Use of Partial Surrogate Endpoints in Integrated Phase II/III Designs

Sally Hunsberger, Yingdong Zhao, and Richard Simon

From the Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute Bethesda MD

Address reprint requests to Sally Hunsberger, PhD, Biometric Research Branch 6130 Executive Blvd, EPN-8120, MSC 7434 National Cancer Institute, Bethesda MD, 20892; phone 301-402-0637; fax 301-4020560; e-mail: sallyh@ctep.nci.nih.gov

ABSTRACT

The traditional oncology drug development paradigm of single arm phase II studies followed by a randomized phase III study has limitations for modern oncology drug development. Interpretation of single arm phase II study results is difficult when a new drug is used in combination with other agents and when progression free survival is used as the endpoint rather than tumor shrinkage. Randomized phase II studies are more informative for these objectives but increase both the number of patients and time required to determine the value of a new experimental agent. In this paper, we compare an integrated phase II/III study design to other study designs to determine the most efficient drug development path in terms of number of patients and length of time to conclusion of drug efficacy on overall survival.

1. Introduction

The clinical development of oncology drugs has traditionally involved three distinct phases, each with its own goal and characteristic design. In phase I the maximum tolerated dose of the drug is determined, the underlying assumption being that higher doses, although more toxic to normal tissue, are more effective for eradicating tumor. Phase II studies attempt to determine whether anti-tumor effect in a particular diagnostic category is sufficient to warrant conducting a phase III clinical trial. Anti-tumor effect has traditionally been evaluated using an endpoint such as tumor shrinkage. Phase II studies are typically single arm studies with 15-40 patients per diagnostic category. Phase III clinical trials are generally large randomized controlled studies with the endpoint being a direct measure of patient benefit, such as survival..

The classic paradigm described above has several limitations for modern oncology drug development. First, successful development of agents that extend survival in patients with cancer has led to the need to study combinations of agents. This makes the design of phase II studies more complex[1] and means that objective responses in single arm phase II studies of combination regimens containing a new drug do not necessarily represent evidence of anti-tumor activity for the drug. To interpret the phase II study one needs a comparison of the activity of the combination containing the new drug to the activity of the regimen given at maximum tolerated doses without the new drug. Such a comparison, if based on prospective randomization would require a much larger sample size than the traditional single arm phase II trial. The limitations of using historical control information for estimating the activity of the control regimen are well documented[2] and even if such information is used, larger sample sizes are required since a comparison is involved[3,4].

The traditional paradigm is also problematic for the development of drugs which may inhibit tumor growth without shrinking tumors. A design based on tumor shrinkage may indicate that a potentially active drug is inactive. As a solution investigators are beginning to use progression free survival (PFS) (defined as time from entry on study to documented progression or death) as an endpoint in phase II studies. It is, however, very difficult to reliably determine whether a new drug extends PFS in a single arm phase II trial. Whereas tumors rarely shrink spontaneously, PFS times often vary widely among patients and determining whether a drug has extended PFS requires the measurement of PFS times for a comparison group of patients who did not receive the drug.

After treatment with active agents, response rates or PFS intervals often vary widely among phase II studies because of variation in patient selection and response measurement. Consequently, single arm phase II studies of combination regimens using tumor shrinkage endpoints or of single agents using PFS endpoints are problematic. Randomized phase II studies of a new regimen containing the drug of interest to a control regimen not containing the drug, can be more reliable but they require larger numbers of patients. This increases both the time and cost of developing drugs. The resource drain from randomized studies during phase II is exacerbated by the fact that the number of studies that need to be performed has increased dramatically. This increase is due to the fact that the number of new agents to be explored has increased and the interest in studying combinations of active agents with and without new agents.

Rubinstein et al[5] discuss the challenges of drug development with molecularly targeted agents. They describe the pitfalls of single arm studies and recommend use of randomized phase II studies where type I error rates are relaxed from the traditional .05 to .20. These issues were also described by Simon et al[6], for therapeutic vaccine studies and by Ratain et al[7]. Ratain et al[8]

used a “randomized discontinuation design” in which 202 patients with metastatic renal cell carcinoma were initially treated with Sorafenib and the 65 patients with stable disease at 12 weeks were randomized to either continue receiving the drug or a placebo. Although this resulted in a relatively small but informative randomized phase II trial, 202 total patients were required.

Because of the tension between the value of randomization in phase II evaluation and the desire to limit the number of patients and duration required for phase II studies, we consider the integrated phase II/III design. With this approach, accrual to a randomized phase II study is designed to continue on into a phase III study if a specified criteria is met. The endpoint used for the phase II evaluation will often differ from that used for the phase III analysis, but data from patients accrued during the phase II study is used in the phase III study. Randomized phase III trials with interim futility analyses are common in practice but generally the same endpoint is used for the interim and final analysis and hence are not phase II/III designs in the sense considered here.

Inoue et al[9], presented a Bayesian phase II/III design in which patients are randomized to an experimental arm or a standard arm and the decision to stop the study early or continue the study is made repeatedly based on simultaneous hypothesis tests of survival and response rates. They compare the efficiency of the design to two independent studies with the first study being a single arm study based on response rates and the second study being a randomized study with survival as the endpoint. In a simulation patterned after a non-small cell lung cancer study, they found the phase II/III design used fewer patients and took less time to complete.

Buaer et al[10] and Proschan and Hunsberger[11] have developed adaptive designs that are very flexible and allow the primary endpoint to be analyzed during the study and used to determine whether the study should continue. In these designs the sample size can also be

readjusted. The framework of the adaptive design allows one to maintain the type I error rate by adjusting the critical value at the end of the study.

In this paper we propose a randomized study design containing two portions. In the first portion of the study evidence of activity is gathered using what we term a "partial surrogate" endpoint such as progression free survival (PFS). This corresponds to the randomized phase II portion of the integrated study. If there is sufficient evidence of activity, accrual continues and enough patients are accrued until the phase III endpoint of survival can be assessed. The initial portion of the study is larger than a single arm phase II study but if the study continues the initial patients are also used to answer the phase III question. Consequently, the phase II/III study can require fewer patients than a sequence of 2 randomized studies (i.e. a randomized phase II study followed by a randomized phase III study).

There has been substantial interest in therapeutics on the potential use of surrogate endpoints for improving the efficiency of clinical drug development. This interest has been enhanced with the development of numerous potential biomarkers of disease status. Establishing that an intermediate endpoint is a valid surrogate of clinical outcome, however, is very difficult[4,12-16]. Because there have been many examples of biologically plausible intermediate endpoints having proven to be invalid surrogates, the criteria for establishing that an intermediate endpoint is a valid surrogate for clinical benefit is stringent. Much of the potential value of using a surrogate endpoint can, however, be achieved by using an unvalidated "partial surrogate." By a partial surrogate we mean an endpoint whose improvement is believed to be necessary but not sufficient for improving clinical outcome. Such an endpoint can be used for the phase II portion of the analysis. If the new treatment does not result in improvement of the partial surrogate, then the study terminates; otherwise the trial is continued to evaluate clinical benefit directly. In the comparisons described in this paper, PFS will be used as the phase II endpoint. The integrated

phase II/III design can be used, however, with many other types of partial surrogate endpoints based on molecular biomarkers or imaging diagnostics.

We discuss several different approaches to phase II/III studies and define metrics for evaluating the approaches with respect to study duration and required numbers of patients. We compare the phase II/III designs to a sequence of two independent randomized studies with the randomized phase II study using PFS as the endpoint followed by a separate randomized phase III study using survival as endpoint if results are promising. We also compare the phase II/III designs to performing a single randomized study with survival as the endpoint, possibly including an interim futility analysis based on survival.

The outline of the paper is as follows. In section 2 we discuss different phase II/III designs along with details of the simulations studies that we performed to evaluate the designs. Section 3 gives the results of the simulation study. Section 4 shows how the integrated design could be useful for drug development in pancreatic cancer. A discussion of the results is presented in section 5.

2. Methods

2.1 Study designs

We now present study designs that will be evaluated in this paper. When presenting designs we use the following notational convention, a subscript of 1 for parameters related to analyses before the final OS comparison and a subscript of o for parameters related to the final OS comparison. The accepted standard of evidence for establishing effectiveness of a treatment is a randomized clinical trial comparing the new treatment to a relevant control and demonstrating statistical significance for OS at a two-sided level of 0.05. Let N be the sample size for a phase III study design based on OS with a two-sided significance level not exceeding

0.05 and power 0.90. All design strategies we consider have the same maximum number of patients N.

The first new approach we consider is a phase II/III study with maximum sample size N. Accrual will be suspended before all patients are accrued with the suspension occurring after time t_1. The patients who have been accrued will be followed for a minimum follow-up time f_1. After t_1+f_1 a comparison of the treated versus control groups based on progression-free survival (PFS) will be performed. If the p-value is less than a specified threshold (α_1), accrual will resume until a total of N patients are accrued. After accruing N patients, follow-up will continue for an additional minimum time f_o. At the end of the study OS will be evaluated on all N patients. This will be referred to as the integrated phase II/III **two-stage** design.

The second type of phase II/III design we consider does not suspend accrual and wait for the follow up of the patients accrued in the first stage (f_1=0). Rather, PFS is simply evaluated after time t_1. This will be referred to as the integrated phase II/III **interim** analysis design. The initial accrual times t_1 for the interim-analysis phase II/III designs are somewhat larger than for the two-stage designs because the follow-up times are shorter for the former; in order to maintain the power for the PFS comparison more patients are needed.

The potential advantage of the phase II/III two-stage design over the phase II/III interim analysis design is that fewer patients may be required. This comes at the expense of a potentially longer study duration. For both designs the advantage over the sequence of independent phase II to phase IIII studies is that patients from phase II are used in the phase III comparison of survival.

These phase II/III designs are compared to three other designs. The first is a single randomized phase III study with OS as the endpoint without any phase II evaluation. This

approach might be used if there is no acceptable phase II endpoint or if the biological rational and pre-clinical development costs are sufficiently great that a phase III trial is warranted. The second approach is a single randomized phase III study with OS as the endpoint but with an interim analysis for futility based on OS. The third approach we consider involves a sequence of two independent studies; a randomized phase II study with PFS as the endpoint followed by a phase III study with OS as the endpoint where the second study is only performed if the first study has a positive result.

2.2 Evaluation Criteria

We compare the five study designs by looking at the efficiency of the designs with respect to length of time to obtain a conclusion on OS, number of patients needed to obtain the result, and power of the design. We now present calculations for the expected sample size, the expected duration of the designs and power.

First consider the single study with no futility analysis. The number of patients in the standard phase III design is always N and the total study time is $N/r_a + f_o$ where r_a is the accrual rate and f_o is the specified minimum follow-up time. The type I error when there is no treatment effect on OS is 0.05 and the power under a specified alternative is always 90% (since there is no futility analysis in the standard design). The type I error level holds regardless of whether there is or is not a treatment effect on PFS since PFS information is not used in a standard randomized phase III design.

For the single study approach with futility analysis based on OS, assume that a single futility analysis is performed at time t_1 when the number of patients accrued is n_1. The time to study completion is t_1 if the futility analysis terminates the study early and t_1+N/r_a+f_0 otherwise.

The expected number of patients and expected time to study completion for the phase III trial can be expressed as

$$E[N]=n_1 + (N-n_1)\ Pr\{continue\}$$

$$E[time]=t_1+(N/r_a- t_1+f_o)Pr\{continue\}$$

where Pr{continue} is the probability that the futility analysis does not terminate the study. The probability of continuing under a specified alternative for the treatment effect on OS can be calculated using equation A2 in appendix A. Note assumptions about treatment effects on PFS have no impact on the power or type I error of this design.

For both of the integrated phase II/III designs, let n_1 denote the number of patients accrued at the time of the PFS analysis. The final analysis of OS is performed when the number of patients is N if the study is not terminated early. The probability of continuing depends, however, on assumptions about the treatment effect on PFS, not on OS. If there is no treatment effect on PFS and the interim analysis is conducted at the level α_1(α_1 is always 1-sided), then the probability of continuing is $(1-\alpha_1)$. The hypothesis of no treatment effect on OS is rejected if the study continues past the interim analysis of PFS and then rejects the null hypothesis on OS at the end of the study. This determines the type I error and the power of the design.

The phase II/III interim analysis design terminates after t_1 months if the PFS difference is not significant at the specified level α_1. If the trial is not terminated based on the PFS comparison, then the total duration is $N/r_a + f_o$ months. The two-stage phase II/III design terminates after $t_1 + f_1$ months if the analysis of PFS is not significant; otherwise it takes $N/r_a + f_1 + f_o$ months. The expected sample size and expected time to completion of study for the integrated phase II/III designs are

$$E[N]=n_1 + (N-n_1)\ Pr\{continue\}$$

$$E[time]=t_1 +f_1+(N/r_a -t_1+f_o)Pr\{continue\}$$

note f_1=0 for the interim analysis phase II/III interim analysis design. The probability of continuing is α_1 under the null hypothesis of no treatment effect on PFS and the probability of continuing is $(1-\beta_1)$, the power for the PFS comparison, under the specified alternative.

For the sequence of independent studies strategy, the number of patients is n_1 if the phase II trial is negative and n_1+ N otherwise. The total expected time until the study completion is t_1 + f_1 if the phase II trial is negative and $t_1+f_1+Nr_a+ f_o$ otherwise. This ignores the time required to organize the phase III trial and obtain approvals of all the participating institutions following completion of the phase II trial. That time may, of course, be quite substantial. The expected sample size and expected time to completion of study for the separate design strategy are

$$E[N]=n_1 + N \Pr\{\text{continue}\}$$

$$E[\text{time}]=t_1 +f_1+(N/r_a+f_o)\Pr\{\text{continue}\}.$$

The probability that the phase II trial is positive depends on the treatment effect on PFS. When there is no treatment effect on PFS, the probability of continuing to the phase III study or the probability that the phase II trial is positive is α_1 where α_1 is set by design. When the treatment effect on PFS is as specified in sizing the phase II trial, the probability that the trial is positive is $(1-\beta_1)$, this is also the probability of continuing to the phase III study under the alternative hypothesis for PFS.

The power of the designs is Pr{continue and reject H_0 at final analysis}. For the single study design the study will always continue to the end so the power is just

$$\Pr\{\text{reject } H_0 \text{ at final analysis}\}=(1-\beta_o).$$

The power for the sequence of two studies is exactly $(1-\beta_1)$ $(1-\beta_o)$ since the studies are independent. For the designs that have interim analyze an approximation for the power is

$$\Pr\{\text{continuing}\}\Pr\{\text{reject } H_0 \text{ at final analysis}\}=(1- \beta_1)\ (1-\beta_o).$$

For the integrated designs this approximation ignores any correlation between PFS and OS. For the single study design with a futility analysis this ignores the correlation in the reanalysis of part of the data. Therefore, the approximation is a lower bound on the power since in both cases the correlation would generally be positive.

We have provided a web based computer program that calculates the approximate expected sample size, expected study duration and power when accrual rates, PFS and OS assumptions are provided (http://linus.nci.nih.gov/brb). The approximations assume no correlation between PFS and OS.

2.3 Description of Simulation

Although approximations for the criteria to evaluate the designs are provided it is also important to evaluate the designs under the more realistic assumptions of correlation. Therefore, we perform simulations to evaluate these criteria under a specified form of correlation between PFS and OS. The correlated PFS and OS values were generated as follows.

The distribution of OS was taken as exponential with median 12 months. The treatment effect for OS is specified by a parameter Δ_o. The treatment effect is created by changing the exponential parameter in the treatment group. The change results in a median survival for the treatment group of 12 Δ_o. For a patient with overall survival value Y_o, the PFS value Y_p $=\min(Y_1,Y_o)$ where Y_1 was generated according to an exponential distribution with median 6 months. We let the effect of treatment on Y_1 be Δ_1. Note that since $Y_p = \min(Y_1,Y_o)$ the treatment effect for PFS is not exactly changed by a factor of Δ_1 and Y_p does not have an exponential distribution. If the medians of Y_1 and Y_o are very different than the correlation is very small and Y_p will have an approximate exponential distribution. In the simulations Δ_1 and Δ_o were varied. All simulations are performed with 10,000 replications.

For the integrated phase II/III designs we consider various threshold p-values for the PFS analysis. We look at designs that stop the study if the p-value is greater than a specified threshold α_1 of .5, .2, .1 or .05. For both combined phase II/III designs, the parameter t_1 is determined so that the interim analysis has a specified power for detecting a treatment effect on PFS of the size postulated, using the designed significance level α_1. We examine the designs with 90% and 95% power at the PFS analyzes. For the two-stage design we let f_1=6 months.

For the design with a futility analysis based on OS we consider two different futility rules; one that continues accrual if the p-value is less than .5 and the other that continues accrual if the p-value is less than .2. The first rule allows accrual to be shorter at the futility analysis, at the expense of continuing the study more often. Two different futility analysis times are considered; one after 1/2 the total number of patients have been accrued and one after 2/3 of the patients have been accrued.

For the sequence of studies strategy we use f_1 = 6 months in our simulations. We set t_1 so that the phase II trial would have either power $(1-\beta_1)$=0.9 or .95 for the postulated treatment effect on PFS with 1-sided α_1=.1.

We considered three situations: (i) No treatment effect on either PFS or OS (global null); (ii) Treatment effect on PFS but not on OS (partial null); (iii) treatment effect on PFS and OS (global alternative). The separate trial strategy and the phase II/III strategies are obviously not suitable for situations when it is anticipated that there may be a treatment effect on OS in the absence of a treatment effect on PFS.

3. Simulation Results

Figure 1 shows a comparison of the five types of designs with regard to expected number of patients and time to completion when the objective is to have 90% statistical power for detecting a hazard ratio of 1.5 for survival and the accrual rate is 10 patients per month. A hazard

ratio of 1.5 corresponds to a 33% reduction in the hazard of death. More detailed results are shown in Table 1 and other simulation results that vary the accrual rate and the size of the treatment effect on PFS and OS are shown in the Appendix B. The separate randomized phase II design and the PFS analysis of the integrated phase II/III designs shown in Figure 1 have 95% power for detecting a hazard ratio on PFS of 2.0, corresponding to a 50% reduction in the hazard of progression or death. Our simulations indicated that designing those analyses for only a 90% power caused a substantial reduction in the power of the survival analysis (see Table 1). The designs shown in Figure 1 have at least 85% power for the survival analysis under the global alternative hypothesis in which the treatment effect on survival has a hazard ratio of 1.5 and the treatment effect on PFS has a hazard ratio of 2.0. Parameters for the futility analysis of the single study design were also selected in order to ensure that the power of the survival comparison did not fall below 85% for the global alternative hypothesis.

For the integrated phase II/III designs, we varied the statistical significance threshold (α_1) for the analysis of PFS over the range 0.05 to 0.50 which resulted in different accrual times (t_1) for the PFS analysis. The accrual time (t_1) was determined based on the significance threshold (α_1) and the desired power of the PFS analysis (90% or 95%). As noted above, the statistical power for detecting an effect on overall survival was improved by having greater power for the interim analysis of PFS although this required that the interim analysis be performed at a later time. The greater power for the survival analysis came therefore at the cost of a larger sample size under the global null hypothesis (no treatment effect on PFS or OS). A more stringent statistical significance threshold for the PFS analysis required a later interim analysis of PFS in order to satisfy the PFS power requirement. For Figure 1, we selected significance threshold α_1=0.2 as this minimized the expected number of patients under the global null for both the 2-stage design and the interim analysis design.

As can be seen in Figure 1, under the global null hypothesis, the integrated phase II/III designs are effective in substantially reducing the development time and number of required patients compared to the strategies of going directly to a phase III trial with OS, even if the latter utilizes futility monitoring on OS. Futility monitoring on PFS is more effective than futility monitoring on OS in this setting because progression events can be observed sooner. The two-stage integrated trial requires fewer patients than the interim-analysis integrated trial, but stopping the study to wait for the PFS data to mature means the study takes longer to complete.

The chance of a false positive conclusion for the separate trials strategy is $\alpha_1\alpha_o$ the product of the significance thresholds for the phase II and phase III trials. This is considerably less than the chance of a false positive conclusion for the single study design. The integrated II/III designs also have significance levels for the OS comparison below .05. In both designs this is due to the screening of studies in the PFS analysis. It would not be acceptable, however, to increase the significance threshold for the phase III trial in the separate trial strategy or the integrated II/III designs to reflect the phase II screening because the reduction in type one error applies to the global null hypothesis, not for hypotheses in which there is a treatment effect on PFS but not on OS. In such cases, the false positive conclusion for OS is $(1-\beta_1)\alpha^*_o$, where the first factor is the power of the phase II trial for detecting a treatment effect on PFS, α^*_o is the significance level for the test of OS where $\alpha^*_o > .05$. If the treatment effect on PFS is large ($\beta_1=0$) but there is no treatment effect on OS, the probability of a false positive conclusion for OS approaches α^*_o. Our assumption in this paper is that we wish to make inferences about the effect of the new treatment on OS. We do not want those inferences to be joint inferences about OS and PFS and we don't want the validity of those inferences to depend on assumptions about PFS

being a true surrogate for OS. Hence, in order to control the type I error for the evaluation of OS, the value of α_o should not exceed the traditional 2-sided 0.05 significance level.

Figure 1 shows results under the global alternative (for the same designs used under the null hypothesis). In this situation it is obviously best to conduct a phase III trial of OS. Including an interim futility analysis of OS serves only to reduce the power from 90% to 87% (see Table 1). The separate trials strategy requires many more patients and a longer time to complete. The duration shown in the table does not include the time needed to organize the phase III trial after the phase II trial is completed. The separate trials strategy shown has statistical power 86%. Although the phase III trial with the separate trials strategy is the same as the phase III trial for the stand-alone strategy, the overall power of the separate trials strategy is reduced since the phase II trial has only 95% power for detecting the treatment effect on PFS. Hence, in 5% of the cases, the phase III trial is not initiated.

The results for the integrated designs under the global alternative are much more efficient than for the separate trials strategy. The integrated designs do not result in an increased number of patients or study duration like the separate trials strategy. The integrated interim analysis design is superior to the two-stage integrated design with regard to study duration when there is a treatment effect on PFS and OS. The power of the integrated designs are 86%.

We also conducted simulations under the partial null hypothesis where there is a treatment effect on PFS but not on overall survival. Table 1 gives these results. Since all of the designs use the same significance threshold for inference about OS in the final analysis, they all provide adequate protection against false positive conclusions about the treatment effect on OS. The combined phase II/III designs are not more efficient than the single study design under the partial null hypothesis because monitoring PFS does not in this case provide useful information about the null effect on OS. The two-stage phase II/III design slows down the process of finding

that there is no treatment effect on OS, whereas the interim-analysis phase II/III design has a more neutral effect. Performing a futility interim analysis on OS is more useful than an interim analysis using PFS, however. The separate trials strategy is inferior here to both the designs that ignore PFS and to the combined phase II/III designs.

In the tables shown in Appendix B we vary the patient accrual rates and the hazard ratios for the alternative hypotheses. The results are qualitatively very similar to those shown in Table 1 and Figure 1. With a more rapid accrual rate, the two types of integrated phase II/III designs differ more with regard to expected number of patients under the global null hypothesis. When the hazard ratios are smaller, larger sample sizes are needed for both PFS and OS analyses.

4. Example

Advanced pancreatic cancer is a disease where the integrated phase II/III design would be useful. From 2004-2006 three negative randomized phase III clinical trials were reported[17-19]. In the clinical trials the addition of either Oxaliplatin, Cisplatin, or Irinotecan to Gemicitabine was studied. All three studies followed single arm phase II studies with promising evidence of activity for the combinations[20-22]. From these three negative studies it is clear that single arm phase II studies of combination regimens in this population of patients are unreliable. It appears that the response endpoint can be influenced merely by the selection of the patients. Thus, there is a strong need for randomized phase II studies rather than single arm phase II studies for such applications.

We examine the saving of time and patients for an integrated II/III interim analysis design with α_1=.2 and power of 95% for PFS. The literature suggests that median OS for this population is 6 months and median PFS is 3 months. We assume a monthly accrual rate of 15 patients. Since this study is for an advanced disease population it is likely that even small OS improvements would be interesting since the drug could then be studied in earlier stages of

disease. Therefore we target an improvement in median PFS of 1.5 months and an median improvement of OS of 1.8 months. This would correspond to hazard ratios of 1.5 and 1.3.

We assume a correlation structure as in the simulations and use simulations to determine the expected sample size, expected study duration and power of the five designs. A randomized phase III study with OS as the endpoint, 90% power and a 2-sided .025 level of significance requires 692 patients or 46.14 months of accrual with 6 months minimum follow up or 52.2 months of total study time. An integrated II/III design with an interim analysis based on a total study size of 692 patients has 87% power, expected sample sizes of 357 and 676 under the global null and alternative hypothesis respectively. The expected duration of study is 25.1 and 50.8 months respectively. If a separate randomized phase II study were instead performed, the expected number of patients under the null and alternative hypothesis would be 369 and 958. The expected duration until conclusion of the benefit of the drug on OS would be 31.2 and 75.54 months under the null and alternative hypothesis. Therefore, the integrated II/III design has an expected sample size similar to a randomized phase II study under the null hypothesis and the expected sample size under the alternative is no larger than performing a single randomized phase III study. Clearly an integrated II/III design is a viable option for clinical studies of advanced pancreatic cancer.

5. Discussion

Initially we weren't sure whether the integrated phase II/III strategy would be effective from a statistical power standpoint. We thought that either the size of the first stage would need to be so large that there would be little savings under the global null or that the total study size would need to be substantially increased in order to make up for the loss in power by stopping studies early. Our investigation showed that both integrated designs performed better than either the single study with OS as the endpoint or a sequence of 2 independent studies.

Of the integrated designs considered, the phase II/III interim analysis design with α_1=.2 and 95% power for the PFS analysis consistently performed the best in terms of study time with only a small increase in sample size. This design allowed the expected time to complete the study to be small because accrual never stopped. In the integrated designs there was also a decrease in the expected total sample size because data from the beginning of the study was used to answer the OS question rather than being used only to answer the PFS question.

The integrated interim analysis designs are more practical than the integrated two-stage designs because they do not require a suspension of accrual. The efficiency of the integrated designs is however dependent on the size of the median time to disease progression. If the median PFS is short, then fewer patients will be required for the PFS analysis because the power of that analysis is generally determined by the number of PFS events, not the number of patients. The efficiency of the integrated phase II/III design also increases as the accrual rate decreases because under the global null the trial can be terminated based on the PFS analysis before too many patients are accrued.

The integrated phase II/III designs may be desirable to pharmaceutical sponsors and regulatory agencies when accelerated approval is of interest. This design would ensure that a randomized phase III trial based on OS was in place at the time that accelerated approval was obtained based on the PFS results. The design would provide pharmaceutical sponsors a well powered, well designed randomized phase II study to obtain accelerated approval based on PFS. The maximum sample size would be smaller than that of a sequence of studies. If accelerated approval were of interest α_1 would generally be set at .05 rather than our recommended .2 level.

The combined phase II/III designs and the separate trial designs are reasonable only if it is expected that improvement of PFS is a necessary, although not sufficient condition, for

improvement in OS. This is weaker than the assumption, however, that PFS is a valid surrogate for OS and is reasonable to expect in many settings. This approach can also be used with partial surrogates other than PFS such as molecular biomarkers or new imaging diagnostics. In this way, new technology can be used to improve the efficiency of clinical development without the risks involved in adopting such measures as full surrogates of clinical outcome.

We have assumed that OS is the primary endpoint for evaluating the new drug. Permitting patients on the control regimen to cross-over to receive the new drug after disease progression makes evaluating survival somewhat problematic. Consequently, it is best to avoid cross-over in design of the trial. In some cases, although avoiding cross-over is not feasible, treatment after progression is of very limited effectiveness and OS will still be a meaningful endpoint.

It is easiest to observe a treatment effect on survival in settings where there are not good salvage regimens, otherwise the effect of the drug under study will be diluted by the other treatments. If the drug has no effect on OS, however, it should demonstrate some other direct clinical benefit in order to be recommended for use. Progression-free survival alone, is often not of direct clinical benefit, particularly for drugs with substantial adverse effects.

We have provided a web based computer program that approximates the expected sample size, expected study duration, and power for all 5 designs studied in this paper (http://linus.nci.nih.gov/brb). Although this program assumes no correlation between PFS and OS the approximation of the savings in sample size or time that could be obtained by using the integrated II/III approach would be adequate to decide whether the design should be used. When designing an integrated phase II/III study we recommend evaluating various sets of parameters. For example the accrual rate should be varied along with the relationship between medians of survival PFS, OS, and the size of the treatment effect on PFS and OS.

Once the parameters of the integrated phase II/III designs have been chosen, implementation is straightforward. The protocol would specify the number of progression events that would be needed for the PFS analysis and the α_1 for stopping the study. The total number of events for the OS analysis would also be specified. After the PFS analysis has been performed typical interim DSMC monitoring based on OS (for efficacy) would be specified in the protocol. The protocol should indicate clearly that early stopping of accrual because of a treatment effect on PFS is not a part of the analysis plan.

With the number and type of new drugs that are being developed today it may be necessary to use new types of designs in the phase II and III setting. We suggest investigators explore the efficiency of integrated phase II/III designs.

Legend

Figure 1. Expected number of patients and expected study duration for the 5 study designs under the null and alternative hypotheses. The patterned bars show the results under the global null and the solid bars show the results under the global alternative.

References

1. Korn EL, Simon R: Using the tolerable-dose diagram in the design of phase I combination chemotherapy trials. J Clin Oncol 11:794-801, 1993

2. Green S, Benedetti J, Crowley J: Clinical trials in oncology, Chapman & Hall, 1997

3. Dixon DO, Simon R: Sample size considerations for studies comparing survival curves using historical controls. J Clin Epidemiol 41:1209-13, 1988

4. Makuch RW, Simon RM: Sample size considerations for nonrandomized comparative studies. Journal of Chronic Disease 33:175-181, 1980

5. Rubinstein LV, Korn EL, Freidlin B, et al: Design issues of randomized phase II trials and a proposal for phase II screening trials. J Clin Oncol 23:7199-206, 2005

6. Simon RM, Steinberg SM, Hamilton M, et al: Clinical trial designs for the early clinical development of therapeutic cancer vaccines. J Clin Oncol 19:1848-54, 2001

7. Ratain MJ, Stadler WM: Clinical trial designs for cytostatic agents. J Clin Oncol 19:3154-5, 2001

8. Ratain MJ, Eisen T, Stadler WM, et al: Phase II placebo-controlled randomized discontinuation trial of sorafenib in patients with metastatic renal cell carcinoma. J Clin Oncol 24:2505-12, 2006

9. Inoue LY, Thall PF, Berry DA: Seamlessly expanding a randomized phase II trial to phase III. Biometrics 58:823-31, 2002

10. Bauer P, Kohne K: Evaluation of experiments with adaptive interim analyses. Biometrics 50:1029-41, 1994

11. Proschan MA, Hunsberger SA: Designed extension of studies based on conditional power. Biometrics 51:1315-24, 1995

12. Prentice RL: Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med 8:431-40, 1989

13. Torri V, Simon R, Russek-Cohen E, et al: Statistical model to determine the relationship of response and survival in patients with advanced ovarian cancer treated with chemotherapy. J Natl Cancer Inst 84:407-14, 1992

14. Fleming TR: Surrogate markers in AIDS and cancer trials. Stat Med 13:1423-35; discussion 1437-40, 1994

15. Ellenberg S, Hamilton JM: Surrogate endpoints in clinical trials: cancer. Stat Med 8:405-13, 1989

16. Buyse M, Molenberghs G, Burzykowski T, et al: The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics 1:49-67, 2000

17. Louvet C, Labianca R, Hammel P, et al: Gemcitabine in combination with oxaliplatin compared with gemcitabine alone in locally advanced or metastatic pancreatic cancer: results of a GERCOR and GISCAD phase III trial. J Clin Oncol 23:3509-16, 2005

18. Heinemann V, Quietzsch D, Gieseler F, et al: Randomized phase III trial of gemcitabine plus cisplatin compared with gemcitabine alone in advanced pancreatic cancer. J Clin Oncol 24:3946-52, 2006

19. Rocha Lima CM, Green MR, Rotche R, et al: Irinotecan plus gemcitabine results in no survival advantage compared with gemcitabine monotherapy in patients with locally advanced or metastatic pancreatic cancer despite increased tumor response rate. J Clin Oncol 22:3776-83, 2004

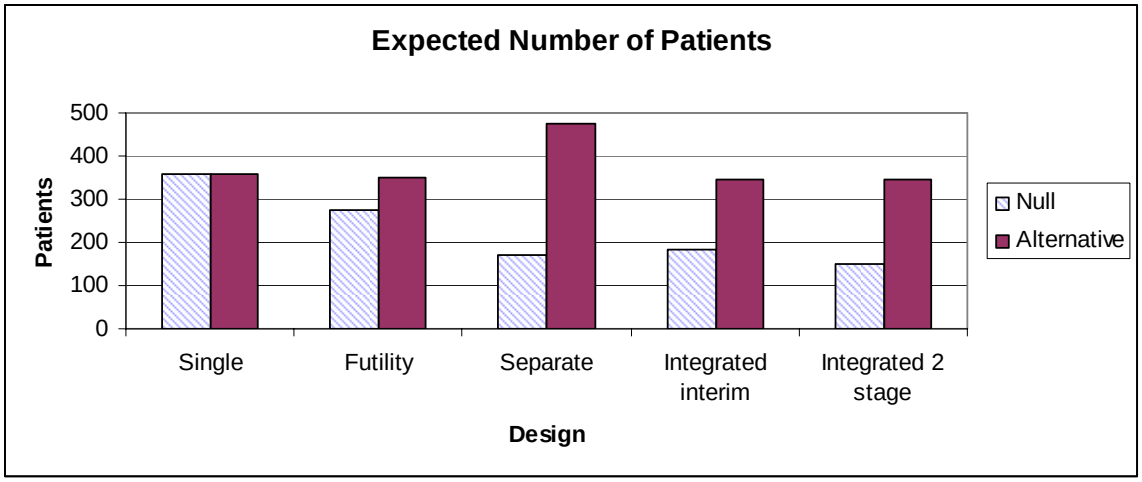
20. Louvet C, Andre T, Lledo G, et al: Gemcitabine combined with oxaliplatin in advanced pancreatic adenocarcinoma: final results of a GERCOR multicenter phase II study. J Clin Oncol 20:1512-8, 2002

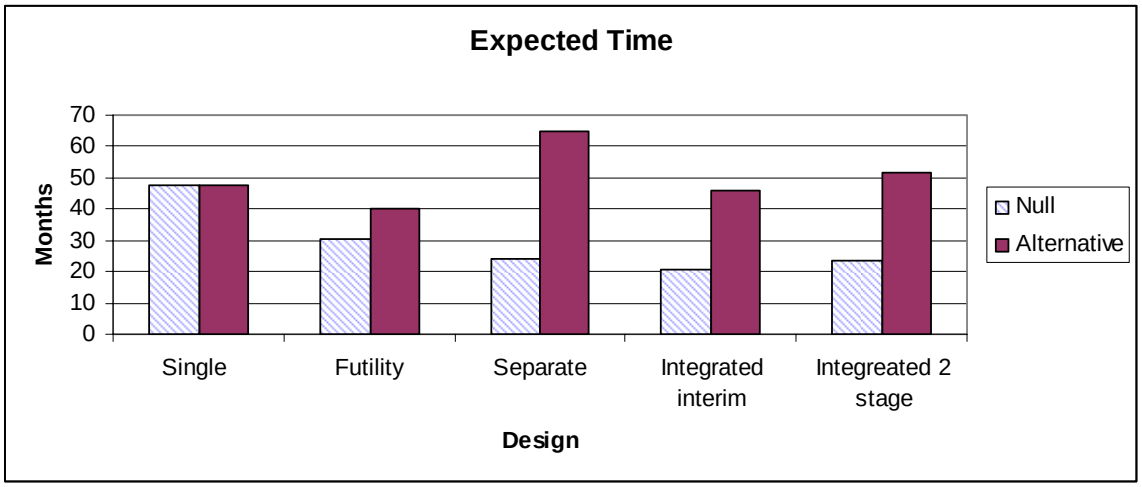
21. Philip PA, Zalupski MM, Vaitkevicius VK, et al: Phase II study of gemcitabine and cisplatin in the treatment of patients with advanced pancreatic carcinoma. Cancer 92:569-77, 2001

22. Wagener DJ, Verdonk HE, Dirix LY, et al: Phase II trial of CPT-11 in patients with advanced pancreatic cancer, an EORTC early clinical trials group study. Ann Oncol 6:129-32, 1995

Table 1: Accrual of 10 patients/month; Data generated according to two exponentials Y_1 with median 6 months and a treatment effect hazard ratio of 2 and Y_2 with median 12 months and a treat hazard ratio of 1.5. Progression was the min(Y_1,Y_2) and survival was Y_2. E[N] is the expected sample size and E[T] is the expected study time. All time is in months.

			Global Null			Partial null			Global Alternative		
	α_1	t_1	Power of Survival Analysis	E[N]	E[T]	Power of Survival Analysis	E[N]	E[T]	Power of Survival Analysis	E[N]	E[t]
Single study	--	35.7	.025	357	47.7	.025	357	47.7	.9	357	47.7
Single study with futility based on overall survival	.2	14.4	.015	186	20.1	.015	186	20.1	.63	286	33.3
		19.1	.018	224	23.5	.018	224	23.5	.75	323	36.7
	.5	14.4	.026	251	28.6	.026	251	28.6	.83	335	39.3
		19.1	.025	275	30.3	.025	275	30.3	.87	348	40.1
Separate Phase II and Phase III											
90% power for PFS (f_1=6)	.1	10.2	.0025	138	21.0	.023	423	59.1	.81	423	59.1
95% power for PFS (f_1=6)	.1	13.4	.0025	170	24.2	.024	473	64.7	.86	473	64.7
Integrated interim with 90% power for PFS analysis (f_1=0)	.05	17.0	.0053	180	18.7	.034	295	37.5	.82	338	44.5
	.1	14.2	.0066	164	17.6	.037	293	37.7	.81	334	44.1
	.2	11.2	.012	163	18.7	.037	294	38.4	.82	332	43.9
	.5	5.9	.027	209	26.9	.043	305	40.4	.81	326	43.3
Integrated two-stage with 90% power for PFS (f_1=6)	.05	12.5	.0022	137	20.3	.030	274	41.1	.81	330	49.6
	.1	10.2	.0057	128	20.0	.032	279	42.3	.82	331	49.9
	.2	7.6	.012	131	21.4	.038	284	43.3	.82	330	49.9
	.5	3.3	.026	195	31.5	.041	298	45.6	.82	328	49.7
Integrated interim with 95% power for PFS analysis (f_1=0)	.05	20.1	.0038	209	21.5	.037	317	40.7	.86	349	46.4
	.1	17.1	.0062	190	20.2	.040	317	41.2	.87	349	46.3
	.2	13.8	.011	183	20.8	.037	319	41.8	.86	346	46.0
	.5	8.1	.025	219	27.9	.044	322	42.7	.85	342	45.6
Integrated two-stage with 95% power for PFS (f_1=6)	.05	15.9	.0038	169	23.5	.038	308	45.8	.87	348	52.2
	.1	13.4	.0068	156	22.9	.041	311	46.7	.87	347	52.2
	.2	9.8	.011	149	23.3	.040	307	46.4	.86	344	51.8
	.5	5.2	.025	205	32.5	.042	323	49.0	.87	344	51.8


Expected Number of Patients
Patients
500
400
300
200
100
0
Single
Futility
Separate
Integrated interim
Integrated 2 stage
Design
Null
Alternative


Expected Time
Months
70
60
50
40
30
20
10
0
Single
Futility
Separate
Integrated interim
Integreated 2 stage
Design
Null
Alternative

Appendix A.

Approximations for computing the performance measures of the integrated phase II/III designs are given below. These were not used in the tabulated results provided in the manuscript.

At the end of the clinical trial the null hypothesis of no treatment effect on overall survival will be tested. If the trial is terminated at the interim analysis, then the null hypothesis is not rejected. The null hypothesis of no effect on PFS is not in itself of primary interest; PFS is used as the basis of the interim analysis but the null hypothesis of interest is that pertaining to OS. The trial is completed if the interim treatment effect on PFS is significant at the one-sided α_1 level. Let $1\text{-}\beta_1(\Delta_1)$ denote the power of the interim analysis for detecting a treatment effect on PFS and α_1.

The probability of rejecting the null hypothesis on OS can be written

$$\begin{aligned} P\{reject \quad H\} &= \int_{continuation} P\{reject \quad H_o \mid D_{\text{int}erim}\} dF\{D_{\text{int}erim}\} \\ &= P\{reject \quad H_o\} P\{continuation\} \qquad \text{(A1)} \\ &= \{1-\beta_o(\Delta_o)\}\{1-\beta_1(\Delta_1)\} \end{aligned}$$

Where the integration is over the continuation region at the interim analysis of PFS and $1\text{-}\beta_o(\Delta_o)$ denotes the power for the survival analysis at the end of the trial when the treatment effect on survival has a hazard ratio of Δ_o. The above approximation ignores the correlation between PFS and OS. Since the correlation will generally be positive, the approximation may somewhat under-estimate the probability of rejection. It should be noted, however, that the simulations were performed based on a model with low correlation between PFS and OS and formula (A1) provides a good approximation to the corresponding quantities in Tables 1-3. For example, under the global null $\beta_o(1)$=.05 and

(A1) equals .025*α_1, in good agreement with the simulated values in the tables for the integrated designs. The power under the global alternative was based on 1-β_o(1.3)=.90 and 1-β_1(1.5) = .90 or .95 for the integrated designs. Hence (A1) gives .81 or .855, in good agreement with the simulated values. For the partial null conditions, the quantities are 1-β_o(1)=.05 and 1-β_1(1.5)=.9 or .95. Hence (A1) gives rejection probability approximations of .045 and .0475 for the integrated designs, depending on whether the interim PFS analysis has 90% or 95% power. This is also in good agreement with the simulated values.

The expected sample size can be written

$$E(N) = N_o P\{contiuation\} + n_1(1 - P\{continuation\})$$

Where N_o denotes the maximum sample size of the phase III trial and n_1 denotes the number of patients accrued at the interim analysis of PFS. For the cases simulated in the tables, N_o=870, 717 and 357. n_1 equals the accrual rate times t_1, the time of the interim analysis shown in the tables.

The distribution of the progression times Y_p are not exponential, but assuming Y_p has an exponential distribution with hazard ratio Δ_1 gives a fairly good approximation of the length of accrual needed for the interim analysis of PFS.(Note, closeness of the approximation of Y_p as an exponential with hazard ratio Δ_1 will depend on how close the medians Y_o and Y_1 are. The farther apart the medians Y_o and Y_1 are, the closer the approximation.) If t_1 is specified then the power at the interim look can be approximated from

$$1/E(D_c)+1/E(D_t)=\left(\frac{\ln(\Delta_1)}{z_{\alpha_1}+z_{\beta_1}}\right)^2 \qquad \text{(A2)}$$

where $E(D_c)$ and $E(D_t)$ denote the expected number of progression events in the control and treatment groups respectively at the interim analysis, z_1 denotes the α_1^{th} percentile of the standard normal distribution and $z_{\beta 1}$ denotes the $\beta_1(\Delta_1)^{th}$ percentile[1]. For the assumed exponential PFS distribution the expected number of events can be computed from

$$E(D_i)=(r_a t_1/2)\left(1-\frac{\exp(-\lambda_i f_1)}{\lambda_i t_1}(1-\exp(-\lambda_i t_1))\right) \qquad \text{(A3)}$$

Where r_a is the accrual rate, λ_i the exponential failure rate for the group (i=control or treatment), and f_1 is the follow-up time for the PFS analysis. For the integrated interim analysis design f_1=0.

References

1. Rubinstein, LV, Gail, MH, and Santer TJ: Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. Journal of Chronic Disease. 34: 469–479, 1981.

Appendix B.

Table B1: Accrual of 29 patients/month; Data generated according to two exponentials Y_1 with median 6 months and a treatment effect hazard ratio of 1.5 and Y_2 with median 12 months and a treat hazard ratio of 1.3. Progression was the min(Y_1,Y_2) and survival was Y_2. E[N] is the expected sample size and E[T] is the expected study time. All time is in months.

			Global Null			Partial null			Global Alternative		
	α_1	t_1	Power of Survival Analysis	E[N]	E[T]	Power of Survival Analysis	E[N]	E[T]	Power of Survival Analysis	E[N]	E[t]
Single study	--	30	.025	870	42	.025	870	42	.90	870	42
Single study with futility based on overall survival	.2	15.0	.017	520	19.1	.017	520	19.1	.70	759	30.6
		20.0	.021	638	22.8	.021	638	22.8	.81	829	32.0
	.5	15.0	.012	652	25.5	.012	652	25.5	.86	841	34.6
		20.0	.028	727	27.1	.028	727	27.1	.89	862	33.6
Separate Phase II and Phase III											
90% power for PFS (f_1=6)	.1	9.0	.0025	348	19.2	.023	1044	52.8	.81	1044	52.8
95% power for PFS (f_1=6)	.1	11.2	.0025	412	21.4	.024	1151	57.1	.86	1151	57.1
Integrated interim with 90% power for PFS analysis (f_1=0)	.05	15.4	.0041	468	16.7	.032	728	33.1	.83	826	39.2
	.1	13.0	.0055	427	15.9	.035	727	33.6	.82	819	39.0
	.2	10.5	.0092	416	16.7	.034	736	34.5	.83	817	39.0
	.5	5.9	.0247	519	23.9	.041	754	36.0	.83	808	38.8
Integrated two-stage with 90% power for PFS (f_1=6)	.05	11.2	.0046	355	18.9	.033	678	37.2	.81	806	44.4
	.1	9.0	.0072	323	18.4	.038	682	37.8	.82	801	44.2
	.2	6.6	.010	323	19.5	.040	757	42.1	.81	799	44.3
	.5	2.8	.024	476	28.4	.042	730	41.0	.82	790	44.0
Integrated interim with 95% power for PFS analysis (f_1=0)	.05	17.8	.0028	534	19.0	.034	781	35.9	.87	851	40.7
	.1	15.4	.0061	490	18.1	.036	778	36.2	.86	847	40.5
	.2	12.6	.011	467	18.5	.042	782	36.9	.87	845	40.6
	.5	7.8	.024	554	25.2	.045	800	38.3	.87	842	40.5
Integrated two-stage with 95% power for PFS (f_1=6)	.05	13.5	.0041	416	21.0	.037	742	40.4	.86	841	46.3
	.1	11.2	.0070	380	20.3	.035	741	40.7	.86	836	46.1
	.2	8.6	.011	369	21.0	.040	758	42.0	.87	837	46.2
	.5	4.3	.028	494	29.0	.047	781	43.5	.87	833	46.1

Table B2: Accrual of 10 patients/month; Data generated according to two exponentials Y_1 with median 6 months and a treatment effect hazard ratio of 1.5 and Y_2 with median 12 months and a treat hazard ratio of 1.3. Progression was the min(Y_1,Y_2) and survival was Y_2. E[N] is the expected sample size and E[T] is the expected study time. All time is in months.

			Global Null			Partial null			Global Alternative		
	α_1	t_1	Power of Survival Analysis	E[N]	E[T]	Power of Survival Analysis	E[N]	E[T]	Power of Survival Analysis	E[N]	E[t]
Single study	--	71.7	.025	717	83.7	.025	717	83.7	.90	717	83.7
Single study with futility based on overall survival	.2	35.9	.019	430	44.2	.79	658	70.8	.79	658	70.8
		47.8	.025	526	53.5	.86	698	73.4	.86	698	73.4
	.5	35.9	.030	541	57.1	.88	703	76.1	.88	703	76.1
		47.8	.028	599	61.9	.90	714	75.3	.90	714	75.3
Separate Phase II and Phase III											
90% power for PFS (f_1=6)	.1	24.5	.0025	317	38.9	.023	890	105.8	.81	890	105.8
95% power for PFS (f_1=6)	.1	29.5	.0025	367	43.9	.024	976	115.0	.86	976	115.0
Integrated interim with 90% power for PFS analysis (f_1=0)	.05	34.0	.0035	359	36.5	.031	592	67.2	.82	679	78.6
	.1	28.0	.0060	322	33.4	.037	591	67.7	.82	674	78.3
	.2	21.0	.011	313	33.8	.035	592	68.2	.82	668	77.7
	.5	11.0	.025	411	47.0	.040	613	71.2	.82	661	77.0
Integrated two-stage with 90% power for PFS (f_1=6)	.05	30.8	.0039	328	39.4	.034	585	72.7	.82	678	84.7
	.1	24.5	.0048	292	36.4	.033	582	72.7	.82	674	84.3
	.2	17.5	.011	287	37.1	.036	587	73.8	.82	667	83.6
	.5	7.4	.024	396	51.6	.040	604	76.3	.82	658	82.8
Integrated interim with 95% power for PFS analysis (f_1=0)	.05	41.0	.0037	427	43.3	.036	641	73.2	.85	701	81.5
	.1	34.5	.0057	384	39.6	.038	639	73.4	.87	700	81.4
	.2	26.0	.011	353	37.7	.039	634	73.2	.86	694	80.8
	.5	15.0	.025	432	49.2	.045	653	75.9	.86	691	80.5
Integrated two-stage with 95% power for PFS (f_1=6)	.05	37.9	.0029	396	46.2	.033	634	78.5	.86	700	87.4
	.1	29.5	.0053	338	41.0	.036	626	78.0	.86	695	86.9
	.2	23.0	.011	327	41.1	.042	631	79.0	.86	693	86.7
	.5	11.5	.026	417	53.7	.046	649	81.5	.86	689	86.3