UCDAVIS HEALTH

Topics in Designing Phase III Clinical Trials

CLINICAL AND TRANSLATIONAL SCIENCE CENTER

Susan Stewart, Ph.D.

Division of Biostatistics

The UC Davis CTSC receives support from the NIH National Center for Advancing Translational Sciences (award TR001860).

Learning objectives

- Understand the objectives and interpretation of **superiority**, **equivalence**, and **non-inferiority** trials
- Know how to approach sample size estimation for **equivalence** and **non-inferiority** trials
- Appreciate the appropriate use of **interim** analyses and stopping rules in clinical trials

Clinical trials classification

- Active control trials: test drug is concurrently compared to an known active drug.
- Possible primary objectives
  - To demonstrate **superiority** of the test drug over the active control
  - To show that the test drug is similar to the active control: **equivalence** trial
  - To verify that the test drug is no worse than the active control: **non-inferiority** trial

Superiority trials

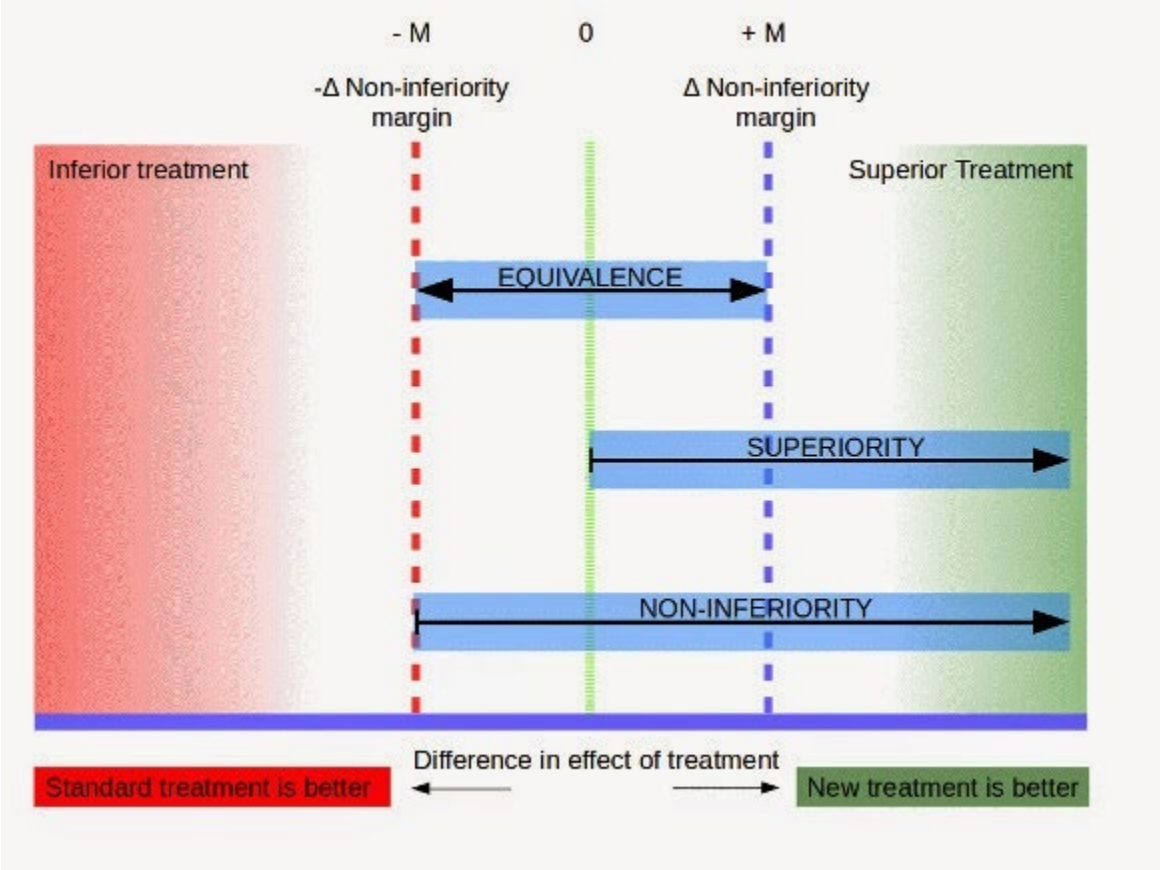
- Primary objective (Chow & Liu): showing that the investigational agent is superior to the comparative agent
  - First, prove that there is a statistically significant difference between the effects of the agents
  - Then show that the difference is in the correct direction

Design and Analysis of Clinical Trials (3rd Ed.)
Chow & Liu, Wiley, 2014

Superiority trials

- The **null hypothesis** (H_0) assumes that there is no difference in outcome between the two groups.
- The **alternative hypothesis** (H_A) assumes that one group has a more favorable outcome than the other.
- The **research hypothesis** is usually the alternative hypothesis.

Clinical Trials Classification



http://www.pvanuden.com/2015/04/equivalence-vs-non-inferiority-vs.html

Superiority trial: hypotheses

- μ_I = intervention group mean
- μ_C = control group mean
- H_0: $\mu_I = \mu_C$
- H_A: $\mu_I \neq \mu_C$
- Reject H_0 if the (1-α) 2-sided confidence interval for $\mu_I - \mu_C$ does not include zero.

Test statistic to compare means: two independent samples

$$z = \frac{\bar{x} - \bar{y}}{\sigma\sqrt{2/n}}$$

- $\bar{x}$ = intervention group mean
- $\bar{y}$ = control group mean
- σ^2 = common variance in each group
- n = sample size in each group

Wittes, Epidemiol Rev, 2002; 24(1):39-53 (eq. 1)

Equivalence and non-inferiority trials

- Equivalence
  - Null hypothesis: effects of the treatments are substantively different
  - Alternative hypothesis: difference between the treatment effects is within the equivalence limits
- Non-inferiority
  - Null hypothesis: the new treatment is substantively less effective than the control
  - Alternative hypothesis: the new treatment is not substantively worse than the control

Requirements

- The control must be demonstrably better than placebo or no treatment
  - Used with dose & formulation proven effective
  - Established for the indication being studied
  - Studies demonstrating benefit should be recent
  - Evidence of benefit must be available to estimate event rate in control group
- Response variable must be sensitive to postulated effects of control and intervention
  - Assay sensitivity: ability to show a difference if one exists

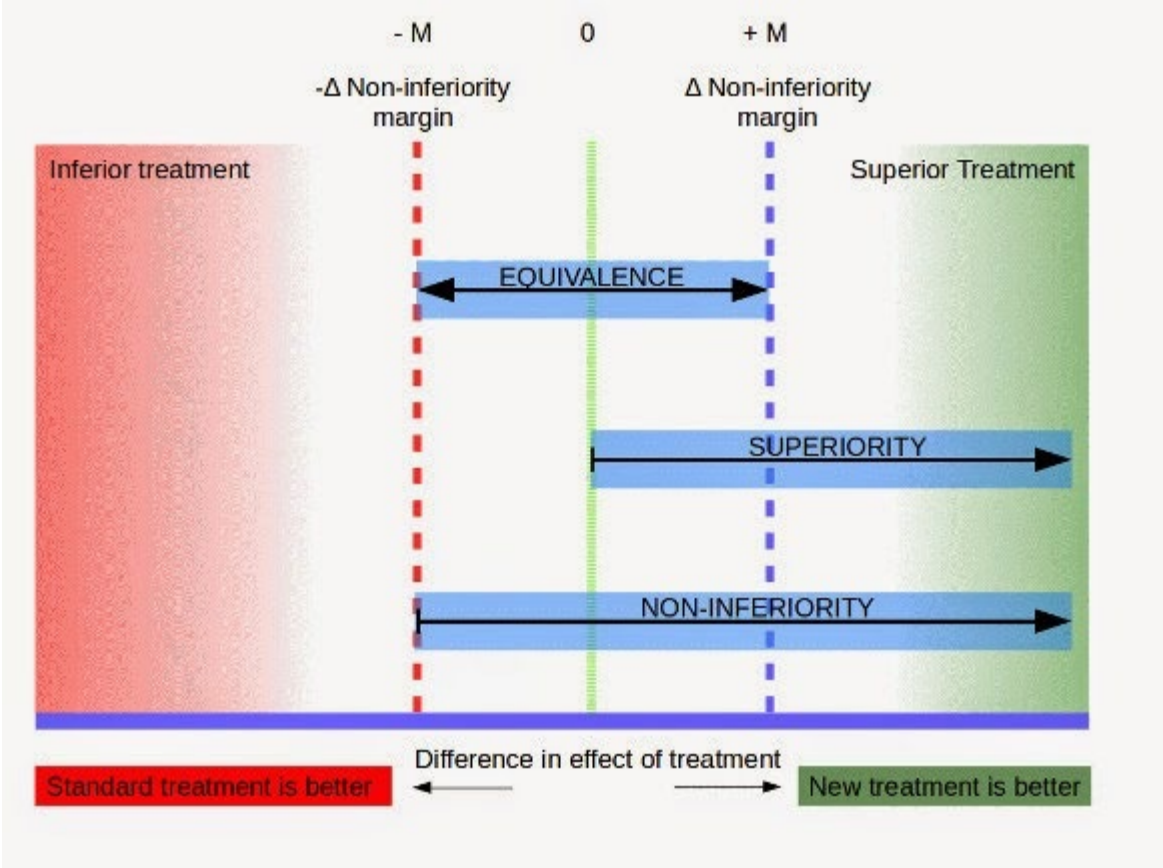
Equivalence limits

- Need to decide how close the effect of the new treatment must be to the control in order to be considered equivalent
  - Or how much smaller the effect can be and still be considered non-inferior
- Can be controversial

Equivalence trials

- The **null hypothesis** (H_0) assumes that the intervention group has an different outcome than the control group.
- The **alternative hypothesis** (H_A) assumes that the outcome of the intervention group is the same as that of the control group, within the **equivalence interval**.

Clinical Trials Classification



http://www.pvanuden.com/2015/04/equivalence-vs-non-inferiority-vs.html

Equivalence trial: hypotheses

- μ_I = intervention group mean
- μ_C = control group mean
- δ_1, δ_2 = equivalence limits
- H_0: $\mu_I - \mu_C \leq \delta_1$ or $\mu_I - \mu_C \geq \delta_2$
- H_A: $\delta_1 < \mu_I - \mu_C < \delta_2$
- Reject H_0 if the (1-2α) 2-sided confidence interval for $\mu_I - \mu_C$ is entirely within the interval $[\delta_1, \delta_2]$

Two one-sided tests (TOST) method

- H_{01}: $\mu_I - \mu_C \leq \delta_1$
- H_{A1}: $\mu_I - \mu_C > \delta_1$

- H_{02}: $\mu_I - \mu_C \geq \delta_2$
- H_{A2}: $\mu_I - \mu_C < \delta_2$

- If both H_{01} and H_{02} are rejected at level α (1-sided), the two treatments are considered equivalent.

Schuirmann, J Pharmocokin Biopharm15:657-680

Example: lymphedema trial

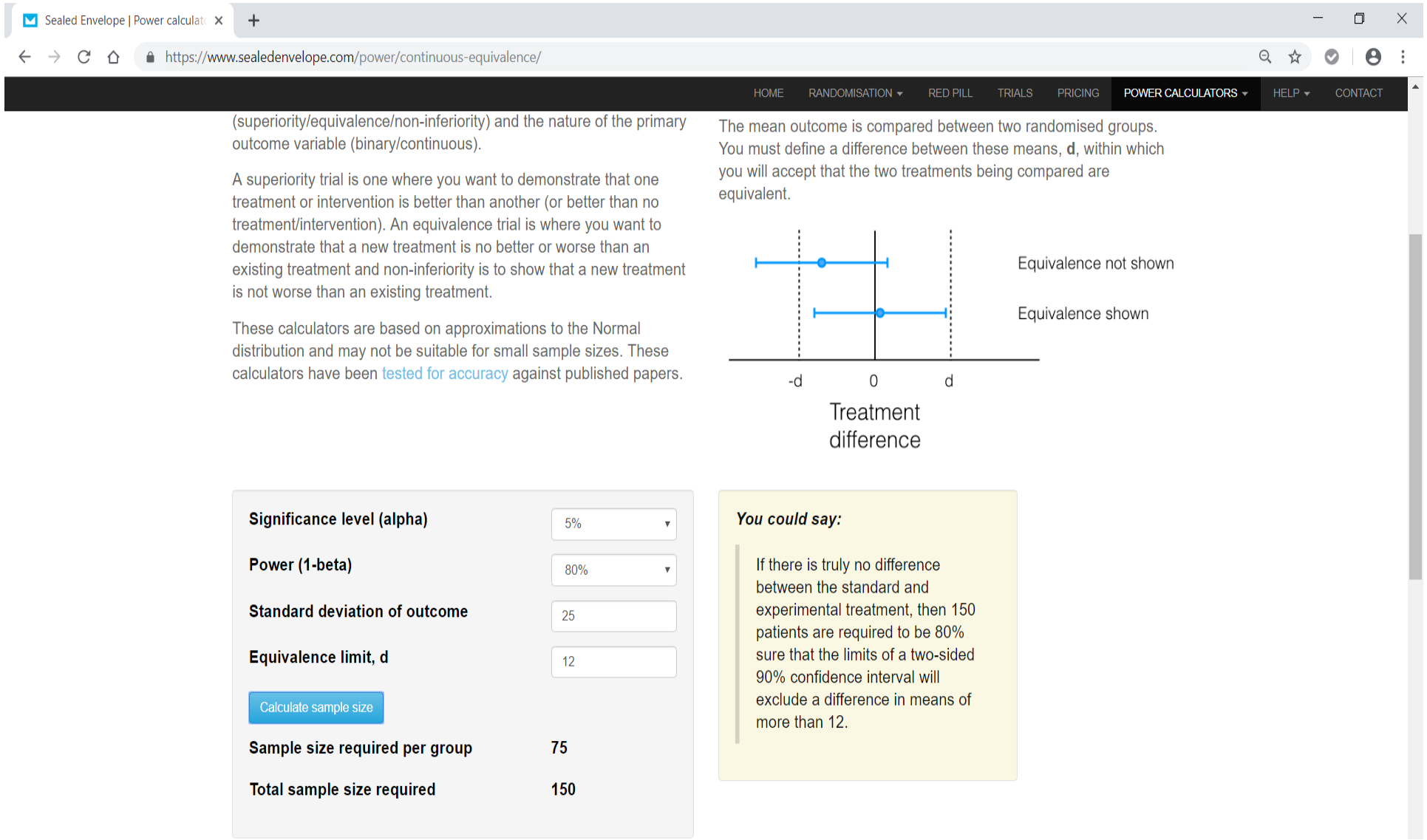
- Randomized, single-blind, equivalence trial testing whether physical therapy is equally effective in treatment of arm lymphedema in breast cancer patients if it includes manual lymphatic drainage or not.
- 4 weeks of treatment with 6-month follow-up
- Primary outcome: percentage volume reduction of arm lymphedema from baseline to 7 months
- Data analysis: ANCOVA with baseline value as covariate

Tambour et al., BMC Cancer 2014; 14:239

Lymphedema trial: assumptions for TOST sample size calculation

- **Level:** 5% (2-sided), or 5% for each 1-sided test
- **Power:** 80%
- **Outcome measure:** % change in lymphedema
- **Equivalence interval:** -12% to +12%
  - "Based on clinically and statistically important differences as well ethical criteria, cost, and feasibility"
- **True mean difference:** 0
- **Standard deviation:** 25%
  - No reason given for this SD
- n=76 patients per study arm=152 total
- Actual planned sample size=160 total

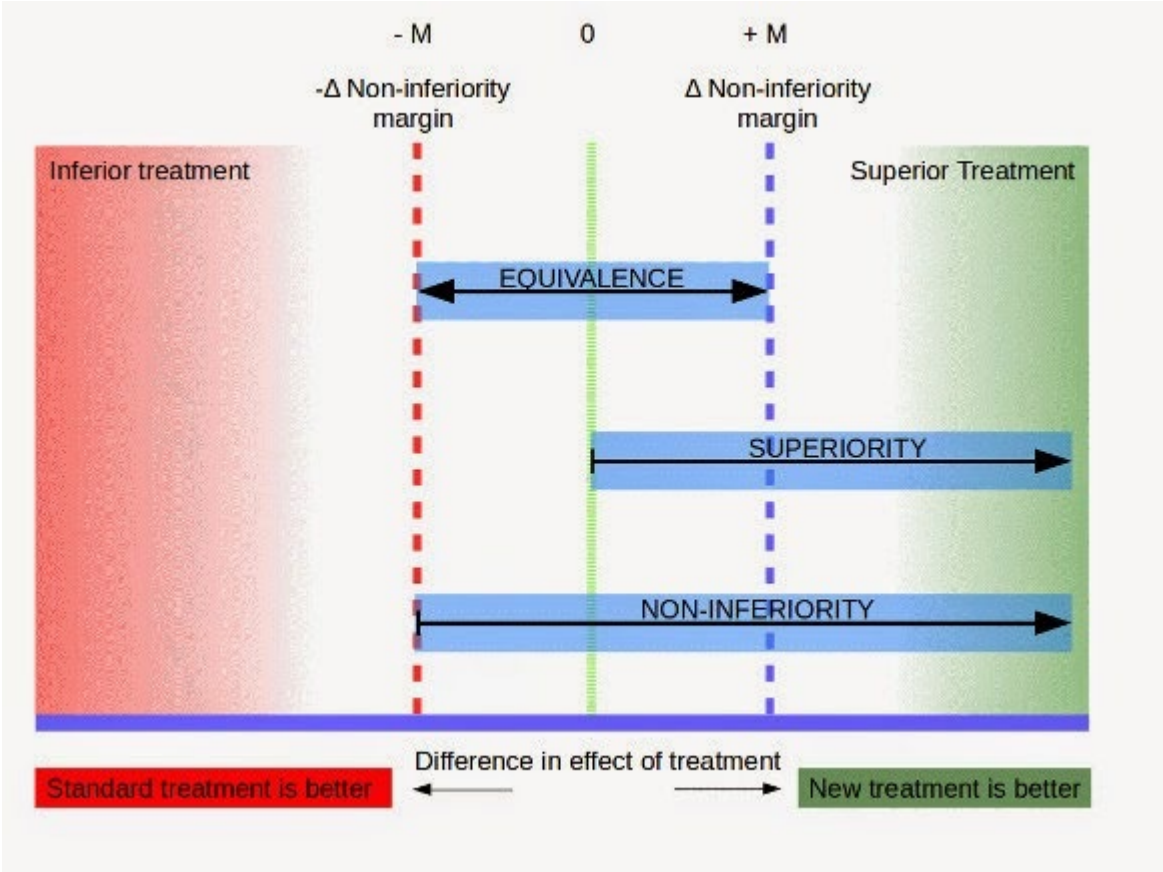
SealedEnvelope.com Equivalence Power Calculation



Non-inferiority trials

- The **null hypothesis** (H_0) assumes that the intervention group has an inferior outcome to the control group.
- The **alternative hypothesis** (H_A) assumes that the outcome of the intervention group is not inferior to that of the control group, within a certain margin.
  - Called the **margin of indifference** or margin of non-inferiority.

Clinical Trials Classification



http://www.pvanuden.com/2015/04/equivalence-vs-non-inferiority-vs.html

Non-inferiority trial: hypotheses

- μ_I = intervention group mean
- μ_C = control group mean
- δ = non-inferiority margin
- H_0: $\mu_I \leq \mu_C - \delta$
- H_A: $\mu_I > \mu_C - \delta$
- Reject H_0 if the (1-α) 2-sided confidence interval for $\mu_I - \mu_C$ is entirely above $-\delta$

Test statistic for non-inferiority trial

$$z = \frac{\bar{x} - \bar{y} - (-\delta)}{\sigma\sqrt{2/n}}$$

- $\bar{x}$ = intervention group mean
- $\bar{y}$ = control group mean
- σ^2 = common variance in each group
- n = sample size in each group
- δ = non-inferiority margin

Non-inferiority trial

- For a continuous variable, the sample size is the same as for a superiority trial, where the detectable difference is the same as the margin of non-inferiority.

Example: Genetic counseling for breast cancer patients

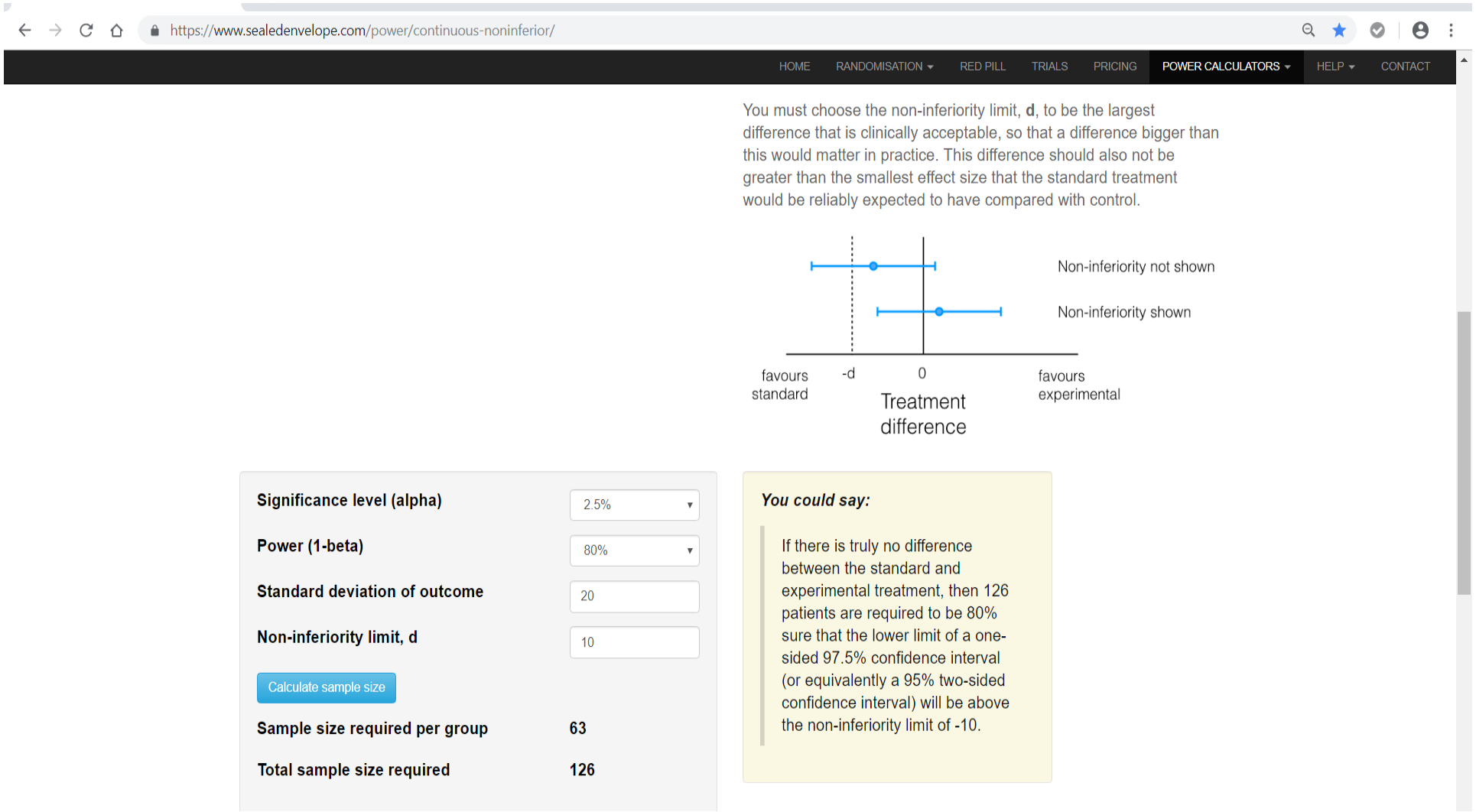
- Randomized non-inferiority trial testing the effect of a brochure vs. in-person counseling about treatment-focused genetic testing on various psycho-social variables and uptake of testing
- Primary outcome: decisional conflict
- Data analysis: linear regression for each outcome measured at 12 months, adjusting for baseline scores

Watts et al., BMC Cancer 2012; 12:320

Genetic counseling: assumptions for sample size calculation

- **Level:** 5% (2-sided)
- **Power:** 80%
- **Outcome measure:** Decisional Conflict Scale
- **Non-inferiority margin:** -10 units
  - Should be +10 units because higher values are bad
  - Corresponds to only 1 of 10 items answered “no”
- **True mean difference:** 0
- **Standard deviation:** 20
  - Deemed “conservative”
- n=64 patients per study arm=128 total
- Actual planned sample size=140 total

SealedEnvelope.com Non-inferiority Power Calculation



Decision for early termination

- Major valid reasons for early termination
  - Serious adverse effects in intervention group
  - Greater than expected beneficial effects
  - Statistically significant difference by the end of the study is unlikely (futility)
  - Problems conducting the study are severe and cannot be corrected
    - Logistical or data quality problems
    - Participant recruitment far behind
  - Question posed by the trial is no longer important

Interim analysis

- Definition: the statistical analysis of results while they are still accumulating

Ludbrook, BMC Medical Research Methodology 2003; 3:15.

Repeated testing for significance

- Monitoring response variables may involve repeated significance testing of accumulating data
- Problem: if the null hypothesis is true and multiple tests are made at the same level of significance alpha, the overall probability of rejecting the null hypothesis > alpha
  - The probability of Type I error will be too high

Adaptive designs

- Purpose (Chow): to give the investigator the flexibility to identify signals or trends of the test treatment without undermining the validity and integrity of the study
- Attractive features
  - Reflects real world medical practice
  - Ethical with respect to safety and efficacy
  - Flexible and efficient
- Concerns
  - Is the p-value or confidence interval correct?
  - Does the trial still answer the original question?

Group sequential methods

- Developed to address the problem of repeated testing
- Assign a critical value (boundary) for the test statistic at each interim analysis
- Ad hoc methods
  - Use a critical value of z=2.6 for interim and final analyses
  - Haybittle-Peto: Use a large critical value, say z=3.0, for interim analyses and use the usual critical value (1.96) at the final analysis
  - Problem: Type I error level is not guaranteed

Calendar time and information time

- Information time is the proportion of the maximum information that has been obtained at a given point in calendar time.
  - Maximum information is obtained when all participants have completed the study.
- At time t, information time is:
  - Time-to-event study: (number of events at time t) / (expected total number of events)
  - Otherwise (generally): (number of participants who have completed the study by time t) / (total planned number of participants)

Group sequential methods

- Assume that interim analyses are conducted at equal information times 1/K, 2/K, etc., for K tests
- Pocock
  - Uses the same critical value for each analysis
  - Critical value is determined so that the probability of Type I error = α if all K tests are performed
- O'Brien-Fleming
  - Uses larger critical values for earlier analyses
  - Critical value =$Z*\sqrt{K/i}$ for the i^{th} analysis, where Z* is chosen so that the probability of Type I error = α if all K tests are performed

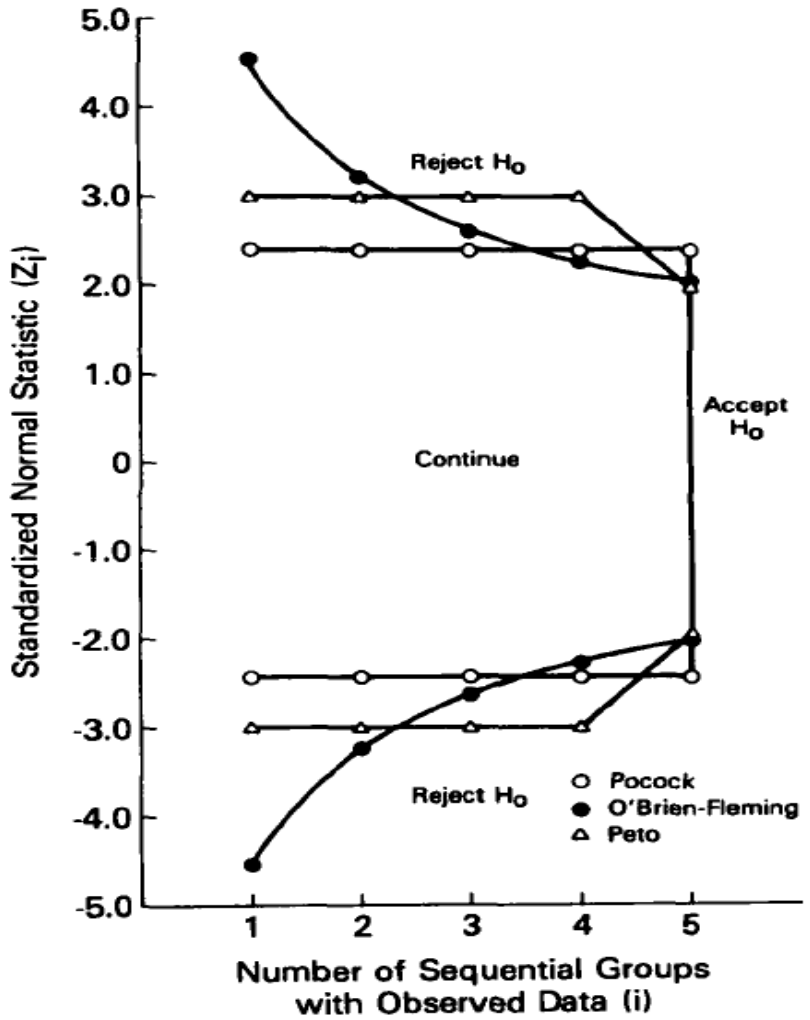


Figure 1. Two-sided 0·05 group sequential boundaries for Pocock, O'Brien–Fleming, and Peto–Haybittle methods for five planned analyses

Stat Med 1994; 13:1341-52

Flexible group sequential methods: spending functions

- Limitations of group sequential methods discussed so far
  - Need to specify number of interim analyses in advance
  - Interim analyses must be evenly spaced in information time
- However, it is possible to specify directly how much of alpha will be allocated to each interim analysis
- Advantage of alpha spending function: number and time of analyses need not be specified in advance
- Can also include beta spending function for futility

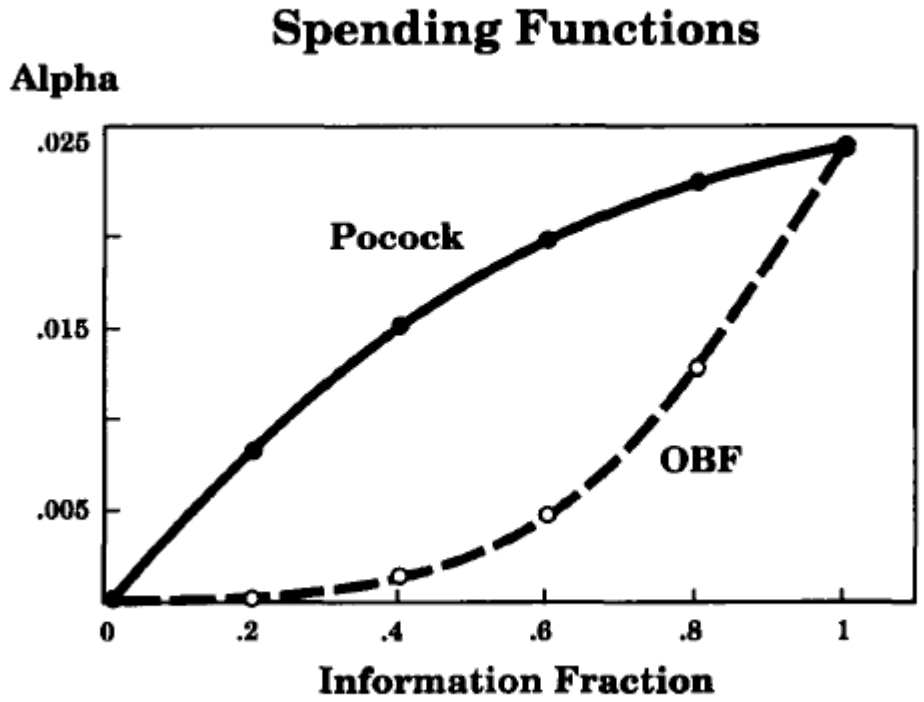


Figure 3. One-sided 0·025 alpha spending functions for Pocock and O'Brien–Fleming type boundaries

Stat Med 1994; 13:1341-52

Applications of group sequential boundaries

- Can be applied to any test statistic that:
  - Can be standardized with a normal distribution (turned into a z-value) and
  - Has independent increments of information between tests
- This includes comparisons of
  - Survival curves (log rank statistic)
  - Means
  - Proportions
  - Slopes

Asymmetric boundaries

- For a superiority trial, the interim analysis boundaries in the direction of benefit may be extreme
  - The benefit must be really, really great to stop the trial early
- However, the boundaries in the direction of harm may be less extreme
  - May not need evidence that harm is really, really great in order to stop the trial early

Sequential Design and Analysis with SAS

- At each stage, run the Use PROC SEQDESIGN to determine boundaries for rejecting and/or accepting the null hypothesis at each stage
- Analysis and use PROC SEQTEST to determine whether to continue to the next stage
- Reference: https://support.sas.com/resources/papers/proceedings09/311-2009.pdf

Example: Genetic Counseling

- Test of intervention to promote genetic counseling among low-income women at high risk for familial breast cancer
  - PI: Rena Pasick (UCSF); R01 CA129096
  - Original sample size: n=144 total—designed to detect 30% vs. 10% uptake of genetic counseling
  - Interim analysis planned with ADDPLAN 6 (Aptiv Solutions, Weston VA) conducted at n=72
  - Interim intervention effect: 35% vs. 5% ($p<0.001$, 1-sided).
  - Trial stopped for efficacy at n=88

Am J Public Health 2016; 106(10):1842-1848

```
ods graphics on;
*Plot of reject/accept boundaries with horizontal axis=sample size;
proc seqdesign
            plots=boundary(hscale=samplesize)
            ;
*1-sided O'Brien-Fleming Design with 2 stages;
*Stop to reject the null hypothesis;
  OneSidedOBrienFleming: design nstages=2
                  method(alpha)=obf
                                alpha=0.025
                                beta=0.20
                                alt=upper stop=reject
                  ;
*Test is comparison of proportions;
*Null proportion=0.1, alternative=0.3;
samplesize model=twosamplefreq(nullprop=0.1
            prop=0.3 test=prop);
*Output boundaries to use in testing;
ods output boundary=bound_prop;
run;
ods graphics off;
run;
```

One-Sided O'Brien-Fleming Design

- Stage 1
  - Enroll 30 participants in each group
  - Compute z-statistic for difference in proportions
  - If $z \geq 2.79651$, reject H_0 and stop
  - If $z < 2.79651$, continue to stage 2
- Stage 2
  - Enroll 30 participants in each group
  - Compute z-statistic for difference in proportions
  - If $z \geq 1.97743$, reject H_0 and stop
  - If $z < 1.97743$, accept H_0 and stop
- Sample size is ~1% larger than fixed sample size design

SAS Output

Boundary Information (Standardized Z Scale)
Null Reference = 0

Stage	Information Level Proportion	Information Level Actual	Information Level N	Alternative Reference Upper	Boundary Values Upper Alpha
1	0.5000	98.87478	59.32487	1.98872	2.79651
2	1.0000	197.7496	118.6497	2.81247	1.97743

Sample Size Summary

Test	Two-Sample Proportions
Null Proportion	0.1
Proportion (Group A)	0.3
Test Statistic	Z for Proportion

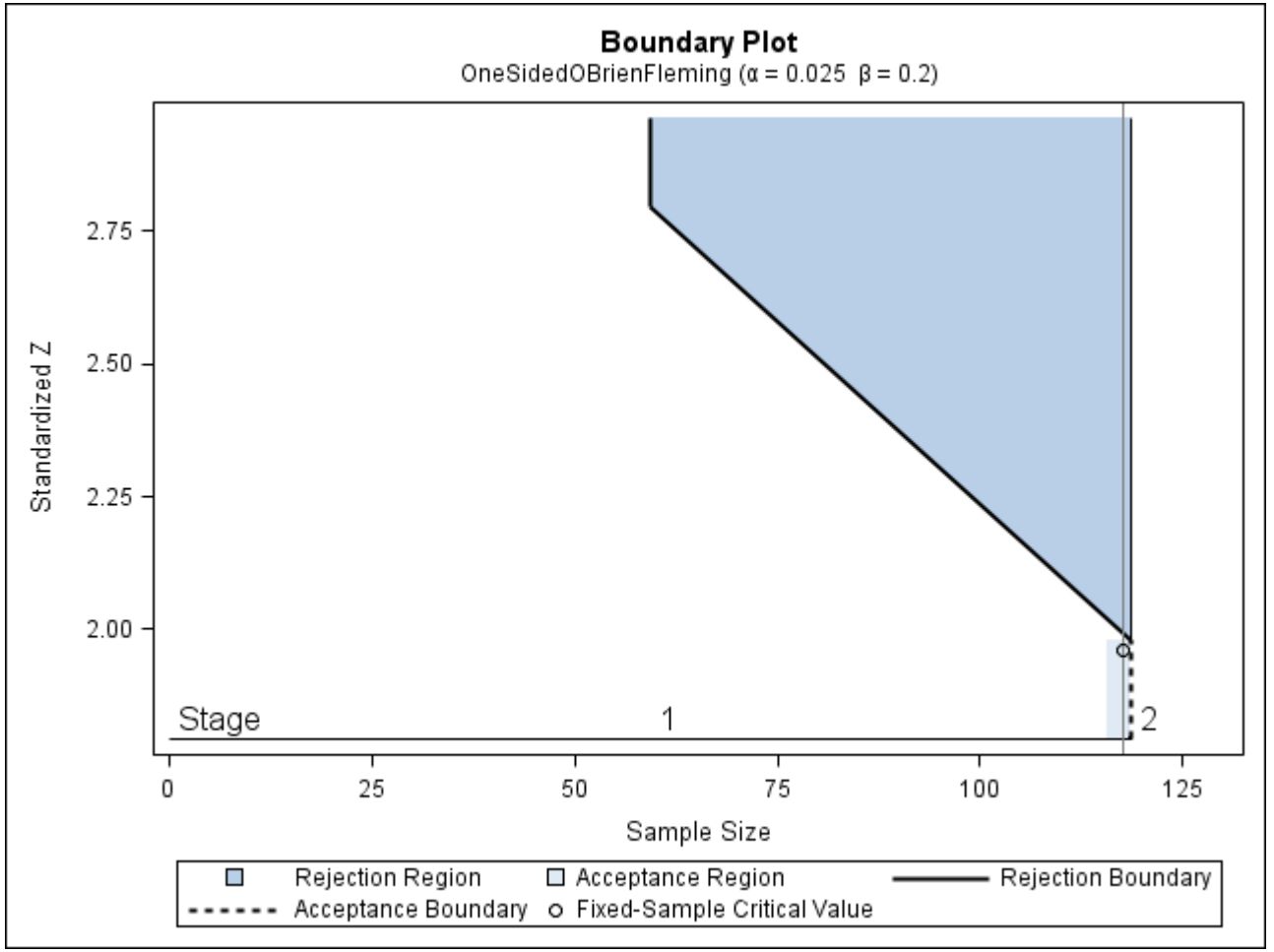
The SEQDESIGN Procedure
Design: OneSidedOBrienFleming

Sample Size Summary

Reference Proportions	Alt Ref
Max Sample Size	118.6497
Expected Sample Size (Null Ref)	118.4965
Expected Sample Size (Alt Ref)	106.2149

Sample Sizes (N)
Two-Sample Z Test for Proportion Difference

Stage	Fractional N	Fractional N(Grp 1)	Fractional N(Grp 2)	Fractional Information	Ceiling N	Ceiling N(Grp 1)	Ceiling N(Grp 2)	Ceiling Information
1	59.32	29.66	29.66	98.8748	60	30	30	100.0
2	118.65	59.32	59.32	197.7	120	60	60	200.0


Boundary Plot
OneSidedOBrienFleming (α = 0.025 β = 0.2)
Standardized Z
2.75
2.50
2.25
2.00
Stage
1
2
0
25
50
75
100
125
Sample Size
Rejection Region
Acceptance Region
Rejection Boundary
Acceptance Boundary
Fixed-Sample Critical Value