Ingram Olkin Forum: Unplanned Clinical Trial Disruptions

Discussion of talks by

Cornelia Kunz and Gary Rosner

on

Bayes and Frequentist Approaches to

Rescuing Disrupted Trials

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

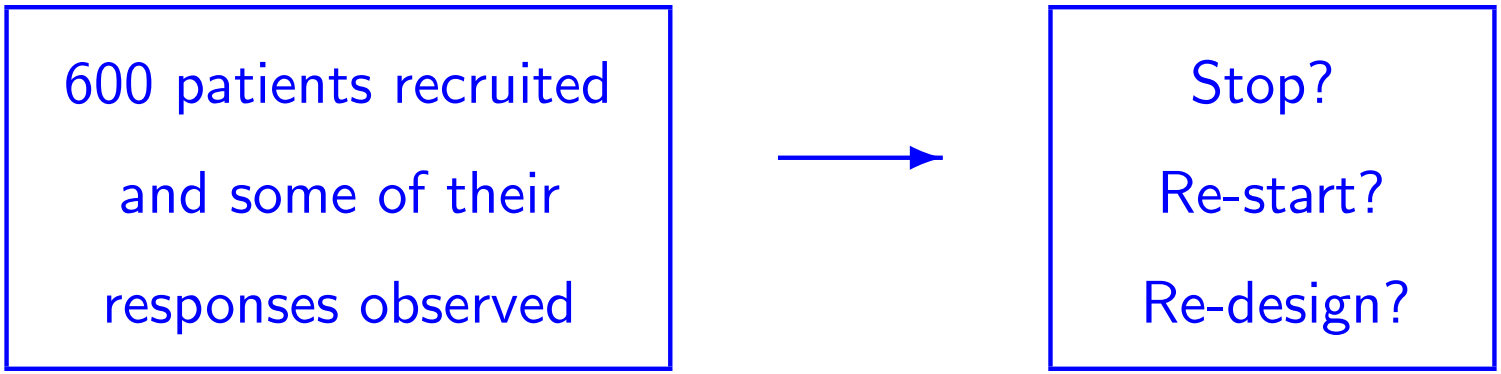
http://people.bath.ac.uk/mascj

27 April 2021

Outline of discussion

We have the example of a trial in Type 2 Diabetes.

The target sample size is 500 patients pre treatment arm, but the trial is paused with 300 patients enrolled on each arm.



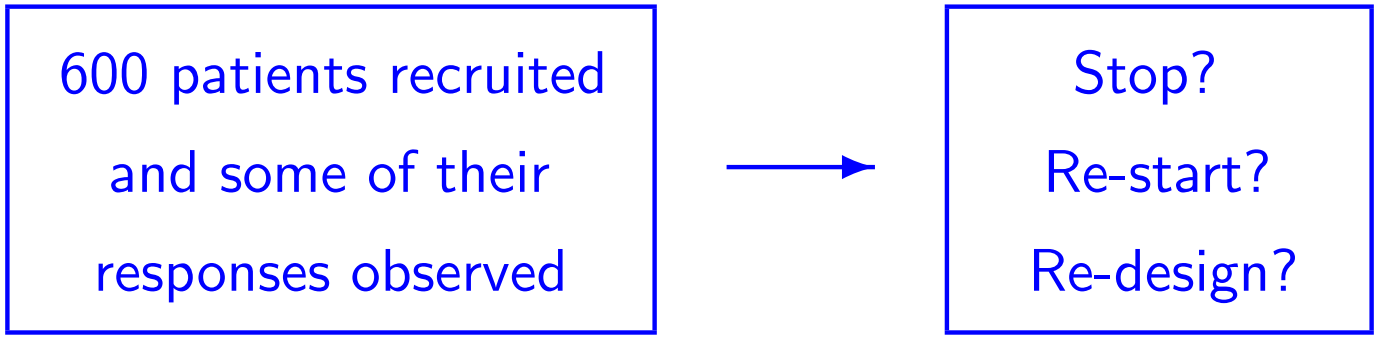
The speakers have considered two cases:

1. Investigators are blinded to the results obtained thus far.
2. Investigators have seen some results on the primary endpoint.

I shall discuss these cases in turn.

Case 1. When investigators are blinded to results

The problem is similar to the initial task of designing the trial.



Suppose investigators had initially considered a design with an interim analysis at the point where disruption occurred.

At this analysis, they might:

a) Stop and declare the new treatment to be superior to control,

b) Stop the trial for futility,

c) Continue the trial, possibly modifying the final sample size.

We can consider Frequentist and Bayes approaches to this design.

Group sequential trial design: Frequentist

Suppose we assume reduction in HbA1c

$$Y_{Ti} \sim N(\mu_T, \sigma^2) \quad \text{for patient } i \text{ on the new treatment,}$$

and

$$Y_{Ci} \sim N(\mu_C, \sigma^2) \quad \text{for patient } i \text{ on the new treatment.}$$

Then the treatment effect is $\theta = \mu_T - \mu_C$.

We wish to test H_0: $\theta \leq 0$ against $\theta > 0$ with

Type I error probability $\alpha = 0.025$,

Power $1 - \beta = 0.9$ at $\theta = \delta = 0.2$,

A fixed sample size trial needs a sample size per treatment arm of

$$n_{fixed} = 2(z_\alpha + z_\beta)^2 \sigma^2 / \delta^2.$$

Group sequential trial design: Frequentist

Dr Kunz notes one can retrospectively decide to apply a group sequential design with analysis 1 at the time of the disruption.

This is a valid approach since the investigators who are to create this design have not yet observed any response data.

How should one choose the stopping boundary — and other prospective features of the trial design?

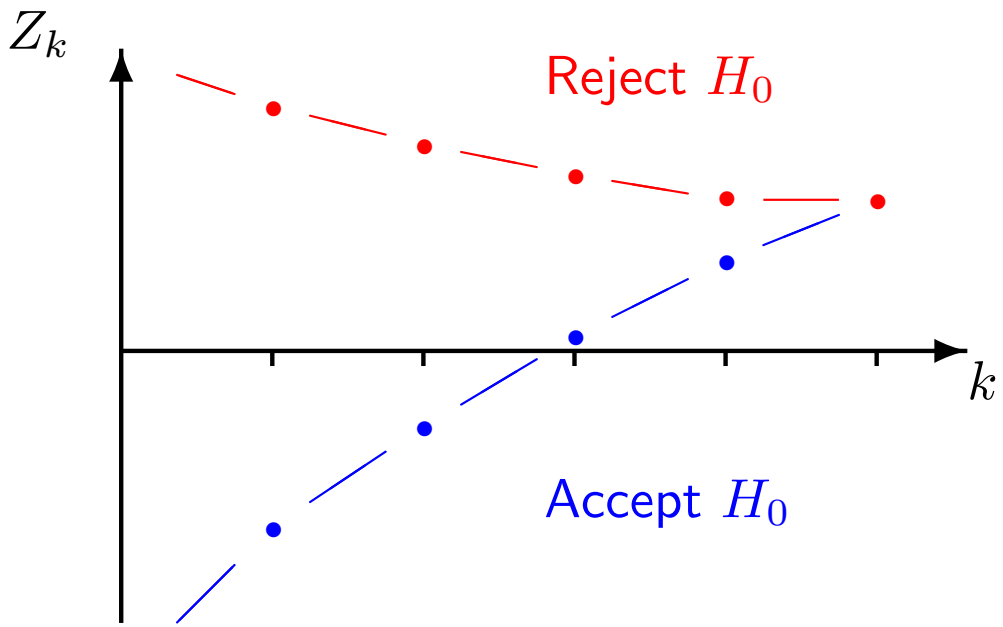
Professor Rosner suggests a futility analysis with early stopping if predictive power is low.

This should not inflate the type I error rate if the only change to the original plan is to halt the trial with a negative result.

How should one choose the prior under which predictive power is calculated — and the threshold for "low" power that implies stopping for futility?

Group sequential trial design: Frequentist

Consider a group sequential design with maximum sample size $n_K = R\, n_{fixed}$ per treatment $(R > 1)$ and K analyses after $n_1, \ldots, n_K$ observations per treatment.



The stopping boundary can be chosen to satisfy error constraints

$$P_{\theta=0}\{\text{Reject } H_0\} = \alpha \quad \text{and} \quad P_{\theta=\delta}\{\text{Reject } H_0\} = 1 - \beta.$$

Indeed, there are many such boundaries.

Group sequential trial design: Frequentist

We can optimise the stopping boundary to minimise a chosen criterion, such as

$$\{E_{\theta=0}(N) + E_{\theta=\delta}(N)\}/2 \quad \text{or} \quad \int w(\theta)\, E_\theta(N)\, d\theta.$$

Here, $w(\theta)$ may reflect the likelihood associated with a particular treatment effect θ and the importance of early stopping under different values of θ.

The "inflation factor" R and maximum number of analyses K can be chosen in view of the reduction in $E(N)$ they can deliver.

Adaptive group sequential designs

In an adaptive design, one can vary the size of each group of observations, depending on the data observed thus far.

Optimal designs can be found in this larger class of procedures (Jennison & Turnbull, *Biometrika*, 2006).

Group sequential trial design: Bayesian

Again, assume reductions in HbA1c follow

$$Y_{Ti} \sim N(\mu_T, \sigma^2) \quad \text{and} \quad Y_{Ci} \sim N(\mu_C, \sigma^2)$$

and $\theta = \mu_T - \mu_C$.

We can specify a prior $\pi(\theta)$ and define expected loss, for example,

$$\begin{aligned} E(L) &= \int_{\theta \leq 0} \pi(\theta)\, P_\theta\{\text{Reject } H_0\}\, \boldsymbol{\lambda_1}(\theta)\, d\theta \\ &+ \int_{\theta > 0} \pi(\theta)\, P_\theta\{\text{Do not reject } H_0\}\, \boldsymbol{\lambda_2}(\theta)\, d\theta \\ &+ \int \pi(\theta)\, \mathbf{c}(\theta)\, E_\theta(N)\, d\theta. \end{aligned}$$

Then, for given $n_1, \ldots, n_K$, we seek a stopping rule that minimises this expected loss.

Group sequential trial design: Bayesian

In a fully Bayes decision theoretic formulation, the prior $\pi(\theta)$ should represent the investigators' beliefs.

Further, $c(\theta)$ denotes the cost of treating a pair of patients on treatment and control arms, while the functions $\lambda_1(\theta)$ and $\lambda_2(\theta)$ represent the financial or medical "cost" of an incorrect decision.

Calibration

In practice a regulator (FDA, EMEA, etc.) will insist that type I error be controlled, so

$$P_{\theta=0}\{\text{Reject } H_0\} \leq \alpha$$

and a sponsor may require a certain power,

$$P_{\theta=\delta}\{\text{Reject } H_0\} \geq 1-\beta.$$

One is then led to modify the prior $\pi(\theta)$ and cost functions $\lambda_1(\theta)$, $\lambda_2(\theta)$ and $c(\theta)$ to satisfy these constraints.

Group sequential trial design: Bayesian

The role of the prior

In a standard Bayesian analysis, the prior has a significant effect on inferences drawn from the data. Hence, the prior can be crucial.

However, with "calibration" to fix $P_\theta(\text{Reject } H_0)$ at $\theta = 0$ and $\theta = \delta$, there is not much scope left for the prior to affect inferences.

But the prior does still determine the properties of a design that are optimised, in our case

$$\int \pi(\theta)\, c(\theta)\, E_\theta(N)\, d\theta.$$

The same phenomenon appears in adaptive designs where rules are put in place to protect type I error, then other design aspects are chosen to optimise an expected utility, averaged over a prior distribution for certain parameters. See Burnett & Jennison (*Statistics in Medicine*, 2021) for adaptive enrichment designs.

Group sequential trial design: Bayesian = Frequentist

It can be shown that the two approaches I have described lead to the same set of optimal trial designs (Complete Class Theorem).

Essentially, the frequentist design's

$$\int w(\theta)\, E_\theta(N)\, d\theta$$

corresponds to the Bayesian design's

$$\int \pi(\theta)\, c(\theta)\, E_\theta(N)\, d\theta$$

and if $w(\theta) = \pi(\theta)\, c(\theta)$, both formulations lead to the same design.

Indeed, computationally, the best way to solve the frequentist problem, with its error rate constraints, is to convert it to an unconstrained Bayes sequential decision problem.

Despite this fact, there are two sets of papers in the literature that work towards these equivalent (essentially identical) designs.

My conclusions

1. One may use either approach — but it is important to get the ingredients of the decision problem right.

2. Optimise conditionally on the "inflation factor" R and number of analyses K, then look to see which values of R and K are most acceptable.

3. Using optimal designs as a benchmark, we see that a simple and efficient choice is an error spending design that spends error probability in proportion to $\mathcal{I}^{\rho}$ where $\mathcal{I}$ denotes information and ρ is in the range 1 to 3.

4. I am not very keen on defining a futility boundary as a function of predictive power as I do not know what the appropriate function should be.

Returning to our disrupted trial . . .

There are some special features of the design for a disrupted trial.

1. Impact of the delay in re-starting the trial

The time that a decision is reached can be just as important as the sample size on termination, particularly for a positive trial outcome. This decision time determines how soon patients can benefit from a new treatment and how long a company can benefit from their patent for a molecule.

Since there will be a delay in re-starting the trial, this implies a greater than usual benefit to stopping at the interim analysis that takes place at the point of disruption.

Thus, one may expect to spend more type I error probability than usual at this analysis.

Returning to our disrupted trial . . .

2. Changes in the post-disruption phase

There may be differences in the patient population or in the treatment effect on patients before and after disruption.

Lockdown affects behaviour and one might expect there to be consequences for diet and exercise, with subsequent effects on HbA1c for patients with Type 2 diabetes.

Also, there could have been disruptions to treatment or key HbA1c measurements may not have been recorded.

Considerations:

Is it still reasonable to test a single hypothesis about treatment effect before and after disruption?

If not, is the only option to stop and analyse pre-disruption data?

Can we analyse post-disruption data as a new, independent trial?

Case 2. When investigators are not blinded to results

Dr Kunz notes it makes no sense to look at blinded or unblinded data in deciding on the procedure to follow for resuming the trial.

However, we can discuss what to do if this has happened.

In this case, protecting the type I error rate is a challenge!

Conservative procedures

Suppose we look at the interim data, choose a sample size for the remainder of the study, then naively analyse the (pooled) final data as if no interim look had occurred.

Dr Kunz refers to a result of Wassmer & Brannath (2016) on the maximum possible increase in type I error rate that can arise in this case. Hence we can conduct the final test at a significance level less than α, chosen so that the real type I error is at most α.

I suspect this is a rather conservative approach.

When investigators are not blinded to results

Conservative procedures

Professor Rosner proposes an interim analysis at which one may stop for futility or continue to the final analysis "as originally planned".

This will not inflate the type I error rate if the only change to the original plan is to halt the trial with a negative result.

However, if the final sample is modified in light of the interim data, the frequentist type I error rate could increase, as seen in Wassmer & Brannath's result.

When investigators are not blinded to results

Another option

Dr Kunz mentions the "conditional error principle".

Here, we calculate

$$\alpha' \;=\; P_{\theta=0}(\text{Reject } H_0 \,|\, \text{Interim data}),$$

assuming the trial proceeds as originally intended.

We can then re-design the remainder of the trial and carry out a level α' test of H_0: $\theta \leq 0$ test, using just the new data.

If this test rejects H_0, we reject H_0 in the overall procedure.

Comment

It may not be straightforward to say precisely what would have happened if "the trial had proceeded as originally intended".

When investigators are not blinded to results

Questions on implementing the conditional error principle

(i) Suppose patient recruitment has been lower than anticipated. Should we assume the final sample size will also be lower than planned or would the trial have been extended to make up the shortfall?

(ii) If the final analysis is a t-test, what value do we assume for the variance in calculating α' ?

(iii) Could investigators work through a number of versions of "the trial proceeding as originally intended" and choose the version that gives the highest value of α' ?

These questions concern the control of a frequentist error rate.

Can a Bayesian approach to inference avoid such issues?