MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

Bayesian Methods for Clinical Trials

Lecture 7: Bayesian methods for interim analyses

Libby Daniells & Pavel Mozgunov & Thomas Jaki

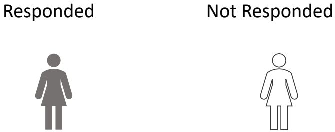
MRC Biostatistics Unit

February 25, 2023

Example

- Consider a **Phase II study**: does a new anti-cancer drug provide any benefits in terms of Response Rate (θ);

- **Binary Outcome**: Response vs No Response;



Responded

Not Responded

- How could we design it using a Bayesian approach?

Bayesian decision-making

- Decision are made based on the **posterior probabilities for the treatment effect** given the clinical trial data;

For example, the trial is claiming **benefit** if

$$\mathrm{Prob}\left[\theta \geq p_0 \mid \mathrm{Data},\ \mathrm{Prior}\right] > c$$

where

- p_0 is an effect threshold
- c is probability threshold (often chosen to control type I error)
- ‘Data” is the observed trial data
- “Prior” is specified prior distribution on the treatment effect.

Bayesian decision-making: Back to the example

- Assume that the planned sample size is $N = 40$ patients;
- The null response rate is $p_0 = 0.30$
- Clinically interesting response rate is $p_1 = 0.50$

The type I error is defined as

$$\mathrm{Prob}\{\mathrm{Probability}\left[\theta \geq p_0 \mid \mathrm{Data},\ \mathrm{Prior}\right] > c | \theta = p_0\}$$

For $c = 0.959$,

$$\mathrm{Prob}\{\mathrm{Prob}\left[\theta \geq 0.3 \mid \mathrm{Data},\ \mathrm{Prior}\right] > 0.959 | \theta = 0.3\} = 0.03134$$

The power is defined as

$$\mathrm{Prob}\{\mathrm{Prob}\left[\theta \geq p_0 \mid \mathrm{Data},\ \mathrm{Prior}\right] > c | \theta = p_1\}$$

Bayesian decision-making: Back to the example

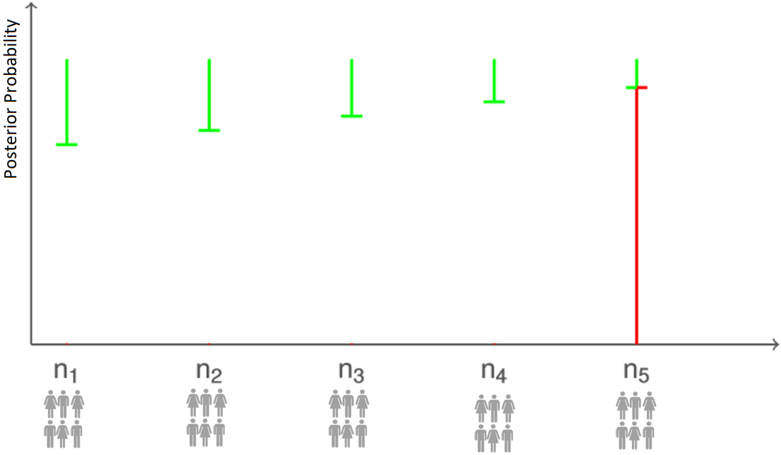
```
# Trial Setting
N<-40
p0<-0.30
p1<-0.50

# Generate data
nsims<-100000
X<-rbinom(n=nsims,size=N,prob=p0)

# Find posterior probabilities
post.prob<-pbeta(p0,shape1=X+1,shape2=N-X+1,lower.tail=F)

# Find Type I error
c<-0.959
mean(post.prob>c)
[1] 0.03289
```

Introducing interim analyses

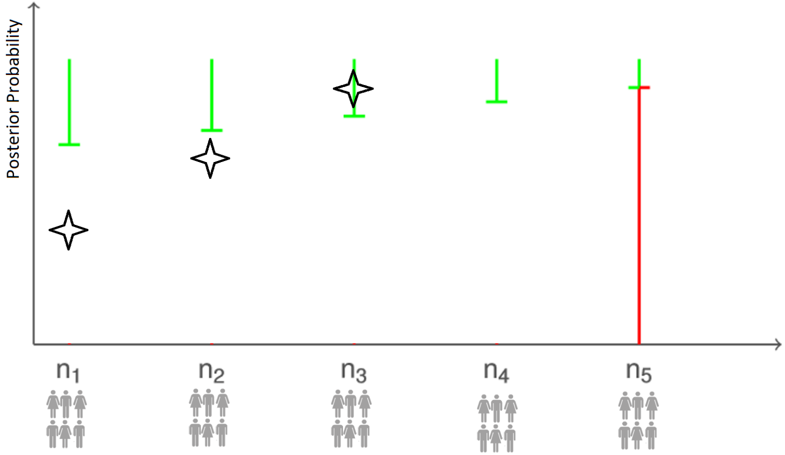
- Assume now we would like to have interim look(s) at the data
- We would like to stop earlier for benefit if there is early evidence supporting this.
- At interim analysis *i* with "Data$_i$" collected in the trial, we claim benefit if

$$\mathrm{Prob}\left[\theta \geq p_0 \mid \mathrm{Data_i},\ \mathrm{Prior}\right] > c_i$$

Group Sequential Trials: efficacy early stopping



Group Sequential Trials: efficacy early stopping



Example: interim analysis

- In our example, assume that we do an interim after 20 patients;
- Early stopping for benefit only.

What would happen if we keep $c_1 = c_2 = 0.959$?

As we are looking at the data twice, we will increase our chance to make a wrong conclusion (e.g. type I error inflation).

```
X1<-rbinom(n=nsims,size=N/2,prob=p0)
X2<-rbinom(n=nsims,size=N/2,prob=p0)
post.prob.1<-pbeta(p0,shape1=X1+1,shape2=N/2-X1+1,lower.tail=F)
post.prob.2<-pbeta(p0,shape1=X1+X2+1,shape2=N-X1-X2+1,lower.tail=F)
c<-0.959
mean(post.prob.1>c | post.prob.2>c)
[1] 0.0646
```

Need to adjust *c* to control the type I error

Futility analyses

In fact, earlier in the development, it can be more common to stop earlier for futility (lack of benefit) rather than benefit;

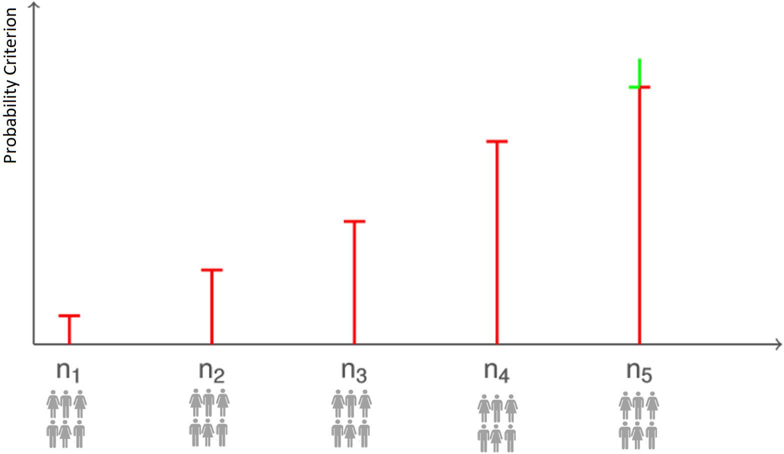
- Futility: trial seems unlikely to achieve its objectives
- Futility IA increasingly used in clinical trial design due to ethical motivations, potentially significant savings
- Adding futility leads to power loss;
- “Stopping when we should” versus “Continuing when we should” are always in conflict and need to be balanced

Futility analysis: Level of aggressivity

Level of aggressivity of the stopping rule allows to minimize one or the other of the two possible incorrect decisions.

- Aggressive futility rule increases chances of (correct) stopping under null, i.e. reduction in type I error
- but also increase chances of (false) stopping under alternative , i.e. reduction in power
- Cautious futility rule decreases chances of stopping under H0 but also increase chances of continuing under H1

Group Sequential Trials



Futility rule based on posterior probability

One way to define the futility stopping is to use the same type of criterion of the posterior probability. Specifically,

- At interim analysis *i* with "Data$_i$" collected in the trial, we stop the trial earlier for futility if

$$\mathrm{Prob}\left[\theta \leq p_0 \mid \mathrm{Data_i},\ \mathrm{Prior}\right] > t_i$$

- For example, in our example, we would like to stop the trial earlier if the posterior probability that the response rate is below 30% is at least 50%

$$\mathrm{Prob}\left[\theta \leq 0.3 \mid \mathrm{Data_i},\ \mathrm{Prior}\right] > 0.50$$

Posterior probability futility: example

Assume that after 20 patients, we have observed 5 responses. Then the probability

$$\mathrm{Prob}\left[\theta \leq 0.3 \mid \mathrm{Data_i},\ \mathrm{Prior}\right] = 0.637 > 0.50$$

so, we would recommend to stop the trial for futility.

```
> post.prob.futility<-pbeta(p0,shape1=x+1,
shape2=N/2-x+1,lower.tail=T)
> post.prob.futility
[1] 0.6372881
```

Assessing the futility rule

While the probability threshold t_i is informative for the decision-making, it might provide limited insight into the actual statistical properties of such an early stopping.

What actually matters are the operating characteristics of the futility rule.

- **Power loss** (%): how much the study power is decreased by imposing a futility rule? (undesirable behaviour);
- **Correct stop under the null** (%): chance that the futility criterion would be reached under the null (desirable behaviour)

Assessing the futility rule: example

To assess the operating characteristics, we can conduct a simulation study to see how likely we are to stop the trial earlier under different hypothesis (null/alternative). In our example,

```
> post.prob.futility<-pbeta(p0,shape1=X1+1,
shape2=N/2-X1+1,lower.tail=T)
> post.prob.efficacy<-pbeta(p0,shape1=X1+X2+1,
shape2=N-X1-X2+1,lower.tail=F)
> c<-0.959 # efficacy threshold from single-stage design
> t<-0.50  # futility threshold

> mean(post.prob.futility<t & post.prob.efficacy>c)
[1] 0.03118 # Type I error

> mean(post.prob.futility>t)
[1] 0.41668 # Probability to stop the trial earlier
```

Assessing the futility rule: example

Design Option	Futility Bound	Type I	P(Stop) (under null)	Power	P(Stop) (under alt)
Fixed	–	3.2%	–	78.6%	–
Adaptive	75%	3.2%	23.6%	78.4%	0.6%
Adaptive	50%	3.2%	41.6%	78.6%	2.1%
Adaptive	25%	2.8%	77.4%	74.4%	13.0%
Adaptive (adj)	25%	5.0%	77.5%	80.2%	13.3%

Futility rule based on Conditional Power

Alternative decision-rules are also commonly used for the futility stopping. These could be easier to communicate as these directly link to the **predicted success of the trial.**

Conditional Power (CP) is the probability that the trial will reach statistical significance at the final analysis if it continues, given the results at the interim analysis, and assuming **certain response** governs the remainder of the trial.

- The observed treatment effect is assumed for the rest of the trials;
- The treatment effect under the alternative effect is assumed for the rest of the trial;
- If $CP_i \leq l_i$ then we stop the trial earlier for futility
- Choose l_i to achieve desirable OCs

Conditional Power: Example

To achieve our criterion for benefit at the final analysis, one needs to observe at least 18 responses in 40 patients

```
> x<-18
> pbeta(p0,shape1=x+1,shape2=40-x+1,lower.tail=F)
[1] 0.9800707 # > critical value = 0.959
```

Assume that we have observed 9/20 responses at the interim.

We need at least 9/20 more to declare benefit at the final

Probability to achieve this under the observed rate is 58.6%

```
> sum(dbinom(x=9:20,size=20,prob=9/20))
[1] 0.5856938
```

Probability to achieve this under 50% (alternative) rate is 74.8%

Conditional Power: Example

To find the critical value for the CP, one can, again, resort to simulations. Specifically, for the considered example

```
CP<-c()
for(i in 1:nsims){
  CP[i]<-sum(dbinom(x=(18-X1[i]):20,size=20,
  prob=(X1[i]/(N/2))))
}
post.prob.efficacy<-pbeta(p0,shape1=X1+X2+1,
shape2=N-X1-X2+1,lower.tail=F)

l<-0.001
> mean(CP>l & post.prob.efficacy>c)
[1] 0.0324 # Type I error
> mean(CP<l)
[1] 0.41393 # Probability to stop early under the null
```

Futility rule based on Predictive Probability of Success (PPoS)

The CP approach allows to assume the response rate for the reminder of the trial. However, the response rate is a **random variable** itself.

- PPoS is the probability that the trial will claim benefit at the final analysis if it continues, given the results at the interim analysis (and prior) - no further assumptions.
- PPoS is obtained by integrating the data likelihood over the posterior distribution;
- If PPoS is below the futility threshold l_i, then the trial is stopped earlier for the futility

PPoS Futility rule: example

For a binary single-arm trial, we already came across the predictive distribution! it is the **beta-binomial** distribution

So, instead of computing $\Pr(X = 9:20 | \theta = 9/20)$, we compute

$$\Pr(X = 9:20 | Data) = \int_0^1 \Pr(X = 9:20 | \theta) \times \pi(\theta | Data) \mathrm{d}\theta.$$

After 9/20 responses, we need at least 9/20 responses in the second stage to declare benefit.

The predictive probability to observe these is 57%

```
library("rmutil")
sum(dbetabinom(9:20, 20, (9+1)/(20+2), 20+2))
[1] 0.5697892
```

To remind, the CP under the observed effect was 59%/75%.

Comparing different futility rules

At a given time point, a futility rule expressed on any particular scale can be transformed to any other.

These scales are more devices for expressing futility rules rather than meaningful ways of quantifying the chance of trial success.

In choosing a futility criterion, we should focus on operating characteristics.

Impact of the Prior Distribution

Assume that there is some **prior knowledge** about the response rate:

1. The most likely response probability is 40%;

2. We are pretty sure (80% sure) that it is between 20% and 65%.

Prior Distribution

Resulting Beta prior distribution is $Beta(3.2, 7.5)$

Design Option	Futility Bound	Type I	P(Stop) (under null)	Power	P(Stop) (under alt)
Uniform	25%	2.8%	77.4%	74.4%	13.0%
$B(3.2, 7.5)$	25%	5.9%	60.9%	84.5%	5.7%
$B(3.2, 7.5)$ (adj)	25%	3.0%	61.0%	77.2%	5.8%

Treatment Effect

- We study the **treatment effect** on the Experimental over the control.
- We measure it via Log Odds Ratio (OR)

$$\theta = \log\left(\frac{p_E \quad \times \quad (1 - p_C)}{(1 - p_E) \times p_C}\right)$$

The trial is stopped **for futility** if

$$\mathrm{Prob}\left[\theta \leq 0.0 \mid \mathrm{Data},\ \mathrm{Prior}\right] > 90\%$$

The trial is stopped **for benefit** if

$$\mathrm{Prob}\left[\theta \geq 0.0 \mid \mathrm{Data},\ \mathrm{Prior}\right] > 95\%$$

Discussion

- A variety of criteria can be used under the Bayesian framework
- Aid with the interpretation/communication of the decisions
- What is important is how this translates into the statistical properties (type I error, power/sample size, probability to stop earlier)
- We still should consider type I error and power for Bayesian procedure and adjust for multiplicity appropriately