PhUSE 2017

Paper AS01

Futile or Not? An Interim Analysis Case Study

Ingrid Franklin, Vermed Ltd, Twickenham, United Kingdom
Rosalind Walley, UCB Pharma, Slough, United Kingdom

ABSTRACT

Optimally, interim analyses are incorporated as part of the original study design. However, unforeseen circumstances may require a protocol to be amended after study initiation to include an interim analysis. It is important to strike a balance between selecting the optimum timepoint to perform the interim analysis based on the operating characteristics of the futility rule, and the logistics of implementation including budget and recruitment forecasts. Given the time constraints for decision-making, conveying information to the study team in a straightforward manner, such that the key elements required to make an informed decision on timing are clear is vital. This case study examines the theoretical and practical processes of incorporating such an interim analysis into a phase II study, and how simulations can be used to identify an appropriate futility rule. Different scenarios, including graphical representation of the futility rule will be presented.

INTRODUCTION

An interim analysis is analysis of data conducted _before_ data collection has been completed and may be incorporated into clinical trials to make an informed decision on whether a study adjustment is appropriate. The International Conference for Harmonisation E9 guideline *'Statistical Principles for Clinical Trials'* (1998)[1] states that a clinical trial should only be stopped if the power is no longer acceptable, which may be due to dropouts or failure to recruit the planned number of subjects, or for ethical reasons that include safety concerns, efficacy or futility.

Early Stopping for Ethical Reasons:

- **Safety:** If a study treatment has a detrimental effect on study participants and there are safety concerns that are considered related to the study treatment, then the trial must be stopped to avoid exposing additional patients to this danger.
- **Efficacy:** If a study treatment has an overwhelmingly positive effect on study participants then stopping should be considered. If stopped, the treatment could be taken straight to submission or the next study could begin sooner, which would also lead to an earlier submission, such that the treatment can get to market and be made available to the patient population sooner.
- **Futility:** If a study treatment has no beneficial effect then the study should be stopped to limit the number of patients that are exposed to an inefficacious drug. Correctly stopping a study for futility will benefit patients by stopping them from being exposed to an inefficacious drug and will benefit the study sponsor by saving them resources that could instead be used on alternative products that may have a greater chance of success.

Including an interim analysis in any trial will affect the operating characteristics of the final analysis and it is therefore optimal to plan the interim analysis at the initial study design stage. This upfront planning will avoid the risk of having an inflated false positive rate or an underpowered study for the final analysis if the study is not stopped at the interim. The interim analysis should therefore be defined in the protocol prior to the start of the trial if possible. Unforeseen circumstances, however, such as slow recruitment, or the unequivocal emergence of remarkable improvement or deterioration in patients, could make implementing an interim analysis while a trial is already ongoing unavoidable and in in this case a protocol amendment must be made to define the interim analysis prior to unblinding the data. Designing and implementing interim analyses in these time critical cases can be a challenge. Presenting the merits and downfalls of the available options to the study team members in a straight forward fashion is key to ensuring the most appropriate path is chosen.

This paper examines a case study where it became necessary to design and implement an interim analysis for futility into an ongoing trial for a rare disease. The role of a statistician in this scenario is discussed, including: what must be considered when creating an appropriate stopping rule; the method for calculating the operating characteristics of the rule by simulation; how to present the possible options to the team in a simplistic manner and the practical aspects involved in implementing the interim analysis.

INTERIM FOR FUTILITY: CASE STUDY

This case study is based on a phase II, double-blind, placebo-controlled, parallel group trial of a rare disease where patients were randomized to either placebo or treatment using a 1:1 ratio. A classical frequentist approach was used for the design, which was based on a 2-sided t-test with 80% power and a 5% false positive rate; target treatment difference and assumed standard deviation were defined using historical literature as 3.8 and 5 respectively and this gave a sample size of 29 patients per arm (58 in total). The primary endpoint was the treatment difference at the end of the treatment period (week 12) of a continuous disease activity score; the higher the score the higher the level of disease activity. The study would be defined as successful after the final analysis if the lower end of the 95% confidence interval is above zero and the null and alternative hypotheses for the study were as follows:

Null hypothesis: *There is no difference in disease activity score between the placebo and the treatment groups at week 12.*

Alternative hypothesis: *There is a difference in disease activity score between the placebo and the treatment groups at week 12.*

At the time of the original study design there were no concerns about recruitment and no expectations that overwhelming efficacy may be found early; therefore, there was no justification for the increased budget that would come with including an interim analysis. However, recruitment turned out to be much slower than forecast and two years into the study only 18 subjects (~30% of the total planned) had been recruited. There were concerns that if recruitment continued at this rate and the treatment was not found to have a statistically significant effect over placebo at the end of the trial, then it would be a tremendous waste of resources. Furthermore, the patients entering the trial would be exposed to an inefficacious drug which in turn would prevent them from receiving a potentially beneficial treatment elsewhere. Senior management requested an interim analysis and the statistician's role was to work with the team to create an appropriate stopping rule.

OPERATING CHARACTERISTICS FOR CONSIDERATION

From a statistical perspective, it is important to be aware when designing an interim analysis, (i) the operating characteristics of that interim analysis and (ii) how the interim analysis may affect the operating characteristics of the final analysis. Of course, the optimum design in terms of keeping the false positive and false negative rates as low as possible may not be the optimum design in terms of practical aspects. The below sections outline what is and is not a concern when designing an interim for futility.

OVERALL FALSE POSITIVE RATE (ALPHA SPENDING): NOT A CONCERN

In this case, the aim was not to test for early signs of efficacy; which meant that alpha spending was not a concern and the overall 5% false positive rate would be maintained at the final analysis. Put another way, there would only be one opportunity to find a statistically significant difference between the placebo and the treatment arms and therefore only one opportunity to falsely declare the drug to be efficacious. Thus there was no requirement for an adjustment to the nominal false positive rate to be used at the end of the study.

OVERALL FALSE NEGATIVE RATE: A CONCERN

Since the study was initially powered without an interim analysis for futility and this was added after the study had started, the overall false negative rate will increase because there would now be two opportunities to find a negative result and declare no statistically significant difference between the placebo and the treatment arms of the study. An increase in the overall false negative rate equates to a reduction in the overall power of the study and it is important to check that the overall power of the study does not drop substantially by including the interim analysis.

NOMINAL FALSE POSITIVE AND NEGATIVE RATE OF INTERIM

It is not worth performing an interim analysis unless it has a reasonably good chance of doing what is needed to help make an informed decision, namely, concluding futility if the drug is inefficacious. For the futility analysis to have a reasonable chance of giving the correct result, the nominal false positive and negative rates must be relatively low. The lower the nominal false negative rate, the lower the overall false negative rate and the less the overall power of the study will drop. Since the study was initially powered at 80%, an industry accepted minimum, it was important that the any drop in the overall power of the study was minimal. In contrast, one might be less concerned about an inflated nominal false positive rate at the interim.

In the context of an interim for futility, we have nominal error rates as defined:

- **False positive rate:** The probability of not stopping an inefficacious drug (not declaring futility when the null hypothesis is true)

- **False negative rate:** The probability of stopping an efficacious drug (declaring futility when the null hypothesis is false)

PROBABILITY OF SUCCESS AT STUDY-END

If the interim analysis indicates futility and the study is stopped, what is the probability that the drug would have been found to be efficacious at the study-end? It may be that the interim analysis indicates futility but the team decides to ignore the result and continue the study anyway. The probability of finding the drug to be efficacious at the final analysis given that the interim has indicated futility should be low and the statistician should present these to the team to limit the risk of poor decision making (i.e. ignoring the futility analysis result!).

CREATING AN APPROPRIATE DECISION RULE

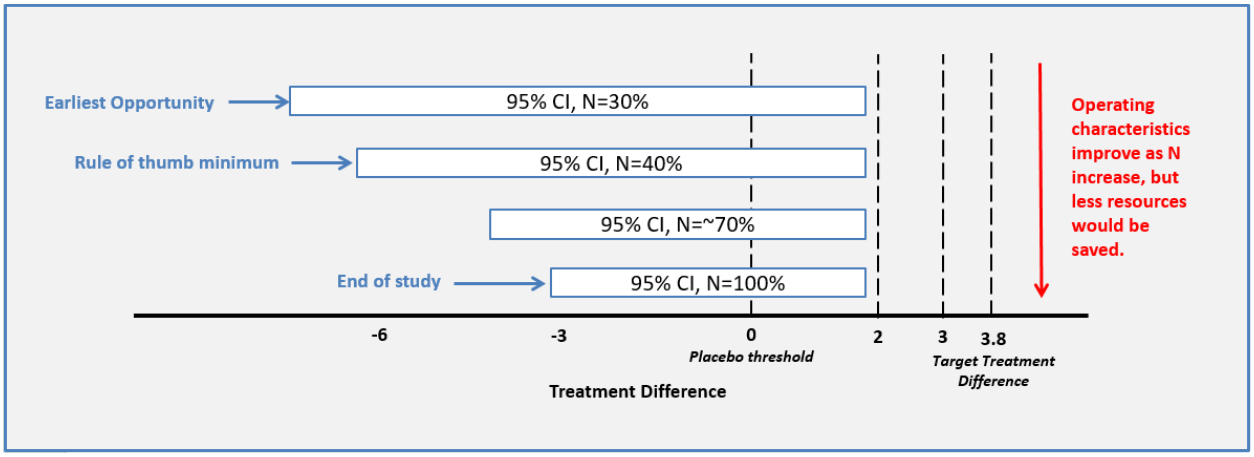
Continuing an inefficacious drug through to study-end is the sponsors risk and the purpose of the decision rule is to give the best possible chance of stopping the study correctly, based on the risk the sponsor is willing to take. The rule formalizes what otherwise is naturally quite a subjective decision and ensures that any subjectivity is defined upfront, rather than after the interim data is revealed, when 'freestyle' decision-making would of course be prone to bias.

If the CI includes the target treatment difference that was defined in the initial study design, there would be no basis to stop the study because there would be a clear possibility that the drug may be found to be efficacious at study-end. A large part of the decision making lay in how stringent the rule should be, that is how poorly performing does the drug need to be at the interim for the team to have no confidence that it may be successful at study-end. The stringency of the rule is related to how far below the target treatment difference the CI must be, for the team to conclude that the chance of success at study-end is so small that the study should be stopped. The closer the threshold value chosen is to the target treatment difference, the better performing the drug must be to pass the interim test, thus the more stringent the rule.

The study team suggested a threshold of 2 as a starting point for defining the decision rule. Although the target treatment difference would, from a statistical point of view, have made sense to look from the outset, it was important to look at the operating characteristics of the rule as suggested by the team, such that we could illustrated why this clinically relevant threshold (of 3.8) would be most appropriate.

Rule option 1: *If the upper limit of the 95% CI at the interim is less than 2, then it is unlikely that the drug will be found to be efficacious at study-end and the study may be stopped for futility at the interim.*

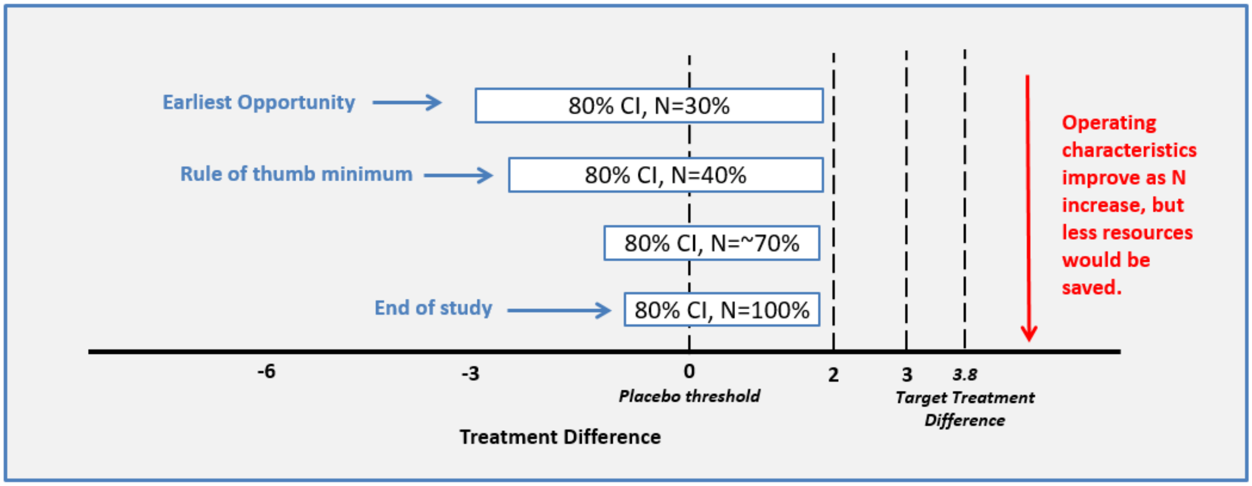
The width of the CI was calculated for different points in the study, starting with the earliest possible timepoint based on recruitment so far. The standard error was calculated based on the standard deviation estimate from historical data that was used in the original study design and the number of subjects at each of the different timepoints. For the purposes of illustrating the rule graphically, the mean difference was fixed such that the upper limit of the CI fell just below the treatment difference threshold that would indicate futility. The 95% CIs were calculated using the formula: *mean difference* ± $z_{0.95}$**standard error for the difference* and the these are displayed graphically below.



A general rule of thumb is that an interim analysis should include at least 40% of the planned total subjects for the operating characteristics to make it worthwhile. However, because of the extremely slow recruitment of this study there was a team consensus that if the properties of the design were reasonable then an earlier interim was preferable. The earliest opportunity in this case would be to perform the analysis at the current recruitment level of N=30% of the total planned number of subjects. The graphic above shows how the variability decreases as the number of subjects included in the interim analysis increases. The operating characteristics also improve with increasing N, but of course the later the interim analysis is performed, the less resources would be saved if the study was stopped. It was important therefore to strike a balance between good operating characteristics and amount of resources that could be saved if the study was stopped.

The above plot shows the 95% CIs at different timepoints in the study for ***Rule option 1*****.** The CIs were calculated for N=30%, 40% and 70%, as well as study-end at N=100%. In all scenarios presented, the mean treatment difference (the middle of each confidence interval) is below the zero threshold; this means that placebo would need to outperform the drug for the team to consider stopping the study. Even with 100% of subjects included in the analysis (not an interim) this would not be an effective rule. It was agreed that this rule was not stringent enough and drugs that were no better than, or only slightly better than placebo could easily be taken forward.

The same scenarios were then presented for an 80% CI (using the formula: *mean difference* ± $z_{0.80}$**standard error for the difference*). The plot below shows that again, this rule is far too stringent and even at N=70%, drugs that are barely outperforming placebo may be taken forward. The team wanted a rule that would give more confidence of success at study-end if the study was not stopped at the interim.

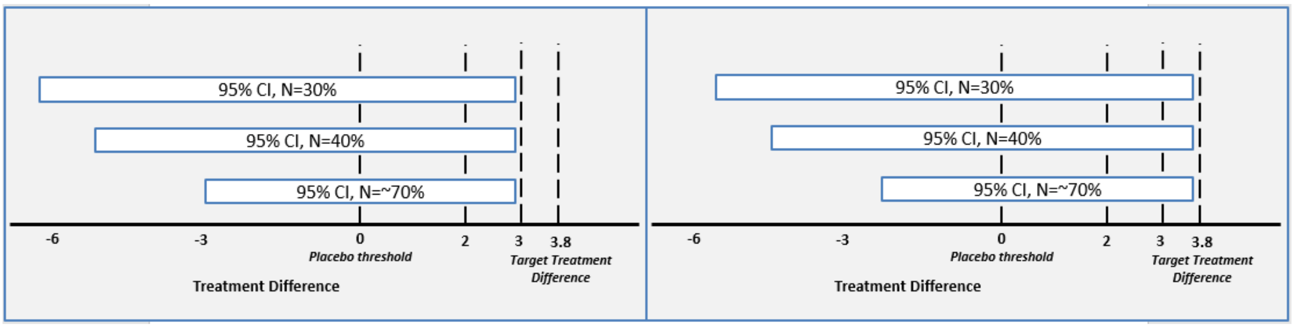


The next logical approach was therefore to increase the treatment difference threshold that the CI must fall below to consider stopping for futility. After considering this the team suggested two further threshold options, (i) some value in between 2 and the target treatment difference and (ii) the target treatment difference itself. This led to the following rule options:

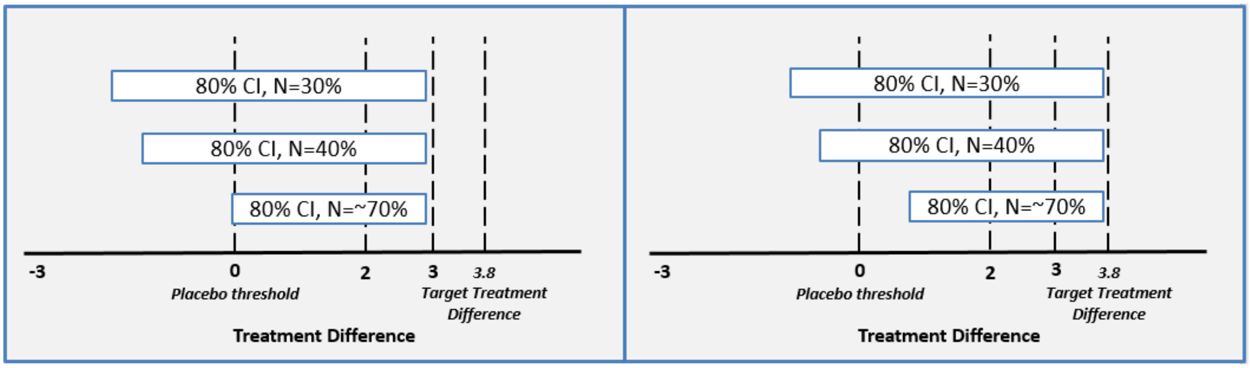
Rule option 2: *If the upper limit of the 95% CI at the interim is less than 3, then it is unlikely that the drug will be found to be efficacious at study-end and the study may be stopped for futility at the interim.*

Rule option 3: *If the upper limit of the 95% CI at the interim is less than 3.8, then it is unlikely that the drug will be found to be efficacious at study-end and the study may be stopped for futility at the interim.*

The team looked at the 95% CIs in relation to ***Rule option 2*** (below left) and ***Rule option 3*** (below right) but the width of these confidence intervals at the early stages of the study, when the interim was being considered, was off-putting because it meant there would be a lower chance of stopping an inefficacious drug. The increased false negative rate that would come with the 80% CI was traded off for the increased probability of stopping an inefficacious drug (as outlined in the next section).



These 80% CIs for ***Rule option 2*** and ***Rule option 3*** (plotted below) were then presented. ***Rule option 3*** was more palatable to the team and the statisticians because the overriding feeling was that based on the extremely slow speed of recruitment and the concern about wasting resources, the drug should need to reach a higher hurdle to be taken forward to study completion.



Rule option 3 using an 80% CI was therefore the rule of choice and the next step was to calculate and present the operating characteristics for this rule at different levels of N, to check whether these were acceptable to the team and to define the most appropriate timepoint to perform the interim analysis.

CALCULATING OPERATING CHARACTERISTICS OF THE RULE

Having decided on a decision rule, namely that if the upper limit of the 80% CI is less than 3.8 this indicates futility, the next task was to calculate and present the operating characteristics to the team. This was done by simulation in R using the following method:

1. Simulate individual patient data for 10,000 studies
2. Calculate mean and variance for each simulated study of the n=100% subjects per arm at primary endpoint. These are the observed summary statistics at the final analysis.
3. Calculate mean and variance for each simulated study of the n=30% subjects per arm at primary endpoint. These are the observed summary statistics at the interim analysis.
4. Calculate CIs for the primary endpoint for each simulated study, for both the interim and the final analysis
5. Define the threshold used in interim analysis rule (3.8)
6. Calculate the overall false positive and false negative rate and the operating characteristics of the interim decision rule
7. Repeat for interim scenarios with different proportion of total N (40% and 70%)

The R code for calculating the operating characteristics of the decision rule (step 6) is below.

```
########################################################################
#   Study with interim - stop if threshold is inconsistent with data   #
########################################################################

#Code Key:
#nsim=Number of simulations
#UCL=Upper confidence limit
#LCL=Lower confidence limit
```

```
#int.UCL.uneff.diff=Simulated mean difference at interim analysis for study with
inefficacious treatment
#int.UCL.eff.diff=Simulated mean difference at interim analysis for study with
efficacious treatment
#fin.UCL.uneff.diff=Simulated mean difference at final analysis for study with
inefficacious treatment
#fin.UCL.eff.diff=Simulated mean difference at final analysis for study with
efficacious treatment


#Calculating the operating characteristics:
#Defining threshold used at interim
threshold<-3.8

#Standard false positive rate and power definitions
#Overall false positive rate: Probability that for an inefficacious drug,
#UCL>threshold at interim AND LCL>0 at study-end
sum((int.UCL.uneff.diff>threshold)*(fin.LCL.uneff.diff>0))/nsim

#Overall power: Probability that for an efficacious drug, UCL>threshold at interim AND
#LCL>0 at study-end
sum((int.UCL.eff.diff>threshold)*(fin.LCL.eff.diff>0))/nsim

#Probability of stopping an inefficacious drug at the interim, probability that for
#an inefficacious drug UCL<threshold
sum(int.UCL.uneff.diff<threshold)/nsim

#Probability of stopping an efficacious drug at the interim, (1- (probability that
#for an efficacious drug UCL>threshold)
1-((int.UCL.eff.diff>threshold)/nsim)
###########################################################################
```

The operating characteristics were calculated based on the possibility of the interim analysis being performed at 30%, 40% and 70% of the total planned N and were presented for both 80% and 95% CIs to see how this affected the rule.

N %	N	CI	Probability of stopping an inefficacious drug at interim	Probability of stopping an efficacious drug at interim
30%	18	95%	32%	3%
30%	18	80%	61%	10%
40%	24	95%	45%	3%
40%	24	80%	72%	10%
~70%	40	95%	66%	3%
~70%	40	80%	87%	10%

From the simulations, the overall false positive rate remained unchanged as expected. The overall false negative rate increased marginally to 23% (the overall power dropped to 77%) for the 80% CI and increased by even less for the 95% CI. This drop in the overall power was small enough to be viewed as negligible. The probability of stopping an inefficacious drug showed a marked increase between the 95% and 80% CI and the probability of stopping an efficacious drug, at 10%, was small enough that the team were happy to move forward with the 80% CI option. It was agreed that although the probability of stopping an inefficacious drug was markedly higher at N=70%, this was simply not practical, because it may well take another 2 years to recruit that many subjects and by that point many more resources would have been wasted and patients would have continued to be exposed to a drug that had little or no effect. At N=40%, the probability of stopping an inefficacious drug was still more than acceptable, however the prevailing opinion was that since this is an internal sponsor risk, a 61% probability of stopping an inefficacious drug was adequate and the interim analysis should be conducted at N=30% to save as many patients and resources as possible should it find the study to be futile.

THE RISK OF IGNORING THE DECISION RULE

To fully understand the rule and its operating characteristics, it is important to understand the probability that the study is successful at study-end, for both efficacious and inefficacious drugs. For us to have confidence in the decision rule, these probabilities should be low; if the probability of success at study-end is high when the interim analysis indicates that you should stop for futility then this is not a good rule. Furthermore, presenting these probabilities to the study team will avoid the risk of the decision rule being ignored when the results are available.

Below is the R code used to calculate the probability of success at study-end if the interim rule is ignored, followed by the output.

```
######################################################################################
#   Study with interim - if you do not stop when threshold is inconsistent with data #
######################################################################################

##This code calculates the probability that you are successful at the study-end given
that the interim rule indicated you should stop study for futility but you continue
anyway.
#Use code key from ‘Calculating Operating Characteristics’ code above
#Probability that for an efficacious drug UCL<threshold at interim AND LCL>0 at study-
end
((int.UCL.eff.diff <threshold)*(fin.LCL.uneff.diff>0))/nsim

#Probability that for an inefficacious drug UCL<threshold at interim AND LCL>0 at
study-end
((int.UCL.uneff.diff<threshold)*(fin.LCL.uneff.diff>0))/nsim

##################################################################################
```

The table below shows the probability of stopping at the interim, as well as the probability of success at study-end if the interim rule is ignored for an efficacious and an inefficacious drug. The probabilities of success at study-end if the interim rule is ignored are low as expected; the probability is extremely low for an inefficacious drug but still very low for an efficacious drug, certainly too low to risk pursuing until study-end ‘just in case’. These probabilities are low because including an interim analysis means that the drug must now pass two hurdles to be considered efficacious and it is more difficult to successfully reach the study-end.

			Probability of stopping at interim for:		Probability of success at study-end, given that the interim analysis indicates that you should stop and you do not for:	
N %	N	CI	Inefficacious drug	Efficacious drug	Inefficacious drug	Efficacious drug
30%	18	80%	61%	10%	0.3%	5.2%

PRACTICALITIES OF IMPLEMENTING THE INTERIM ANALYSIS

FORMAL DOCUMENTATION

The interim rule has been chosen and the operating characteristics of the rule are acceptable to the team. The interim analysis must now be implemented and there are several formal steps involved that must be documented. Firstly, the analysis must be defined up front and so a *protocol amendment* must be made. The amendment should include an overview of analysis to be performed, the decision rule and the operating characteristics that correspond to that decision rule. Secondly, the analysis will need defining in detail in a specific *interim statistical analysis plan (SAP)*. The interim SAP will also include the decision rule and its corresponding operating characteristics, as well as outputs that are required for the interim analysis, which will usually be a subset of the final outputs. In addition, the interim SAP will contain the standard information required in an analysis plan to ensure that the tables, figures and listings for the analysis outputs are programmed as intended.

The study was designed as a double-blind study, therefore specific personnel that will have access to the unblinded data at the stage of the interim analysis will need documenting. In order to avoid bias, the unblinded personnel should not be involved in the day-to-day running of the study. Unblinded statisticians and programmers should be different from those that have been working on the ongoing study and restricted access areas will need creating such that only the unblinded team can access the unblinded data and results.

TIMELINES AND BUDGET

If a clinical research organization has been outsourced to work on the study, then the timelines must be established such that there is a realistic estimation of how long it will take to clean and program the data from the time of the data cut-off; this is the role of the statistician. The interim analysis will of course bring increased budget implications, but this would usually be dealt with by the project manager.

PRESENTING THE RESULTS

The statistician is responsible for collating the results into an ‘easy-to-interpret’ format for the unblinded senior management team to review and decide whether to stop the study for futility. This would usually be a set of slides that contain the key results used to inform the audience of crucial information needed to make an informed decision. The unblinded senior management team are unlikely to be familiar with the detail of the study, so it is important to provide an overview that includes the study design, introduces the primary endpoint as well as any additional variables that will be reviewed and outlines the purpose of the interim analysis and the decision rule that has been created. The recruitment forecast may well have an influence on the final decision, so the current recruitment status and number of subjects that will be saved if the study is stopped early should be presented, as well as the estimated time for reaching study-end if the study is not stopped for futility. The main body of the slides will be used to present the results from the futility analysis and the conclusion, which is whether the results indicate that the study should be stopped for futility.

CONCLUSION

Ideally, an interim analysis would be planned into a study at the initial design stage so that false positive and negative rates could be controlled from the forefront. However, certain scenarios, for example slow recruitment, can lead to the need for an interim analysis to be introduced into an already ongoing study, with the aim of stopping the study if the likelihood of success at study-end is found to be unacceptably low. The ‘unacceptably low’ probability is the sponsor’s risk and is defined by the study sponsor in the creation of the interim analysis decision rule. Stopping the study early if the drug is unlikely to be found to be efficacious at study-end will save the study sponsor time and money that could be used more valuably elsewhere, and saves patients from being exposed to an inefficacious drug for an extended period. Once an interim has been requested, the statistician should work with the clinical study team to create design options that enable the most appropriate decision rule and timepoint to be identified, such that the operating characteristics and implementation practicalities of the interim analysis strike the most appropriate balance for the scenario. Once the decision rule and timepoint have been chosen, the statistician should prepare the formal documentation including a protocol amendment, an interim SAP and a document that identifies the unblinded personnel that will view the results of the interim analysis. Timelines and budget must be agreed with the partner organisation if there is one involved, or internally if not. It is then the role of the statistician to put together and present slides that contain the key results from the interim analysis that are easy-to-interpret, such that the unblinded senior management team can make an informed decision on whether the study should be stopped for futility.

REFERENCES

[1] International Conference for Harmonisation, ICH Harmonised Tripartite Guideline, Statistical Principles for Clinical Trials E9, Current Step 4 version, dated 5 February 1998, viewed 8th August 2017, http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf

ACKNOWLEDGMENTS

Rosalind Walley (UCB Pharma) for her work on the design of the interim analysis for this study.

DISCLAIMER

This research was funded by UCB Pharma. Rosalind Walley is a UCB Pharma employee and holds stock and stock options. Ingrid Franklin is an employee of Veramed who were funded by UCB Pharma for statistical services.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Ingrid Franklin
Veramed Ltd
5th Floor Regal House, 70 London Road, Twickenham TW1 3QS
Work Phone: +44 (0) 203 696 7240
Email: ingrid.franklin@veramed.co.uk
Web: http://veramed.co.uk