Title: Comparing Oncology Clinical Programs by Use of Innovative Designs and Expected Net Present Value Optimization: Which Adaptive Approach Leads to the Best Result?

Authors:

Tom Parke, BSc, Berry Consultants, Abingdon, Oxfordshire, United Kingdom

Olga Marchenko, PhD, Advisory Services Analytics, Quintiles, Durham, NC, United States

Vladimir Anisimov, PhD, School of Mathematics and Statistics, University of Glasgow, UK

Anastasia Ivanova, PhD, Department of Biostatistics, University of North Carolina at Chapel Hill, NC, USA

Christopher Jennison, PhD, Department of Mathematical Sciences, University of Bath, Bath, United Kingdom

Inna Perevozskaya, PhD, Statistical Research and Consulting Center, Pfizer, Inc., Collegeville, PA, United States

Guochen Song, DRPH, Advisory Services Analytics, Quintiles, Durham, NC, United States

Correspondence Author:

Tom Parke

e-mail: tom@berryconsultants.net, Mobile: +44 7801 520250, Office: +44 330 113 6183

Acknowledgements: Dr. Ivanova's work was supported in part by the NIH grant P01 CA142538.

ABSTRACT

Designing an oncology clinical program is more challenging than designing a single study. The standard approaches have been proven to be not very successful during the last decade; the failure rate of Phase 2 and Phase 3 trials in oncology remains high. Improving a development strategy by applying innovative statistical methods is one of the major objectives of a drug development process. The oncology sub-team on Adaptive Program under the Drug Information Association Adaptive Design Scientific Working Group (DIA ADSWG) evaluated hypothetical oncology programs with two competing treatments and published the work in the Therapeutic Innovation and Regulatory Science journal in January, 2014. Five oncology development programs based on different Phase 2 designs, including adaptive designs, and a standard two parallel arm Phase 3 design were simulated and compared in terms of the probability of clinical program success and expected Net Present Value (eNPV). In this article we consider eight Phase2/Phase3 development programs based on selected combinations of five Phase 2 study designs and three Phase 3 study designs. We again used the probability of program success and eNPV to compare simulated programs. For the development strategies we considered, the eNPV showed robust improvement for each successive strategy, with the highest being for a three-arm response adaptive randomization design in Phase 2 and a group sequential design with 5 analyses in Phase 3.

Keywords: Adaptive Trial Design, Decision Analysis, Expected Net Present Value, Optimizing Drug Development, Probability of Success, Group Sequential Design, Modeling and Simulation.

1. INTRODUCTION

Oncology drug development presents unique challenges with many of the studies failing in the Phase 2 and Phase 3 stages, despite increasing costs of research and development. The reasons for such failures are multi-dimensional, including both clinical and statistical aspects, such as the lack of a well-defined disease target, uncertainty about mechanisms of action of the investigational drug, insufficient correlation between surrogate and clinical endpoints and inefficient choices for clinical trial designs. In this paper we attempt to address the latter, aiming to formally optimize a clinical program on the basis of its Net Present Value (NPV). We look at various trial design choices for the clinical program and evaluate their expected NPV (eNPV) using simulations. This work is built on the efforts of Adaptive Program Oncology sub-team (under DIA ADSWG) and it extends the framework of Marchenko et al. (2013) who utilized eNPV to optimize study design choices in a hypothetical oncology clinical development program. Their work, in turn, originated from a larger effort undertaken by Adaptive Program Working Group under DIA ADSWG who first introduced the concept of optimizing a clinical program design on the basis of NPV. Examples of the application of that strategy to case studies in diabetes and neuropathic pain can be found in Antonijevic et al. (2013) and Patel et al. (2012), respectively.

Adaptive designs present particularly attractive options for such an exercise. Not only do they allow dynamic learning about accumulating treatment information without compromising the integrity of the trial, but they also promote usage of formal quantitative approaches to evaluate tradeoffs between the quality of information accumulated in a clinical trial and its cost, thus

promoting more efficient trial design choices. While there has been extensive work done on application of adaptive designs at the individual trial level, particularly in oncology, evaluation of their usage on the clinical program level remains challenging. The primary reason is that evaluating design choices for multiple stages of a clinical program introduces additional complexity, e.g., competing designs need to be compared on the basis of Type I error, power, precision, cost and duration; and some of these metrics may push the decision making in different directions. Furthermore, the comparison between designs also depends heavily on assumptions made about the treatment effect. Choosing NPV as a unified metric incorporating all these factors helps to alleviate some of the complexity. Another aspect-dependency of the NPV on the unknown treatment effect can be dealt with by taking an expectation of NPV (i.e., eNPV) over a possible range of treatment effect values with probability weights assigned according to some reasonable scenarios.

The original simulation project of Marchenko et.al (2013) considered 5 hypothetical clinical development programs in pancreatic cancer with two competing candidate investigational treatments. The objective of the Phase 2 trial was to pick the best treatment to take forward; Phase 3 then further evaluated the selected compound. The design of Phase 3 was a single stage design, while Phase 2 designs consisted of single and multi-stage designs with adaptive features of increasing complexity. The programs were compared using the eNPV. The key finding was that additional investment in the complexity of Phase 2 almost always paid off in terms of the eNPV and the most notable increase in the eNPV was in moving from the program evaluating 2 candidate compounds via separate trials in Phase 2 to the program using a 3-arm Phase 2 trial of both compounds with a shared control. This paper takes a more in-depth look into this

comparison, adding an adaptive design option for Phase 3, similar to the approach of Jennison (2011). We consider five Phase 2 designs as in Marchenko et.al (2013) and add Phase 3 designs with 2 and 5 analyses. We also use the Phase 2 results to compute the sample size for the Phase 3 trial and optimize design parameters for the Phase 3 (e.g., Go/No-Go decision criteria, and timing of the first interim analysis). The comparison is carried out on the basis of the probability of program success, measured by the power of the Phase 3 study, and the eNPV, which is calculated using simulations. As in Marchenko et.al (2013) we consider a pancreatic cancer indication with overall survival (OS) as the primary endpoint and use the same scenarios for enrollment rates and treatment effects. Our objectives were to understand how the addition of adaptive features to the Phase 3 trial affects the best choice of design for the Phase 2 trial and to identify the best Phase 2/Phase 3 development program.

2. METHODS

Simulation of Phase 2

The simulation framework consisted of a Phase 2 simulator (FACTS), which was used to generate 5,000 simulated trials for each scenario, for each type of Phase 2 trial design evaluated, at each Phase 2 sample size investigated. A particular development program was then simulated by loading the corresponding set of Phase 2 results into R and simulating the post Phase 2 decision and the subsequent Phase 3 trial. The post Phase 2 decisions were limited to a) whether to proceed to Phase 3 or not, b) if so which treatment to take to Phase 3 (if 2 had been tested), and c) what the size of the Phase 3 trial should be.

Simulation of Phase 3

Possible designs for Phase 3 included a single stage design and group sequential designs (GSDs) with either 1 or 4 interim analyses (i.e., 2 or 5 analyses including the final analysis).

Efficacy Scenarios

The primary endpoint was overall survival. We express the efficacy of a new drug (ND) in terms of its hazard ratio (HR) relative to control (e.g., the standard of care). We considered 5 scenarios for each ND, where the possible "true" values for the HR were: 1, 0.9, 0.8, 0.7, and 0.6. We gave prior weights to how likely these scenarios were to occur. The weights were based on raising the assumed HR to a power and re-normalizing. We selected a power that gave a combined probability that the treatment was ineffective (HR of 1 or 0.9) of ~80% which yielded probabilities of 52.9%, 27%, 12.7%, 5.4% and 2% respectively for the HRs 1, 0.9, 0.8, 0.7 and 0.6. When a second treatment was included in the program we retained the above probabilities for the HR of the first treatment but selected a power that gave a probability of ~90% that the second treatment was ineffective, yielding probabilities of 66.4%, 23.7%, 7.5%, 2% and 0.4% respectively; in weighting scenarios with two new treatments, we assumed the values of HR for the first and second treatments to be independent. Thus for programs where one treatment arm was being tested, there were five scenarios simulated, and where two treatment arms were being tested, 25 scenarios were simulated.

Formula for Net Present Value

The value and commercial viability of a drug are measured by NPV which is computed as the total revenue from the drug minus the total cost of getting the drug to the market. As mentioned in Marchenko et al. (2013), total revenue depends on price, market share, population size, and years of revenue. We compute NPV as:

NPV = discounted total revenue – cost of Ph2 – discounted cost of Ph3,

where Ph2 and Ph3 denote Phase 2 and Phase 3, respectively. In turn,

Cost of a Ph2 or Ph3 clinical trial = number of subjects × cost per subject + trial-duration × development overhead costs (per unit time).

Total revenue and cost of Phase 3 were discounted using an annual compound Discount Rate (DR) of 9%[1]. For simplicity just the discount to the start time of the revenue or cost was used.

We assume that 12 years of effective patent[2] life remain at the beginning of Phase 2. Some duration terms will now be defined:

[1] This was used because it is typical of discount rates used in industry for project planning; subsequently it was realised that this includes an element for "risk of project failure", as our calculations have expressly addressed the principal risk of failure in the decision tree this rate should probably be lower, closer to the "risk free cost of capital", which at the time of writing is ~2.5% (US 20 year treasury rate). A change to the discount rate would apply to all the designs and it seems extremely unlikely that it would change the final ranking.

2 The full patent life for a drug is 20 years, in using a time of 12 years we are taking into account firstly the time that will already have elapsed between taking out the patent and getting to the start of the execution of the Phase 2 trial, and secondly the likelihood of significantly diminished revenue in the last years of the drug's patent life from the introduction of competing compounds with the same method of action ("me-too" drugs).

T1 = Ph2 duration = Ph2 accrual time + 6 months follow-up,

T2 = time from end of Ph2 to beginning of Ph3 = 6 months,

T3 = Ph3 duration = Ph3 accrual time + 6 months follow-up,

T4 = time from end of Ph3 to product launch =12 months.

The revenue that would be likely to be achieved should the drug development be successful was calculated assuming that the maximum annual revenue for the current treatment in this setting was $204M. We supposed the current treatment had increased median expected life expectancy by 1.15 months from 4.76 to 5.91 months and the maximum annual revenue for a new drug would be proportional to the increase in median lifetime over 4.76 months, with earnings being 50% more than those of the current treatment if it were twice as effective, i.e., if the median life expectancy were extended to $4.67 + 2 \times 1.15 = 6.97$ months[3]. For our scenarios with HRs of 1, 0.9, 0.8, 0.7 and 0.6 this gives peak revenues of $153M, $240M, $350M, $490M and $677M respectively. Note that the Null scenario (HR of 1) has a positive potential peak revenue equal to 75% of that of the current treatment. This reflects the expectation that a new drug comparable to the current treatment would get some revenue, but not as much. However, our Phase 3 trial is designed to test superiority in comparison with the standard treatment, so the probability of being licensed and achieving this revenue is very low (one-sided Type 1 error is set to 0.025).

[3] There are huge uncertainties inherent in estimating future revenue even assuming successful registration, it is important to remember that the primary concern here is to develop a yardstick for comparing development strategies with different time, cost and risk trade-offs, and not financial forecasting.

We make the further assumption that the average annual revenue over the effective patent lifetime after the launch of a new treatment is one half of the peak annual revenue. Under these assumptions, a successful Phase 3 trial will lead to

Total revenue = $[(\$204M \times (0.75) \times (S-4.76)/(1.15)) \times (T-D)/2] \times (1-DR)^{D}$, where

S = true median survival time for the new drug, in months,

T = time to patent expiry from beginning of Ph2 (12 years),

DR = discount rate (0.09),

D = development time from beginning of Phase 2 to launch = T1+T2+T3+T4.

Also, we have

Total Ph2 cost = N2×C1+T1×C2, where

N2 = number of subjects in Ph2,

C1 = cost per subject ($20,000),

C2 = development overhead ($10M yearly), and

Total Ph3 cost = $[N3 \times C1+(T2+T3) \times C2] \times (1-DR)^{T1+T2}$, where

N3 = number of subjects in Ph3.

It is assumed that Phase 2 patients can be enrolled at the rate of 10 per month and Phase 3 patients at the rate of 20 per month.

To use the revenue and cost formulas, all durations (except for S, true median survival time, in months) are converted to years.

For the adaptive program, based on experience in the working group, an additional time of three months for trial planning and design, and further costs of $0.5M for additional operational logistics and $0.5M for additional materials were added to the Phase 2 time and costs.

The expected NPV is a probability-weighted average of the associated costs and revenues, taking into account the probability distribution of the underlying degree of efficacy of the new drugs and the probabilities of reaching each stage of clinical development and product approval (conditional on the degree of efficacy).

Thus, in all five programs explored here, if positive results are obtained in Phase 2, the selected treatment will be further investigated in a single[4] two-arm Phase 3 clinical trial. If the Phase 3 trial uses a group sequential design, then a fixed number of interim analyses occur at pre-specified intervals based on the number of deaths that have been observed. If the p-value meets the success or futility criteria for that interim analysis, the trial stops with the appropriate decision. Otherwise, patients are followed up until the maximum target number of deaths is observed and the final p-value is compared to the final success threshold.

[4] With pancreatic cancer the burden of unmet medical need is so high, that it is believed that a single Phase 3 trial will be acceptable to regulatory authorities

The interim stopping boundaries and final success threshold were set to control the overall one-sided Type 1 error rate at the level of 0.025[5]. We have assumed that once statistical significance is attained in Phase 3, a new-drug application will be submitted for regulatory review and approval to market the drug − and sales of the new drug will commence 12 months later.

The Phase 2 Statistics

In the Phase 2 designs the decision criteria were based on a Bayesian calculation of the probability that the HR of the treatment compared to the control was < 1. This calculation was made under a Bayesian exponential model in which the prior distribution for λ, the weekly hazard rate (h) on the control arm, was assumed to be a gamma distribution with mean 0.027 and weight of 1 (a weak prior with a mean corresponding to a median survival time of 4.76 months with an effective sample size of 1 observation). The hazard rate for each new drug was assumed to be of the form $h = \lambda e^{\theta}$, where the log hazard ratio θ has an independent prior distribution for each new drug, assumed to be a normal distribution with mean 0 and standard deviation 5.

In designs with an option to drop an arm at an interim analysis, either treatment arm was dropped if the posterior probability that the hazard ratio was less than one was below some preset threshold: $Pr(HR<1) < C_{Drop}$. Values for C_{Drop} were used, capped to limit the 'erroneous drops' to less than 1 percentage point when the treatment had a HR of 0.7. This cap was determined by simulation. The post Phase 2 decision logic was expanded to accommodate the arm dropping. If both arms were dropped, then the Phase 2 was futile and no Phase 3 was run. If one arm was

[5] Note the stopping boundaries were set assuming that the futility stopping boundary was non-binding, it is believed that the FDA prefers this more conservative assumption.

dropped, the trial continued as a single arm Phase 2 trial. If neither arm was dropped, then at the end of the trial the arm with the smallest estimated HR was selected and its estimated HR and posterior probability Pr(HR<1) were used for the post Phase 2 decisions.

In designs with adaptive allocation, at each interim analysis the probability of each treatment being the best was calculated. Posterior estimates of λ, θ_{ND1} and θ_{ND2} , where ND1 and ND2 denotes the two doses of a new drug, were calculated using an MCMC Metropolis Hastings algorithm and the posterior probability of ND1 being best was estimated by the proportion of MCMC samples with $h_{ND1} < h_{ND2}$. After an initial fixed allocation of 24 patients to the control arm and 20 patients to each treatment arm, five out of every eight patients were randomized to ND1 and ND2 in proportion to the square root of the posterior probability that the drug had the better efficacy, weighted by the expected reduction in variance from adding one more subject to that arm. Thus, the randomization weighting for dose *x* at an interim analysis was given by

$$< \boldsymbol{i} > \boldsymbol{\omega_x} \propto \sqrt{\frac{\mathbf{P(ND_x\ is\ best)} * \mathbf{VAR(\theta_x)}}{\mathbf{n_{NDx}} + \mathbf{1}}} \; </\boldsymbol{i} </\boldsymbol{i} >>$$

and these weights were updated at each interim analysis. The remaining 3 subjects were randomized to the control arm.

At the end of the trial the treatment with the smallest estimated HR was selected and its estimated HR and the posterior probability Pr(HR<1) were used for the post Phase 2 decisions.

The 5 Different Phase 2 Designs

We simulated Phase 2 trials with 5 different designs:

P2.2a.f *A 2 arm (control and 1 treatment) Phase 2 trial with fixed, equal allocation to each arm. At the end of the trial the Bayesian posterior probability that the HR of OS on the treatment arm compared to control was less than 1 was calculated. If this probability was greater than a predetermined threshold, then the trial was successful and ND1 was investigated in a Phase 3 trial with the sample size based on the estimate of the HR. We simulated 5 HR scenarios: 1, 0.9, 0.8, 0.7 and 0.6.*

P2.3a.f *A 3-arm (control and 2 treatments) Phase 2 trial with fixed, equal allocation to each arm. At the end of the trial, we calculated for each arm the posterior probability that the HR of OS of the new drug compared to control was less than 1. If at least one of these probabilities exceeded a predetermined threshold, then the trial was deemed to be successful and the treatment with the lower estimated HR was investigated in a Phase 3 trial with the sample size based on the estimate of the HR. We simulated all 25 combinations of the 5 HRs for each treatment.*

P2.3a.1i *A 3-arm Phase 2 trial with one interim analysis at which either or both ND1and ND2 can be dropped (if both are dropped the trial stops for futility) or the trial can be stopped early for success. If a treatment is dropped, the overall trial size is decreased (any future subjects that would have been recruited into the dropped arm are no longer recruited). Simulations were run with different timings for the interim analysis (in terms of number of events) in order to be able to select the optimal timing. The decision to drop either treatment was based on whether the*

Bayesian posterior probability that the HR of OS on the treatment arm compared to control was less than 1, was below a threshold.

P2.3a.mi *Same as P2.3a.1i but with multiple interim analyses (every 20 events) at which the trial could stop for success or futility or drop an arm.*

P2.3a.ad *Same as P2.3a.mi but at each interim analysis response adaptive randomization (RAR) was used, modifying the allocation proportion between the two treatment arms (ND1 and ND2) to favor the apparently more effective arm, if the trial had not already stopped for success or futility.*

The Phase 3 Designs

Three Phase 3 designs were evaluated in various combinations with the Phase 2 designs.

P3.0i *A single stage two arm study with the sample size based on the estimated HR of the treatment arm to be tested.*

P3.1i *A group sequential two arm study with 1 interim analysis at which the trial could stop for success or futility.*

P3.4i *A group sequential two arm study with 4 interim analyses at which the trial could stop for success or futility.*

In each case, following Phase 2, the required sample size for the Phase 3 was calculated using ssizeCT.default() from the R powerSurvEpi package. The maximum sample size for the Phase 3 trial was calculated based on the estimate of HR from Phase 2 limited by a minimum expected

HR (to avoid dangerously small Phase 3 trials when the random fluctuation in events occurring in the Phase 2 trial had dramatically favoured the new treatment), and a maximum permitted HR above which the treatment's efficacy was deemed too poor to be worth continued development and would have required either an unfeasibly large Phase 3 trial or an underpowered one.

The minimum expected HR was a design parameter that we optimized, the maximum HR was fixed at 0.9 for all the programs.

For the group sequential Phase 3 trial design we calculated Phase 3 boundaries based on the maximum expected number of events. The maximum expected number of events was based on the maximum trial size and the proportion of subjects for which events would be observed, this latter being a tuning parameter for the design of the program.

Interim analyses would be equally spaced (in terms of number of events observed) after the first interim analysis, the timing of which (as a proportion of the maximum expected number of events) was a tuning parameter for the design of the program.

We used an error-spending group-sequential design (Jennison and Turnbull, 2000, Ch. 7). The stopping boundary is chosen so that the cumulative Type 1 error rate at analysis k (under log(HR)=0) is $\alpha\,(I_k / I_{max})^2$. Here I_{max} is a target maximum amount of information – one quarter of the maximum number of subjects in the sample population that we expect to have events during the trial, since the information for a log(HR) is Number-of-events/4. Similarly the cumulative Type 2 error at analysis *k* (under log(HR)=δ) is $\beta\,(I_k / I_{max})^2$. At the end of the trial, the number of observed events may not be exactly equal to the planned maximum: in this case,

the final boundary point is calculated (separately for each individual simulated trial) to give total cumulative Type 1 error equal to α (see Jennison and Turnbull, 2000, Sec. 7.3.a). For each simulated Phase 2 trial the subsequent Phase 3 trial is designed based on the estimated HR at the end of Phase 2 and the following program design parameters:

- The minimum expected HR,
- The target maximum information – one quarter of the expected number of subjects for whom events will be observed,
- The minimum information – the fraction of the maximum information that must be observed for the first interim analysis to be carried out, the remaining interim analyses being equally spaced up to the maximum information,
- The required Phase 3 Type 2 error β (power = $1 - \beta$) at the observed HR.

These parameters were selected independently for each development program, optimized to maximize the expected NPV over the different treatment effect scenarios.

Two parameters were fixed for all programs:

- The maximum HR for which a Phase 3 trial will be run = 0.9,
- The required one sided α for the Phase 3 trial = 0.025.

The Programs Evaluated

Not all combinations of Phase 2 and Phase 3 design were evaluated:

- The only combination of Phase 2 with P3.0i was P2.2a.f. This combination of the two simplest trial designs (which still represent a very common program design, possibly a majority of program designs) forms a baseline against which the others can be compared, and in particular a reference point for the program that combines the two most complex trial designs to show how much overall value could be generated by deploying sophisticated types of trial design.

- P2.2a.f and P2.3a.f were evaluated combined with both P3.1i and P3.4i.

- P2.3a.1i, P2.3a.mi and P2.3a.ad (the more complex types of Phase 2 design) were only combined with P3.4i, after earlier evaluations confirmed that P3.4i was always superior to P3.1i.

Thus we evaluated the following programs:

R0:	P2.2a.f+P3.0i	Two-arm fixed Phase 2, no interim analyses in Phase 3
R1:	P2.2a.f+P3.1i	Two-arm fixed Phase 2, 1 interim analysis in Phase 3
R2:	P2.2a.f+P3.4i	Two-arm fixed Phase 2, 4 interim analyses in Phase 3
R3:	P2.3a.f+P3.1i	Three-arm fixed Phase 2, 1 interim analysis in Phase 3
R4:	P2.3a.f+P3.4i	Three-arm fixed Phase 2, 4 interim analyses in Phase 3

R5: P2.3a.1i+P3.4i Three-arm Phase 2 with one interim analysis, 4 interim analyses in Phase 3

R6: P2.3a.mi+P3.4i Three-arm Phase 2 with many interim analyses, 4 interim analyses in Phase 3

R7: P3.3a.ad+P3.4i Three-arm Phase 2 with many interim analyses and Response Adaptive Randomization (RAR), 4 interim analyses in Phase 3.

Optimization of Program Parameters

The aim of the exercise is to compare the notional value of each program design so we can both rank them and also have a benchmark for addressing the question "is the additional complexity of this design over that one likely to be worthwhile?". There are a number of risks to this exercise of comparing designs that we try to mitigate by this optimization stage. Firstly it is unlikely that one set of parameter settings would be suitable for all designs, and using just one set could unfairly penalize some designs compared to others. Secondly, if we decide to manually set the parameters individually for each design, it is not clear how we should set these parameters − and it is difficult to judge whether we have set them fairly. How do we avoid consciously or unconsciously favoring a design we might prefer? And how do we demonstrate that we have not favored one? There can also be interplay between the phases, for instance, a design that uses a larger Phase 2 should provide a better estimate of the effect size for use in the Phase 3 sample size calculation, leading to a more efficient Phase 3 requiring less conservative parameters for the design of Phase 3. An automatic process of optimizing the parameters for each design to

maximize the eNPV for that design appears to be the best solution. This is the most relevant form of optimization as the eNPV will be the "score" by which we judge the programs. The advantage of eNPV is that it allows us to combine into a single figure the various operating characteristics of a drug development program – its cost, the time it takes, its probability of success and the value of the treatment selected and taken to market.

There was a third and unexpected benefit of the automated optimization, the optimized parameter values shed light on the consequences of the design assumptions and eNPV parameters (essentially helping us to "debug" them) and differences in the optimized parameter values from one design to another, shed light on the different characteristics of the designs.

Some of the parameters required to calculate the eNPV may have large uncertainty, e.g., the expected peak revenue and Phase 3 accrual rate. For the purposes of this paper we only use eNPV as a tool to evaluate design options, optimizing the program design parameters to ensure a level playing field. We note though that the same approach could be extended, using further simulation over the uncertainties in the assumptions to understand the range of eNPV and the robustness of different program design parameter choices.

Each program was optimized as follows

- First a set of simulation results were generated for the Phase 2 trials, by running simulations over the different scenarios with HRs of 1, 0.9, 0.8, 0.7 and 0.6 if one treatment is being developed, or the 25 combinations of these HRs for the two treatments if two are being developed. For each Phase 2 design, simulations were run for each combination of sample size

and timing of the interim analysis (or the first of 4 interim analyses) to be evaluated. (In some cases simulations were run at additional values to ensure that parameters giving a maximum eNPV had been evaluated).

- Each set of simulation results was then evaluated for a set of specific values of the Phase 3 design parameters: the Phase 2 success threshold, the Minimum expected HR[6] (equivalent to specifying a minimum Phase 3 size), the Maximum information (the Phase 3 sample size multiplied by the proportion of subjects who have events and divided by 4), the Minimum information (the timing of the first interim analysis) and the required Type 2 error rate for the Phase 3 trial.

- Given the Phase 2 simulation results and the Phase 3 design parameters it is possible to calculate (1) whether Phase 2 was a success (and thus we "go" to Phase 3), (2) the appropriate Phase 3 trial size (which involves the observed HR), (3) the Phase 3 stopping boundaries, and (4) the actual probability of futility or success at each interim analysis or the final analysis in Phase 3, based on the "true" HR of the selected treatment in the scenario being simulated. The timing of the interim analyses of the Phase 3 trial was sampled based on stochastic models of patient recruitment and closed-form expressions for predicting counts of events (Anisimov, 2011) and this helped reduce the time of computations very substantially.

- The above computations allow us to calculate the cost, expected revenue and probability of each possible outcome:

[6] Low HRs are good, and the purpose of the "Minimum Expected Hazard Rate" was to set a floor on the estimate of the HR from Phase 2 that was used to calculate the Phase 3 sample size. This corresponds to a caution in the clinical team when designing the Phase 3 when the Phase 2 results look "too good to be true".

- These results are then averaged over all the simulations for each scenario and then averaged over all the scenarios, weighting the values by the probability of each scenario.

- Finally the eNPV is adjusted to penalize "reckless" programs that are willing to proceed to Phase 3 even when the likelihood of success is low. This occurs because in this model with these assumptions the size of the expected revenue is much greater than the cost of Phase 3. So in addition to maximizing the eNPV we set an additional constraint that the desired average probability of success in Phase 3 across all the scenarios should be 50%[7]. In our first attempt we set the eNPV of any program that failed to meet this criteria to zero, but this interfered with the optimizers search for the best parameters. It created a 'cliff' in the eNPV values and large "flat" area of value zero. So instead we penalized any design with a success rate in Phase 3 of less than 50% by the ratio of successful to unsuccessful Phase 3 divided by 50%. Thus we scaled the program eNVP by min(1, P3_success_ratio/0.5). This was a sufficiently strong penalty that programs with a success ratio less than 50% were rarely optimal, and programs with a success ratio much less than 50% never were.

- Conventional optimisation methods for multi-parameter models either assume a derivative is available or that evaluation of the function is not too expensive (in time or computing capacity). Here neither is true, but we only had a small number of parameters to explore and the eNPV changes quite slowly – the surface is smooth − thus the parameter space

[7] This was a somewhat arbitrary target, and might be too high, in particular in an indication of high, unmet medical need. In the context of a large pharmaceutical company it might represent a threshold required to justify investment in this development program rather than in an alternative drug that could be developed, or it might represent a corporate risk aversion. We think that if this number were adjusted up or down to 60% or 40% while changing he absolute eNPV figures, it would be unlikely to change the ranking of the designs.

could be searched by a relative unsophisticated "binary chop" method (see the Appendix for details).

Optimizing the adaptive programs R5-7 was trickier because there were now additional Phase 2 trial design parameters, and the effect of these could not be evaluated without re-running the simulations of the Phase 2 trials, so the exploration of effect of these parameters was less comprehensive.

R5: Introduced an interim analysis in the Phase 2 trial with the possibility of dropping one or both arms at the interim analysis or stopping for success, so the additional parameters were the timing of the interim analysis in the Phase 2 trial and the threshold at which to drop an arm or declare success and go to Phase 3.

R6: Introduced multiple interim analyses (every 20 events observed) at which one or both arms could be dropped. The additional parameters were the timing of the first interim analysis in the Phase 2 trial and the threshold at which to drop an arm (at the first or any subsequent interim analysis) or to declare success and go to Phase 3.

R7: Used multiple interim analyses, but rather than dropping an arm, adjusted the randomization ratio to favor the better performing treatment until the trial stopped for futility or success. The additional parameters were the timing of the first interim analysis in the Phase 2 trial and the threshold at which to stop for futility or declare success and go to Phase 3.

Rather than devote a large amount of computational effort to exploring combinations of values for the 4 design parameters for the Phase 2 trial (sample size, timing of interim analyses or first

interim analysis, threshold to stop for futility or drop an arm, threshold to stop for success), we derived the values for the thresholds simply on the basis of the Type 1 error and power in Phase 2 and we then explored combinations of interim analysis timing with sample sizes that had been optimal for the similar fixed trial (R4).

The stopping decisions were based on the posterior probability of the HR of the best performing treatment being less than 0.9. The Type 1 error in the scenario where the HR was 1 varied between 11% and 26% and the typical decision threshold after optimization was 65-75%. That is, if the posterior probability that the HR of the treatment arm versus control (or best treatment arm if two were being tested) being better than 0.9 was 65-75% or better, then the Phase 2 trial was judged a success and the program proceeded to Phase 3. For revenue, one of the most important scenarios was the scenario where one treatment had a HR of 1 and the other a HR of 0.7. For simplicity we focused on the power of this scenario which in the fixed programs was typically around 78%.

By a little trial and error we determined that if the final Phase 2 success threshold was Pr(HR < 1) > 0.75 we could use an early stopping threshold of Pr(HR < 1) < 0.25 for futility or arm dropping and Pr(HR < 1) > 0.9 for early success and preserve the Type 1 error and power characteristics of the fixed programs. We then explored a small set of combinations of sample size and timing of the first interim analysis.

Table 1 shows the eNPV at the optimized parameter values for program R5 at different Phase 2 sample sizes and different possible timings (in terms of events observed) of the interim analysis. What is striking is how the eNPV does not vary. There is clearly enough flexibility in the choices

of the Phase3 parameters that they can compensate for slight differences in the operating characteristics of the Phase 2 trial. For the sake of comparison with the other designs we select the smaller of the sample sizes and the later time for the interim analysis.

For programs R6 & R7, because of the many interim analyses compared to program R5's single interim analysis, the thresholds for deciding early futility and success need to be more conservative, but both designs take a lot longer to simulate. In order to determine reasonable initial values for the early stopping thresholds, the Phase 2 trials for programs R6 and R7 were simulated without early stopping and the results analyzed to predict the consequences of different possible stopping thresholds.

The FACTS software we used to simulate the Phase 2 trial has a built in graphing facility that calculates contours of equal proportions of successful simulations for combinations of early and final success threshold (see Figures 2 & 3).

Null scenario - contours of equal Type 1 error HR = 0.7 scenario - contours of equal power

We select the Phase 2 success thresholds for which the Type 1 error rate is similar to that demonstrated through simulation – here declaring success in the null scenario – to roughly those levels that we saw in the optimized fixed programs, about 16%. As this is in Phase 2, analytical control of Type 1 error is not required, nor do the levels of control need to be as stringent as those used when testing the null hypothesis in a confirmatory trial.

Superimposing the contour of 16% success rate in the Null scenario on the scenario with HRs 1 and 0.7 (a key alternative scenario) we see it crosses the contours of the highest probability of success at the left hand end – where the threshold for early stopping for success is 0.9375 and the threshold for success at the end (if the trial has not stopped early) is 0.75.

On the graphs, plotting contours of equal proportions of simulations that are futile, we look at the futility rate (Type 2 error) in our representative alternate scenario when the final threshold is 0.75 (final futility is the opposite of final success, so the same threshold is used, but futility is when the posterior probability the HR < 0.9 is below the threshold).

Null scenario - contours of equal Type 1 error HR = 0.7 scenario - contours of equal power

It appears that we can use an early stopping for futility threshold of about 0.2 without significantly increasing the futility rate in this alternate scenario, but we would expect to see the futility rate rise if the threshold were much higher.

This then gives the early stopping thresholds: 0.9375 for success, 0.2 for futility.

Conventional Sample-Size Perspective

As a point of reference, from a conventional-sample-size perspective, suppose no program-wide criteria (such as eNPV) and no adaptations had been applied in the design of the clinical program, so the scheme is like our program R1. Then, in order to have a statistical power of 80% in Phase 2 and 90% in Phase 3, assuming exponentially-distributed deaths, a median OS of 5.91

months in the control arm, and an HR of 0.7, a sample size of 180 patients (90 per arm) would have been needed in Phase 2 and 400 patients (200 per arm) in Phase 3.

3. RESULTS

Programs with Fixed Phase 2 trials

The comparison of programs R0-R4 was relatively straight forward, the 2 types of fixed Phase 2 trial (with one or two treatments arms) were simulated with sample sizes in the range of 80 to 320, and the optimal program parameters were derived for each one, and the probability was weighted, and the net present value was calculated for each one.

It can be seen that having a group sequential Phase 3 is clearly better (~$40M) than a fixed Phase 3 (R0) and having more interim analyses (R3 & R4) is somewhat better (~$15M) than having just 1 (R1 & R2).

Unlike in the previous paper of Marchenko et al. (2013), at this point we see no advantage in testing 2 treatments in Phase 2 compared to testing just one. This was a consequence of having reduced the weightings for the second treatment's successful scenarios so the expectation of success is less than that for the lead compound. The reduced weightings not only make it less likely to be successful but also less likely to be **very** successful. As a result, in the programs that test 2 treatments in fixed Phase 2 trials the advantage of "having a second shot at goal" is almost exactly offset by the additional time and cost of testing the second compound.

Some observations on these results:

- The eNPV at the optimized parameter values varies very smoothly with Phase 2 sample size when the Phase 3 is fixed (R0), but fluctuates more when the Phase 3 trial is group sequential. This is because the interim analyses in the Phase 3 acts somewhat like the go/no-go decision after Phase 2, thus changes in the Phase 2 sample size can be compensated for with the other parameters.

- The eNPV at the optimized parameter values varies smoothly when 2 treatments are being tested, presumably because the power to correctly select between the two treatments relies solely on the Phase 2 trial data and adjusting the Phase 3 parameters can no longer compensate for changes in Phase 2 sample size.

- The variability in the eNPV from unimodal curve is however quite small – $2-4M.

- There is nevertheless a clear "maximal eNPV" region for the Phase 2 sample size, below which smaller trials loose power faster than they gain value by being faster and above which larger trials loose more value by taking longer than they gain by the increase in power or accuracy of determining the treatment effect in order to make a better choice of sample size for Phase 3.

- While there is a region of quite similar eNPV for different Phase 2 sample sizes, the expected performance of the program differs with the different sample sizes. With the smaller sample sizes, the program is quicker so the NPV if the program is successful is higher, but compared to the program with larger Phase 2 the chance of success is lower.

- The optimal Phase 2 size of 180-200 when testing 1 treatment, or 210-240 when testing 2 treatments, was surprisingly independent of whether the Phase 3 was fixed or group sequential, or how many interim analyses the group sequential Phase 3 design used.

We see a steady increase in eNPV as the complexity of the program increases and, while at each step the increment is relatively modest, the final design R7 has an eNPV ~65% greater than that of the simplest program: testing 1 arm in a fixed Phase 2 trial, followed by a fixed Phase 3 trial.

Lastly we performed a relatively simple robustness test on these results. We re-evaluated each of the programs adjusting one of the values used in the assumptions. From other work we expected that the parameters to which the eNPV would be most sensitive would be the expected peak value (PV) and the Phase 3 recruitment rate (P3RR). We simply re-evaluated each of the programs varying these parameters by +/- 20%.

Our key conclusion from these results is that the rankings of the different programs do not change as we vary these key assumptions. For the scenarios we considered, Program R7 was the best choice. It had the highest eNPV and the second highest probability of success. Program R4 had slightly higher probability of success, but lower eNPV. Program R0 was the worst, in terms of probability of success as well as the eNPV. Overall, the eNPV increased as the complexity of designs increased. It has been seen that having a group sequential Phase 3 trial was clearly better than a fixed Phase 3 trial and having more interim analyses was somewhat better than having just one. A Phase 3 GSD can help reduce Phase 2 sample size because we do not have to rely on an accurate estimate of the true HR when choosing the Phase 3 sample size – but we do still need to select the correct treatment in Phase 2.

In practice there are further areas of robustness checking that should be carried out, in particular:

- The choice of early stopping thresholds in designs R6 and R7

- The choice of the actual Phase 3 parameters optimized not for specific values of the eNPV parameters as here, but over the potential range of their values.

4. CONCLUSION

Traditionally, optimization of the clinical development program has been done at an individual-study level. Though extremely useful, optimization at an individual-study level does not always lead to an optimal clinical program or improved chance of success for the compound under investigation. Therefore, in our work, we aimed to optimize a clinical program part that includes both Phase 2 and Phase 3 studies. This is a second paper written by the oncology sub-team on Adaptive Program in which we evaluate hypothetical oncology programs and compare them using probability of program success and eNPV. As in Marchenko et al (2013), we assumed that there are two compounds which have graduated from Phase 1 and ready for further development in advanced pancreatic cancer. In this paper we explored the impact of using group sequential design in Phase 3 and less optimistic expectations that the two compounds would be successful. In this setting the value of using a group sequential design in Phase 3 was clear (a roughly 35% increase in eNPV) and the advantage of using an Adaptive Phase 2 design slightly reduced (only about 15% further increase in eNPV). However, there was almost no benefit from simultaneously testing this poor second best treatment in the Phase 2 unless the Phase 2 was adaptive. We note that the benefit of the adaptive Phase 2 increased the more adaptive it was and

that the benefit was robust across changes of assumptions regarding market value and Phase 3 recruitment rate. We also note that the adaptive Phase 2 was able to deliver this increase in value despite minimizing its opportunities to do so through having an adaptive Phase 3, only two treatments to choose between and a reduced expectation that the second treatment would be effective.

It has been suggested to us that in this setting – with the same endpoint being used in Phase 2 and Phase 3 – exploring a seamless design would be interesting design option to explore. The additional revenue from savings in time between Phase 2 and 3 would almost certainly exceed any additional cost of development, and the cost of having to invest in Phase 3 ahead of Phase 2. It would be interesting to see how low the initial expectation of success (in the terminology of this paper the relative weighting of the scenarios with treatment ratio < 0.9) could go before the eNPV of the seamless design was no greater than that of the equivalent separate stages. While such an investigation is beyond the scope of this paper, it could well be the basis of an interesting paper in its own right.

High performance computers have facilitated widespread advances in the development of computational algorithms, statistical modeling, and simulations. These advancements helped to increase the use of Bayesian and hybrid designs in clinical trials that add to flexibility and interim decision making based on accumulated data in the trial and knowledge from outside the trial. Computer simulations are widely used to facilitate decision making under uncertainty and to compare different designs under consideration. When comparing individual study designs, typical metrics include sample size, power (or probability of success), duration of study, etc.

under different scenarios. To make comparisons simple, we decided to use eNPV that allows one to combine various operating characteristics and understand the investment and gain better. The NPV formula given in the Methods section was derived from consultation with our marketing colleagues. In addition to eNPV, we also reported the probability of program success. We feel this evaluation technique might become increasingly important as the industry moves to more quantitative decision making methods.

Compared to the paper of Marchenko et al (2013), we found reduced advantage in Phase 2 adaptive designs testing two experimental treatments rather than one because we significantly reduced the weightings of the scenarios where the second drug was effective. This leads to an important conclusion that investigators need to consider carefully what they know and what they expect if they are to gain the most from process optimization. For simplicity, we considered discrete treatment effects in terms of hazard ratios. The idea can be easily generalized to use a distribution of the treatment effect. Such distribution can be derived from available data, preclinical or clinical. Also, time and space limited us to eight designs; there are certainly many other candidate programs that one could compare and the framework we have put together can easily accommodate that.

REFERENCES

1. Anisimov V. Predictive event modelling in multicenter clinical trials with waiting time to response, *Pharmaceutical Statistics*, 10, iss. 6, 2011, 517-522.

2. Antonijevic Z, Kimber M, Manner D, Burman CF, Pinheiro J, Bergenheim K. (2013) Optimizing drug development programs: Type II diabetes case study. *Therapeutic Innovation & Regulatory Science; 47,3: 363-374.*

3. Antonijevic Z, Editor (2015). Optimization of Pharmaceutical R&D Programs and Portfolios, Design and Investment Strategy. *Springer*

4. Jennison C. Talk at ADAPT conference on Adaptive Clinical Trials, Philadelphia, September 2011 "Effective design of Phase II and Phase III trials: an over-arching approach": *http://people.bath.ac.uk/mascj/talks_2011/cj_ADAPT.pdf.*

5. Jennison C and Turnbull B. (2000). "Group Sequential Methods with Applications to Clinical Trials", Chapman & Hall.

6. Marchenko O, Miller J, Parke T, Perevozskaya I, Qian J, Wang Y. (2013) Improving Oncology Clinical Program by Use of Innovative Designs and Comparing Them via Simulations *Therapeutic Innovation & Regulatory Science* 2013 47: 602.

7. Patel N, Bolognese J, Chuang-Stein C, Hewitt D, Gammaitoni A, Pinheiro J. Designing Phase 2 Trials Based on Program-Level Considerations: A Case Study for Neuropathic Pain. *Drug Information Journal*. 2012; 46,4: 439-454.

Appendix: An example of the Optimization Method

Tables A1, A2 and A3 show the search for Program R2 when just treatment ND1 is tested in a fixed Phase 2 trial with N=200, followed by a group sequential Phase 3 trial with 4 interim

analyses. For the optimisation initial values were chosen that were thought to represent values a clinical team might plausibly choose. The parameters being optimized are:

P2 Go The required threshold for the posterior probability Pr(HR <1). If Pr(HR <1) > P2, we go to Phase 3. Initial value: 0.8

P3 Beta When calculating the sample size for the Phase 3 trial, the required power is (1 – P3_Beta). When the Phase 3 used a group sequential design, the P3 Beta is also used in deriving the stopping boundaries. Initial value: 0.2

Min Exp HR The Phase 3 sample size was calculated using the point estimate for the Hazard Ratio from the Phase 2. This parameter sets a floor on how low that value was allowed to be. Thus for the expected effect size in the sample size calculation we used max(HR estimate from Phase 2, Min Exp HR). Initial value: 0.6

Max Factor This for the purposes of Phase 3 sample size calculation and group sequential boundary calculation is the proportion of subjects in Phase 3 that are expected to have events. Initial value: 0.8

Min Factor This was used to allow the timing of the interim analysis if only one, or the first interim analysis if more than one, in the group sequential trial. Conventionally in group sequential trials the interim analyses are equally spaced, in terms of information, through the trial. But it was not clear to us that that was necessarily optimal and we thought it would be interesting to include this as a parameter to be optimised. Initial value: 0.5.

The **Maximum HR** parameter was not optimised, it was fixed at 0.9: a point estimate of above 0.9 meant not running a Phase 3. This was justified on the grounds that the resulting Phase 3 would be so large that the expected revenue was inevitably very small. The required Alpha (type I error) for Phase 3 was also fixed, in this case at 0.025 (one-sided) as is typically required by regulators.

The scheme is based on the idea of a multi-dimensional "binary chop". Upper and Lower boundaries are set for each parameter, and at each step we check parameter values half way between the current value and the boundaries yielding "High" and "Low" alternatives to the current value of the parameter. The eNPV of the design is tested with each parameter set in turn to its High and Low values. A new set of parameter values is then chosen by replacing the current value with the "High" or "Low" alternative if that gave a higher eNPV than the current parameter values. If changing all the parameters in one step yields an eNPV that is lower than that at the current value then that new set of parameter values is rejected and just a single parameter is changed – the one yielding the greatest increase in the eNPV over that of the current parameters. If none of the High or Low alternative values yields an increase in eNPV then the Upper and Lower boundaries are temporarily brought in to half the distance from the current value and the process repeated until either a maximum number of reductions has been tried or a set of parameter values yielding a higher eNPV was found. More sophisticated optimisation packages exist but when they yielded bizarre results it was difficult to determine why, hence the development of a simple method that could be debugged, that exploits the fact that our parameters have boundary values, minimises the number of evaluations required, but relies on eNPV as a function of these parameters being broadly smooth – which it seems to be.

In the following tables that show an example of optimization, we show the values of the parameters "current" at the start of the step, and below them the program eNPV at those parameters. Either side of the "current" values we show the "Low" and "High" alternate values of that parameter to be tested and in the columns outside those, the program eNPV if that parameter were to takes its low or high value and the other parameters kept their current values. The new values for the next step are obtained by replacing the current values by their low or high value if that lead to a higher program eNPV[8].

In program R2, the Phase 3 sample size depends on Phase 2 data, in particular, the posterior expectation of HR after Phase 2. Table A4 shows the Phase 3 sample sizes for various expected HRs that result from the optimized parameters shown in Table A3. Note that it has been determined that it is optimal to set a floor on the value of the expected HR (Min Exp HR) of 0.8.

Figures A1 and A2 show the stopping boundaries for the 4 interim analyses of the Phase 3 trial when the expected HR is 0.8 (or less) and the probabilities of crossing the success boundaries for different true HRs.

Table A5 shows the summary probability, cost and revenue figures for this program with these parameters. The columns are "Pr P2 fail", "Pr P3 fail" "Pr P3 succ" are the probabilities of the 3 program outcomes – failing at the end of Phase 2, failing at the end of Phase 3 and succeeding at the end of Phase 3. "P2 Cost", "P3 (fail) Cost" and "P3 (succ) Cost" are the average costs of Phase2, when Phase 3 fails and Phase 3 when Phase 3 is successful. "Rev" is the average

[8] Unless that particular combination gave a lower program eNPV, in which case the single substitution that gave the highest eNPV is used. If there was no such value the size of the steps to the higher and lower parameter were reduced until they were within some delta of the current value and then the optimization search stops.

revenue if successful, "eNPV" is the expected Net Present Value, "Scen wt" is the scenario weighting and "Wt'd eNPV" is the eNPV after scaling by the scenario weighting. Table A6 shows the different optimized parameter values for R2 with different Phase 2 sample sizes.

Table A7 shows the optimized parameter values for each of the programs using a fixed Phase 2 design at the optimal Phase 2 sample size for that program.

Here we have just illustrated the method by its application to R2. The same method was applied to optimize all the designs, with the number of parameters to optimize varying between designs. However, some adaptive designs had parameters that impacted the execution of the Phase 2 part, and these could only be explored by running simulations at different values and optimizing the remaining "Phase 3" parameters for each set. Since these Phase 2 parameters were much more expensive to optimize, the optimization was performed at a small set of discrete values and the parameter space explored manually, whilst the exploration of the Phase 3 parameters could be performed automatically.

However it must be stressed that the purpose of the optimization was to allow fair comparison of the different designs, not to produce actual optimal parameters for the design itself. In deriving such optimal designs, one would need to allow for the enormous uncertainty in some of the parameters of the eNPV model – particularly those around expected revenue and Phase 3 accrual – and this is another, different, though equally interesting, problem from the one of choosing between different design strategies.

Table 1. Table of optimized eNPV for Program R5 at different Phase 2 sample sizes and different timings of the interim analysis

	Time of first interim analysis		
Phase 2 (events)			
Sample Size	**100**	**120**	**140**
240	244	244	245
270	244	244	245
300	245	244	245

Table 2. Optimized parameter values for programs R5-R7 at Phase 2 sample sizes that maximize eNPV

				Parameter Values				
Program	**P2 Sample Size**	**First Interim analysis**	**eNPV $M**	**P2 Go**	**P3 Beta**	**Min Exp HR**	**Max Factor**	**Min Factor**
R5: P2.3a.1i+P3.4i	240	140	245	0.75	0.26	0.72	0.58	0.13
R6: P2.3a.mi+P3.4i	270	120	250	0.75	0.54	0.74	0.58	0.13
R7: P3.3a.ad+P3.4i	330	120	276	0.75	0.54	0.74	0.58	0.13

Table 3. Scenario weighted eNPV at the optimized parameter settings for different sizes of fixed Phase 2 trial for programs (testing a single treatment) R0, R1 & R2.

	Weighted eNPV $M at different Phase 2 Sample Sizes									
Program	**80**	**100**	**120**	**140**	**160**	**180**	**200**	**140**	**280**	**320**
R0: P2.2a.f+P3.0i	137	145	152	159	163	166	166	157	144	129
R1: P2.2a.f+P3.1i	173	180	194	204	199	204	206	199	186	171
R2: P2.2a.f+P3.4i	187	195	209	216	215	219	222	216	203	188

Table 4. Scenario weighted eNPV at the optimized parameter settings for different sizes of fixed Phase 2 trial for programs (testing two treatments) R3 & R4.

	Weighted eNPV $M at different Phase 2 Sample Sizes						
Program	**150**	**180**	**210**	**240**	**270**	**300**	**330**
R3: P2.3a.f+P3.1i	192	201	210	207	205		
R4: P2.3a.f+P3.4i	209	217	225	224	221	221	218

Table 5. Probability of Success and eNPV for all eight programs.

	Results		
Program	**eNPV $M**	**Pr(Success)**	**P2 Type 1 error**
R0: P2.2a.f+P3.0i	167	11.4%	11.4%
R1: P2.2a.f+P3.1i	207	14.6%	11.6%
R2: P2.2a.f+P3.4i	210	15.0%	11.6%
R3: P2.3a.f+P3.1i	222	16.8%	18.3%
R4:	225	19.7%	26.4%

P2.3a.f+P3.4i			
R5: P2.3a.1i+P3.4i	245	18.3%	17.2%
R6: P2.3a.mi+P3.4i	250	18.3%	17.6%
R7: P3.3a.ad+P3.4i	271	19.4%	15.5%

Table 6. The eNPV of the various programs as we vary the value of 2 of the key assumptions, the expected peak value and the recruitment rate during Phase 3.

	eNPV $M								
Program	**Median**	**P3RR+**	**P3RR-**	**PV+**	**PV+, P3RR+**	**PV+, P3RR-**	**PV-**	**PV-, P3RR+**	**PV-, P3RR-**
R0: P2.2a.f+P3.0i	167	180	145	201	219	176	130	142	114
R1: P2.2a.f+P3.1i	207	222	185	250	269	224	163	175	145
R2: P2.2a.f+P3.4i	210	226	187	254	274	227	165	179	147

R3: P2.3a.f+P3.1i	222	238	199	269	289	242	175	188	157
R4: P2.3a.f+P3.4i	225	241	204	273	292	247	178	191	161
R5: P2.3a.1i+P3.4i	245	263	221	297	318	268	193	207	174
R6: P2.3a.mi+P3.4i	250	266	227	302	322	275	197	210	179
R7: P3.3a.ad+P3.4i	271	290	244	328	351	296	214	229	192

Table A1. First step in the optimization of the program parameters

	Parameter Values				
Parameter	**eNPV at the Low value $M**	**Low**	**Current**	**High**	**eNPV at the High value $M**
P2 Go	**190**	**0.65**	0.80	0.90	111
P3 Beta	**164**	**0.10**	0.20	0.225	150
Min Exp HR	148	0.35	0.60	**0.80**	**168**
Max Factor	**154**	**0.65**	0.80	0.90	153
Min Factor	**153**	**0.25**	0.50	0.75	149

153

eNPV at the current parameter

values $M

Table A2. Second step in the optimization of the program parameters

	Parameter Values				
Parameter	**eNPV at the Low value $M**	**Low**	**Current**	**High**	**eNPV at the High value $M**
P2 Go	201	0.58	**0.65**	0.83	167
P3 Beta	203	0.05	0.10	**0.18**	**220**
Min Exp HR	202	0.45	**0.80**	0.90	76
Max Factor	**214**	**0.58**	0.65	0.83	213
Min Factor	**217**	**0.13**	0.25	0.63	169

eNPV at the current parameter values $M	213

Table A3. Seventh and final step in the optimization of the program parameters

	Parameter Values				
Parameter	**eNPV at the Low value $M**	**Low**	**Current**	**High**	**eNPV at the High value $M**
P2 Go	221	0.58	**0.65**	0.83	220
P3 Beta	222	0.05	**0.175**	0.18	222
Min Exp HR	222	0.45	**0.80**	0.90	221
Max Factor	222	0.58	**0.575**	0.83	222
Min Factor	222	0.13	**0.125**	0.63	222
			222		
	eNPV at the current parameter				

values $M

Table A4. Example Phase 3 Sample sizes in program R2 for different expected HRs

	Expected HR										
	0.8	**0.81**	**0.82**	**0.83**	**0.84**	**0.85**	**0.86**	**0.87**	**0.88**	**0.89**	**0.90**
P3 Sample Size	1096	1228	1382	1566	1788	2056	2386	2796	3316	3990	4878

Table A5. Table of operating characteristics of R2 for the different scenarios

Scenario HR	Pr P2 fail	Pr P3 fail	Pr P3 succ	P2 Cost	P3 (fail) Cost	P3 (succ) Cost	Rev	eNPV	Scen wt	Wt'd eNPV
1	0.88	0.11	0.00	7	17	19	411	-8	0.529	-4
0.9	0.68	0.22	0.09	7	19	19	663	46	0.27	13
0.8	0.38	0.11	0.51	7	20	18	1,125	549	0.127	70
0.7	0.13	0.01	0.87	7	18	16	1,887	1,614	0.054	87
0.6	0.02	0.00	0.98	7	12	14	2,936	2,862	0.02	57

Sum	222

Table A6. Optimized parameter values for different Phase 2 sample sizes for program R2

		Parameter Values				
Phase 2 sample size	**eNPV $M**	**P2 Go**	**P3 Beta**	**Min Exp HR**	**Max Factor**	**Min Factor**
80	187	0.65	0.24	0.80	0.58	0.125
100	195	0.70	0.18	0.80	0.57	0.125
120	209	0.67	0.21	0.80	0.54	0.125
140	216	0.65	0.21	0.80	0.54	0.125
160	215	0.70	0.18	0.80	0.57	0.125
180	219	0.67	0.18	0.80	0.57	0.180
200	222	0.65	0.18	0.80	0.58	0.125

240	216	0.63	0.22	0.80	0.53	0.125
280	203	0.60	0.22	0.80	0.54	0.125
320	188	0.58	**0.21**	0.80	0.54	0.125

Table A7. Optimized parameter values for programs R0-R4 at Phase 2 sample sizes that maximize eNPV

			Parameter Values				
Program	**P2 Sample Size**	**eNPV $M**	**P2 Go**	**P3 Beta**	**Min Exp HR**	**Max Factor**	**Min Factor**
R0: P2.2a.f+P3.0i	180	166	0.67	0.22	0.74	0.99	0.13
R1: P2.2a.f+P3.1i	200	206	0.65	0.23	0.78	0.65	0.40
R2: P2.2a.f+P3.4i	200	222	0.65	0.18	0.80	0.58	0.13
R3: P2.3a.f+P3.1i	210	210	0.75	0.23	0.80	0.61	0.38
R4: P2.3a.f+P3.4i	210	225	0.65	0.23	0.80	0.65	0.25

Figure 1. The simulated decision tree

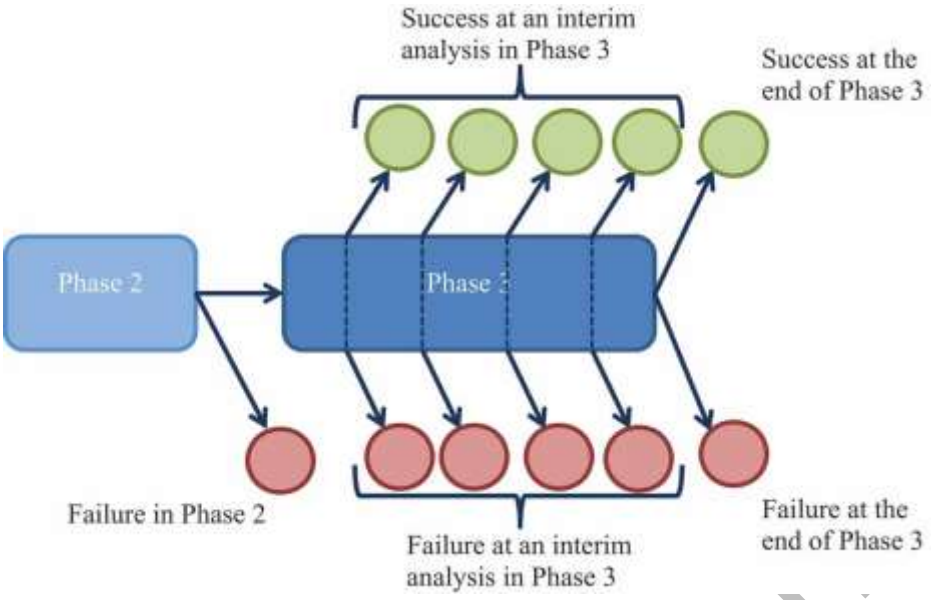

Figure 2. Setting the success thresholds

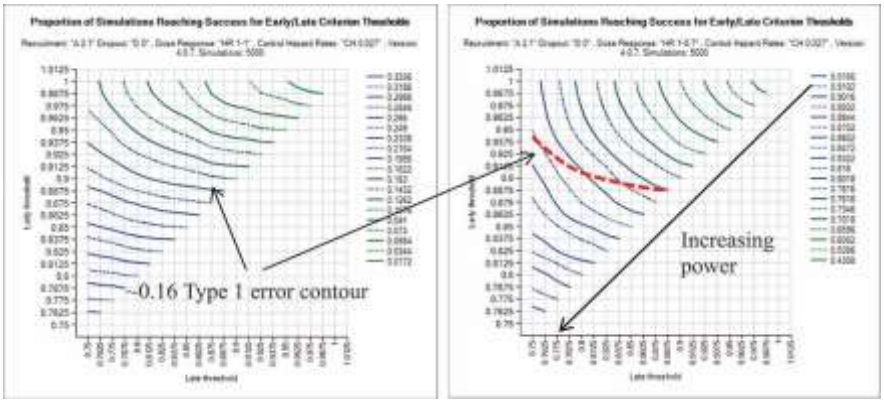

Figure 3. Setting the futility thresholds

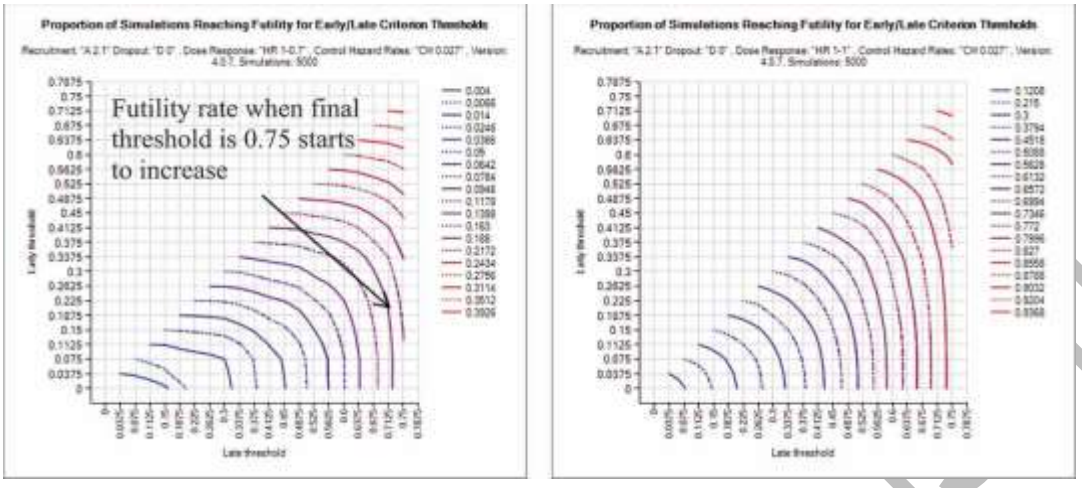

Figure 4. Expected Net Present Value for Programs R0-R4 at different Phase 2 sample sizes

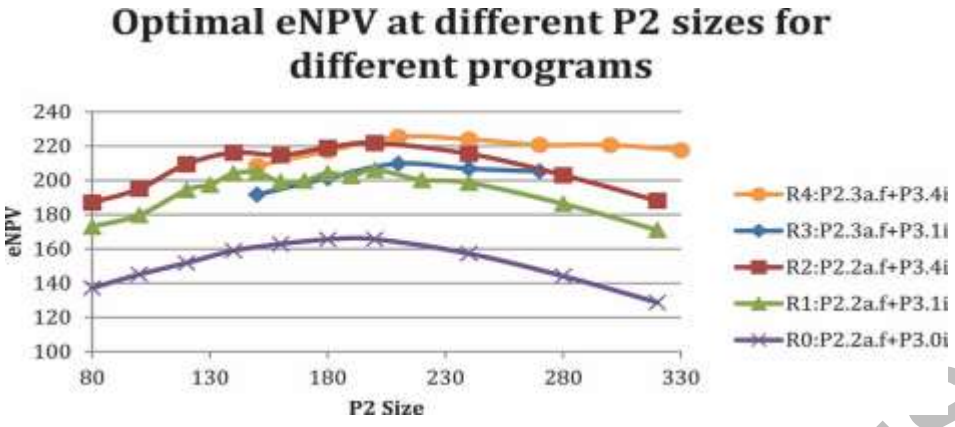

Figure A1. The Phase 3 stopping boundaries for the N=1096 case

Figure A2. The probability of stopping at the interim analyses for different true HRs