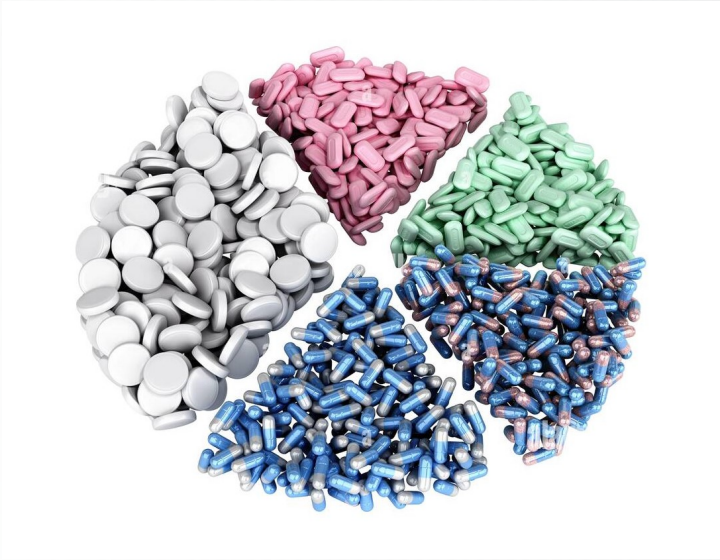
Introduction to Medical Statistics 2026

Oxford University Clinical Research Unit
March 23-27, 2026

Ronald Geskus and the biostatistics crew
Oxford University Clinical Research Unit
Hospital for Tropical Diseases,
Ho Chi Minh City, Viet Nam

Part IX

Study Designs: RCTs; Sample Size Calculation



Contents:

1. Study designs
2. Experimental versus observational studies
3. Randomized controlled trials
4. Sample size calculation

Criteria for reviewers (PLOS neglected tropical diseases)

Methods:

1. Are the objectives of the study clearly articulated with a clear testable hypothesis stated?
2. Is the study design appropriate to address the stated objectives?
3. Is the population clearly described and appropriate for the hypothesis being tested?
4. Is the sample size sufficient to ensure adequate power to address the hypothesis being tested?
5. Were correct statistical analysis used to support conclusions?
6. Are there concerns about ethical or regulatory requirements being met?

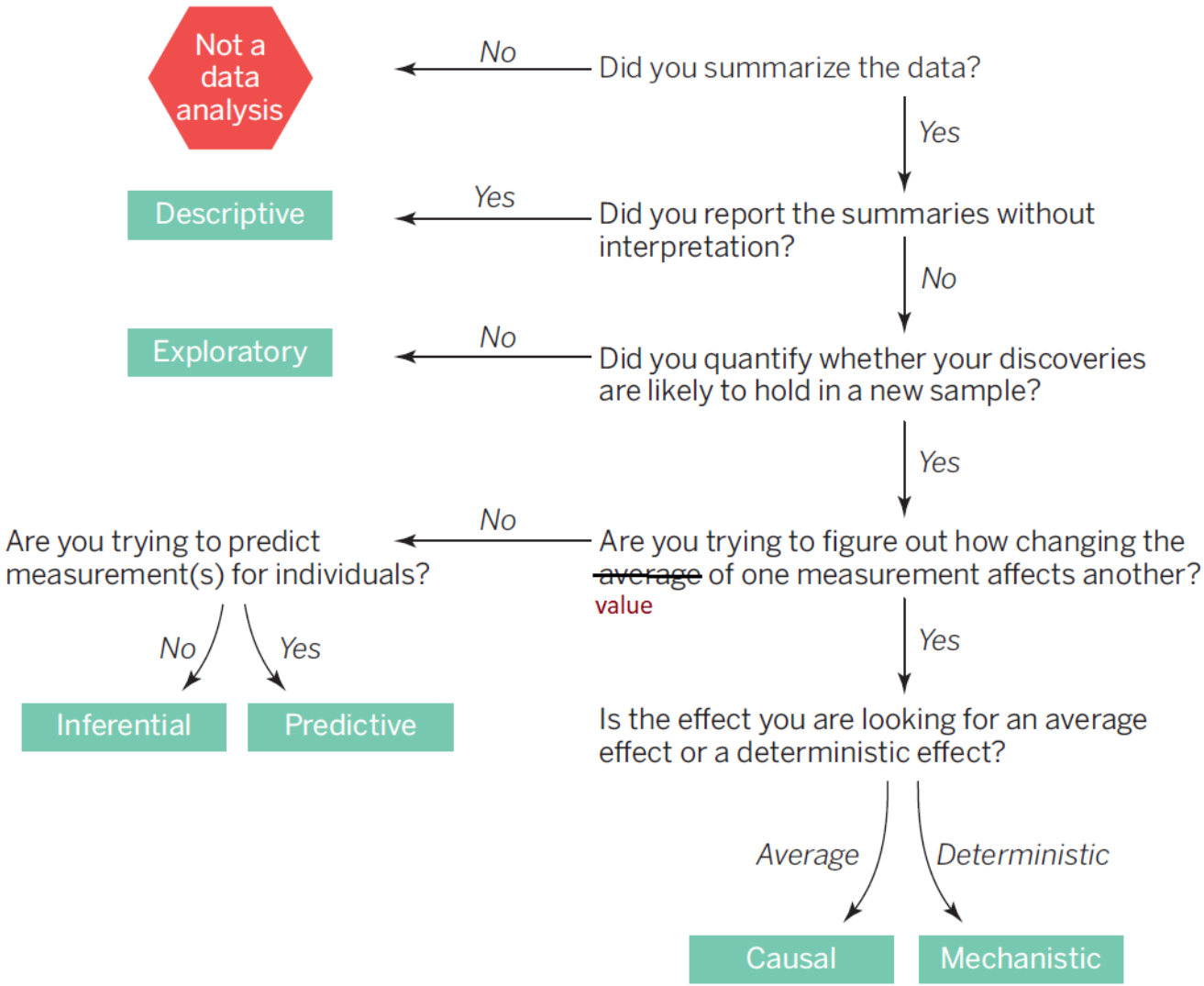
Criteria for reviewers (PLOS neglected tropical diseases)

Methods:

1. Are the objectives of the study clearly articulated with a clear testable hypothesis stated?
2. Is the study design appropriate to address the stated objectives?
3. Is the population clearly described and appropriate for the hypothesis being tested?
4. Is the sample size sufficient to ensure adequate power to address the hypothesis being tested?
5. Were correct statistical analysis used to support conclusions?
6. Are there concerns about ethical or regulatory requirements being met?

Outline

Data analysis flowchart



What is your question?

- Study question: descriptive, exploratory, inferential, predictive, causal

What is your question?

- Study question: descriptive, exploratory, inferential, predictive, causal
- Causal
  - effect of intervention (treatment) or exposure
  - understanding disease etiology and biological mechanisms

  Tool: hypothesis tests.
  Decision criteria: p-value, confidence interval

What is your question?

- Study question: descriptive, exploratory, inferential, predictive, causal
- Causal
  - effect of intervention (treatment) or exposure
  - understanding disease etiology and biological mechanisms

  Tool: hypothesis tests.
  Decision criteria: p-value, confidence interval
- Inferential
  - find most important risk factors for the outcome, often based on p-values
  - no explicit causal hypotheses; association instead of effect
  - suggest possible mechanisms based on the results
  - no strong conclusions can be drawn; findings need to be validated in separate studies

What is your question?

- Study question: descriptive, exploratory, inferential, predictive, causal
- Causal
  - effect of intervention (treatment) or exposure
  - understanding disease etiology and biological mechanisms

  Tool: hypothesis tests.
  Decision criteria: p-value, confidence interval
- Inferential
  - find most important risk factors for the outcome, often based on p-values
  - no explicit causal hypotheses; association instead of effect
  - suggest possible mechanisms based on the results
  - no strong conclusions can be drawn; findings need to be validated in separate studies
- Predictive: prognosis/diagnosis

Predictive questions

- Obtain accurate predictions for the individual case
- "Accurate" via measures of predictive value instead of p-values
  - calibration: on average accurate (unbiased)
    "average Tmax in HCMC on March is 34 °C*"*
  - discrimination: distinguish between low and high risk cases (AUC, Brier score)
    "this March 31, Tmax in HCMC will be 32 °C*"*
- Proper validation of predictive value important
  Calibration and discrimination with new data may be much worse, due to overfitting or different populations
- Use statistical regression models or machine learning techniques

Criteria for reviewers (PLOS neglected tropical diseases)

Methods:

1. Are the objectives of the study clearly articulated with a clear testable hypothesis stated?
2. Is the study design appropriate to address the stated objectives?
3. Is the population clearly described and appropriate for the hypothesis being tested?
4. Is the sample size sufficient to ensure adequate power to address the hypothesis being tested?
5. Were correct statistical analysis used to support conclusions?
6. Are there concerns about ethical or regulatory requirements being met?

Outline

Types of studies designs

Observational ⟷ **Experimental**

Researcher collects data without influencing course of events

Researcher influences course of events and studies effect of the intervention

Prospective ⟷ **Retrospective**

Data collected forwards in time from the start of the study

Data refer to past events and is acquired from existing sources

Longitudinal ⟷ **Cross-sectional**

Study changes over time, observations taken on more than one occasion

Observations taken only once

Experimental: trial

- Prospective study comparing the causal effect of a novel intervention against a control intervention
- Intervention
  - therapy/drug
  - medical device, decision support tool
  - behavioral intervention/training
- Control
  - placebo
  - standard of care
- Prospective
  - all study procedures (including study hypotheses and planned analyses) defined before the start of the study
  - participants followed from a well-defined “baseline” point (intervention)

Experimental: trial

- Prospective study comparing the causal effect of a novel intervention against a control intervention
- Intervention
  - therapy/drug
  - medical device, decision support tool
  - behavioral intervention/training
- Control
  - placebo
  - standard of care
- Prospective
  - all study procedures (including study hypotheses and planned analyses) defined before the start of the study
  - participants followed from a well-defined “baseline” point (intervention)

Randomized Controlled Trial (RCT)

- Random assignment of participants to intervention or control
- Control other factors
  - equality of patient care (except for randomized intervention)
  - equality of endpoint assessment
- Blinding (if possible)
  - patients
  - investigator/physician
  - both (double-blind)
- Special issues of experiments in humans
  - safety
  - compliance
  - ethics

Randomized Controlled Trial (RCT)

- Random assignment of participants to intervention or control
- Control other factors
  - equality of patient care (except for randomized intervention)
  - equality of endpoint assessment
- Blinding (if possible)
  - patients
  - investigator/physician
  - both (double-blind)
- Special issues of experiments in humans
  - safety
  - compliance
  - ethics

Criteria for reviewers (PLOS neglected tropical diseases)

Methods:

1. Are the objectives of the study clearly articulated with a clear testable hypothesis stated?
2. Is the study design appropriate to address the stated objectives?
3. Is the population clearly described and appropriate for the hypothesis being tested?
4. Is the sample size sufficient to ensure adequate power to address the hypothesis being tested?
5. Were correct statistical analysis used to support conclusions?
6. Are there concerns about ethical or regulatory requirements being met?

Types of studies designs

Observational ⟷ **Experimental**

Researcher collects data without influencing course of events

Researcher influences course of events and studies effect of the intervention

Prospective ⟷ **Retrospective**

Data collected forwards in time from the start of the study

Data refer to past events and is acquired from existing sources

Longitudinal ⟷ **Cross-sectional**

Study changes over time, observations taken on more than one occasion

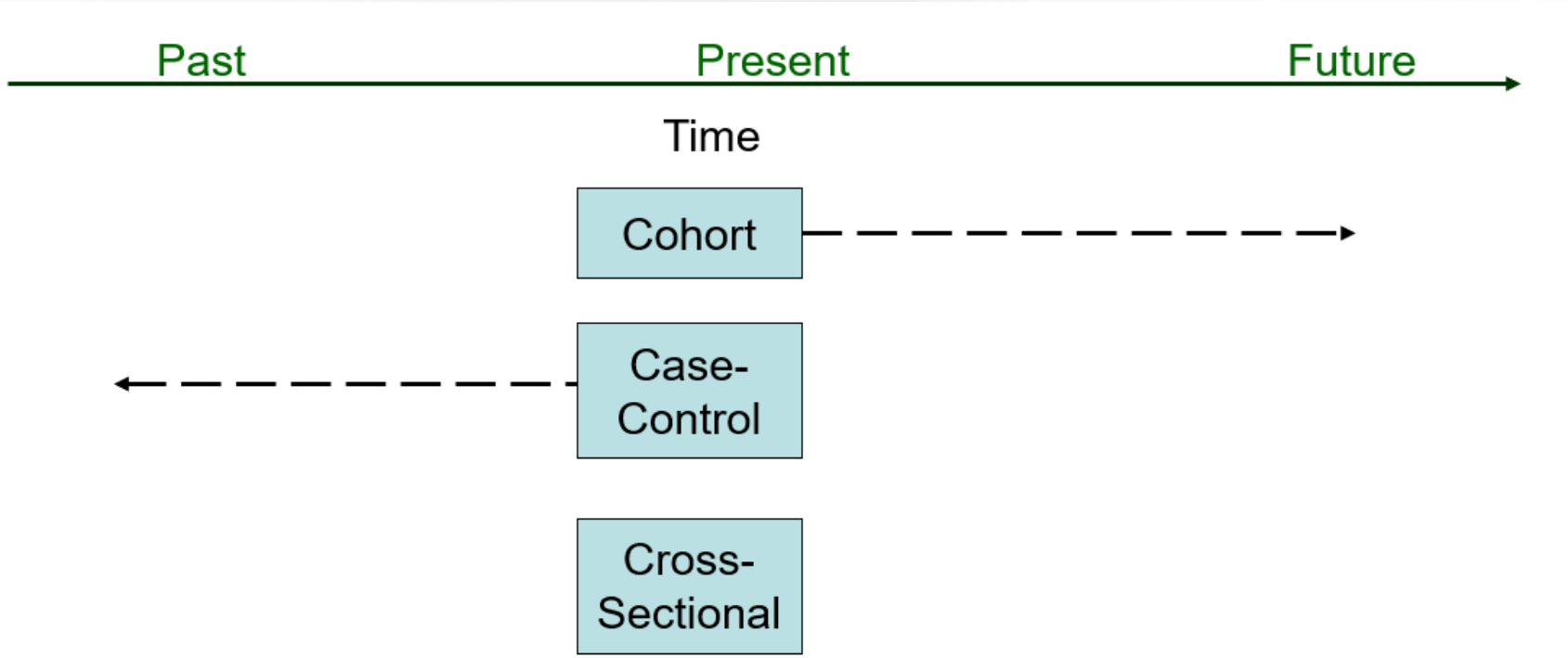
Observations taken only once

Observational studies

- To answer causal questions if exposure or intervention cannot be randomized
- Examples:
  - relation between passive smoking and lung cancer
  - relation between radiation exposure and miscarriage
  - relation between alcohol drinking and suicide
- “The aim of the observational study should be to arrive at the same conclusions that would have been obtained by an experimental trial” Gray-Donald and Kramer (1988)

Observational study designs

- Main study designs:
  - cohort study (longitudinal)
  - case-control study
- Cross-sectional studies
  - usually inferential
  - we don't know what came first, reverse causation possible



Bias in causal observational studies

- Lack of randomization $\rightarrow$ assessment of causal factors in observational studies may be biased: confounding
- Smoking and lung cancer: an observed association could be due to
  - smoking causes lung cancer (causation)
  - cancer (or a pre-cancerous condition) is a factor inducing cigarette smoking (reverse causation)
  - both smoking and lung cancer are caused by a specific genotype (confounding)
- The design of observational studies needs careful epidemiological and causal considerations

Statistical methods to control confounding

- Avoid confounding by design $\rightarrow$ randomization (RCT)
- Adjust for confounding in the analysis or the design
  - stratification
  - regression
  - matching

  May require use of advanced statistical methods
- Confounders can only be adjusted for in the analysis if the information on the confounder has been collected, i.e. we cannot adjust for "unmeasured confounding"

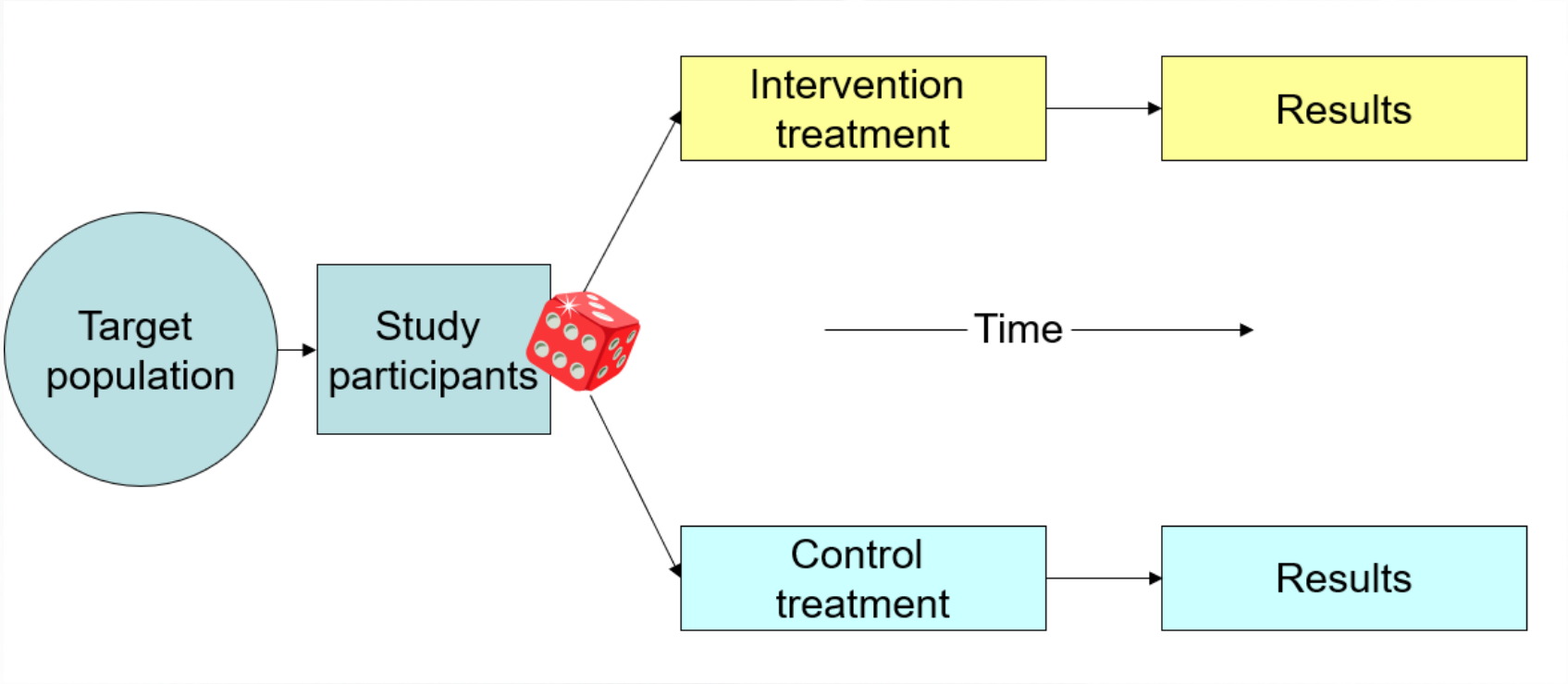
Outline

Study questions and study designs
- Study designs

Randomized controlled trial
- Some practical advice for study design

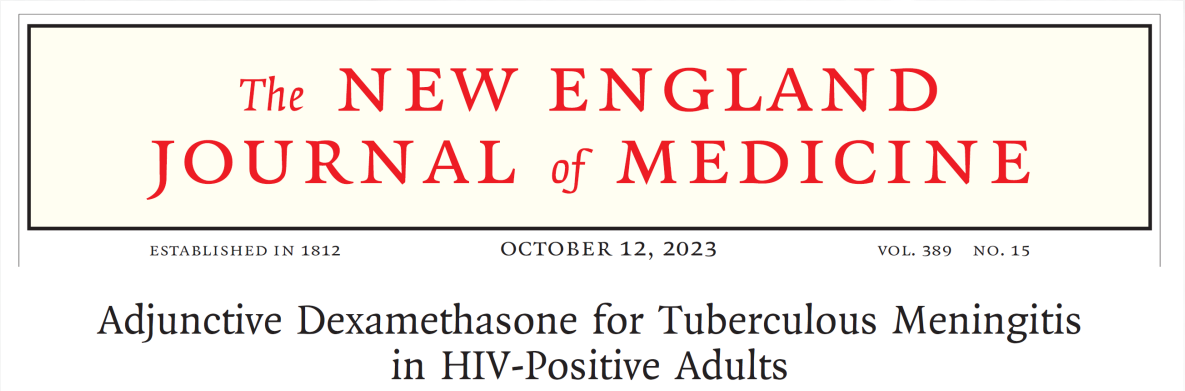
Sample Size Calculation
- Estimating a single proportion
- Comparing two independent groups (of equal size)

Structure of an RCT

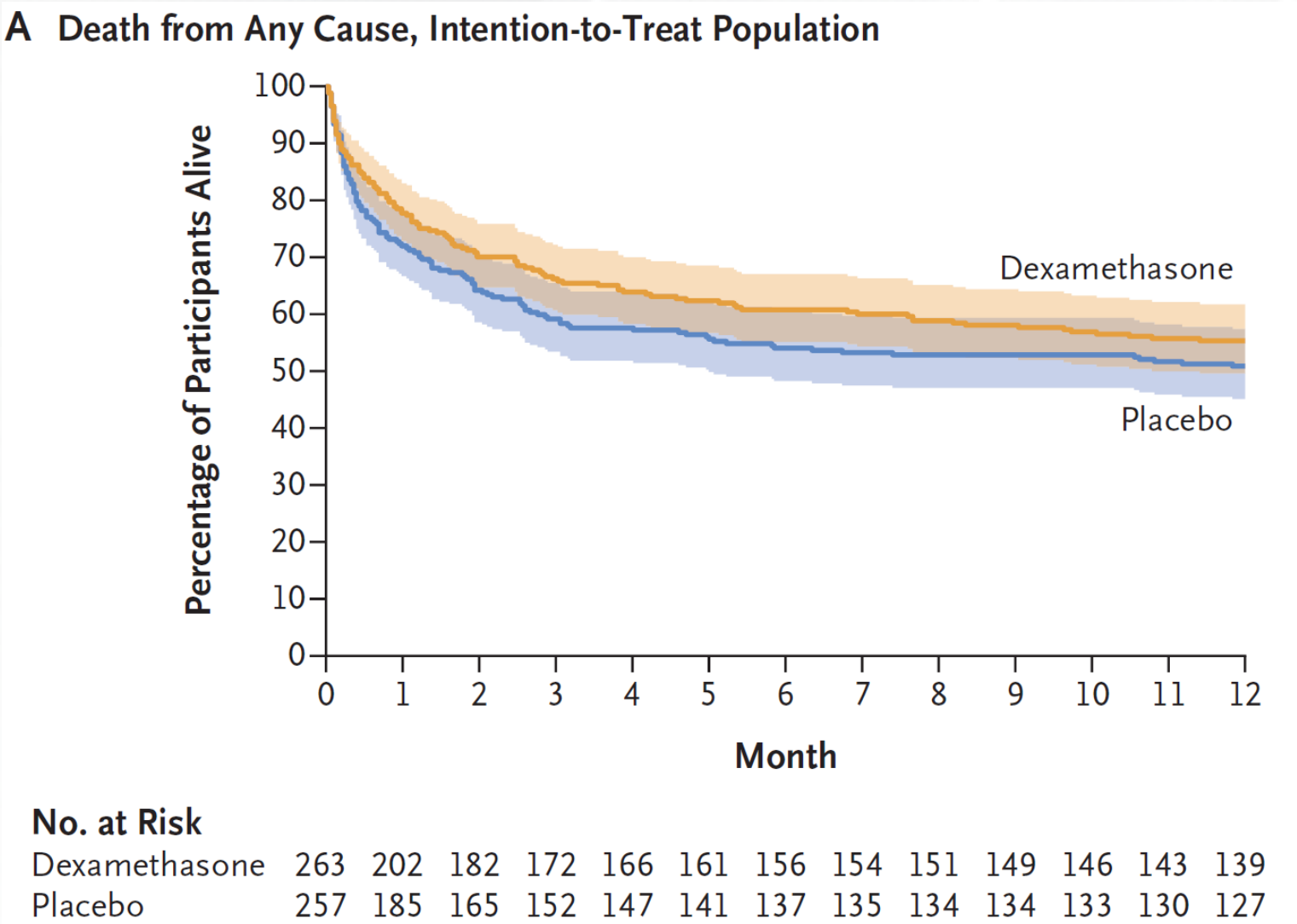


Randomization

- Chance decides about the assignment of the patients to their treatment group
- Advantages of randomization
  - (on average) produces treatment groups which are comparable with respect to known and unknown risk factors
  - avoids investigator bias in the allocation of patients
  - avoids bias due to confounding factors
- Randomization is the only known method with the above characteristics



- Objective: is dexamethasone a safe and effective addition to anti-TBM treatment
- Treatment: 6 (MRC I)/8 (MRC II or III) weeks of adjunctive dexamethasone (N=263) versus placebo (N=257)
- Endpoint: overall survival within 12-months from randomisation


A Death from Any Cause, Intention-to-Treat Population
Percentage of Participants Alive
100
90
80
70
60
50
40
30
20
10
0
0 1 2 3 4 5 6 7 8 9 10 11 12
Month
Dexamethasone
Placebo
No. at Risk
Dexamethasone 263 202 182 172 166 161 156 154 151 149 146 143 139
Placebo 257 185 165 152 147 141 137 135 134 134 133 130 127

ICH E9(R1) addendum on estimands

Precise definition of the treatment effect reflecting the clinical question posed by a given clinical trial objective
Precise specification of five attributes:

1. treatment conditions

2. target population (e.g. eligibility criteria)
3. variable (outcome, endpoint)
4. population-level summary (ratio, difference)

5. other intercurrent events, occurring after treatment initiation
   - treatment switch or discontinuation, additional medication, rescue medication
   - terminal events (death, leg amputation for diabetic foot ulcers)

ICH E9(R1) addendum on estimands

Precise definition of the treatment effect reflecting the clinical question posed by a given clinical trial objective
Precise specification of five attributes:

1. treatment conditions
   dexamethasone versus placebo
2. target population (e.g. eligibility criteria)
3. variable (outcome, endpoint)
4. population-level summary (ratio, difference)
   hazard ratio: $0.85(0.66, 1.10)$; $P = 0.22$
5. other intercurrent events, occurring after treatment initiation
   - treatment switch or discontinuation, additional medication, rescue medication
   - terminal events (death, leg amputation for diabetic foot ulcers)

ICH E9(R1) addendum on estimands

Precise definition of the treatment effect reflecting the clinical question posed by a given clinical trial objective
Precise specification of five attributes:

1. treatment conditions
   dexamethasone versus placebo
2. target population (e.g. eligibility criteria)
3. variable (outcome, endpoint)
4. population-level summary (ratio, difference)
   hazard ratio: $0.85(0.66, 1.10)$; $P = 0.22$
5. other intercurrent events, occurring after treatment initiation
   - treatment switch or discontinuation, additional medication, rescue medication
   - terminal events (death, leg amputation for diabetic foot ulcers)

Criteria for reviewers (PLOS neglected tropical diseases)

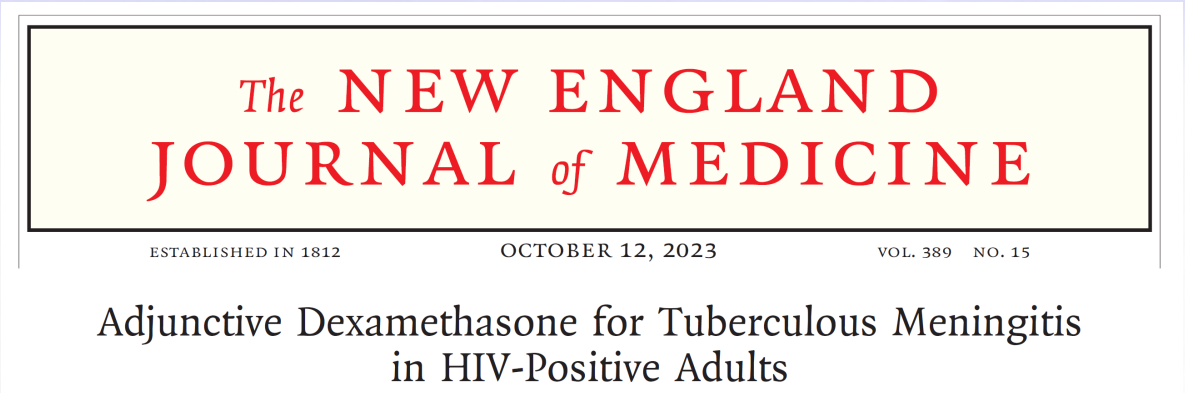
Methods:

1. Are the objectives of the study clearly articulated with a clear testable hypothesis stated?
2. Is the study design appropriate to address the stated objectives?
3. Is the population clearly described and appropriate for the hypothesis being tested?
4. Is the sample size sufficient to ensure adequate power to address the hypothesis being tested?
5. Were correct statistical analysis used to support conclusions?
6. Are there concerns about ethical or regulatory requirements being met?

ICH E9(R1) addendum on estimands

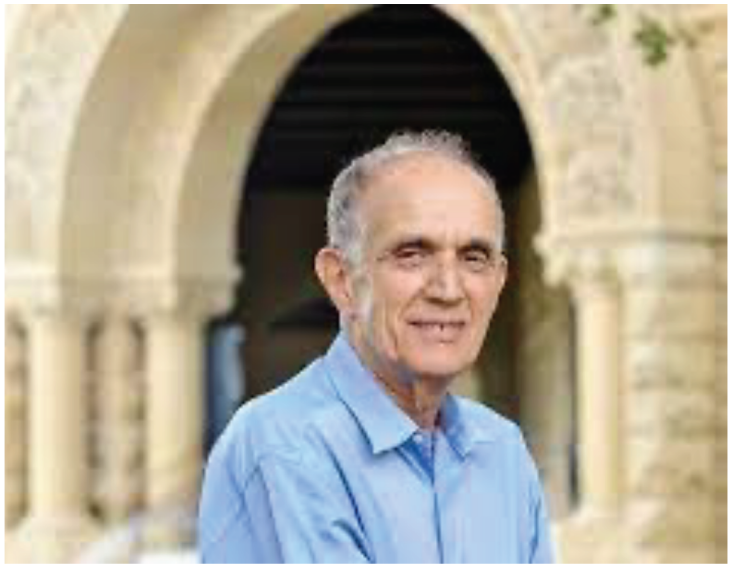
Precise definition of the treatment effect reflecting the clinical question posed by a given clinical trial objective
Precise specification of five attributes:

1. treatment conditions
   dexamethasone versus placebo
2. target population (e.g. eligibility criteria)
3. variable (outcome, endpoint)
4. population-level summary (ratio, difference)
   hazard ratio: $0.85(0.66, 1.10)$; $P = 0.22$
5. other intercurrent events, occurring after treatment initiation
   - treatment switch or discontinuation, additional medication, rescue medication
   - terminal events (death, leg amputation for diabetic foot ulcers)



- Objective: is dexamethasone a safe and effective addition to anti-TBM treatment
- Treatment: 6 (MRC I)/8 (MRC II or III) weeks of adjunctive dexamethasone (N=263) versus placebo (N=257)
- Endpoint: overall survival within 12-months from randomisation
- Intercurrent events:
  - switch to open-label dexamethasone (n=138; 26.5%)
  - $<$7 days randomised study drug for reason other than death (n=22; 4.2%)
  - $<$30 days of anti-tuberculosis drugs for reason other than death (n=7; 1.3%)

Why haven't research questions been defined sufficiently precisely?



Efron is the second winner of International prize in Statistics Foundation

"There could not be worse experimental animals on earth than human beings;

- they complain,
- they go on vacations,
- **they take things they are not supposed to take**,
- they lead incredibly **complicated lives**, and, sometimes,
- they **do not take their medicine**."

(Efron, 1998)

Intercurrent events may break randomization

Main types of analysis

- Intention-to-treat (ITT)
  - includes all randomized patients
  - participant is analyzed according to the randomized treatment group (regardless of the actual treatment given during the study)
  - usually main analysis

  Describes efficacy of an intervention in “real life” (where patients stop treatments, switch to other treatments etc)

Main types of analysis

- Intention-to-treat (ITT)
  - includes all randomized patients
  - participant is analyzed according to the randomized treatment group (regardless of the actual treatment given during the study)
  - usually main analysis

  Describes efficacy of an intervention in “real life” (where patients stop treatments, switch to other treatments etc)
- Per protocol (PP)
  - includes all patients who “followed” the protocol (no major violation of inclusion/exclusion criteria, treated according to the randomized treatment arm, completed follow-up)
  - often used, but incorrectly: randomization is lost if they are excluded based on what happens during follow-up
  - proper correction for loss of randomization may require use of advanced statistical methods

ACT HIV trial: analyses

- Intention-to-treat population: dexamethasone versus placebo, including possible switch to open-label dexamethasone
  - switch to open-label dexamethasone if after clinical and neuroradiological review the attending physician believes a patient's neurological deterioration is due to tuberculoma
  - mentioned as limitation in discussion
- Per-protocol population: exclude
  - negative confirmatory HIV test (n=2; 0.4%)
  - received $>$6 days of TB drug before enrollment (n=1; 0.2%)
  - $<$7 days randomised study drug for reason other than death (n=22; 4.2%)
  - $<$30 days of anti-tuberculosis drugs for reason other than death (n=7; 1.3%)

ACT HIV trial: analyses

- Intention-to-treat population: dexamethasone versus placebo, including possible switch to open-label dexamethasone
  - switch to open-label dexamethasone if after clinical and neuroradiological review the attending physician believes a patient's neurological deterioration is due to tuberculoma
  - mentioned as limitation in discussion
- Per-protocol population: exclude
  - negative confirmatory HIV test (n=2; 0.4%)
  - received $>$6 days of TB drug before enrollment (n=1; 0.2%)
  - $<$7 days randomised study drug for reason other than death (n=22; 4.2%)
  - $<$30 days of anti-tuberculosis drugs for reason other than death (n=7; 1.3%)
- No correction for switch to open-label dexamethasone. Reviewer on incomplete PP-analysis: "Use an approach that appropriately accounts for making comparison between non-randomized groups." Not easy!

Outline

(RCT) protocol development

- Describe study design
- Define target population and treatment groups
- Formulate clear study hypotheses and study outcomes
  - define primary outcome and hypothesis
  - secondary outcomes (not too many)
  - additional exploratory analyses
  - avoid fishing expeditions
  - preference for objective “hard” outcomes
- Ensure key outcomes are accurately collected in the Case Report Form
- Clearly describe planned key analyses, with effect measure
- Give sample size justification

During the study

- Ensure high data quality
- Accurate capture of key endpoints
- Minimize the amount of missing data
- Write a separate statistical analysis plan (as early as possible)
  - Detailed description of all planned statistical analyses
  - Tests and analysis methods used
  - Derivation rules for complex endpoints based on the case report form
  - How to deal with missing data etc.
- Clean the data prior to the analysis (consistency checks etc.)

Outline

Criteria for reviewers (PLOS neglected tropical diseases)

Methods:

1. Are the objectives of the study clearly articulated with a clear testable hypothesis stated?
2. Is the study design appropriate to address the stated objectives?
3. Is the population clearly described and appropriate for the hypothesis being tested?
4. Is the sample size sufficient to ensure adequate power to address the hypothesis being tested?
5. Were correct statistical analysis used to support conclusions?
6. Are there concerns about ethical or regulatory requirements being met?

Importance

- Question: How many samples/patients do I need to include in my study to show an effect with reasonable certainty?
- Critical issue in the planning of an RCT (also seen as important in other studies)
  - ethical aspects (for studies of an intervention)
    - study large $\rightarrow$ too many patients exposed to the risk of the intervention
    - study too small $\rightarrow$ study has not enough power to detect clinically important differences
  - economical aspect (any study)
    - resources and time are wasted
- Based on assumptions with respect to true value of parameter $\rightarrow$ often a fairly wild guess

Based on primary outcome analysis

- Variable measured during study to answer the primary study question.
  - must be clearly defined upfront (in the study protocol)
- Example
  - ACT-HIV: death within 12 months

Outline

Confidence interval for a proportion

- Given an observed proportion $\hat{p} = x/n$, an approximate 95% CI for the population proportion π is given by

$$\hat{p} \pm 1.96 \times \text{SE} = \hat{p} \pm 1.96 \times \sqrt{\hat{p}(1-\hat{p})/n}$$

- The half-width ("precision") of the confidence interval is

$$1.96 \times \sqrt{\hat{p}(1-\hat{p})/n}$$

Recipe for sample size calculation

- Make a realistic assumption about how large the population proportion π might be
  (if nothing is known, chose $\pi = 0.5$ which is conservative)
- Choose a target precision
- Solve the formula

$$\text{precision} = 1.96 \times \sqrt{\pi(1-\pi)/n}$$

for the sample size n:

$$n = 3.84 \times \pi(1-\pi)/\text{precision}$$

Example

- Setting: HIV-positive adults hospitalized at HTD with tuberculous meningitis (TBM)
- Goal: determine the 9-month mortality in this population
  - mortality is assumed to be around 0.5 (50%)
  - target precision is 0.1
  - $n = 3.84 \times \frac{0.5 \times 0.5}{0.1^2} = 96$
- Need to enroll at least 96 patients

Outline

Steps in hypothesis testing

Effect exposure **E** on disease **D**

$$\mathbf{E} \longrightarrow \mathbf{D} \quad \text{in population}$$

1. Define null hypothesis H_0 (usually "no effect, $\beta = 0$") and alternative hypothesis H_1 ("$\beta \neq 0$") for the population
2. Plan study and collect data
3. Calculate summary statistics
4. Calculate test test statistic *TEST* based on sample data
5. Calculate p-value: probability that *TEST* exceeds some value if H_0 were true
6. Draw conclusion on whether to reject H_0

Decisions

Decision	H_0 true (no effect)	H_0 false (effect)
H_0 not rejected	undecided	type II error
H_0 rejected	type I error	correct

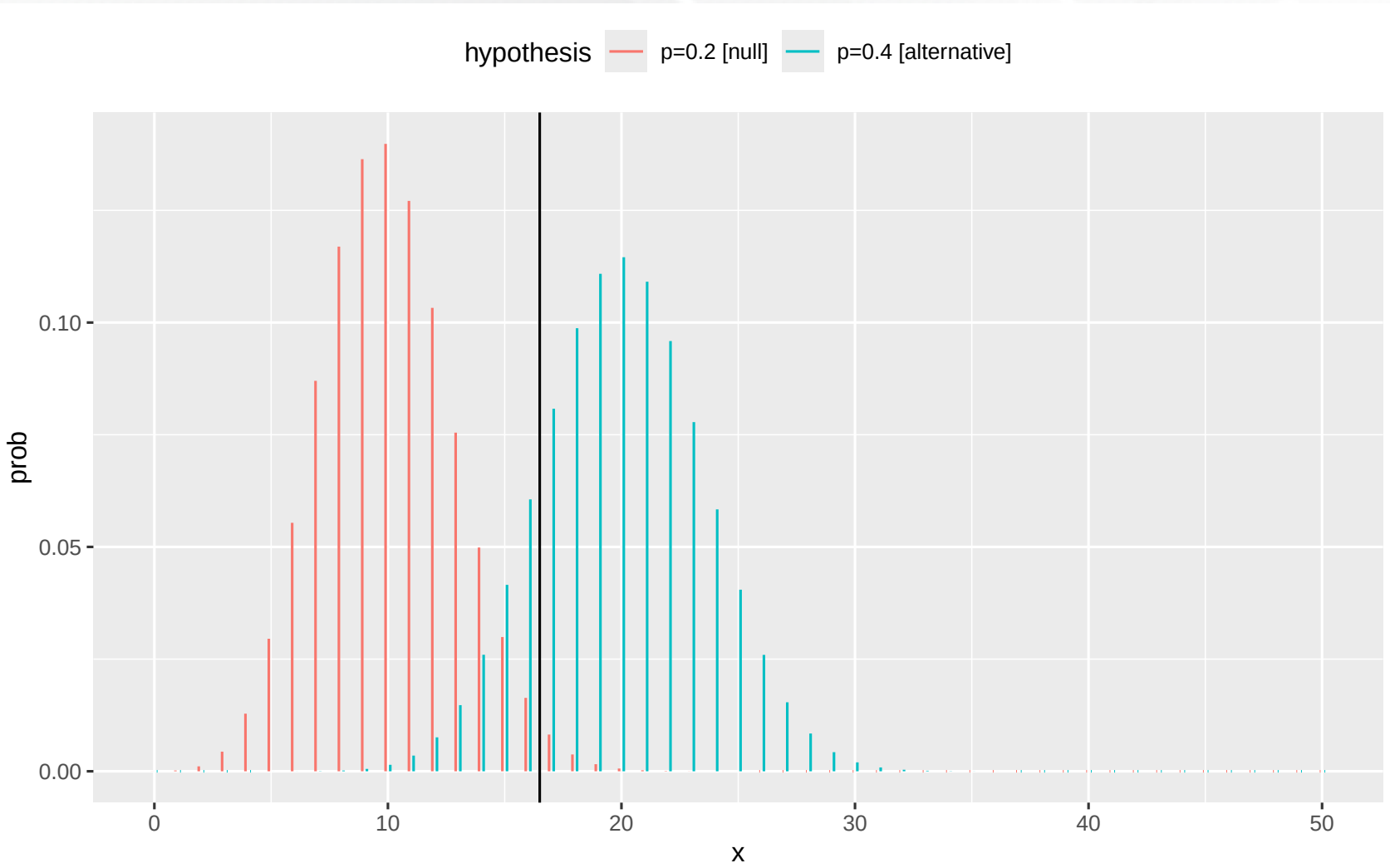
- TYPE I error: reject H_0 while it is true
  - maximum probability type I error determined in advance via significance level α (typically $\alpha = 5\%$ or $\alpha = 1\%$)
  - p-value $\leq \alpha$: reject H_0, observed difference unlikely due to chance
  - p-value $> \alpha$: do not reject H_0. Does not imply that H_0 is true
- TYPE II: do not reject H_0 while it is not true
  - probability of type II error depends on effect size and sample size
  - choose appropriate sample size based on chosen power (typically: power $= 80\%$ or power $= 90\%$) and hypothesized magnitude of effect size
  - power: $1 -$ probability of type II error

Decisions

Decision	H_0 true (no effect)	H_0 false (effect)
H_0 not rejected	undecided	type II error
H_0 rejected	type I error	correct

- TYPE I error: reject H_0 while it is true
  - maximum probability type I error determined in advance via significance level α (typically $\alpha = 5\%$ or $\alpha = 1\%$)
  - p-value $\leq \alpha$: reject H_0, observed difference unlikely due to chance
  - p-value $> \alpha$: do not reject H_0. Does not imply that H_0 is true
- TYPE II: do not reject H_0 while it is not true
  - probability of type II error depends on effect size and sample size
  - choose appropriate sample size based on chosen power (typically: power $= 80\%$ or power $= 90\%$) and hypothesized magnitude of effect size
  - power: $1 -$ probability of type II error

Power visualized



Sample size calculation (numeric outcome)

- Assume (hypothesize)
  - difference in population means between the groups: δ
  - standard deviation of outcome around the group means: σ

  $\rightarrow$ standardized effect size: $\Delta = \frac{\delta}{\sigma}$
- Choose
  - significance level α
  - power
- Simplified formula
  - $\alpha = 5\%$, power 80%: sample size per group is $n \approx \frac{16}{\Delta^2}$
  - $\alpha = 5\%$, power 90%: sample size per group is $n \approx \frac{21}{\Delta^2}$

Example RCT

- Test of a blood pressure lowering drug versus placebo
- Primary endpoint: Lowering of blood pressure one month after randomization (= date of intake of the drug)
- Assumptions:
  - Data in both arms are approximately normally distributed with known $\sigma = 10$mm Hg
  - We want to have sufficient power to detect a 5 mm Hg larger reduction in the intervention arm compared to placebo ($\delta = 5$)
  - $\alpha = 5\%$, 90% power
  - How many patients required?
- Formula gives $\Delta = 0.5 \rightarrow n = 84$ per group. In total, 168 patients need to be randomized

Sample size calculation (binary outcome)

- Assume (hypothesize)
  - Probability of the outcome in the control group: π_1
  - Probability of the outcome in the intervention group: π_2
- Choose significance level α and power
- *n* per group and $\alpha = 5\%$:

π_1	π_2	*n*(80%power)	*n*(90%power)
0.05	0.1	435	582
0.1	0.2	199	266
0.2	0.4	82	109
0.3	0.6	42	56
0.05	0.15	141	188
0.2	0.3	294	392
0.3	0.4	356	477
0.4	0.5	388	519
0.05	0.25	49	65
0.1	0.3	62	82
0.3	0.5	93	124
0.4	0.6	97	130

Example (RCT)

- Comparison of a novel chemotherapy compared to the standard of care in cancer
- Primary endpoint: Proportion of patients showing a complete tumor response
- Assumptions: $\pi_1 = 0.2$, $\pi_2 = 0.4$
  Choose $\alpha = 5\%$, power 90%
- Table gives $n = 109$ per group $\rightarrow$ $N = 218$ in total

Sample size calculation (binary outcome)

- Assume (hypothesize)
  - Probability of the outcome in the control group: π_1
  - Probability of the outcome in the intervention group: π_2
- Choose significance level α and power
- *n* per group and $\alpha = 5\%$:

π_1	π_2	*n*(80%power)	*n*(90%power)
0.05	0.1	435	582
0.1	0.2	199	266
0.2	0.4	82	109
0.3	0.6	42	56
0.05	0.15	141	188
0.2	0.3	294	392
0.3	0.4	356	477
0.4	0.5	388	519
0.05	0.25	49	65
0.1	0.3	62	82
0.3	0.5	93	124
0.4	0.6	97	130

oucru

UNIVERSITY OF OXFORD

wellcome

Typical sample size justification in a study protocol

- The study is powered for the primary endpoint, i.e. the proportion of patients showing a complete tumor response in each study arm
- Based on earlier studies, about 20% of patients achieve a complete tumor response in the control arm
- An increase in the primary endpoint by 20% (from 20% to 40%) due to the intervention was judged as both realistic and clinically relevant
- 218 patients (109 per arm) are required to detect such an increase with 90% power at the two-sided 5% significance level
- To account for some potential loss-to-follow-up and protocol violations, a total of 240 patients (120 per arm) will be randomized.

Discussion

- Determination of sample size critical issue in study planning and essential part of protocol (including justification of the assumptions on which the calculation is based)
- However, true effect size is uncertain at best (from earlier studies), or completely unknown in worst case
- Clinical and feasibility considerations also affect the final sample size of a study
- Sample size formulae exist for many situations/tests and there exists specialized software
  - Basic calculations in R (`power.t.test`, `power.prop.test`)
  - R task view on power and sample size calculations
  - `https://github.com/vubiostat/ps`
  - via simulation (see practical)