Adaptation Based on Blinded Data

Discussion of talks by:

Janet Wittes,

Ekkehard Glimm,

Tobias Mütze, Heinz Schmidli & Tim Friede,

Michael Proschan & Martha Nason

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

10th Int. Conference on Multiple Comparison Procedures

University of California, Riverside, June 2017

Type I error rate inflation

The key issue addressed in the presentations is how to avoid inflation of the Type I error rate.

This may occur when sample size is modified in response to:

- An interim estimate of response variance,
- An interim estimate of treatment effect,
- Other interim information that could be a surrogate for the treatment effect.

I shall illustrate these points in examples of adaptive designs with sample size re-estimation.

Ch. 14 of *Group Sequential Methods with Applications to Clinical Trials* (Jennison & Turnbull, 2000) contains similar examples.

Type I error rate inflation

Case 1: Normal response with unknown variance, σ^2.

A trial compares a new treatment with control.

The treatment effect, θ, is the difference in mean responses.

We test H_0: $\theta \leq 0$ vs $\theta > 0$ with type I error rate set at $\alpha = 0.025$.

A provisional sample size is set to meet a specified power requirement, assuming an initial estimate of σ^2.

At an interim analysis, an estimate $\widehat{\sigma}^2_{(1)}$ is calculated from the interim data and total sample size is re-calculated using this estimate in the sample size formula.

Low values of $\widehat{\sigma}^2_{(1)}$ lead to low total sample sizes.

High values of $\widehat{\sigma}^2_{(1)}$ lead to higher total sample sizes which dilute the excess in this estimate over the true σ^2.

Overall, this produces a downwards bias in the final estimate of σ^2.

The role of asymptotic theory

The final estimate of σ^2 is used in the t-test for a treatment difference: the downwards bias in this estimate of σ^2 increases the probability of rejecting H_0: $\theta \leq 0$.

Using an **unblinded** estimate of σ^2 at the interim analysis, based on 20 to 30 degrees of freedom, **the type I error rate rises from 0.025 to around 0.028**.

Both Glimm and Proschan & Nason mention asymptotic results.

As the number of degrees of freedom for estimating σ^2 at the interim analysis tends to infinity, we can suppose $\widehat{\sigma}^2_{(1)} \approx \sigma^2$ and the source of type I error inflation disappears.

In reality, we still need to know whether our sample size is large enough for this theory to apply, so computations for small sample sizes remain important.

The role of asymptotic theory

In other types of adaptive design, one wishes to maximise the benefits of adaptation by adapting as early as possible.

For example, in a multi-arm study, it is desirable to drop a poorly performing treatment before further subjects are randomised to that treatment.

In an adaptive enrichment design, if recruitment is to be restricted to a subpopulation, making this change early will give the greatest increase in the number of subjects from this subpopulation.

So, by design, such adaptation decisions are liable to be made when there is substantial noise in the data — and asymptotic theory may not be helpful.

Unblinded estimation of σ^2

Back to Case 1: Normal response with unknown variance, σ^2.

Suppose we calculate $\widehat{\sigma}^2_{(1)}$ from pooled data.

With n_1 observations per treatment and $\widehat{\theta}_{(1)}$ and $s^2_{(1)}$ the usual unblinded estimates of θ and σ^2, the blinded estimate of σ^2 is

$$\widehat{\sigma}^2_{(1)} = \frac{(2n_1 - 2)\, s^2_{(1)} + (n_1/2)\, \widehat{\theta}^2_{(1)}}{2n_1 - 1}.$$

Remarkably, conditional on knowing the total sum of squares

$$(2n_1 - 2)\, s^2_{(1)} + (n_1/2)\, \widehat{\theta}^2_{(1)},$$

the interim t-statistic

$$\frac{\widehat{\theta}_{(1)}}{\sqrt{(2\, s^2_{(1)}/n_1)}}$$

has the standard $t_{(2n_1-1)}$ distribution under $\theta = 0$.

Unblinded estimation of σ^2

The second stage sample size, n_2, is chosen based on the pooled variance estimate $\widehat{\sigma}^2_{(1)}$, which is independent of

$$t_1 = \frac{\widehat{\theta}_{(1)}}{\sqrt{(2\, s^2_{(1)}/n_1)}}$$

The final t-statistic is of the form

$$\tilde{t} = \frac{a\,\widehat{\theta}_{(1)} + b\,\widehat{\theta}_{(2)}}{\sqrt{(c\, s^2_{(1)} + d\, s^2_{(2)})}} \qquad (1)$$

where a/c is fixed but b and d depend on n_2 (with $b/d \approx a/c$).

It is observed that a final test based on $\tilde{t}$ has type I error rate very close to the target α (Friede & Miller, *Applied Statistics*, 2012).

It would be nice to prove mathematically that (1) implies type I error rate should be closer to α than in the unblinded case.

Type I error rate inflation

Case 2: Binary response.

The primary endpoint is a binary outcome with success probabilities p_1 and p_2 on the new treatment and control, respectively.

To attain power at a specified treatment difference, $p_1 - p_2 = \delta$, the required sample size depends on the sum of variances

$$p_1(1-p_1) + p_2(1-p_2) \;\approx\; 2\,\bar{p}\,(1-\bar{p}),$$

where $\bar{p} = (p_1 + p_2)/2$.

Sample size can be modified in response to the interim estimate of $\bar{p}_{(1)}$ and when $\bar{p}$ is away from 0.5 there is, typically, a small inflation of the type I error rate for similar reasons to Case 1.

However, the estimate of $\bar{p}$ is the same whether data are blinded or unblinded — so estimating $\bar{p}$ from blinded data does not protect against inflation of the type I error rate.

Sample size modification in response to $\widehat{\theta}_{(1)}$

Case 3: Normal response (known variance) with sample size modification in response to $\widehat{\theta}_{(1)}$

It is well understood that such sample size modification can lead to type I error rate inflation.

Here is a simple example:

We plan a study with total sample size n but conduct an interim analysis after observing $n/2$ observations.

We compute the test statistic Z_1 based on the interim data, and

If $Z_1 > 1$, continue as planned,

If $Z_1 < 1$, increase the total sample size to $1.5\,n$.

Naively, we analyse the final set of data as if it were from a fixed sample size trial.

Sample size modification in response to $\widehat{\theta}_{(1)}$

The adaptive increases of sample size has a systematic effect.

Suppose $\theta = 0$, so the null hypothesis is true:

> If early data give a low $\widehat{\theta}$, we increase the total sample size and dilute these data.
>
> If early data give a high $\widehat{\theta}$, we retain the original sample size.

As a consequence, the actual type I error rate $=$ **0.0287**.

Suppose the sample size is modified in response to a secondary endpoint that is correlated with the primary endpoint.

Wittes suggests this should not have a substantial effect on type I error rate. Is that so?

Sample size modification based on a secondary endpoint

We can adapt our example:

Plan a study with total sample size n. Conduct an interim analysis after observing $n/2$ observations of a secondary endpoint.

Compute the standardised statistic $\tilde{Z}_1$ for the secondary endpoint based on the interim data, and

If $\tilde{Z}_1 > 1$, continue as planned,

If $\tilde{Z}_1 < 1$, increase the total sample size to $1.5\,n$.

Analyse the primary endpoint as if the final data were from a fixed sample size trial.

If the secondary endpoint has mean 0 and correlation ρ with the primary endpoint, we have

ρ	0.4	0.6	0.8	1
Error rate	**0.0261**	**0.0269**	**0.0280**	**0.0287**

Sample size modification based on the observed event rate

Wittes discusses the case of a study with a survival endpoint where investigators believe they know the hazard rate on the control arm.

Suppose the (blinded) overall event rate is low: this suggests there is improved survival on the new treatment.

Let θ be the log hazard ratio between new treatment and control.

Can sample size modification based on overall event rate be akin to modification based on $\widehat{\theta}_{(1)}$ — and so inflate type I error rate?

I wanted to run simulations to assess possible type I error inflation.

Simulating under $\theta = 0$, suppose we see a low overall hazard rate. This cannot tell us $\theta > 0$!

Under $\theta = 0$, the distribution of $\widehat{\theta}_{(1)}$ given the total number of events at an interim analysis is (approximately) normal, mean zero.

I conclude: No grounds for type I error rate to be inflated here.

Using external information on nuisance parameters

Mütze proposes a meta-analysis of "similar" trials to give a prior distribution for a nuisance parameter, such as response variance.

There have been similar proposals to use historical data on a control treatment to reduce numbers on a new trial's control arm

A key feature of the meta-analysis is to capture the variability seen in previous studies of the mean response on control.

Applying the same principles to historical data on a nuisance parameter, why not simply trust the results of the meta-analysis? If you do trust these results, why "robustify" the prior by mixing it with a more dispersed distribution?

An important distinction: Historical data on the mean response on control are used in testing for a treatment effect, while historical data on nuisance parameters guide the data collection process.

Some concluding comments

Much of the discussion in the talks has been about whether the "conventional" analysis of the final set of data is valid.

The conventional analysis can be "calibrated" by conducting the final hypothesis test at level $\alpha' < \alpha$ where α' is chosen so that inflation raises the actual type I error rate to at most α.

An alternative solution is to use a combination test.

Typically, the power of a combination test is about the same as that of a calibrated version of the conventional test.

The combination test loses some efficiency through not using the sufficient statistic.

However, the conventional test loses power since calibration is driven by the worst case scenario and type I error is controlled conservatively in other situations.