease population or tumor type ("the umbrella").[30,31] Patients with a molecular target are randomized to an experimental and a control arm in a phase II trial. These trials can also include single-arm designs followed by a randomized phase II and/or III trial.[45] Unlike basket trials that involve several tumor types within a molecular substudy or basket, often employing single-arm trials and response endpoint, umbrella trials focus on a single tumor type and several molecular targets, more commonly using randomized controlled trials within each molecular substudy.

Multi-arm multi-stage trials (MAMS), with several experimental arms and multiple adaptive stages with intermediate endpoints at early stages (e.g., PFS, biomarker, or response) and a definitive endpoint (OS) at later stages, are also being used in recent trials.[46] Early stages in MAMS trials share some of the characteristics of phase II trials. One key benefit of the MAMS concept is its efficiency, as multiple treatments can be assessed at once, and the incorporation of adaptive decision-making allows the pruning of ineffective therapies and thus strengthening more successful strategies.

Summary

The array of phase II designs continues to evolve, with new concepts entering into the already extensive list of potential options for phase II and hybrid trials that traverse the classical stages. Randomization and inclusion of a concurrent control arm are highly recommended in phase II trials, particularly for time-to-event endpoints or biomarker-guided designs. In selected scenarios, such as rare diseases, the use of historical controls or real-world data (RWD) may be justified. Disease response has been a traditional endpoint for single-arm phase II trials with new agents. For randomized trials, PFS has been preferred in contemporary trials with molecularly targeted trials or combination therapies. In some situations, such as a disease with a poor prognosis, OS is an appropriate endpoint in a phase II trial. There exist new designs integrating phases of a clinical trial, such as seamless phase II/III, MAMS, or hybrid phase I-II/III designs. Planning of trials involving these complex designs demands statistical expertise given all the non-trivial elements involved in it (e.g., arm selection, multiple testing, group sequential tests, etc.).

Funding Sources:

National Cancer Institute NRG Oncology Statistical and Data Management Center grant (U10CA180822)

Glossary

Concurrent control

Group of patients concurrently randomized to a control arm in the same clinical trial as the experimental arms.

Historical controls

Group of patients treated with the standard therapy in a separate study. Historical controls must meet some requirements to be deemed appropriate.

Non-comparative randomized clinical trial

A trial that includes randomization to several experimental arms but not to a concurrent control arm. These trials are powered to compare each experimental arm against historical controls, so they are not structured to statistically compare the experimental arms with each other.

Controlled randomized clinical trial

A trial that includes randomization to a concurrent control arm and one or more experimental arms.

Progression-free survival (PFS)

In randomized trials, generally, PFS is the time from randomization to disease progression or death of any cause.

Disease-free survival (DFS)

In randomized trials, generally, DFS is the time from randomization to disease recurrence/failure or death of any cause. DFS is also called relapse-free survival (RFS).

iomarker-guided design

A design that incorporates one or more biomarkers to determine eligibility or treatment assignment in a clinical trial

Non-inferiority trial

A clinical trial to assess whether a new intervention is not worse than the standard of care (control arm), as determined by a non-inferiority margin.

Surrogate endpoint

In a broad sense, a surrogate is a clinical endpoint, laboratory measure, or a physical sign that is intended as a substitute for the clinical endpoint of interest. The most commonly adopted statistical definition of a surrogate is based on the Prentice criteria.[47,48] Example of surrogate endpoints used for drug approvals can be found here.[49]

References

1. U.S. Food and Drug Administration (FDA). Step 3: Clinical Research. The Drug Development Process https://www.fda.gov/patients/drug-development-process/step-3-clinical-research.
2. Bornkamp B, Pinheiro J & Bretz F MCPMod : An R Package for the Design and Analysis of Dose-Finding Studies. J. Stat. Soft 29, (2009).
3. Saville BR & Berry SM Efficiencies of platform clinical trials: A vision of the future. Clinical Trials 13, 358–366 (2016). [PubMed: 26908536]
4. Rovin L 22 Case Studies Where Phase 2 and Phase 3 Trials Had Divergent Results. 44 (2017).
5. Iasonos A & O'Quigley J Design Considerations for Dose-Expansion Cohorts in Phase I Trials. JCO 31, 4014–4021 (2013).
6. Yan F, Thall PF, Lu KH, Gilbert MR & Yuan Y Phase I–II clinical trial design: a state-of-the-art paradigm for dose finding. Annals of Oncology 29, 694–699 (2018). [PubMed: 29267863]
7. Zhou Y, Lin R, Kuo Y-W, Lee JJ & Yuan Y BOIN Suite: A Software Platform to Design and Implement Novel Early-Phase Clinical Trials. JCO Clinical Cancer Informatics 91–101 (2021) doi:10.1200/CCI.20.00122. [PubMed: 33439726]
8. Brown SR et al. Designing phase II trials in cancer: a systematic review and guidance. Br J Cancer 105, 194–199 (2011). [PubMed: 21712822]
9. Dhani N, Tu D, Sargent DJ, Seymour L & Moore MJ Alternate Endpoints for Screening Phase II Studies. Clin Cancer Res 15, 1873–1882 (2009). [PubMed: 19276273]
10. Eisenhauer EA et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). European Journal of Cancer 45, 228–247 (2009). [PubMed: 19097774]
11. Gill S & Sargent D End Points for Adjuvant Therapy Trials: Has the Time Come to Accept Disease-Free Survival as a Surrogate End Point for Overall Survival? The Oncologist 11, 624–629 (2006). [PubMed: 16794241]
12. Del Paggio JC et al. Evolution of the Randomized Clinical Trial in the Era of Precision Oncology. JAMA Oncol 7, 728 (2021). [PubMed: 33764385]
13. Sherman EJ et al. 1914MO Randomized phase II study of radiation therapy and paclitaxel with pazopanib or placebo: NRG-RTOG 0912. Annals of Oncology 31, S1085 (2020).
14. U.S. Department of Health and Human Services. Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics. https://www.fda.gov/media/71195/download (2018).
15. Beitler JJ et al. Final results of local-regional control and late toxicity of RTOG 9003: a randomized trial of altered fractionation radiation for locally advanced head and neck cancer. Int J Radiat Oncol Biol Phys 89, 13–20 (2014). [PubMed: 24613816]
16. Wilson MK, Karakasis K & Oza AM Outcomes and endpoints in trials of cancer treatment: the past, present, and future. The Lancet Oncology 16, e32–e42 (2015). [PubMed: 25638553]
17. Stallard N, Whitehead J, Todd S & Whitehead A Stopping rules for phase II studies. Br J Clin Pharmacol 51, 523–529 (2001). [PubMed: 11422011]
18. Rubinstein L Phase II design: history and evolution. Chinese Clinical Oncology 3, 7–7 (2014). [PubMed: 25842085]
19. Fleming TR One-Sample Multiple Testing Procedure for Phase II Clinical Trials. Biometrics 38, 143 (1982). [PubMed: 7082756]
20. Simon R Optimal two-stage designs for phase II clinical trials. Controlled Clinical Trials 10, 1–10 (1989). [PubMed: 2702835]
21. Thorlund K, Dron L, Park JJ & Mills EJ Synthetic and External Controls in Clinical Trials – A Primer for Researchers. CLEP Volume 12, 457–467 (2020).

22. Franklin JM et al. When Can Nonrandomized Studies Support Valid Inference Regarding Effectiveness or Safety of New Medical Treatments? Clinical Pharmacology & Therapeutics n/a,.

23. Ghadessi M et al. A roadmap to using historical controls in clinical trials – by Drug Information Association Adaptive Design Scientific Working Group (DIA-ADSWG). Orphanet J Rare Dis 15, 69 (2020). [PubMed: 32164754]

24. Lim J et al. Minimizing Patient Burden Through the Use of Historical Subject-Level Data in Innovative Confirmatory Clinical Trials: Review of Methods and Opportunities. Drug Inf J 52, 546–559 (2018).

25. Grossman SA, Schreck KC, Ballman K & Alexander B Point/counterpoint: randomized versus single-arm phase II clinical trials for patients with newly diagnosed glioblastoma. Neuro-Oncology 19, 469–474 (2017). [PubMed: 28388713]

26. Foster JC, Freidlin B, Kunos CA & Korn EL Single-Arm Phase II Trials of Combination Therapies: A Review of the CTEP Experience 2008–2017. J Natl Cancer Inst 112, 128–135 (2019).

27. Grayling MJ, Dimairo M, Mander AP & Jaki TF A Review of Perspectives on the Use of Randomization in Phase II Oncology Trials. JNCI: Journal of the National Cancer Institute 111, 1255–1262 (2019). [PubMed: 31218346]

28. Stewart DJ Randomized Phase II Trials: Misleading and Unreliable. JCO 28, e649–e650 (2010).

29. Farlow JL, Birkeland AC, Swiecicki PL, Brenner JC & Spector ME Window of opportunity trials in head and neck cancer. JCMT 2019, (2019).

30. Park JJH, Hsu G, Siden EG, Thorlund K & Mills EJ An overview of precision oncology basket and umbrella trials for clinicians. CA A Cancer J Clin 70, 125–137 (2020).

31. Simon R Critical Review of Umbrella, Basket, and Platform Designs for Oncology Clinical Trials: Review of umbrella, basket, and platform trial designs. Clin. Pharmacol. Ther 102, 934–941 (2017). [PubMed: 28795401]

32. Rubinstein LV et al. Design Issues of Randomized Phase II Trials and a Proposal for Phase II Screening Trials. JCO 23, 7199–7206 (2005).

33. Yom SS et al. Reduced-Dose Radiation Therapy for HPV-Associated Oropharyngeal Carcinoma (NRG Oncology HN002). JCO 39, 956–965 (2021).

34. Freidlin B & Korn EL Biomarker enrichment strategies: matching trial design to biomarker credentials. Nat Rev Clin Oncol 11, 81–90 (2014). [PubMed: 24281059]

35. Freidlin B, McShane LM, Polley M-YC & Korn EL Randomized Phase II Trial Designs With Biomarkers. JCO 30, 3304–3309 (2012).

36. Hu C & Dignam JJ Biomarker-Driven Oncology Clinical Trials: Key Design Elements, Types, Features, and Practical Considerations. JCO Precision Oncology 1–12 (2019) doi:10.1200/PO.19.00086.

37. Mehta C, Schäfer H, Daniel H & Irle S Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints. Statist. Med. 33, 4515–4531 (2014).

38. Renfro LA, Mallick H, An M-W, Sargent DJ & Mandrekar SJ Clinical trial designs incorporating predictive biomarkers. Cancer Treatment Reviews 43, 74–82 (2016). [PubMed: 26827695]

39. Wieand S, Schroeder G & O'Fallon JR Stopping when the experimental regimen does not appear to help. Stat Med 13, 1453–1458 (1994). [PubMed: 7973224]

40. Viele K et al. Use of historical control data for assessing treatment effects in clinical trials. Pharmaceut. Statist. 13, 41–54 (2014).

41. Non-Inferiority Clinical Trials to Establish Effectiveness Guidance for Industry. 56.

42. Maca J, Bhattacharya S, Dragalin V, Gallo P & Krams M Adaptive Seamless Phase II/III Designs —Background, Operational Aspects, and Examples. Drug Information J 40, 463–473 (2006).

43. Friede T et al. Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: An application in multiple sclerosis. Statist. Med. 30, 1528–1540 (2011).

44. Korn EL, Freidlin B, Abrams JS & Halabi S Design Issues in Randomized Phase II/III Trials. JCO 30, 667–671 (2012).

45. Kaplan R et al. Evaluating Many Treatments and Biomarkers in Oncology: A New Design. J Clin Oncol 31, 4562–4568 (2013). [PubMed: 24248692]

46. Millen GC & Yap C Adaptive trial designs: what are multiarm, multistage trials? Archives of Disease in Childhood - Education and Practice 105, 376–378 (2020).

47. Prentice RL Surrogate endpoints in clinical trials: Definition and operational criteria. Statistics in Medicine 8, 431–440 (1989). [PubMed: 2727467]

48. Heller G Statistical controversies in clinical research: an initial evaluation of a surrogate end point using a single randomized clinical trial and the Prentice criteria. Ann Oncol 26, 2012–2016 (2015). [PubMed: 26254442]

49. U.S. Food and Drug Administration (FDA). Table of Surrogate Endpoints That Were the Basis of Drug Approval or Licensure. https://www.fda.gov/drugs/development-resources/table-surrogate-endpoints-were-basis-drug-approval-or-licensure.

Highlights

- Phase II trials are designed to assess if a treatment has sufficient signal of activity or other meaningful benefit to warrant further investigation in a definitive phase III trial. It is common to include assessment of toxicity and other key secondary endpoints such as quality of life (QOL), biomarkers, locoregional control, and distant metastasis, among others.

- Evidence from a phase II trial should not be used to change medical practice, except for situations in which a definitive phase III trial is not feasible.

- The sample size in phase II trials is typically small to moderate, ranging from tens to hundreds.

- In phase II trials, one-sided 0.05–0.20 are typical values for the type I error probability and at least 80% for the power.

- Disease response is the preferred primary endpoint for single-arm phase II trials, particularly with single cytotoxic agents. For randomized trials, progression-free survival (PFS) and disease-free survival (DFS) capture disease progression and survival status. Overall survival (OS) can be used as the primary endpoint in phase II trials but is typically limited to diseases with poor prognoses. In some trials with radiation therapy, locoregional control is used as the primary endpoint.

- In phase II non-inferiority trials, the primary endpoint often relates to the most important benefit for patients of the new treatment (e.g., toxicity, quality of life, etc.)

- Randomized phase II trials with a control arm are preferred. Single-arm designs are an option for trials with novel agents with promising activity, rare diseases, lack of standard treatments, salvage settings, or the absence of reliable historical controls. However, a modest upward drift in modern versus historical control rates can increase the risk of concluding that an ineffective treatment warrants further investigation incorrectly (i.e., false positive).

Do's and Don'ts

- Do consult with an expert in biostatistics about the features of potential trial designs: endpoints, implications, advantages and disadvantages, and feasibility. Typically, the choice of a trial design requires considerable deliberation among the stakeholders.
- Clearly define all endpoints in the trial. For instance, when using PFS the definition of disease progression must be clearly stated in the protocol, and results' publications, with the details about what constitutes a PFS failure and the acceptable methods to assess disease progression.
- Do not be overly optimistic about the targeted treatment effect size. The main problem with being overly optimistic is that if the actual treatment effect size is smaller (e.g., HR closer to 1), the study will be underpowered with the undesirable consequence of abandoning a promising therapy with a more realistic treatment effect.
- Do not assume a single-arm trial is the best option for rare diseases. Think about the standard of care outcomes in this population, and projected accrual rates. There are examples of randomized controlled trials in rare diseases, such as NRG/RTOG 0912 and NRG/RTOG 1008 (NCT01220583).
- Do randomize and include a concurrent control arm whenever feasible, particularly for time-to-event endpoints. Otherwise, if historical controls are the most viable alternative, make sure you the trial design team understands the natural history of the disease and critical details of historical data, such as the patient characteristics, available treatments and staging, disease assessment definitions, follow-up schedules, etc.
- Do remember that the ultimate goal of the phase II trial is to identify treatment regimen(s) that have the best opportunity of showing benefit in a subsequent definitive phase III trial.

Case vignette

It is now established that patients with p16-positive oropharyngeal cancer (OPC) and minimal smoking history have a uniquely favorable prognosis. These markedly improved outcomes have motivated a variety of de-intensification therapies to reduce toxicity and improve quality-of-life without compromising locoregional control. The NRG-HN002 phase II clinical trial randomized patients with favorable p16-positive OPC into two parallel non-comparative regimens of reduced-dose intensity-modulated radiation therapy (IMRT) with or without cisplatin (IMRT 6 weeks [60 Gy] plus concurrent cisplatin 40 mg/m^2 and IMRT 5 weeks [60 Gy]). The main aim of this trial was to select an experimental arm to compare against the standard of care (IMRT plus cisplatin 100 mg/m^2 every three weeks) in a subsequent non-inferiority (NI) definitive trial. The investigators needed to determine an acceptable de-intensification regimen to motivate a phase III trial while ensuring the quality-of-life improvement was sufficient to justify the risk of a de-intensification trial in a curable disease.

Table 1.

Sample size and characteristics of potential designs for NRG-HN002 (1:1 randomization). One-sided alpha of 0.10 and 80% power per comparison

	Primary endpoint	# of arms	# of patients per arm (overall)	Accrual duration (months)	Trial duration (months/ years)
NCRT (NRG-HN002)[a]	2-yr PFS	2	136 (272)	18.1	42.1 / 3.5
CRT with NI design[a]	2-yr PFS	3	206 (618)	41.2	65.2 / 5.4
CRT with NI design[b]	PFS	3	180 (540)	36.0	65.4 / 5.4

NCRT: Non-comparative randomized trial; CRT: Controlled, randomized trial.

[a] 24 months of follow-up after accrual closure for primary endpoint (2-yr PFS) completion (i.e., all patients with at least two years of follow-up)

[b] 29 months of follow-up after accrual closure for primary endpoint (PFS) completion (i.e., the required number of events for the final analysis based on 61 PFS events/comparison).