Defining Good Guidelines for Futility Stopping based on Conditional Power or Predictive Power

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

Department of Mathematical Sciences,
University of Bath

PSI Conference, Amsterdam, 2024

Group Sequential Designs

Group sequential designs allow a clinical trial to be terminated

For efficacy:

When there is overwhelming evidence that the new treatment is effective,

For futility:

When it is clear the trial is unlikely to reach a positive conclusion.

In retrospective analyses of 72 ECOG cancer studies, Rosner & Tsiatis (*Statist. in Med.*, 1989) found that, if group sequential stopping rules had been applied, early stopping (mostly to accept H_0) would have occurred in around 80% of cases.

Many clinical trials have a formal stopping rule for efficacy but informal guidelines for futility stopping. Why are these issues treated differently — and is this a good idea?

Outline of talk

1. Defining an efficacy stopping rule through an error spending function.

2. Using conditional power and predictive power to guide stopping for futility — and why this can be problematic.

3. Error spending designs with an efficacy boundary and a non-binding futility boundary.

4. Efficient guidelines for using conditional power or predictive power in deciding whether to stop for futility.

1. An efficacy stopping rule

Consider a Phase 3 clinical trial comparing a new treatment against a standard.

Let θ denote the "effect size", a measure of the improvement in the new treatment over the standard.

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

Rejecting H_0 allows us to conclude the new treatment is superior.

We allow type I error probability α for rejecting H_0 when it is true.

We specify power $1 - \beta$ as the probability that H_0 should be rejected when $\theta = \delta$.

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

Joint distribution of estimates

Reference: Ch. 11 of *Group Sequential Methods with Applications to Clinical Trials*, Jennison & Turnbull, 2000.

Let $\widehat{\theta}_k$ denote the estimate of θ based on data at analysis k.

The information for θ at analysis k is

$$\mathcal{I}_k = \{\text{Var}(\widehat{\theta}_k)\}^{-1}, \quad k = 1, \ldots, K.$$

Canonical joint distribution of $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$

In many situations, $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$ are approximately multivariate normal,

$$\widehat{\theta}_k \sim N(\theta, \mathcal{I}_k^{-1}), \quad k = 1, \ldots, K,$$

and

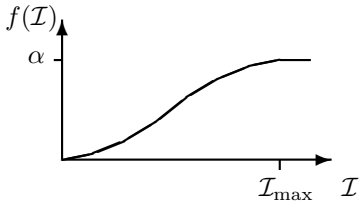
$$\text{Cov}(\widehat{\theta}_{k_1}, \widehat{\theta}_{k_2}) = \text{Var}(\widehat{\theta}_{k_2}) = \mathcal{I}_{k_2}^{-1} \quad \text{for } k_1 < k_2.$$

Error spending tests (J & T, Ch. 7)

When the sequence $\mathcal{I}_1, \mathcal{I}_2, \ldots$ is unpredictable, a group sequential design must adapt to observed information levels.

Lan & DeMets (1983) introduced "error spending" tests of H_0: $\theta = 0$ against $\theta \neq 0$.

Maximum information design with error spending function $f(\mathcal{I})$



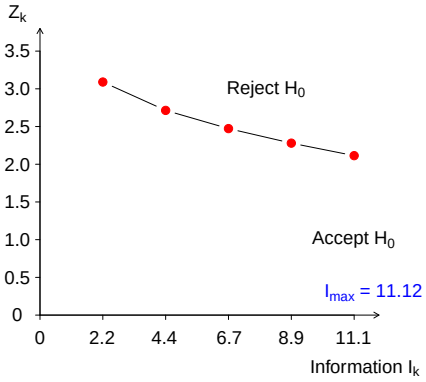
The boundary at analysis k is set to give cumulative type I error probability (under $\theta = 0$) equal to $f(\mathcal{I}_k)$.

If the target information, $\mathcal{I}_{\max}$, is reached without rejecting H_0, then H_0, is accepted.

Trial with efficacy boundary only

A test with 5 planned analyses, type I error probability $\alpha = 0.025$, power 0.9 if $\theta = \delta = 1$, and type I error spending function

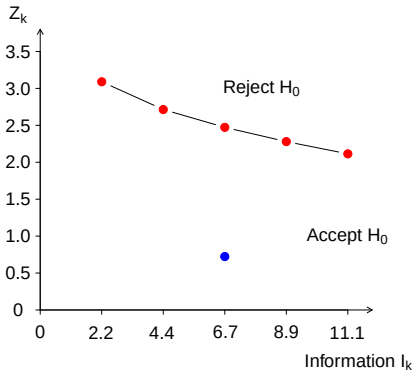
$$f(\mathcal{I}) = \min\{(\mathcal{I}/\mathcal{I}_{\max})^2, 1\}\,\alpha.$$



Fixed sample size design has $\mathcal{I} = 10.5$.

2. Using conditional power

Suppose the trial has reached analysis 3, $\hat{\theta}_3 = 0.28$ and $Z_3 = 0.72$.



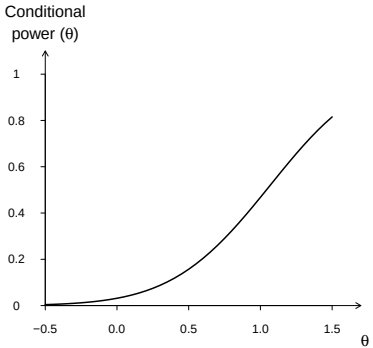
One may ask

"How likely is it that the trial's final outcome will be positive?"

Using conditional power

We can compute the “conditional power function”,

$$P_\theta\{H_0 \text{ will be rejected at analysis 4 or 5} \mid Z_3 = 0.72\}.$$

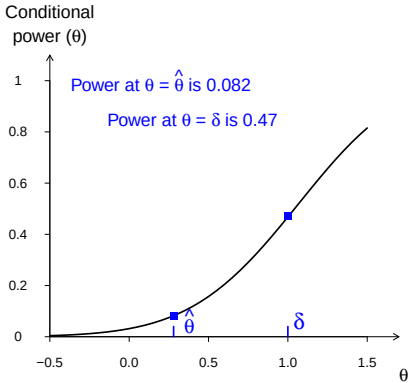


However, we do not know the true value of θ.

Using conditional power

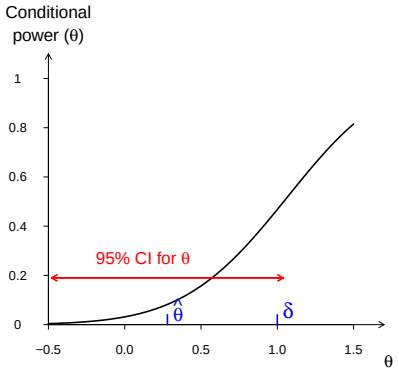
One may focus on conditional power if θ equals the current estimate, $\hat{\theta}_3 = 0.28$.

Or, one might focus on conditional power if $\theta = \delta = 1$, the value used in the power calculation.



Using conditional power

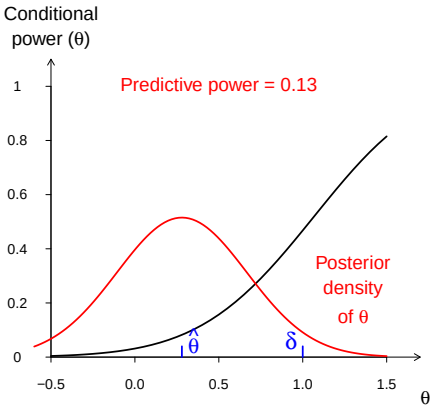
It is important to remember that an interim estimate of θ has high variance.



Adopting a Bayesian approach, one can integrate conditional power over a posterior distribution to obtain a “predictive power”.

Using predictive power

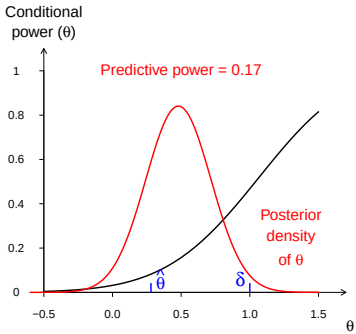
It is common to assume a flat (improper) prior for θ in calculating predictive power.



In this case the posterior distribution of θ is $N(0.28,\, 0.39^2)$.

Using predictive power

Given the high variance of the interim estimate $\hat{\theta}_3$, the choice of prior can have a significant impact on the predictive power.



Under the more reasonable prior $\theta \sim N(0.6,\, 0.3^2)$, the posterior distribution of θ is $\theta \mid \hat{\theta}_3 \sim N(0.48,\, 0.24^2)$, and predictive power rises from 0.13 to 0.17.

Using predictive power

Once you have calculated your chosen conditional power or predictive power, the question remains:

How high should the conditional probability of success be to justify continuation of the trial?

One needs to balance

- The benefits from saving resources in this study and moving on to conduct trials for other promising therapies,
- The risk of stopping the current trial prematurely when it would have gone on to produce a positive result.

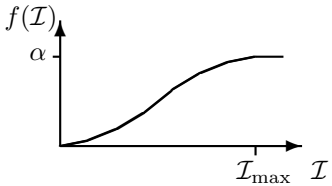
Decision making is hard when conditional power is low but there is a non-negligible chance the trial may still succeed.

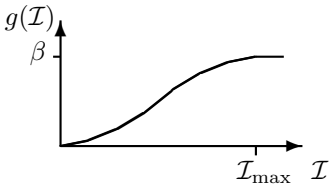
3. Error spending tests

For a one-sided test of H_0: $\theta \leq 0$ against $\theta > 0$ with

Type I error probability α at $\theta = 0$ and Type II error probability β at $\theta = \delta$,

and both efficacy and futility boundaries, we need two error spending functions.





Type I error probability α is spent according to the function $f(\mathcal{I})$, and type II error probability β (under $\theta = \delta$) according to $g(\mathcal{I})$.

Treating the futility boundary as "non-binding", we calculate Type I error probabilities ignoring the futility boundary.

Error spending tests

Recall, we want a group sequential test of H_0: $\theta \leq 0$ vs $\theta > 0$ with

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

$$P_{\theta=\delta}\{\text{Accept } H_0\} = \beta,$$

Analyses at $\mathcal{I}_k = (k/K)\,\mathcal{I}_{\max}, \quad k = 1, \ldots, K.$

If we specify α, β, δ, K and $\mathcal{I}_{\max}$, we can find the stopping rule that minimises

$$\sum_i w_i E_{\theta_i}(\mathcal{I}) \quad \text{or} \quad \int w(\theta) E_\theta(\mathcal{I})\, d\theta.$$

See:

Barber & Jennison (*Biometrika*, 2002),

Öhrn (*PhD thesis, University of Bath*, 2011),

Jennison & Turnbull (*Kuwait J. Science*, 2013).

Error spending tests

Barber & Jennison (2002) and Öhrn (2011) observe that group sequential tests with error spending functions of the form

$$f(\mathcal{I}) \;=\; \min\{(\mathcal{I}/\mathcal{I}_{\max})^{\rho_1}, 1\}\,\alpha \quad \text{(type I error)}$$

and

$$g(\mathcal{I}) \;=\; \min\{(\mathcal{I}/\mathcal{I}_{\max})^{\rho_2}, 1\}\,\beta \quad \text{(type II error)}$$

have high efficiency for a variety of optimality criteria.

Values of ρ_1 and ρ_2 determine $\mathcal{I}_{\max}$ and, hence, the trial's maximum sample size.

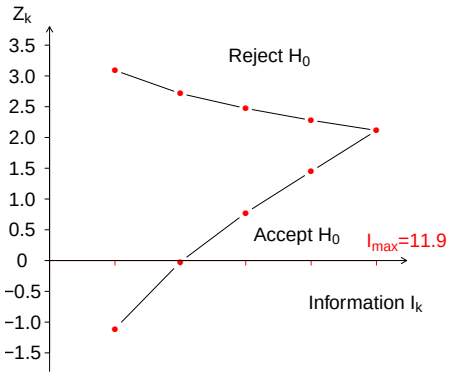
The resulting designs are efficient for this value of $\mathcal{I}_{\max}$.

The monitoring committee can treat the (non-binding) futility boundary as a guideline, allowing them to consider safety data or secondary endpoints in deciding whether to stop for futility.

A non-binding futility boundary

A design with 5 planned analyses, type I error probability $\alpha = 0.025$, power 0.9 when $\theta = \delta = 1$, and type I and II error spending functions

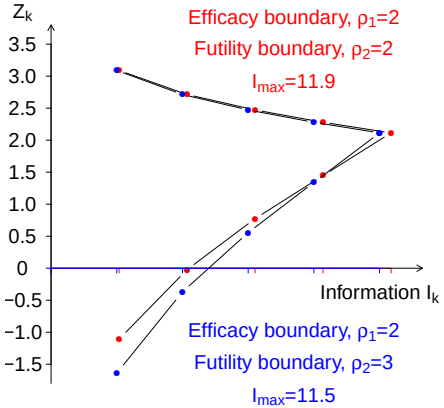
$$f(\mathcal{I}) = \min\{(\mathcal{I}/\mathcal{I}_{\max})^2, 1\}\,\alpha, \quad g(\mathcal{I}) = \min\{(\mathcal{I}/\mathcal{I}_{\max})^2, 1\}\,\beta.$$



A non-binding futility boundary

Contrast: A test with type I and type II error spending functions

$$f(\mathcal{I}) = \min\{(\mathcal{I}/\mathcal{I}_{\max})^2, 1\}\,\alpha, \quad g(\mathcal{I}) = \min\{(\mathcal{I}/\mathcal{I}_{\max})^3, 1\}\,\beta.$$



Savings from early stopping

Tests with $\alpha = 0.025$, $1 - \beta = 0.9$ and 5 equally spaced analyses.

Values of $\mathcal{I}_{\max}$ and $E_\theta(\mathcal{I})$, expressed as a percentage of $\mathcal{I}_{fix}$.

		$E_\theta(\mathcal{I})$			
Design	$\mathcal{I}_{\max}$	$\theta=0$	$\theta=0.5$	$\theta=1.0$	$\theta=1.5$
E: $\rho_1 = 2$ only	106	105.2	96.7	70.5	46.8
E: $\rho_1 = 2$, F: $\rho_2 = 2$	113	59.2	80.9	70.6	48.2
E: $\rho_1 = 2$, F: $\rho_2 = 3$	109	64.1	83.0	70.1	47.5

E: Efficacy boundary, F: Non-binding futility boundary

Comparing $\rho_2 = 2$ and $\rho_2 = 3$: With $\rho_2 = 2$, the maximum sample size is larger but expected sample size at low values of θ is lower.

The number of analyses and design parameters ρ_1 and ρ_2 can be chosen to give acceptable values of $\mathcal{I}_{\max}$ and $E_\theta(\mathcal{I})$.

4. Conditional & predictive power

The futility boundaries of the error spending designs we have just presented can be described in terms of conditional or predictive power.

Error spending design with: $\rho_1 = 2$ (efficacy boundary), $\boldsymbol{\rho_2 = 2}$ **(futility boundary)**

	Analysis, k			
	1	2	3	4
Conditional power under $\theta = \delta$	0.55	0.41	0.51	0.39
Conditional power under $\theta = \hat{\theta}^{(k)}$	0.00027	0.037	0.096	0.13
Predictive power, improper, flat prior	0.021	0.072	0.14	0.16
Predictive power, prior $\theta \sim N(0.6, 0.3^2)$	0.16	0.14	0.19	0.17

Conditional & predictive power

The futility boundaries of the error spending designs we have just presented can be described in terms of conditional or predictive power.

Error spending design with: $\rho_1 = 2$ (efficacy boundary), $\boldsymbol{\rho_2 = 3}$ **(futility boundary)**

	Analysis, k			
	1	2	3	4
Conditional power under $\theta = \delta$	0.48	0.33	0.39	0.30
Conditional power under $\theta = \hat{\theta}^{(k)}$	0.000005	0.0087	0.043	0.084
Predictive power, improper, flat prior	0.0049	0.028	0.077	0.11
Predictive power, prior $\theta \sim N(0.6, 0.3^2)$	0.10	0.080	0.12	0.12

Conditional & predictive power

Suppose a trial steering committee favours the futility boundary given by a particular error spending design.

However, the monitoring committee insists on looking at conditional power or predictive power when deciding whether to stop for futility.

The steering committee can present thresholds for conditional power or predictive power that correspond to their error spending futility boundary as the default values for futility stopping, e.g., for $\rho_2 = 2$,

Stop if conditional power under $\theta = \delta$ is less than 0.45

or

Stop if predictive power for prior $\theta \sim N(0.6, 0.3^2)$ is less than 0.16.

Any departure from a rule with these thresholds should be justified by additional information on safety or secondary endpoints.

Conclusions

It is common practice for trials to take an informal approach to stopping for futility.

Decision making is often guided by conditional power calculations.

Just how one should use “conditional power” or “predictive power” in deciding whether to stop a trial is unclear.

We can create a group sequential design with a non-binding futility boundary.

The ρ-family of error spending designs provides efficient procedures.

The futility boundary of such a design can be described in terms of conditional power under $\theta = \delta$ or predictive power under a specific prior.

The monitoring committee can be provided with this futility boundary as a guide — but still use their discretion in deciding when to stop the trial.

References

Barber, S. and Jennison, C. (2002). Optimal asymmetric one-sided group sequential tests. *Biometrika*, **89**, 49–60.

Jennison, C., and Turnbull, B. W. (1999). Group Sequential Methods with Applications to Clinical Trials. *CRC Press*.

Jennison, C., and Turnbull, B. W. (2013). Interim monitoring of clinical trials: decision theory, dynamic programming and optimal stopping. *Kuwait Journal of Science*, **40**, 43–59.

Lan, K.K.G. and DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, **70**, 659–663.

Öhrn, C.F. (2011). *Group Sequential and Adaptive Methods — Topics with Applications to Clinical Trials.* Chapter 4: Group sequential designs with non-binding futility boundaries. PhD thesis, University of Bath.

Rosner, G.L. and Tsiatis, A.A. (1988). The impact that group sequential tests would have made on ECOG clinical trials. *Statistics in Medicine*, **8**, 505–516.

Savings from early stopping

Tests with $\alpha = 0.025$, $1 - \beta = 0.9$ and **2** equally spaced analyses.

Values of $\mathcal{I}_{\max}$ and $E_\theta(\mathcal{I})$, expressed as a percentage of $\mathcal{I}_{fix}$.

		$E_\theta(\mathcal{I})$			
Design	$\mathcal{I}_{\max}$	$\theta = 0$	$\theta = 0.5$	$\theta = 1.0$	$\theta = 1.5$
E: $\rho_1 = 2$ only	103	102.2	97.9	80.5	59.68
E: $\rho_1 = 2$, F: $\rho_2 = 2$	106	70.7	89.2	80.8	60.7
E: $\rho_1 = 2$, F: $\rho_2 = 3$	104	75.5	91.5	80.4	60.0

E: Efficacy boundary, F: Non-binding futility boundary

Savings from early stopping

Tests with $\alpha = 0.025$, $1 - \beta = 0.9$ and **3** equally spaced analyses.

Values of $\mathcal{I}_{\max}$ and $E_\theta(\mathcal{I})$, expressed as a percentage of $\mathcal{I}_{fix}$.

		$E_\theta(\mathcal{I})$			
Design	$\mathcal{I}_{\max}$	$\theta = 0$	$\theta = 0.5$	$\theta = 1.0$	$\theta = 1.5$
E: $\rho_1 = 2$ only	104	103.6	97.1	75.0	52.3
E: $\rho_1 = 2$, F: $\rho_2 = 2$	109	64.4	84.8	75.3	53.5
E: $\rho_1 = 2$, F: $\rho_2 = 3$	106	69.5	86.9	74.8	52.8

E: Efficacy boundary, F: Non-binding futility boundary