Interim analyses and early stopping in clinical trials : classical and adaptive methods

Christopher Jennison

Dept of Mathematical Sciences,

University of Bath, UK

Astra Zeneca,

Mölndal, Sweden

September 2006

Plan of talk

1. Design of clinical trials

2. Interim monitoring

3. Group sequential analysis of the primary response

    Repeated hypothesis testing

        Example: BHAT study

    Lower “futility” boundary

4. Adaptive designs

    Nuisance parameters

    Mid-study design changes

5. Seamless transition between trial phases

6. Conclusions

1. Design of randomised clinical trials

A trial is conducted to compare treatments, e.g., a new treatment against the current standard or a placebo.

The effect size, θ, is a measure of the improvement in the new treatment over the standard.

We shall test the null hypothesis H_0: $\theta \leq 0$ against the alternative $\theta > 0$.

Rejecting H_0 allows us to conclude the new treatment is better than the standard.

Define in the trial protocol:

Treatment,

Patient population,

Primary endpoint.

Design of clinical trials

Specify:

Allowable type I error probability (of rejecting H_0 when it is actually true).

Desired power, the probability of (correctly) rejecting H_0 when $\theta = \delta$.

Here, δ is, typically, the minimal clinically significant treatment difference.

With a statistical model for patients' responses, we can find the sample size needed to meet these type I error and power requirements.

Recruit the required subjects, N per treatment, say, observe their responses and analyse the data.

2. Interim monitoring

Data and Safety Monitoring Board (DSMB) monitors:

Rate of recruiting subjects,

Treatment administration and compliance,

Proper measurement of response,

and takes actions as appropriate.

The DSMB also looks at:

Safety variables,

Primary and secondary endpoints.

Information on primary responses could indicate an early conclusion of the trial is desirable.

3. Group sequential analysis of the primary endpoint

Consider a trial where the final analysis will reject H_0 if

the P-value (one-sided) is < 0.025,

or, equivalently, the Z-value is > 1.96.

Suppose investigators examine interim data half way through the trial and calculate $Z = 2.1$ ($p = 0.02$)

— can they stop the trial at this point and report a positive conclusion?

The answer is NO:

To maintain credibility, the study has to be conducted according to the protocol.

Repeated testing of a hypothesis increases the likelihood of a type I error by giving more opportunities to reject H_0.

The need for special methods

Multiple looks at accumulating data can lead to over-interpretation of interim results.

Overall type I error rate when
applying repeated significance tests
at $\alpha = 5\%$ to accumulating data

Number of tests	Error rate
1	0.05
2	0.08
3	0.11
5	0.14
10	0.19

Pocock (1983) *Clinical Trials*, Table 10.1,

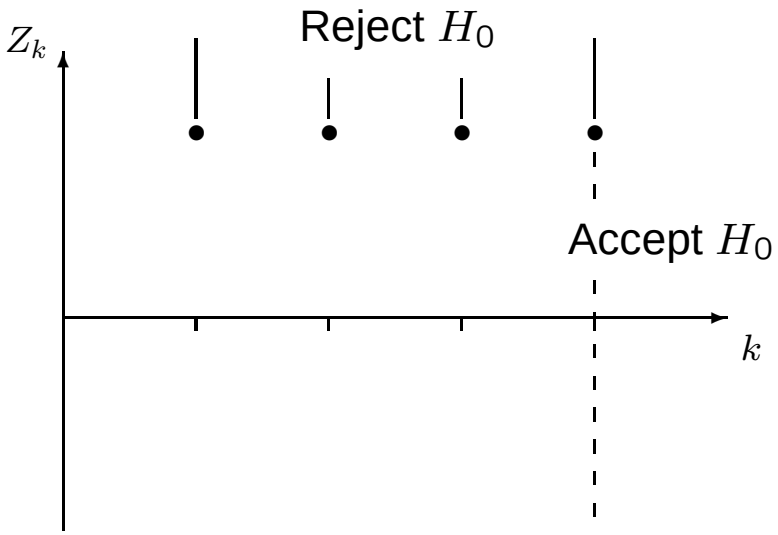
Armitage et al. (*JRSS, A*, 1969), Table 2.

Group sequential tests

A stopping rule can be defined that gives overall type I error probability $\alpha = 0.025$.

A number of interim analyses is specified and a stopping boundary is created.

A Pocock boundary:



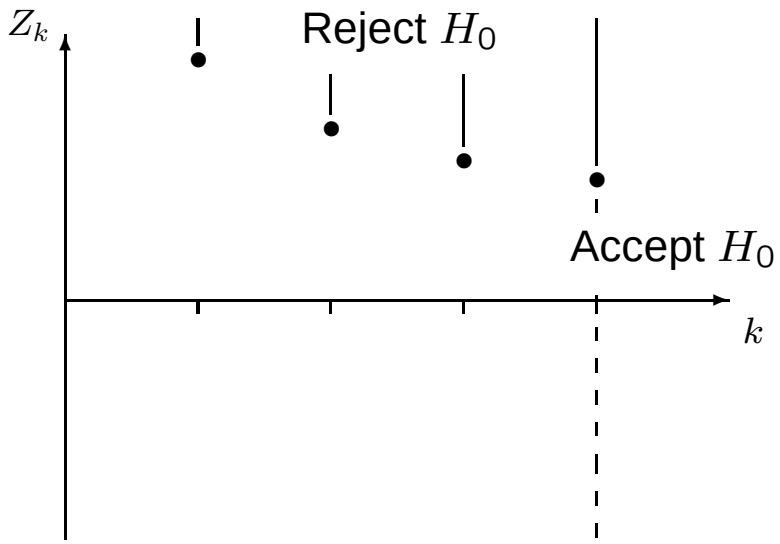
At each analysis k, calculate the Z-statistic Z_k and stop, with rejection of H_0, if $Z_k > c$.

The critical value c is chosen so that the overall probability of a type I error is $\alpha = 0.025$.

Group sequential tests

Other shapes for stopping boundaries are possible.

An O'Brien & Fleming boundary:



The trial stops to reject H_0 at analysis k if $Z_k > c_k$, where the critical values c_k decrease with k.

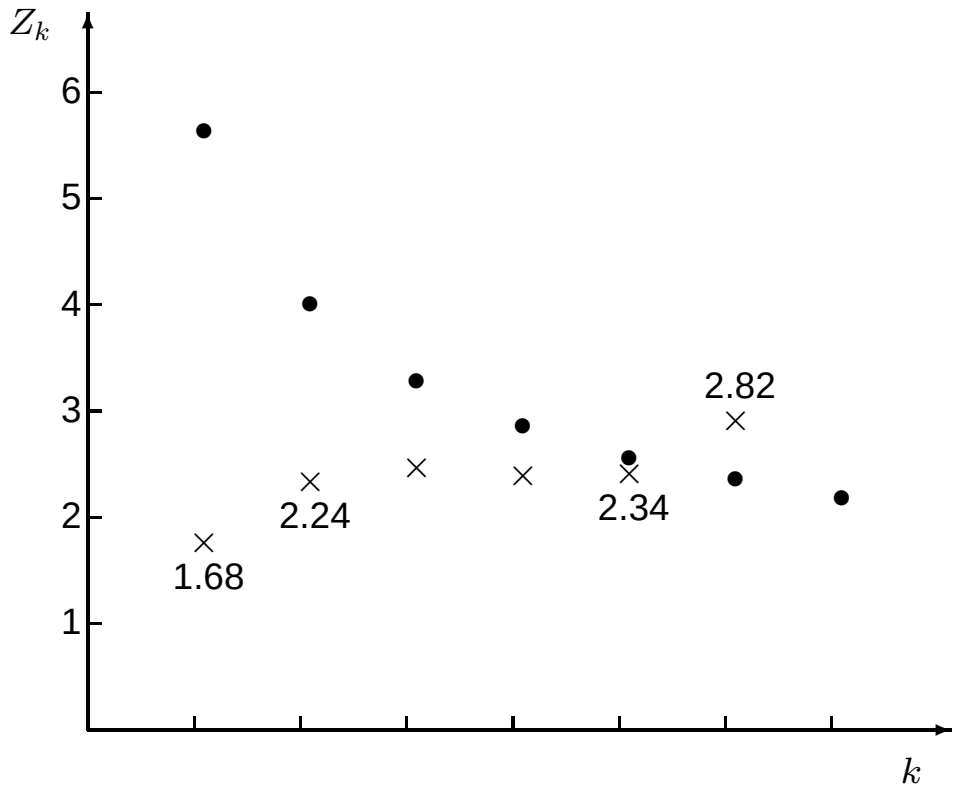
These designs allow early stopping for a positive outcome with proper allowance for multiple analyses.

Example: BHAT trial

DeMets et al. (*Controlled Clinical Trials*, 1984) report on the Beta-Blocker Heart Attack Trial.

This double-blind trial compared propanolol with placebo.

The trial stopped after the 6th of 7 planned analyses.



The BHAT trial

The BHAT trial is an early example of a sequential trial.

The monitoring rule was informal — investigators appear to have been surprised by the success of the new drug and the stopping boundary was created retrospectively.

Nowadays, regulators require sequential analyses to be declared in the protocol *before* a study gets under way.

The U.S. *Federal Register* (1985) published regulations for New Drug Applications. These were elaborated in a Guideline (*FDA*, 1988):

> "all interim analyses . . . should be described in full even if treatment groups were not identified. The need for statistical adjustment because of such analyses should be addressed."

Regulatory requirements

The FDA guidelines were updated in the *Federal Register* (1998) as

"E9 Statistical Principles for Clinical Trials".

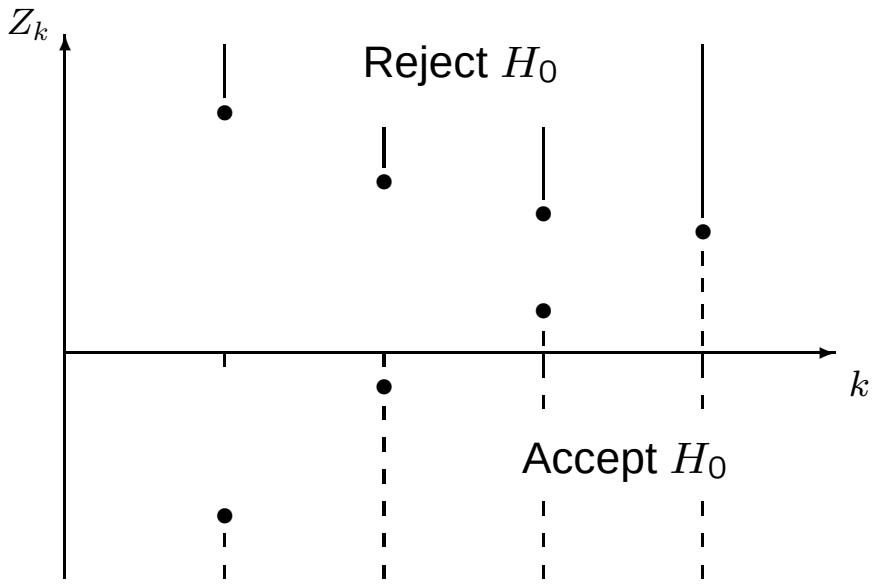
This was prepared under the auspices of the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH).

The document lists recommendations for statistical principles and methodology applied to clinical trials in the pharmaceutical industry.

It advocates use of ***group sequential designs*** and gives detailed recommendations for trial conduct, including ***trial monitoring, interim analysis, early stopping, sample size adjustment*** and the role of an ***independent data and safety monitoring board*** (DSMB).

Group sequential tests: stopping for futility

We can add a lower boundary to stop a trial when there is little chance of it reaching a positive conclusion.



Rosner & Tsiatis (*Statistics in Medicine*, 1989) carried out retrospective analyses of 72 cancer studies of the U.S. Eastern Co-operative Oncology Group.

If group sequential stopping rules had been applied, early stopping would have occurred to accept H_0 in up to 80% of cases (depending on the stopping boundary used).

Scope of group sequential tests

Response type

Normal,

Binary,

Longitudinal data (repeated measures),

Survival data.

Early stopping

For a positive outcome, rejecting H_0,

for a negative outcome, accepting H_0.

Need to specify

Number of interim analyses,

Shape of stopping boundary.

Benefits of group sequential tests

Benefits

Reduce cost and conserve resources,

With a positive outcome, proceed rapidly to the next stage or to submit a New Drug Application.

Requirements

Software – several packages are available.

Overall

For a trial design with 5 analyses, where a fixed sample study would need 100 observations, a group sequential design will require, on average, 60 to 70 observations.

4. Adaptive designs

“Adaptive” methods allow modifications to the design of a clinical trial during the course of the trial.

There are many reasons to consider such modification:

Adaptation to external factors

Changes in the clinical setting or economic background

Following withdrawal of a competing treatment, a smaller treatment effect becomes of clinical interest.

An improved financial position means sponsors can invest more in this trial.

In both these cases, the sample size will need to be increased in order to improve power — but the type I error probability must be protected.

Motivation for adaptive designs

Adaptation to internal factors

Nuisance parameters affecting sample size

In-study estimates of response variance show need for higher sample size to achieve the intended power.

Overall failure rates in a survival study are low: greater accrual and longer follow-up are required.

Sub-group analyses

Treatment appears to benefit a particular sub-group: investigators wish to re-define the target population.

Change of endpoint

An alternative endpoint appears more informative: wish to re-define the primary endpoint.

Pre-specified adaptations

Adaptivity can be incorporated in the study protocol.

Example: Attaining power when sample size depends on unknown parameters

a) Normal response with unknown variance.

Sample size required to give a particular power is proportional to the response variance σ^2.
At interim points, estimate σ^2 from data available thus far and revise the target sample size accordingly.

b) Survival data

Power depends on the total number of failures observed.

As accrual and overall failure rates are seen, the accrual period and length of follow-up can be extended to ensure the target will eventually be met.

Flexible adaptation

Reference: Bauer & Köhne (*Biometrics*, 1994)

Flexible methods must maintain the type I error rate.

To do this, the trial is split into two stages and a P-value is calculated from the data in each stage:

Under H_0: no treatment difference, P-values p_1 and p_2 have their usual distribution (uniform on 0 to 1) ***even if Stage 2 of the trial is re-designed in the light of outcomes in Stage 1.***

This allows p_1 and p_2 to be combined in a test of H_0 with the specified type I error probability, α.

Uses of flexible adaptation

Flexible methods can be used:

To modify treatment definition,

To modify the primary endpoint,

To focus on a particular sub-population of patients,

To increase sample size in order to improve power.

Some caution is needed here since a “positive” outcome rejects the *combined* null hypothesis of the two stages.

E.g., if treatment is modified, you really want to show the treatment *as finally defined* is superior to control — but data come from both versions of the treatment.

This issue can be addressed, but there is sometimes a cost in terms of the total sample size required.

Adaptation to increase power

Methods exist to increase sample size and hence improve power, while maintaining the type I error rate.

It may be reasonable to do this in response to external developments or information on secondary endpoints and safety data within a study.

However:

Responding to data on the primary endpoint during the study is exactly what group sequential tests do. They adjust the final sample size by continuing to sample or terminating the study at each interim analysis.

Well-designed group sequential tests govern the final sample size very efficiently, while achieving type I error and power requirements: there is no need to look beyond classical methods for this purpose.

5. Seamless transition between trial phases

New drugs proceed through phases of testing:

Phase II	Testing on the target population and comparison against control,
Phase IIb	Finding the optimal dose, and comparison against control,
Phase III	Final stage of testing

Typically, each phase starts afresh with its own protocol, regulator’s approval, etc.

This approach allows the new phase to be designed with knowledge of all previous testing.

Changes can be made to the target patient population, the treatment definition, or the way it is administered.

A major disadvantage is time lost waiting between phases.

Seamless transition

Methods used in adaptive designs to combine data from two stages of a single trial can be applied to combine Phase IIb and Phase III trials.

The key null hypothesis to test is that concerning the treatment (dose level, etc.) selected for Phase III — so the statistical methods must do exactly this.

The final analysis can make use of responses in Phase IIb on the version of the treatment chosen for Phase III.

This use of Phase IIb data improves statistical efficiency.

Benefits of this extra efficiency depend on the relative sizes of Phase IIb and Phase III trials: planning the two phases together may also lead to better choice of their sample sizes.

Seamless transition

The major gain:

The big gain in combining two phases of testing comes from eliminating the “white space” between phases.

The requirements to achieve this:

The main practical requirement is a combined protocol, providing rules for treatment choice and all other aspects of the Phase III study, given information from Phase IIb.

If sponsors are to be blinded, these rules must be comprehensive and clear so they can be implemented by an independent monitoring group — this sounds a lot to ask, but experience shows it can be done.

6. Conclusions

1. Group sequential methods offer reductions in sample size and speedier conclusions in all forms of clinical trial.

There is comprehensive theory and methodology and software to help design and implement the methods.

2. Adaptive methods allow modifications during the course of a study.

This can be according to rules set in the protocol or more flexibly, in response to unanticipated events.

3. Combining phases, particularly Phase IIb and Phase III, offers a more rapid procedure with the potential to bring a successful treatment to market sooner.