Lecture 5: Noninferiority Designs

Learning Objectives:

- Recognize when to use a noninferiority design, based on statements of the research goals
- Specify the parameter of interest in a noninferiority study design, including its values under H_0 and H_A
- State three conditions for the validity of a noninferiority trial design
- Understand how failure of the assumptions underlying the noninferiority study design can lead to
    - inflation of the type-1 error rate
    - rejecting H_0 when it is true leads to use of an inferior treatment
- Understand 1-to-1 relationship between between α levels of test statistics and confidence intervals
- Describe one way to define a noninferiority margin

Lecture 5 - 'Noninferiority Trial'

A randomized clinical trial in which

- an experimental treatment is compared with an active-control treatment
- adequate efficacy is based on the Experimental arm being "as good as" (noninferior to) the Control

Motivations:

- If an effective standard-of-care therapy (C) is available, it may be difficult to justify a placebo-controlled trial.
- The experimental agent (E) may provide advantages other than enhanced effectiveness that would make E an attractive alternative option for patients.
    - Shorter course; Lower pill burden; Lower cost
    - Fewer adverse events . . .
- Investigators can use E in place of C without meaningful loss of effect on primary outcome

Example:

In resource-poor settings, can family/friend be trained to monitor vital signs of patients with severe sepsis as effectively as health-care workers?

Outcome = 7-day mortality → π = arm-specific rate on day 7
[Get oriented: Do we hope to lower or raise the outcome? How about 7-day survival?]

Intervention arms:

- Control = 100% of "Vitals" monitoring handled by (busy) nursing staff
- Experimental = 75% of "Vitals" monitoring handled by trained family/friend

Superiority Trials

Learning by analogy . . . To date we've discussed study designs to detect superiority (*E* more effective than *C*) –

- Usually we test a two-sided alternative against the null hypothesis of no treatment effect
  (N.B. δ_0 is the boundary between the parameter spaces consistent with H_0 and with H_A. Note slight notation shift!)

$$H_0: \delta = \delta_0 \text{ versus } H_A: \delta \neq \delta_0, \text{ where } \delta = \pi_E - \pi_C \text{ and } \delta_0 = 0$$

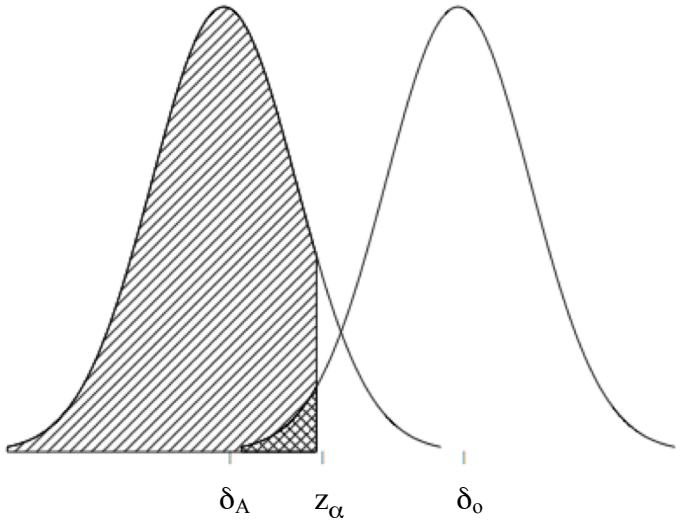
- We could specify a one-sided alternative, seeking a difference consistent with *E* more effective than *C*
  (N.B. Direction of alternative below reflects example above. Once H_A is defined, H_0 is its complement.)

$$H_0: \delta \geq \delta_0 \text{ versus } H_A: \delta < \delta_0, \text{ where } \delta = \pi_E - \pi_C \text{ and } \delta_0 = 0$$

  - Analysis at the end of the trial: Test statistic now omits absolute value bars, reflecting that H_0 is only rejected by outcomes in one direction (given by the example above). (Compare with Lecture 4, p.4.)

    (i) $Z_{obs} = \frac{d - \delta_0}{\sqrt{2\sigma/n}}$, *where* $d = \bar{x}_E - \bar{x}_C$ and $\bar{x}_E = \frac{1}{n_E}\sum x_i$. Reject H_0 if $z_{obs} < z_\alpha$.

    (ii) Estimate mean (90% CI) for δ. 90% upper confidence bound must be $< \delta_0$.
    (N.B. 90% CI is consistent with the 1-sided test; could use 95% CI but it's inconsistent with test.)



Probability functions of standard normal statistics; H_A on left, H_0 on right. $z_{\alpha/2} = -z_{1-\alpha/2}$ and $z_\alpha = -z_{1-\alpha}$, etc, due to symmetry. (N.B. Assume δ_0 and δ_A are standardized.)

Noninferiority Trials

- *Always* test a one-sided alternative

- The direction of noninferiority (experimental "as good as" than control) is the same as the direction of superiority (experimental better than control)

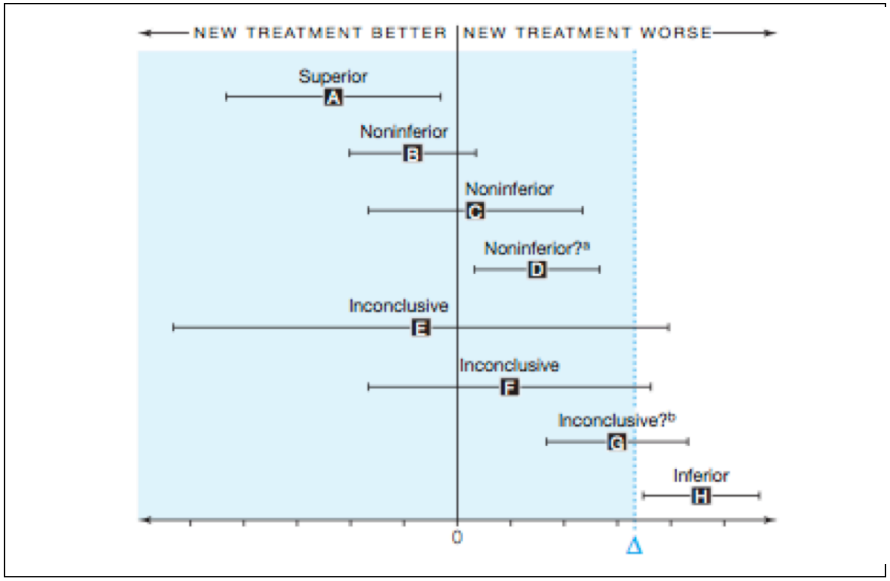
$$H_0: \delta \geq \delta_0 \text{ versus } H_A: \delta < \delta_0, \text{ where } \delta = \pi_E - \pi_C \text{ and } \delta_0 = \textbf{?}$$

- Must define the threshold (i.e., the value of δ_0) that specifies the boundary between the parameter spaces consistent with H_0 and with H_A. → We call this threshold the "noninferiority margin"

[Get oriented: For our example, should the threshold be negative or positive?

What value would you choose? How is it interpreted?]

  - Analysis at the end of the trial: Exactly as above; the value of null-alternative threshold changes but the notation doesn't.
  - Diagram shows: $\delta_A = 0$ (typical default choice), the difference we need adequate power to detect, and $\delta_0 = \Delta$, the NI margin. (Δ is another author's notation; I sometimes denote the value of δ_0 by $\delta_{EC} = \pi_E - \pi_C$, to help make a point ... see following.) *(Recall δ_0 and δ_A are "beliefs" about δ vs true unknown value we're trying to estimate.)*



NI trials are only valid under three conditions . . .

1. There is reliable and reproducible evidence of the effect of the control arm on the outcome, based on historical studies (i.e., $\delta_{CP} = \pi_C - \pi_P$ is clinically and statistically significant).

2. The planned noninferiority trial conforms as closely as possible to the design of the studies that showed the effect of the control arm, δ_{CP}, in terms of:

    - Patients: disease definition, study populations
    - Intervention: control regimen of interest; concomitant medications
    - Outcome: endpoint definitions and timing of collection(s) within the study period

3. The selected noninferiority margin, δ_0 (also denoted δ_{EC}), is closer to 0 than the effect of the control compared with placebo, δ_{PC} , and rules out a clinically relevant inferiority of E to C.

Validity Criterion 1:

There is reliable and reproducible evidence of the effect of the control arm on the outcome, based on historical studies (i.e., $\delta_{CP} = \pi_C - \pi_P$ is clinically and statistically significant).

- Efficacy of an intervention is defined relative to a placebo – such as $\delta_{CP} = \pi_C - \pi_P$. (Can also calculate $\delta_{PC} = \pi_P - \pi_C$.)
  In the NI trial, there is no placebo arm. We can only estimate $\delta_{EC} = \pi_E - \pi_C$.
  We use indirect inference to conclude that the experimental intervention is effective:

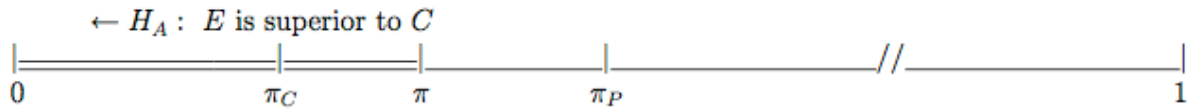
$$\begin{aligned} \delta_{EP} &= \delta_{EC} - \delta_{PC} \\ &= (\pi_E - \pi_C) - (\pi_P - \pi_C) \\ &= \pi_E - \pi_P \end{aligned}$$

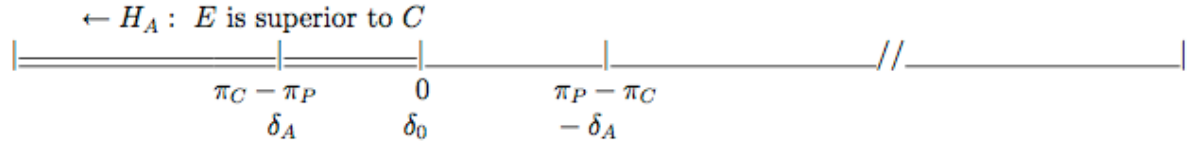
- Note that the historical evidence, δ_{CP}, and the current evidence, δ_{EC}, both cite π_C. The value of π_C must remain constant across time. This is called the "constancy assumption." We need the estimate used in our calculations to be unbiased, precise (narrow CI), and reproducible across trials.

> *"If the effectiveness of the active control is at all in question, or if one holds that historical data are not relevant to the setting of current trials, investigators either should use another control that satisfies these criteria or should choose a design other than NI, because NI trials are entirely based on having relevant historical evidence that establishes substantial effect of the control drug." (Powers, 2010)*

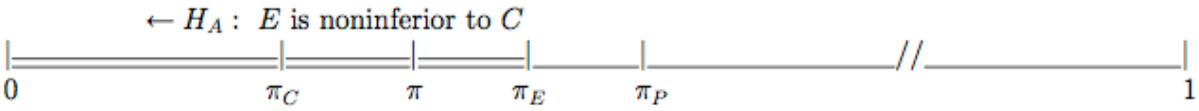
Notation for setting where low outcome rates are desired

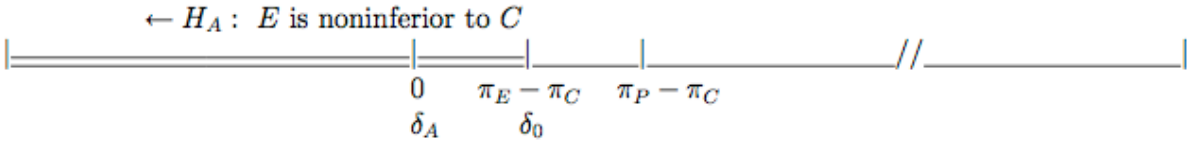
Superiority design: Ranges of response rates (upper) and their differences (lower), where $\delta_0 = 0$, $\delta_A = \pi_C - \pi_P$ ($\delta_A < 0$), and π = mean of π_C and π_P.





Noninferiority design: Ranges of response rates (upper) and their differences (lower), where $\delta_A = 0$, $\delta_0 = \pi_E - \pi_C$ ($\delta_0 > 0$), and π = mean of π_E and π_C.





Note: The definitions of π, δ_0, and δ_A differ across designs. Both alternative hypotheses are in the same direction, with rejection of H_0 supporting approval of the new (C and E, respectively) intervention. "Superiority" is a stronger trial finding than "noninferiority."

Validity Criterion 2:

The planned noninferiority trial conforms as closely as possible to the design of the studies that showed the effect of the control arm, δ_{CP}, in terms of study design (Patients, Intervention, and Outcome).

Concern: The magnitude of δ_{CP} may be overestimated, due to placebo-controlled evidence may be limited, or past trials may have been biased. → *Carrying forward an exaggerated effect, δ_{CP}, could magnify the type-I error rate of the NI trial*.

What constitutes evidence?

Example: Can we justify extrapolating the effects of the antimicrobials compared with nonspecific treatment in the early literature (sulfonamides and penicillin) to the active control drugs used in current NI trials? (EG: Similar mode of action?)

1. Ideal evidence of patient benefit: A comprehensive systematic review of randomized, double-blinded, placebo-controlled superiority trials that have appropriate endpoints and analyses that are generalizable to the relevant patient population.

2. Weak evidence: One or two statistically significant results, based on short-term studies of limited size studying surrogate endpoints rather than overall patient benefit.

3. Alternative sources: (i) Case series of patients receiving C versus historical or concurrent series of patients receiving supportive care alone. (ii) Studies in which every other enrolled subject was administered C vs. P.

4. Buyer Beware!! Evidence in treatment guidelines may be a poor source of data on which to base trial design.

- Cardiology guidelines: Only 11% of recommendations were based on "Level A" evidence (RCT adequate for FDA approval); 48% were based on "Expert Opinion."
- Infectious disease guidelines: Only 15% of recommendations were based on "Level I" evidence (at least one RCT)

→ **Evaluate evidence critically**.

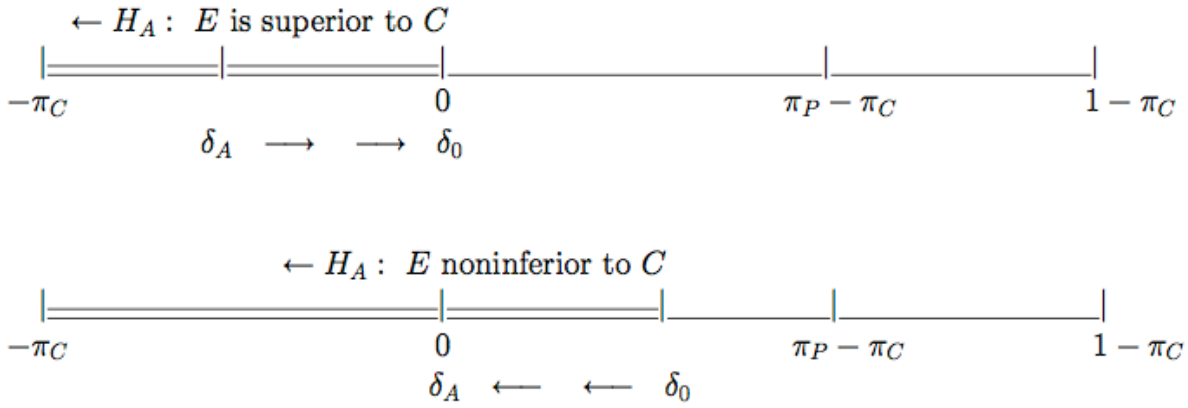
Validity Criterion 3:

The selected noninferiority margin, δ_0 (also denoted δ_{EC}), is closer to 0 than the effect of the control compared with placebo, δ_{PC}, and rules out a clinically relevant inferiority of E to C.

Concern: Biases that make interventions appear to be more similar ($\delta = 0$) produce . . .

- False-negative (FN) results in superiority trials → Supports H_0: Continue using current SOC.
- False-positive (FP) results in noninferiority trials → Worse error: Supports H_A: Wrongly begin using an inferior treatment.

Why so?



Validity of Statistical Test

Notation for setting where low outcome rates are desired

In general:

Decision Rule	True State H_0	True State H_A
Do not reject H_0 (Negative trial)	TN	FN
Reject H_0 (Positive trial)	FP	TP

Superiority design:

Decision Rule	True State $\delta_0 = 0$	True State $\delta_A = \Delta\ (< 0)$
"No difference" Do not reject H_0	TN	FN β
"Difference" Reject H_0	FP α	TP Power

Noninferiority design:

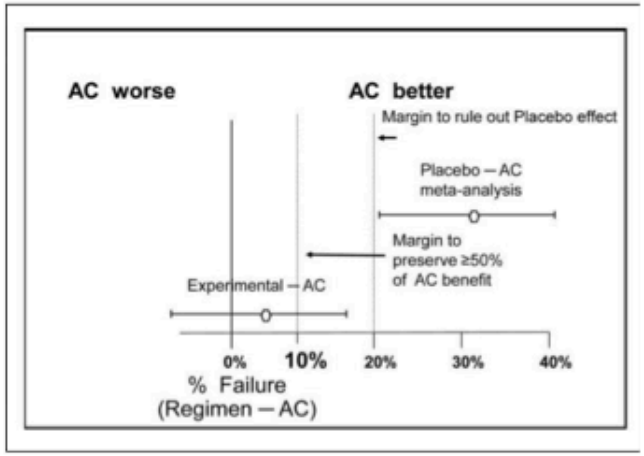
Decision Rule	True State $\delta_0 = \Delta\ (> 0)$	True State $\delta_A = 0$
"Difference" Do not reject H_0	TN	FN β
"No difference" Reject H_0	FP α	TP Power

One way to define the Noninferiority Margin: *Preserve a percentage of Control's efficacy*

Example using anti-infective drugs to treat CAP:
The value of δ_{PC} is used to define the noninferiority margin.

NONINFERIORITY TRIAL	Failure
New Quinolone (EXP)	38 / 150 (25%)
Penicillin (AC)	30 / 150 (20%)
(EXP − AC)	95% C. I. : (−5%, **15%**)

PENICILLIN TRIAL	Failure
Placebo (PLA)	87 / 175 (50%)
Penicillin (AC)	35 / 175 (20%)
(PLA − AC)	95% C.I. : (20%, 40%)



Left panel:

- (lower) Historical effect of penicillin (active control):

$$\delta_{PC} = \pi_P - \pi_C = 0.50 - 0.20 = 0.30 \text{ (95\% CI, 0.20 to 0.40)} \rightarrow p < 0.05$$

- (upper) Projected effect of "new quinolone" in the planned NI trial:

$$\delta_{EC} = \pi_E - \pi_C = 0.25 - 0.20 = 0.30 \text{ (95\% CI, -0.05 to 0.15)} \rightarrow \text{CI excludes 0.15 – what was margin?}$$

Right panel:

- Values of δ_{PC} and projected δ_A (± CI) are used to define the noninferiority margin, δ_0.
- Less conservative value of $\delta_0 = 0.20$; more conservative value of $\delta_0 = 0.10$. (Source: Fleming & Powers, 2008)

A Case Study

A clinical investigator sought biostatistical input on the design a randomized controlled trial. The goal was to study the difference between a new single-drug regimen and a standard 4-drug regimen in proportions of patients with active TB who are not cured at 1 year.

The advantage of the experimental treatment is a shorter course – 2 months, compared with 6 months on the control regimen – which may result in

1. fewer treatment-related adverse events due to shorter course
2. if new regimen is toxic, routine monitoring for hepatotoxicity should prevent long-term harm
3. greater compliance and less development of treatment resistance
4. (especially in developing countries) less reliance on personnel and other resources

The published failure rate for the Control regimen of active TB is about 5% ($\pi_C = 0.05$).

The investigators would consider the new regimen sufficiently effective if up to 10% more patients on Arm E than Arm C fail during a 12-month follow-up period ($\delta_0 = \delta_{EC} = 0.10; \pi_E = 0.15$).

They also anticipate that the new treatment may be as effective as the standard regimen ($\delta_A = 0$).

Implications for Study Design, using *PICO*:

- Patients – At baseline, eligible patients have active TB
- Intervention groups – Experimental: single-drug regimen; Control: 4-drug regimen
- Comparison – Estimate difference in proportions, $\delta_{EC} = \pi_E - \pi_C$. This research question calls naturally for a randomized trial with a one-sided noninferiority design because PI is willing to trade off some efficacy for other advantages.

  Concerns: How was the effectiveness of the Control established and quantified?

  Is this answer known for the proposed patient population?
- Outcome variable: One year after randomization, cure status (yes / no)

Parameters:

- $\pi_C = 0.05$.
- Implied study design: Noninferiority, because willing to allow E slightly less effective than C: $\pi_E = 0.15$.
- Best anticipated case: $\pi_E = 0.05$ where $\delta_A = \pi_E - \pi_C = 0.05 - 0.05 = 0$ (optimum result under H_A)
- Implied hypotheses are one-sided:

  $H_0: \delta \geq \delta_0$, where $\delta_0 = \pi_E - \pi_C = 0.15 - 0.05 = 0.10$ ("noninferiority margin")

  $H_A: \delta < \delta_0$ (Again, small values of δ_{EC} support H_A.)

Data Collection and Analysis Plan

Covariate of interest is intervention group (binary variable):

- $X_i = 1$ if Experimental (E), $i = 1, \ldots, n_E$ patients
- $X_i = 0$ if Control (C), $i = 1, \ldots, n_C$ patients

Event of interest is treatment failure (binary outcome variable):

- $Y_i = 1$ if not cured, $Y_i = 0$ if cured at 1 year follow-up, $i = 1, \ldots, N = n_C + n_E$.

$\hat{\pi}_E = \frac{Y_{1E}+Y_{2E}+\cdots+Y_{n_E E}}{n_E}$ estimates the proportion of E patients who fail; $\hat{\pi}_C$ is analogous; $d = \hat{\pi}_E - \hat{\pi}_C$

Analysis Plan:

Investigators want the Decision Rule (test statistic) to be based on the difference in failure rates (as opposed to odds ratio or relative risk).

A normally distributed test statistic that doesn't force equal allocation to groups (i.e., depends on the allocation ratio, γ):

$$z_{obs} = \frac{d - \delta_0}{\sqrt{\hat{\pi}_E(1-\hat{\pi}_E) + \hat{\pi}_C(1-\hat{\pi}_C)/\gamma}}$$

H_A: If E is noninferior to C, the failure rate will be lower in the Experimental group. Small values of $d = \hat{\pi}_E - \hat{\pi}_C$ (and small values of z_{obs}) provide evidence in support of H_A.

→ *Need to find sample size to detect* $\delta_A - \delta_0 = 0 - 0.10 = -0.10$.

(N.B. *This difference,* $\delta_A - \delta_0$*, also applies to sample size calculations for superiority trials. Because* δ_0*=0 in those cases, we often simplify the notation and just show* δ_A *in the formula.*)

In general, for binary data,

$$n_C = \frac{[z_{1-\alpha}\sqrt{\sigma_0^2} + z_{1-\beta}\sqrt{\sigma_A^2}]^2}{(\delta_A - \delta_0)^2},$$

and

$$N = (1+\gamma)n_C, \text{ where } \gamma = n_E/n_C \text{ is the allocation ratio.}$$

Specifically,

Superiority trial: $\delta_0 = 0 \rightarrow \pi_E = \pi_C = \pi$ *when* H_0 *is true*

$$n_C = \frac{\left[z_{1-\alpha}\sqrt{\pi(1-\pi)\left(1+\frac{1}{\gamma}\right)} + z_{1-\beta}\sqrt{\pi_C(1-\pi_C) + \pi_E(1-\pi_E)/\gamma}\right]^2}{(\delta_A - \delta_0)^2}$$

Noninferiority trial: $\delta_A = 0 \rightarrow \pi_E = \pi_C = \pi$ *when* H_A *is true*

$$n_C = \frac{\left[z_{1-\alpha}\sqrt{\pi_C(1-\pi_C) + \pi_E(1-\pi_E)/\gamma} + z_{1-\beta}\sqrt{\pi(1-\pi)\left(1+\frac{1}{\gamma}\right)}\right]^2}{(\delta_A - \delta_0)^2}$$

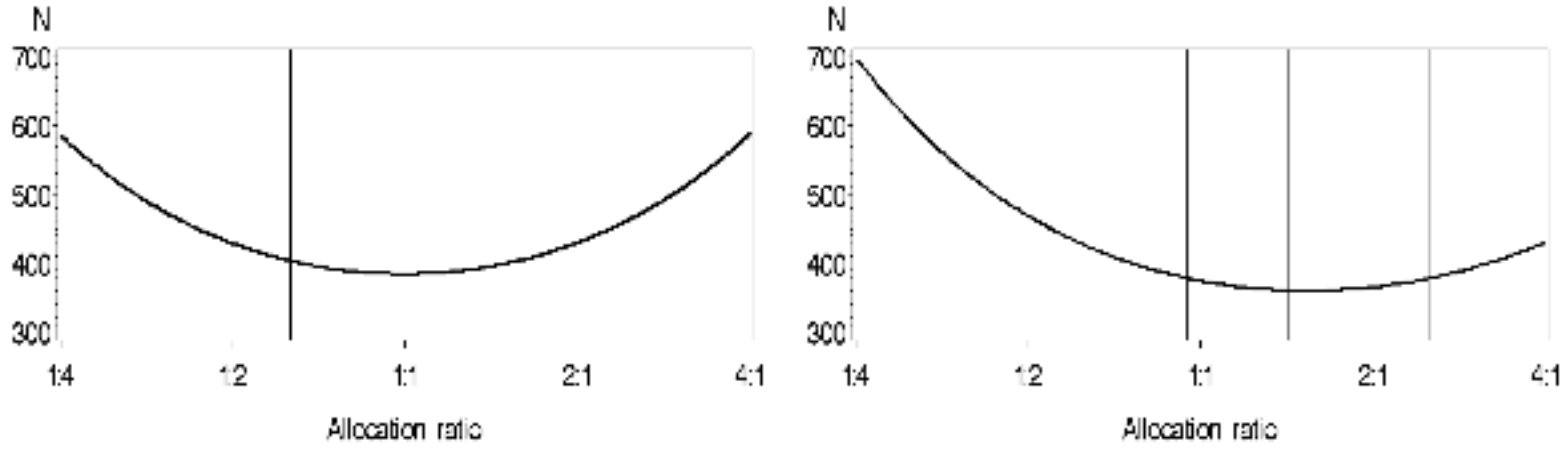
Find Allocation ratio

Superiority Trial: In most cases, $\gamma = 1$ (ie, $n_E = n_C$) minimizes $N = n_E + n_C \rightarrow$ Finding the value of γ that minimizes *N* can be calculated by hand.

Noninferiority Trial: In most cases, $\gamma \neq 1$ (ie, $n_E \neq n_C$) minimizes $N = n_E + n_C \rightarrow$ Finding the value of γ that minimizes *N* cannot be calculated by hand; it requires software (iterative solution).

Example: Plots of overall sample size (N) against allocation ratio (γ) for (left) Superiority design ($\delta_A = -.10$) and (right) Noninferiority design ($\delta_0 = +.10$).

N is at its minimum between the inner vertical reference lines and within 5% of its minimum between the outer verticals.



Also, see that n_C calculation relies on π and recall that

$$\pi = \left(\frac{\gamma}{1+\gamma}\right)\pi_E + \left(\frac{1}{1+\gamma}\right)\pi_C$$

Return to TB example:

Investigator specified $\delta_0 = 0.10$ (noninferiority margin) and $\delta_A = 0$; $\pi_C = 0.05$.

For one-sided $\alpha = 0.05$ and $power = 0.90$, use software at http://www.epibiostat.ucsf.edu/biostat/joan to find:

$\pi = 0.088$ and $\gamma = 1.67$ … that is, 5 E-participants for every 3 C-participants.

Solve for sample sizes:

$$n_C = \frac{\left[z_{1-\alpha}\sqrt{\pi_C(1-\pi_C)+\pi_E(1-\pi_E)/\gamma} + z_{1-\beta}\sqrt{\pi(1-\pi)\left(1+\frac{1}{\gamma}\right)}\right]^2}{(\delta_A-\delta_0)^2}$$

$$= \frac{\left[1.645\sqrt{(0.05)(0.95)+(.15)(.85)/1.67} + 1.28\sqrt{(.088)(1-.088)\left(1+\frac{1}{1.67}\right)}\right]^2}{(0-0.10)^2}$$

$$= 108 \text{ Control patients}$$

$$N = (1+\gamma)n_C$$

$$= (1+1.67)108 = 286 \text{ Total patients}$$

$$n_E = N - n_C = 286 - 108 = 178 \text{ Experimental patients}$$

Summary of NI Trial Concepts

When designing noninferiority trials, must keep π_P (Placebo response rate) in mind.

- Real interest is in effect of E relative to P (e.g., $\delta_{EP} = \pi_E - \pi_P$) but can only get at this indirectly.
- In setting of failure rates and comparison of E with an active control arm, we expect to infer:

  $0 < \pi_C < \pi_E$ ($< \pi_P$) from a noninferiority trial … and $0 < \pi_E < \pi_C$ ($< \pi_P$) from a superiority trial

- We assume we have past evidence that $\delta_{PC} = \pi_P - \pi_C > 0$ (i.e., that C is safe & effective). If C has become standard of care, cannot ethically withhold it from patients.
- We try to assess how the within-arm estimate of π_C in the current trial compares with past evidence.
  - If it's too different, we cannot be sure that $\pi_E - \pi_P < 0$.
  - Best if we have evidence that $\delta_{PC} = \pi_P - \pi_C > 0$ was gathered under identical circumstances (patient population, trialist expertise).
  - To allow for trial-to-trial variability, the noninferiority margin, $\delta_0 = \delta_{EC}$, should not be too close to the lower bound of δ_{PC} (e.g., halfway between this value and δ_A).
- Usual alternative of interest is $\delta_A = 0$ because, if H_A is true, we have evidence that Arm E is as effective as Arm C: we gain its advantages without losing any efficacy.

References for this lecture:

- Pocock SJ (2003), The pros and cons of noninferiority trials. Fundamental & Clinical Pharmacology, 17: 483–490.

Reading:

- CONSORT guidance for noninferiority trials