Research Article
Open Access
CrossMark

A Bayesian single-arm design using predictive probability monitoring

Volume 7 Issue 4 - 2018

Patrick D Mitchell
Early Clinical Development, US

Correspondence: Patrick D Mitchell, Early Clinical Biometrics, Early Clinical Development, IMED Biotech Unit, AstraZeneca, Boston, USA, Email patrick.mitchell@astrazeneca.com

Received: July 17, 2018 | **Published:** July 26, 2018

Abstract

Background: Following first in man dose finding oncology trials, a dose expansion is conducted to determine assess anti-tumor activity. Historically, dose expansions have been single arm trials which may include one interim analysis for the purpose of determining futility. These designs required a fixed number of patients be exposed to experimental treatments before an interim analysis was conducted. Alternatively, single arm designs with more frequent assessments of efficacy can be adopted. The proposed design uses posterior probabilities as the basis for decisions and can incorporate prior information in addition to predictive probability as a means of continuous monitoring.

Methods: Simulation methods are used to evaluate design operating characteristics including average sample size E(N). Trials with a maximum of N=50 patients are assessed, with an interim at N=30 and predictive power monitoring from a specified start patient until the interim. Two target success levels are evaluated. A series of true effects from distributions with varying levels of information (equivalent N=0 to 20) are assessed.

Results: The characteristics of the design show a decrease in the average sample size with lower True Effect means as well as higher probability of early stopping.

Conclusion: For the single arm designs evaluated, adding predictive power monitoring on average will require fewer patients while maintaining acceptable decision risks. With the introduction of predictive probability monitoring before an interim analysis starting at some point in the sample before an interim analysis there can be a measurable savings in the case of poor performing compounds.

Keywords: early development trial design, futility, decision criteria, probability, patients

Introduction

Typically oncology programs begin development in humans using dose escalation trials with the purpose of determining the maximum tolerable dose. This or some lower dose is then tested in a series of trials in order to determine the anti-tumor potential of the drug along with other objectives such as mechanism of action. Following this determination, trials are usually conducted to establish the comparative efficacy of the experimental drug with an established treatment or standard of care.

In early development of oncology compounds, single arm designs are commonly used as the initial means to assess anti-tumor activity. Over the years, several variations on the approach to single arm designs have been used. While these single arm designs can be relatively small compared to designs used in later phase development, there can be a relatively long time commitment required to complete them. As a means to address the time required, interim analyses have been incorporated in order to allow the possibility of efficacy related decisions to be made prior to observing responses in all patients initially planned. Several variations on single arm designs with interim assessments have been used which incorporate hypothesis testing strategies. Stephanie Green presents a comprehensive overview of a series single arm designs in Crowley & Hoering.[1] Methods based on error spending functions, conditional power, predictive power and parameter-free predictive power are suggested by Jennison & Turnbull.[2] These are more flexible than more traditional single arm designs.

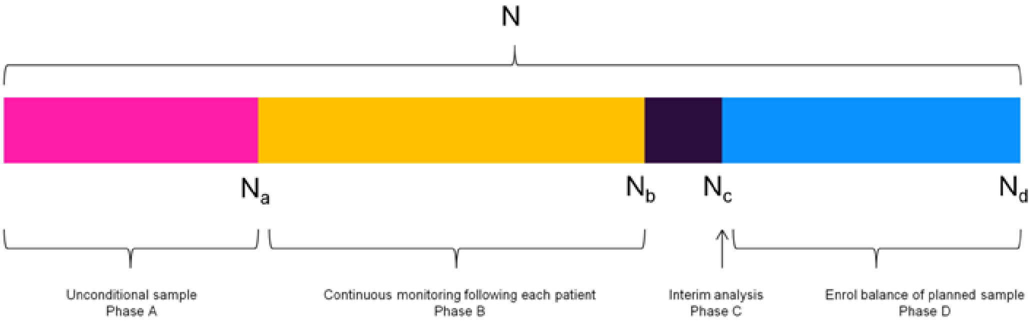
One facet common to many of these designs is that there is a fairly large fraction of the total planned sample that needs to be observed before the first opportunity to evaluate the performance of the compound. While this is still an improvement on the overall time commitment, the trial is designed to run for at least as long as it takes to observe all patients whose results will be included in the interim. In some cases this is balanced by having an early assessment but with relatively few patients to declare sufficient activity levels. With some designs, trialists may be faced with having to decide between either halting recruitment while the last patients are being assessed as part of an interim or allowing recruitment to continue on a possibly ineffective treatment while waiting for interim results. In contrast to the methods described above, Berry et al.,[3] offer a Bayesian approach to the single arm design. Thall[4] presents a single arm design incorporating Bayesian decision criteria with continuous monitoring of efficacy. Thall, Simon & Estey[5] propose a method to monitor single arm trials using Bayesian continuous monitoring of both an efficacy and a toxicity variable using an historically informed prior. Chen & Lee[6] propose a single arm design using simulated annealing (Figure 1).

Proposed design

The proposed design is illustrated in Figure 1 below.

This design is a single arm design in which patients either respond to an experimental treatment or not after being followed for a fixed amount of time. Patients will be enrolled in the trial in four sequential phases. The overall sample will be $N=N_a+N_b+N_c+N_d$ patients. The responses of the individual patients to treatment will be denoted by Y_i (i=1 … N). The chance that patient i responds is θ. The behavior of θ is described in the section below on "True Effect Distribution". N_a patients will be enrolled and evaluated unconditionally based on efficacy response. Once the efficacy response is observed in patient

N_a, trial monitoring using predictive probability including all N_a patients begins. If predictive probability that the proportion of patients responding is at least τ falls below some level φ then the trial will be permanently stopped for futility. If the trial is not stopped then the efficacy response for the first patient in Phase b (Patient N_a+1) will be observed, predictive probability will then be recalculated for all N_a+1 patients and the same criterion will be applied (if predictive probability > φ then continue to the next patient. Otherwise, the trial will be stopped.). The observation of efficacy response in each patient enrolled during Phase b followed by the calculation of predictive probability will continue for all patients in this phase for as long as predictive probability remains above φ. If the trial continues through all patients planned in Phase b, then following the observation of the response of the next patient (Phase c), an interim analysis with both futility and administrative facets including all patients up through Phase c will be conducted using methods described in Frewer.[9] This interim will include two criteria: a 'stop' criterion (S) where the trial will stop if the proportion of patients who respond to treatment is equal to or below this level and a 'go' (G) criterion where other actions associated with more robust efficacy could be taken. If the trial does not stop, as a result of the interim analysis, then the balance of the planned N_d patients, will be enrolled and a final analysis including all N patients will be conducted. Temporary stopping is not assumed during the trial. This analysis will be similar to the interim in that the proportion of patients who respond will be evaluated using two criteria ('stop' and 'go').



Design assumptions

There are several values that are required in order to complete the design:

N: Total sample

N_a: Number of patients to be enrolled in the first phase of enrolment. These form the initial sample needed to estimate the reponse to treatment. A small number can be included here but overly small numbers in this phase could lead to a large proportion of trials inappropriately stopped before the interim analysis.

N_b: The patients to be included in the predictive probability monitoring procedure on a patient by patient basis.

N_c: The number of patients following the end of phase b that will be included in the interim analysis.

N_d: Patients required to complete the planned sample (depends on the previous three values).

τ: Threshold for overall trial success based on all N patients in terms of the primary response variable. This can be set to coincide with a specific decision target at the end of the trial or be adjusted to yield acceptable operating characteristics.

φ: Futility probability: Predictive probability level below which the chance that the trial will yield a proportion of patients responding is small enough to stop for futility.

S and G: 'Stop' and 'go' criteria used to evaluate the proportion of patients responding at set points during the trial. These criteria are a means to map the results of this trial to specific actions that can be taken either in the trial itself or elsewhere in the clnical program.

Prior distribution

Throughout this paper, non-informative priors will be assumed. It is possible to generalize this method to include more informative priors. While the selection of a prior other than one that is non-informative is beyond the scope of this paper, if there is enough information in patients with the disease under study available that information can be used to construct an empirical prior for the true response proportion.

True Effect Distribution

The number of responses in the sample will follow a Binomial distribution with true response parameter θ. A natural choice for the True Effect distribution of θ is Beta with parameters α+1 and β+1. These can be selected as functions of the mean proportion of responders and an equivalent sample size Morita[10] as follows:

$\alpha=N_e Pr$

$\beta=N_e - N_e Pr$

Where N_e is the strength of the True Effect distribution in terms of a sample size equivalent which will allow for rational values (ranging from 0, equivalent to a Uniform (0,1) to large values which yield distributions where almost all of the density is within a small neighborhood of the mean). Pr is the mean of the True Effect distribution. When evaluating the opearting characteristics of the design, N_e should be no more than 10% of the overall sample size. A range of Pr values (selected to cover the parameter space) are used to assess design performance.

Predictive probability based monitoring

Predictive power is presented in Jennison & Turnbull[2] and is similar to the expression shown below. The important difference is that predictive power implies that the target is a function of a critical

Citation: Mitchell PD. A Bayesian single-arm design using predictive probability monitoring. *Biom Biostat Int J.* 2018;7(4):299‒309.
DOI: 10.15406/bbij.2018.07.00222

value. Predictive probability allows for the selection of a target other than the critical value. The specifics of the application to non-comparative trials with binomial endpoints are described here.

For a single arm trial with N planned patients and an interim analysis to occur following N_i patients ($N_i < N$), where n_i of the N_i patients included in the interim respond and we need to observe n_s ($n_s > n_i$) patients in order to declare the trial a success then the predictive probability or the chance to observe at least n_s patients out of N patients (at the end of the trial) is:

$$\sum_{k=0}^{N_i}\left[P\left(X_1=k\mid n_i/N_i\right)\left(1-P\left(X_2\le\left(\left(n_s-n_i\right)-1\right)\mid\mu_k\right)\right)\right]$$

$X_1 \sim \text{Binomial}(N_i, n_i/N_i)$

X_1 ranges from 0 to N_i and is the possible number of successes observed at an interim with N_i total observations.

$X_2 \sim \text{Binomial}(N–N_i, \mu_k)$

X_2 is the number of succeses that remain following the interim necessary to achieve the target at the end of the trial.

The left hand factor in the expression being summed above is a vector of probabilities that k responses are observed assuming that the observed response proportion is actually true. The right hand factor is the probability that at least the balance of the required responses needed is observed in the remaining patients to be observed in the planned sample.

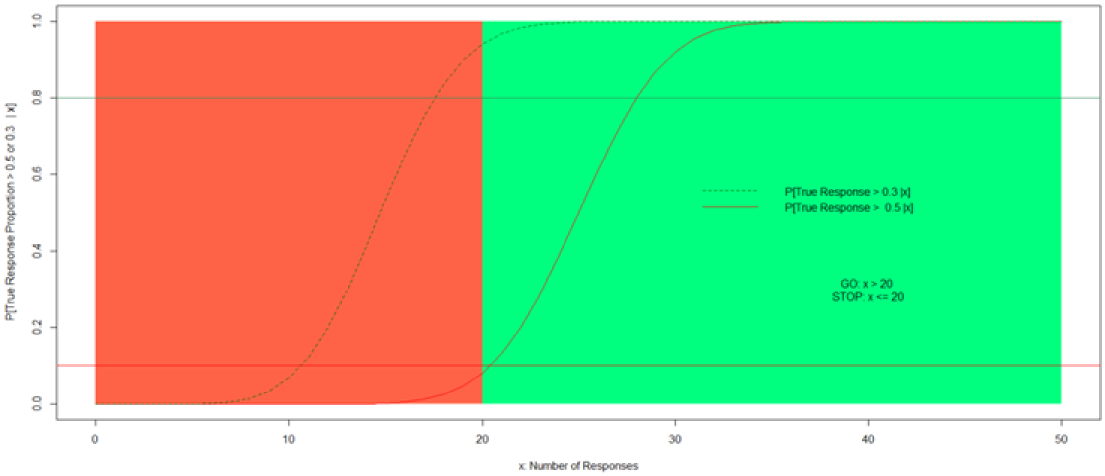
μ_k is typically equal to k/N_i. Exceptions are made for the values k=0 and k=N_i. For these values, $0.5/N_i$ and $(N_i–0.5)/N_i$ are used instead in order to avoid having to use actual 0 and 1 values for μ_k. This is done in order to include good coverage of the parameter space while not completely ignoring the contribution to predictive probability of observing 0 responses in the first n_i patients. The difference between this and using the unadjusted μ_k values is very small and can be used with minimal difference in calculating predictive probability.

The functional form of predictive probability described here closely resembles assurance as described by O'Hagan.[7] Assurance has been presented as a means to assess probability of success in larger comparative studies when there is information from smaller, earlier comparative studies available. Here, the philosophy of learning from accumulating data is the same. The important difference is that predictive probability is being assessed on an emerging sample of an ongoing trial so that the correlation between the interim results and the final is maintained.

Decision criteria

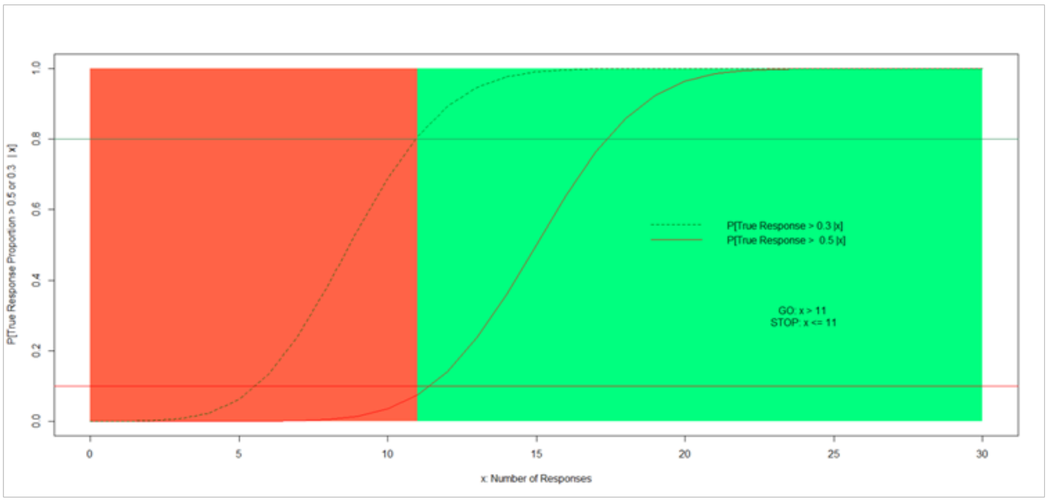
The decision criteria used in this design are not set directly but are a function of two values selected by the project and are a reflection of the performance necessary to support decisions to go forward in development or to stop development. These are the target value (TV) and lower reference value (LRV) as described by Lalonde.[8] In this design, a go decision will be made if the posterior probability that the true response proportion is at least LRV is at least DC_LRV. If the posterior probability that the true response proportion is at least the TV is no more than AR_TV then a STOP decision is made. In cases where both conditions are true, a STOP decision is made. Here the DC_LRV is the desired confidence that the true response proportion is at least as high as the LRV given a GO decision. The AR_TV is the acceptable risk that the true response proportion is at least as high as the TV given a decision to STOP.

For the final analysis (N=50) the posterior probability that the true response proportion is at least 0.3 is at least 0.8 when there are >= 18 patients of 50 responding. The posterior probability that the true response proportion is at least 0.5 is no more than 0.1 when there are no more than 20 patients of 50 responding. Since the go criterion would be lower than the stop criterion, by convention, the sample space is partitioned into GO and STOP regions according to the stop criterion (Figure 2).



Similar criteria are set for the interim analysis following N=30 patients are illustrated below: (Figure 3).

Citation: Mitchell PD. A Bayesian single-arm design using predictive probability monitoring. *Biom Biostat Int J.* 2018;7(4):299–309. DOI: 10.15406/bbij.2018.07.00222



Methods

Operating characteristics of overall design

The operating characteristics of this design are assessed using simulation methods. 10000 simulated trials are generated with each set of values. All simulated trials will include observations beyond those that arrive at STOP decisions up through the planned total sample size. The random seed was set to 125. All simulations were conducted using R v3.1.2.

Example trial designs

For this example, trials will have a total sample size (N) of 50 patients. The numbers of patients enrolled in each of the 4 phases are as follows: N_a=12, N_b=17, N_c=1 and N_d=20. The target criteria TV and LRV are 0.5 and 0.3 respectively. The AR_TV (false stop) and DC_LRV (1-false go) are 0.1 and 0.8 respectively. τ is set to 0.52. φ is set to 0.05. A second example is presented with τ=0.4.

A series of True Effect distributions will be evaluated including means (Pr) of 0.2, 0.4, 0.6 and 0.8 and equivalent sample sizes (N_e) of 0, 0.5, 1, 5, 10 and 20. In each simulated trial, the true proportion of patients responding will be sampled from the True Effect distribution based on the combination of true proportion and equivalent sample size values being evaluated.

Average sample size, fraction of trials that result in STOP and GO decisions overall, as well as the fraction of trials stopping early that would have otherwise led to GO decisions at the final analysis and the fraction of trials with GO decisions at the interim analysis which ultimately result in STOP decisions at the final analysis will be presented.

Operating characteristics of predictive probability monitoring

In a separate simulation study, the operating characteristics of predictive probability monitoring are examined. 10000 iterations will also be generated for each combination of True Effect Mean and starting sample size for predictive probability monitoring. True Effect distributions with means ranging from 0.05 through 0.8 with equivalent sample sizes of 1 and 20 will be used. The values for τ and φ that were used to assess the overall trial design will also be used here. Predictive probability monitoring starting as few as 4 patients up through 20 patients will be examined. Consistent with the overall design, predictive probability monitoring will end following the 29th patient assuming that the trial is not stopped before observing that patient's response. If predictive probability remains above 0.05 in the simulated trial, then the 30th patient will automatically be included in that trial.

All items will be presented as a function of True Effect mean and sample size equivalent. The average sample size at which predictive probability falls below φ will be presented. In cases where predictive probability does not fall below φ before the 29th patient, 30 patients will be assumed. The proportion of simulated trials where a "reversal" is observed will also be presented. A reversal occurs when predictive probability has fallen below φ only to rise above φ at some point later in the simulated trial. The proportion of simulated trials for which a reversal is observed and the observed proportion of patients responding based on the total sample is at least τ will also be presented. This will be called "Meeting Target following a reversal" from this point forward.

Results

This section will be divided into two subsections. The first will address the operating characteristics of the whole trial design. The second will focus on the behavior of predictive probability alone.

Design operating characteristics

As mentioned above, simulation methods are used to evaluate the various costs (in terms of patients to be enrolled) and losses (in terms of the chance to arrive at incorrect conclusions). Table 1 & Table 2 below present the average number of patients and the fraction of simulated trials that arrive at GO and STOP decisions. The average sample size is the average sample size that leads to a STOP decision across all simulated trials under a given collection of true values and design parameters.

In both cases, the strength of the True Effect distribution in terms of sample size equivalent has an impact on the average sample size over the range of True Effect mean values. When the True Effect distribution is non-informative (N_e=0), the average sample size is insensitive to the value of the True Effect mean. While the N_e=0 case is not particularly informative of operating characteristics, it is included for completeness to more fully explore the behavior across a spectrum of N_e values. This is expected, since theta will be drawn from a Uniform (0,1) distribution regardless of the mean. For the other sample size equivalents > 0 there is an increase in the average sample size with increasing True Effect mean. This is expected, since trials are less likely to stop early if the treatment effect is good and less variable. The average sample size approaches the maximum trial size (N=50) with higher True Effect mean values. Average sample size with higher True Effect mean values is closer to the maximum sample size with higher sample size equivalents, again as would be expected. Average sample size also approaches the minimum number of patients to be enrolled for the lower True Effect mean values also getting closer to the minimum sample size with higher sample size equivalents. The proportion of simulated trials reaching a GO decision when using a Uniform True Effect distribution is close to 55% for all True Effect mean values where τ=0.52 and close to 58% where τ=0.4. When considering non-Uniform True Effect distributions, the fraction of trials reaching a GO decision increases with increasing True Effect mean. The rate at which this fraction increases also increases with increasing sample size equivalent. This mirrors the pattern of the average sample size, as expected.

Table 1 Average sample size with success target tau= 0.52 and PP monitoring start time N_a=12

Ne	True effect mean	E(N)	% Trials deciding	
			GO	STOP
0	0.8	34.1	54.6	45.4
	0.6	33.9	53.9	46.1
	0.4	34.3	55.3	44.7
	0.2	34.1	54.6	45.4
0.5	0.8	37.5	63.6	36.4
	0.6	35.5	58.1	41.9
	0.4	33.3	51.8	48.2
	0.2	30.8	45.3	54.7
1	0.8	40.1	70.5	29.5
	0.6	36.6	60.5	39.5
	0.4	32.8	49.8	50.2
	0.2	29.1	40.1	59.9
5	0.8	47.2	91.2	8.8
	0.6	40.5	70.6	29.4
	0.4	31.2	43.8	56.2
	0.2	21.1	17.7	82.3
10	0.8	49	96.7	3.3
	0.6	43.2	78.2	21.8
	0.4	30.4	40.5	59.5
	0.2	17.8	9	91
20	0.8	49.7	99	1
	0.6	44.8	83.1	16.9
	0.4	29.7	36.6	63.4
	0.2	15.4	3.4	96.6

Table 2 Average sample size with success target tau= 0.4 and PP monitoring start time N_a=12

Ne	True Effect Mean	E(N)	% Trials deciding	
			GO	STOP
0	0.8	36.9	57.8	42.2
	0.6	36.6	57.1	42.9
	0.4	37	58.1	41.9
	0.2	36.8	57.3	42.7
0.5	0.8	40.2	66.7	33.3
	0.6	38.5	61.3	38.7
	0.4	36.5	55.3	44.7
	0.2	34	48.6	51.4
1	0.8	42.7	73.6	26.4
	0.6	39.6	64.2	35.8
	0.4	36.3	53.4	46.6
	0.2	32.6	43.5	56.5
5	0.8	48.5	93.7	6.3
	0.6	44.1	75.6	24.3
	0.4	36.3	49.5	50.5
	0.2	25.5	20.8	79.2
10	0.8	49.6	98.2	1.8
	0.6	46.5	83.7	16.3
	0.4	36.4	47.2	52.8
	0.2	22.3	11.4	88.6
20	0.8	49.9	99.7	0.3
	0.6	47.8	89.1	10.8
	0.4	36.7	44.3	55.7
	0.2	19.5	4.6	95.4

Table 3 & Table 4 below, present the chances to stop trials at various points in the trial. For each True Effect mean and sample size equivalent, the NOIA column presents the proportion of trials arriving at a STOP decision in a single arm trial with N=50 patients with no interim analysis. The IA+PP column presents the proportion of trials arriving at a STOP decision either before the interim analysis (Pre IA), at the interim analysis or before (IA or before), or at any time (Final or before). Stopping at the final analysis here is actually observing results that pessimistic in nature where stopping the development of the compound would likely be considered.

With a Uniform True distribution (N_e=0) the chance of stopping the trial before reaching the interim analysis is about 0.44 for τ=0.52. There is no increase in the proportion of trials being stopped during the predictive probability monitoring phase or following the interim analysis for τ=0.52. This suggests that the IA for early futility stopping is not required in this case, since if you continue after patient 29 you will always continue after patient 30, i.e. the number of responses to continue after patient 29 with the predictive probability rule is greater than or equal to the number of responses to continue after patient 30 with the established decision criteria. When stopping at the final analysis is included, very few additional trials are stopped compared to the stopped at the interim or before. In all cases there is some increase

DOI: 10.15406/bbij.2018.07.00222

in the proportion of trials reaching a stop decision including interim analyses and predictive probability monitoring compared to the corresponding single arm design not including interim assessments.

A large fraction of the trials reaching a STOP decision appear to reach this decision before enrolling the number of patients needed for the interim analysis.

Table 3 Probability of stopping at various stages in simulated trials

Timing of STOP decision	True Effect Mean	Final N = 50, Begin PP at N= 12 interim at N= 30, predictive probability target 0.52											
		Ne = 0		Ne = 0.5		Ne = 1		Ne = 5		Ne = 10		Ne = 20	
		No IA	IA + PP	No IA	IA + PP	No IA	IA + PP	No IA	IA + PP	No IA	IA + PP	No IA	IA + PP
Final or before	0.8	0.409	0.454	0.318	0.364	0.249	0.295	0.053	0.088	0.012	0.033	0.002	0.01
	0.6	0.417	0.461	0.374	0.419	0.342	0.395	0.218	0.294	0.138	0.218	0.081	0.169
	0.4	0.403	0.447	0.43	0.482	0.449	0.502	0.477	0.562	0.492	0.595	0.512	0.634
	0.2	0.412	0.454	0.499	0.547	0.546	0.599	0.771	0.823	0.87	0.91	0.946	0.966
IA or before	0.8	0	0.44	0	0.349	0	0.28	0	0.082	0	0.031	0	0.01
	0.6	0	0.444	0	0.403	0	0.377	0	0.276	0	0.203	0	0.157
	0.4	0	0.434	0	0.465	0	0.48	0	0.533	0	0.56	0	0.588
	0.2	0	0.438	0	0.53	0	0.578	0	0.795	0	0.881	0	0.943
Pre IA	0.8	0	0.44	0	0.349	0	0.28	0	0.082	0	0.031	0	0.01
	0.6	0	0.444	0	0.403	0	0.377	0	0.276	0	0.203	0	0.157
	0.4	0	0.434	0	0.465	0	0.48	0	0.533	0	0.56	0	0.588
	0.2	0	0.438	0	0.53	0	0.578	0	0.795	0	0.881	0	0.943

Table 4 Probability of stopping at various stages in simulated trials predictive probability target 0.4

Timing of STOP decision	True Effect Mean	Final N = 50, Begin PP at N= 12 interim at N= 30 predictive probability target 0.4											
		Ne = 0		Ne = 0.5		Ne = 1		Ne = 5		Ne = 10		Ne = 20	
		No IA	IA + PP	No IA	IA + PP	No IA	IA + PP	No IA	IA + PP	No IA	IA + PP	No IA	IA + PP
Final or before	0.8	0.409	0.422	0.318	0.333	0.249	0.264	0.053	0.063	0.012	0.019	0.002	0.003
	0.6	0.416	0.429	0.374	0.387	0.342	0.358	0.218	0.244	0.138	0.163	0.081	0.108
	0.4	0.403	0.419	0.43	0.447	0.449	0.466	0.476	0.505	0.492	0.528	0.512	0.557
	0.2	0.412	0.427	0.499	0.514	0.546	0.565	0.771	0.792	0.87	0.886	0.946	0.954
IA or before	0.8	0	0.389	0	0.3	0	0.231	0	0.051	0	0.015	0	0.002
	0.6	0	0.395	0	0.349	0	0.32	0	0.2	0	0.128	0	0.081
	0.4	0	0.386	0	0.408	0	0.419	0	0.44	0	0.452	0	0.455
	0.2	0	0.393	0	0.476	0	0.518	0	0.732	0	0.825	0	0.905
Pre IA	0.8	0	0.334	0	0.246	0	0.179	0	0.035	0	0.008	0	0.002
	0.6	0	0.342	0	0.29	0	0.259	0	0.136	0	0.074	0	0.043
	0.4	0	0.33	0	0.34	0	0.346	0	0.335	0	0.322	0	0.305
	0.2	0	0.335	0	0.405	0	0.444	0	0.628	0	0.718	0	0.802

Trials where this threshold is set a bit lower τ=0.4 are a contrasting example. With a Uniform True distribution the chance to stop the trial before is about 43%. In these trials, there is some additional stopping observed attributed to the interim analysis. Consistent with the previous case, most of the early stopping happens prior to the IA.

Compared to designs that do not include interim analyses the increase in the probability of stopping in designs that include interim analyses ranges from 0.001-0.122 accross all cases examined. The increased risk of stopping with the more complex designs vs the designs without interims is ~0.013 in cases with good performance (True Effect Mean=0.8) and ~0.029 with poor performance (True Effect Mean=0.2). There is not a large difference in stopping risk between the two designs.

Additional losses can be observed between the interim and final analyses. Here the interim analysis will include stopping triggered during the predictive probability monitoring phase of the trial. Table 5 & Table 6 present the proportion of trials reaching STOP and GO decisions at the final analysis, where no interim analysis is conducted,

Citation: Mitchell PD. A Bayesian single-arm design using predictive probability monitoring. *Biom Biostat Int J.* 2018;7(4):299–309.
DOI: 10.15406/bbij.2018.07.00222

the proportion of trials where consistent decisions were made at both the interim and final, and the proportion of trials where inconsistent decisions were made between the interim and final.

In an average of ~93% of simulated trials using τ=0.52, stop or go decisions made at interim are consistent with what would be made if those trials are continued through to the final analyses. Of the trials which result in inconsistent decisions, the majority of those are due to stopping early followed by a positive result at the final analysis. Trials with inconsistent decisions between the interim and final analyses are observed more frequently with higher sample size equivalents (N_e=5, 10 and 20) where the True Effect mean was 0.4 and 0.6. This is expected, since the true effects are close to either the threshold of 0.52 or the interim and final go boundaries of 12/30 (0.4) and 21/50 (0.42).

Table 5 Losses Occurring between interim and final analyses predictive probability target 0.52 N_a=12

Ne	True Distribution Mean	No IA		Consistent decision		Inconsistent decision	
		STOP at final	GO at final	STOP at both interim and final	GO at both interim and final	STOP at interim, GO at final	GO at interim, STOP at final
0	0.8	0.409	0.591	0.394	0.546	0.045	0.014
	0.6	0.417	0.584	0.4	0.539	0.044	0.016
	0.4	0.403	0.597	0.39	0.553	0.044	0.013
	0.2	0.412	0.588	0.395	0.546	0.042	0.016
0.5	0.8	0.318	0.682	0.302	0.636	0.046	0.015
	0.6	0.374	0.626	0.357	0.581	0.045	0.017
	0.4	0.43	0.57	0.413	0.518	0.052	0.017
	0.2	0.499	0.501	0.481	0.453	0.049	0.017
1	0.8	0.249	0.751	0.234	0.705	0.046	0.015
	0.6	0.342	0.658	0.324	0.605	0.053	0.018
	0.4	0.449	0.551	0.428	0.498	0.053	0.022
	0.2	0.546	0.454	0.524	0.401	0.054	0.022
5	0.8	0.053	0.947	0.048	0.912	0.035	0.005
	0.6	0.218	0.782	0.2	0.706	0.076	0.018
	0.4	0.477	0.524	0.448	0.438	0.086	0.029
	0.2	0.771	0.229	0.743	0.177	0.052	0.028
10	0.8	0.012	0.988	0.011	0.967	0.021	0.002
	0.6	0.138	0.862	0.124	0.782	0.079	0.015
	0.4	0.492	0.508	0.457	0.405	0.103	0.035
	0.2	0.87	0.13	0.841	0.09	0.04	0.029
20	0.8	0.002	0.998	0.002	0.99	0.008	0
	0.6	0.081	0.919	0.069	0.831	0.088	0.012
	0.4	0.512	0.488	0.466	0.366	0.122	0.046
	0.2	0.946	0.054	0.923	0.034	0.02	0.022

Table 6 Losses Occurring between interim and final analyses predictive probability target 0.4 N_a=12

Ne	True Distribution Mean	No IA		Consistent decision		Inconsistent decision	
		STOP at final	GO at final	STOP at both interim and final	GO at both interim and final	GO at interim, STOP at final	STOP at interim, GO at final
0	0.8	0.409	0.591	0.328	0.578	0.013	0.033
	0.6	0.416	0.583	0.336	0.571	0.013	0.034
	0.4	0.403	0.597	0.323	0.581	0.016	0.032
	0.2	0.412	0.588	0.329	0.573	0.015	0.034
0.5	0.8	0.318	0.682	0.24	0.667	0.016	0.033
	0.6	0.374	0.626	0.285	0.613	0.013	0.038
	0.4	0.43	0.57	0.333	0.553	0.017	0.039
	0.2	0.499	0.501	0.399	0.486	0.015	0.038
1	0.8	0.249	0.751	0.174	0.736	0.015	0.033
	0.6	0.342	0.658	0.253	0.642	0.016	0.038
	0.4	0.449	0.551	0.339	0.534	0.017	0.047
	0.2	0.546	0.454	0.436	0.435	0.019	0.047
5	0.8	0.053	0.947	0.029	0.937	0.01	0.013
	0.6	0.218	0.782	0.124	0.756	0.026	0.044
	0.4	0.476	0.524	0.323	0.495	0.028	0.065
	0.2	0.771	0.229	0.619	0.208	0.022	0.06
10	0.8	0.012	0.988	0.004	0.982	0.006	0.004
	0.6	0.138	0.862	0.064	0.837	0.025	0.036
	0.4	0.492	0.508	0.308	0.472	0.035	0.076
	0.2	0.87	0.13	0.712	0.114	0.017	0.061
20	0.8	0.002	0.998	0.001	0.997	0.001	0
	0.6	0.081	0.919	0.03	0.892	0.027	0.028
	0.4	0.512	0.488	0.284	0.443	0.045	0.102
	0.2	0.946	0.054	0.799	0.046	0.009	0.05

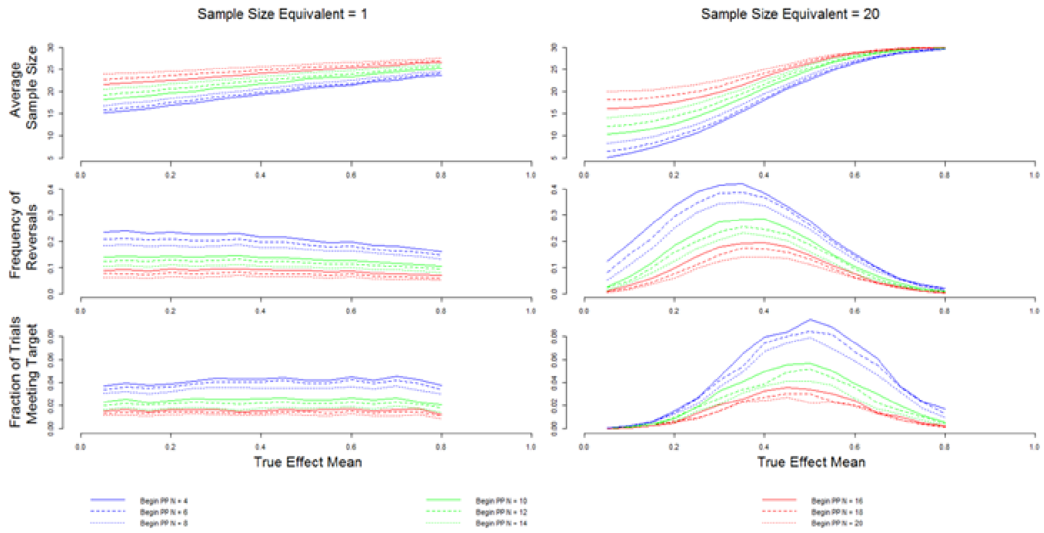
However, when τ=0.4, in about about 88% of simulated trials stop or go decisions made at the final analysis were consistent with those made at the interim. The type of inconsistency observed was STOP at interim followed by a GO at final in more simulated trials than the reverse. Higher frequencies remain with higher sample size equivalents at True Effect means of 0.4 and 0.6. However, these occur with lower frequency than in similar cases where τ=0.52. This likely reflects the closer alliance between the threshold of 0.4 and the interim and final go boundaries of 0.4 and 0.42.

Predictive probability operating characteristics

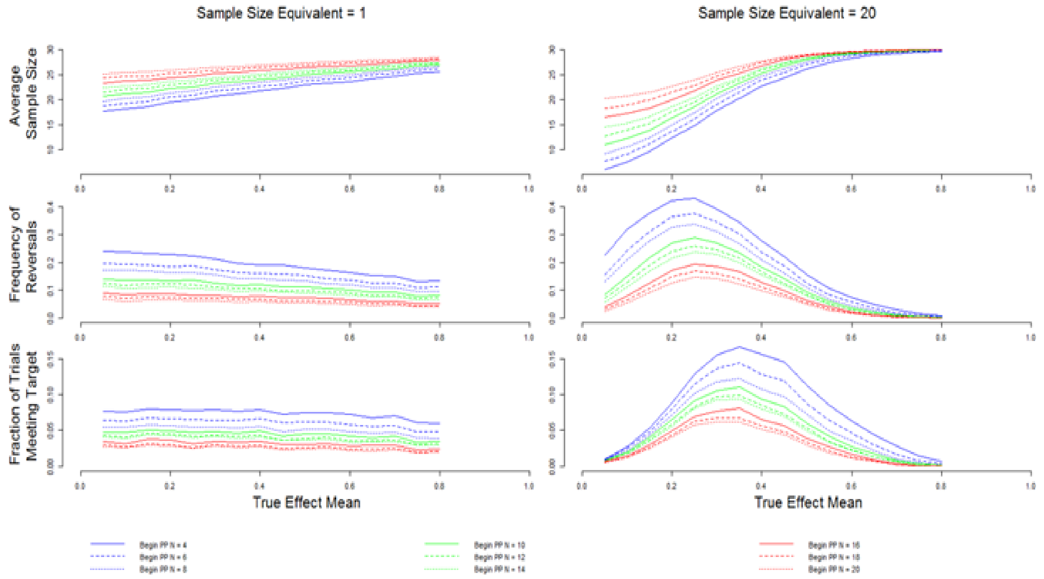
Figure 4 presents the average sample size (Max N=30), the fraction of simulated trials in which a reversal is observed, and the fraction of trials meeting target following a reversal (Figure 4).

There is a clear difference in all measures between the sample size equivalents N_e=1 and N_e=20 cases. In the N_e=20 case there are clear areas where the average sample size rises quickly to approach the maximum sample size as the True Effect Mean increases. The increase in average sample size does increase in the N_e=1 case but the increase appears to be close to linear. For both the frequency of reversals and the fraction of trials meeting target following a reversal, with N_e=20 there is a clear maximum frequency over the range of tested Mean True Effect values. For N_e=1 in these measures, the frequency of trials is only slightly affected by the choice of the True Effect Mean. These findings are most likely due to the high variability in the True Effect distribution associated with N_e=1.

Turning to the effect of the timing of the start of predictive probability monitoring, the average sample size behaves in a relatively predictable way with a steady increase with increasing True Effect mean in all cases with average sample sizes remaining close to the sample size at which predictive probability monitoring begins in the case of lower values of the True Effect mean. The highest proportion of trials with at least one reversal occurs with True Effect Means ranging from about 0.35–0.4 in most cases for N_e=20. The maximum proportion of trials with at least one reversal ranges from ~13% to ~43%. The Maximum fraction of trials meeting target following a reversal occurs with True Effect Means in a neighborhood of 0.5. The maximum fraction of trials meeting target following a reversal ranges from 0.03–0.10. It is clear that this frequency declines with increasing numbers of patients observed before beginning predictive probability monitoring. There is also a diminishing return on this observation with the difference between the Begin PP N=4 and Begin PP N=6 larger than the corresponding difference between Begin PP N=18 and Begin PP N=20. A similar observation regarding diminishing returns can be made for the proportion of trials for which a reversal was observed in monitoring phase also meeting target following 29 patients.



The results of similar simulations generated for the case where the target response proportion τ=0.4 are presented in Figure 5.



The differences between these results and those presented in the previous figure are more noticeable in the trials simulated with N_e=20. As with τ=0.52, the average sample size approaches the maximum sample size with higher True Effect Means. In the case of τ=0.4 this limit is reached more with lower True Effect Means. The maximum frequency of reversals is observed with True Effect Means in the neighborhood of 0.25. The maximum fraction of trials meeting target following a reversal was observed with True Effect means in a neighborhood of 0.35. Predictably, a higher target response proportion leads to fewer trials with reversals that also reach the end of the predictive probability monitoring phase.

Discussion

The main benefit of this design is that there are additional opportunities to stop a trial where the compound under study is not meeting the performance targets necessary to continue development. With higher True Effect means, stopping during the predictive probability phase leading up to an interim will happen less often. Earlier potential to stop is balanced by the increase in the chance to stop early with encouraging true treatment effects. This potential loss is present with any design that includes interim analyses. While there is no adjustment to account for this, if the increase in the risk of early stopping with encouraging treatment effects is acceptable, then this design will require fewer patients on average than similarly designed trials without the predictive probability monitoring. In this case an increase of on average <3% risk of early stopping can be balanced by the savings of ~14 patients (on average over all cases considered) compared to a design that does not include any interim analysis. The cost of monitoring the trial using predictive probability would be slightly more than only analyzing the data at a fixed time in the trial. Since a single arm trial would by nature not be blinded, response data can be transmitted in an ongoing basis. Since the sample space for the number of responses out of the number of total patients is finite for fixed N_i (and start and stop values for the predictive probability monitoring phase), the calculation of predictive probability can be done before the trial begins. The resulting time it would take to calculate predictive probability would not increase analysis time following each patient. So, in terms of time and effort, there would be no cost above what is done in the usual single arm design. That said, waiting until a fixed point in a single arm trial after a large portion of patients have been enrolled will likely result in higher costs especially in cases where there is poor performance of the compound. Trial size that is driven mainly by the observed proportion of responders can also allow drug development resources to be directed to or away from programs when using this design.

Inconsistency phenomenon True Effect Means near TV/LRV

In many cases regardless of the strength of the True Effect used to evaluate the operating characteristics, decisions made at or before interim will be consistent with those that are made at the final analysis. There is however an increase in the chance to observe a decision that is inconsistent at or before the interim and final when the True Effect mean is near the TV/LRV. This chance appears to increase with the sample size equivalent used (See Table 5 & Table 6 where the True Effect mean is either 0.4 or 0.6. Recall that the TV and LRV are 0.5 and 0.3 in the example). This would seem to suggest a risk in using this design when there is reason to select higher sample size equivalents when the True Effect mean is in the neighborhood of the targets set for the development plan.

Guidance on setting tau

There are two considerations regarding the choice of tau. The first is that while tau can relate to the choice of particular decision criteria at the end of the trial, in some cases this can lead to unacceptable operating characteristics. In particular, setting tau too high can result in an unacceptably high probability of premature stopping. Setting tau higher than the target of interest is not recommended and will lead to discarding what would otherwise be good candidates for further development.

Effect of setting tau

In the example τ is set according to the observed proportion of responders required out of the full planned sample that will lead to a non-stop decision at that point. When comparing Table 3 (τ=0.52) and 4 (τ=0.4) the losses Pre IA and IA or before are the same when τ=0.52. This suggests that there may be a point in any similar design where for τ of at least some value, that all of the futility related decision making could be made based on the predictive probability monitoring. An administrative look could then be incorporated if there is a desire to include a program decision that allows acceleration of planning or other activity within the program but outside the trial. However, setting τ too high can lead to an increase of inconsistent early stopping. So, while τ=0.52 was chosen as the non-stop criterion at the final analysis, it is not automatically set in this way. As with the values of any of the parameters, the operating characteristics of the corresponding design should be carefully understood before trial conduct.

TV/LRV and the lack of amber results

With the decision framework that is used in these examples there are only two decision categories (GO/STOP). Results falling into one of these two categories suggest a clear basis for either continuing development or stopping development respectively. With designs that include fewer patients or where the TV and LRV are set closer together, there can be a third decision category between the GO and STOP categories (CONSIDER). Results falling into this category are not by themselves enough to decide. This third category has typically meant that additional information from the trial should be considered in order to support a clear STOP or GO decision.[8,9] Other uses of this category have been considered.

The effect of the timing of beginning predictive probability monitoring

Selecting the beginning of the predictive probability monitoring phase of the design does have an impact on some of its operating characteristics. The effect on average sample size is obvious and more pronounced in trials which terminate early since a certain number of patients are observed before the beginning of the predictive probability monitoring phase. Beyond this, average sample sizes rise to the maximum planned sample size with increasing True Effect mean. The chance of observing a reversal increases with fewer patients observed before the predictive probability monitoring phase begins. However requiring as few as 4 patients before beginning the predictive power monitoring phase can result in a relatively large number of reversals ~41% - 43% in the examples presented, the chance that trials with reversals also meet target is between 10% and 16%. Further in this design there another decision point at the end of the trial that would have to be met before a compound would proceed further controlling the probability of an ineffective compound proceeding in development with borderline interim results. Though it is not something that was explored explicitly here, simulations are the best way to examine the effect of the start of the predictive probability phase of the trial on the overall risk of stopping the trial at various phases (looking specifically at the difference between the chance to stop without interim analysis and the corresponding chance to stop before IA, at IA or before or at any time in the trial).

Practical implications

In order for this method to be successfully implemented, strict adherence to the design as planned is necessary. Changes to the timing of the start of any of the phases within the design are possible but would require simulations to fully understand the changes in the risks from the base case. This is necessary in order to avoid missing reversals in the response proportion. A reversal is suggestive of a STOP decision and should be taken seriously since the true mean proportion of patients responding is likely much lower than may otherwise be suggested when ignoring the reversal. This can lead to higher costs and more patients exposed to an otherwise ineffective compound than necessary. It is also important to note that good ongoing data management is vital to the success of this design in particular when considering the possibility of stopping during the phase of the design concerned with the assessment of predictive probability. In particular, timely data entry and cleaning will quickly inform the possibility of a stop decision early in predictive probability monitoring. Changes in response due to data cleaning can increase the risk of incorrect decisions early and as an extension have unpredictable effects on average sample size. As in all study designs, several assumptions are made for the purposes of planning the trial. In cases where more patients are recruited than originally planned, then as part of the procedures described in Frewer (2016) the criteria should be re-estimated with the current assumptions.

Multiplicity concerns

It is the case that in these examples there are 20 tests (predictive probability after patients 12–29 an interim at patient 30 and the final at patient 50) being conducted. Were this trial to be designed using frequentist methods, there would be quite a number of adjustments made along with stopping boundaries specified for each of the 20 assessments. Using Bayesian methodology and simulations, there is a much simpler way to address the problem. The operating characteristics are fully characterized over a number of True Effect distributions in terms of both sample size equivalents and range of True Effect mean. Since predictive probability is calculated using an algorithm, there is no need to calculate the individual stopping boundaries and include those in a protocol ahead of time. The algorithm is simply run following each patient until either a decision to stop the trial is made or until the planned number of patients in the predictive probability monitoring phase is reached. Even though simulations are used along with an algorithm to estimate the chance of inappropriate early stopping, it is important to note that with more interim analyses (or longer periods of predictive probability monitoring), the risk of inappropriate stopping will increase. Multiple simulations may be needed to select design parameters that lead to an acceptable level of risk.

Limitations with these examples

Clearly the examples presented here are selected to illustrate some of the properties of the design. In the examples presented here, True Effect mean and sample size equivalent are varied in order to examine those effects. Choice of program targets (TV, LRV), trial success target τ, futility threshold φ, Total sample and the number of patients enrolled in each of the phases of the trial (N, N_a, N_b, N_c, N_d) are all parameters that have been kept constant here for the purposes of brevity and presentation of results in a manageable number of dimensions. It is important when considering design options that simulations are run to understand the operating characteristics of each option. With this in mind, the R code used to simulate the design and the corresponding monitoring chart will be made available upon request.

Conclusions

This design is an additional option for single arm designs where there is a desire to stop early for futility in a flexible way. With this design, fewer patients will be exposed to ineffective compounds than would be exposed using either a single arm trial with either no or a fixed interim. As with all designs the operating characteristics should be carefully understood before conducting the corresponding experiment. In the case of compounds that show more activity, more patients will be exposed to the compound in anticipation of progressing into later phase development. Predictive probability can be used as an effective means to continuously monitor the chance of success in an ongoing single-arm trial.

Acknowledgements

I would like to thank Claire Watkins, Sarah Bujac, Jayne Marshall, James Matcham and Andy Stone for their reviews and thoughts leading to this publication.

Conflict of interest

Author declares that there is no conflict of interest.

References

1. Crowley J, Hoering A. *Handbook of Statistics in Clinical Oncology*. 3rd ed. USA: CRC Press; 2012.
2. Jennison C, Turnbull BW. *Group Sequential Methods Applications to Clinical Trials*. USA: Chapman & Hall/CRC Press; 2000.
3. Berry SM, Carlin PC, Lee JJ, Müller M. *Bayesian Adaptive Methods for Clinical Trials*. USA: CRC Press; 2011.
4. Thall PF, Simon R. Practical Bayesian Guidelines for Phase IIB Clinical Trials. *Biometrics*. 1994;50(2):337–349.
5. Thall PF, Simon RM, Estey EH. New Statistical Strategy for Monitoring Safety and Efficacy in Single–Arm Clinical Trials. *Journal of Clinical Oncology*. 1996;14(1):296–303.
6. Chen N, Lee JJ. Optimal Continuous–Monitoring Design of Single–arm Phase II Trial Based on the Simulated Annealing Method. *Contemp Clin Trials*. 2013;35(1):170–178.
7. O'Hagan A, Stephens JW. Campbell MJ. Assurance in Clinical Trial Design. *Pharmaceutical Statistics*. 2005;4(3):187–201.
8. Lalonde RL, Kowalski KG, Hutmacher MM, et al. Model–based Drug Development. *Clinical Pharmacology & Therapeutics*. 2007*;*82(1):21–32.
9. Frewer P, Mitchell P, Watkins C, et al. Decision Making in Early Clinical Drug Development. *Pharmaceutical Statistics*. 2016;15(3):255–263.
10. Morita S, Thall PF, Müller P. Determining the Effective Sample Size of a Parametric Prior. *Biometrics*. 2008;64(2):595–602.

Citation: Mitchell PD. A Bayesian single-arm design using predictive probability monitoring. *Biom Biostat Int J*. 2018;7(4):299–309.
DOI: 10.15406/bbij.2018.07.00222