MSMS 304 - Biostatistics
Study Designs

Ananda Biswas

Last updated : November 3, 2025

Contents

1   Introduction

♠ <u>Definition</u> : A study design is a specific plan for conducting the study, which allows the investigator to translate the **conceptual** hypothesis into an **operational** one.

A study design is like a blueprint or a roadmap - it tells us how to collect data, from whom, when, and how to analyse it so we can answer our research question clearly and correctly.
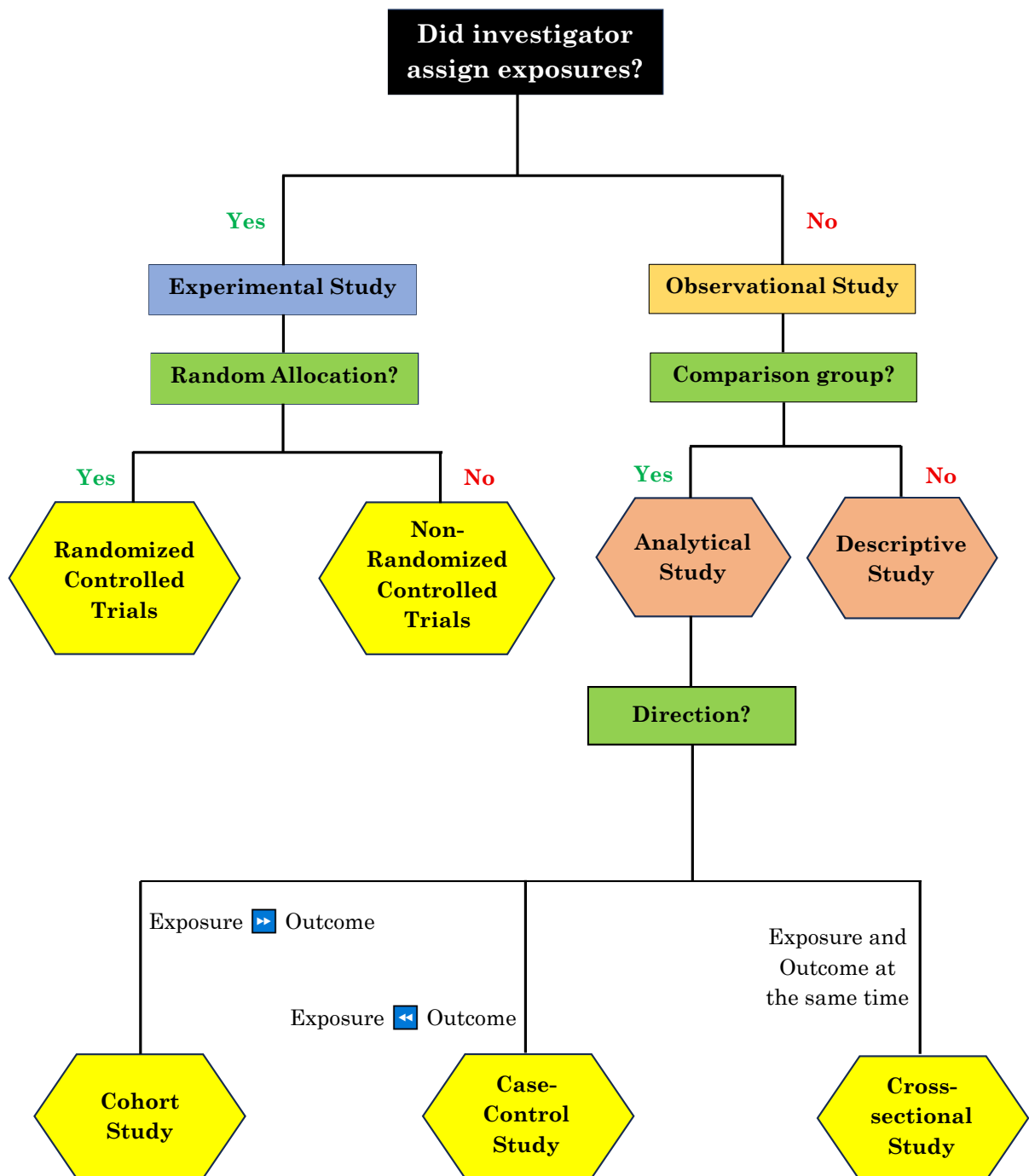
🛈 The type of study design we should choose depends on

- What the research question / objective is,

- Time available for study,

- Resources available for the study,

- Is the disease common or rare ?

- Type of outcome of interest.

Choosing an established design gives you a huge head start in analysis and eliminating biases.

• Existing designs to identify and investigate factors in disease are displayed in the following figure.

Did investigator assign exposures?

Yes

No

Experimental Study

Observational Study

Random Allocation?

Comparison group?

Yes

No

Yes

No

Randomized Controlled Trials

Non-Randomized Controlled Trials

Analytical Study

Descriptive Study

Direction?

Exposure ▶▶ Outcome

Exposure ◀◀ Outcome

Exposure and Outcome at the same time

Cohort Study

Case-Control Study
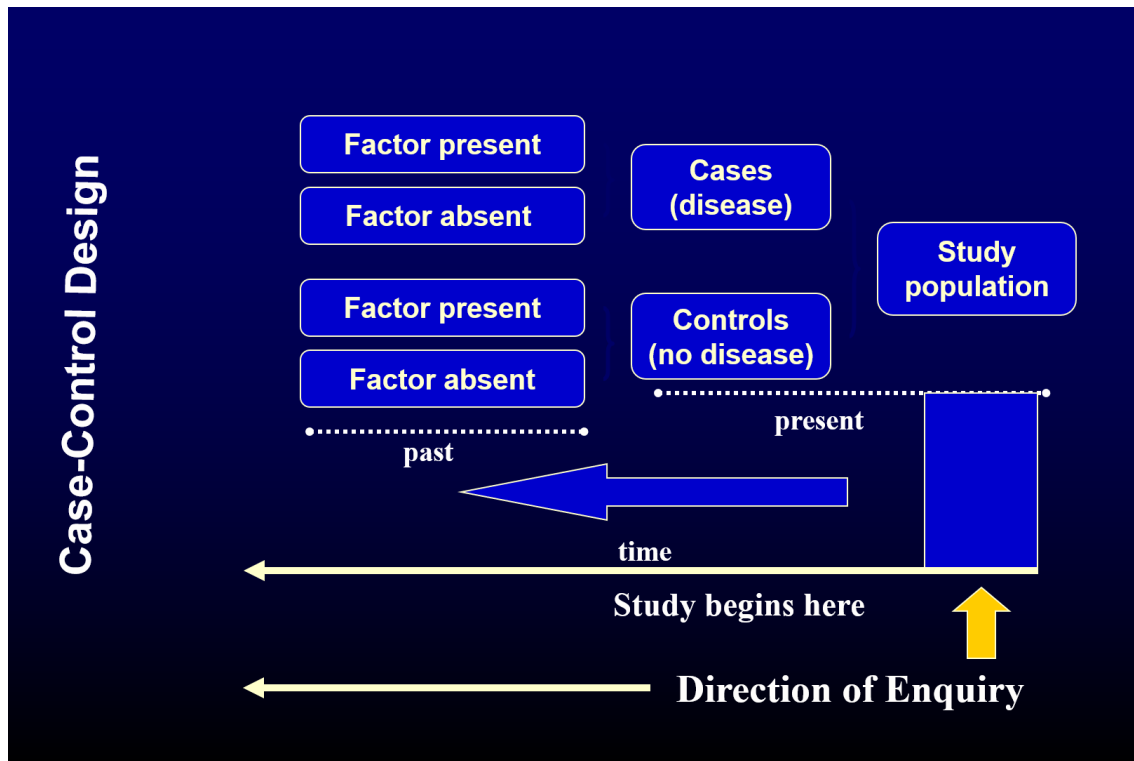
Cross-sectional Study

2 Case-Control Design

A basic question in Analytic Epidemiology is *whether a disease is linked to an exposure.* Case-control studies come handy while answering this question.

♠ <u>Definition</u> : A case-control study can be formally defined as an observational study design in which individuals with a particular outcome or disease (the cases) are compared to individuals without that outcome (the controls) to identify and evaluate factors (exposures) that may be associated with the outcome.

A flowchart of case-control study can be depicted as follows :

Several steps involved in a case-control study are as follows :

1. Start with diseased (case) versus non-diseased (control).

2. Matching

3. Measurement of exposure

4. Analysis and interpretation

⚠ Case-control studies are efficient when outcome is *rare.*

2.1 Selection of Cases

We can consider

- Hospital patients

- Clinic patients

- Patients from disease registries (*e.g.* cancer registries)

as our cases.

2.2 Selection of Controls

Controls come from the same population as the cases, but ***they do not have the disease of our interest***. They are similar to cases in all aspects but for they are free of disease. Sources of controls are hospital controls, relatives, neighbourhood controls, general population etc. Number of controls may be one, two, three or four depending on the size of the cases.

2.2.1 Matching

Matching is defined as the process of selecting the controls so that they are similar to the cases in certain characteristics, such as age, race, sex, socio-economic status, and occupation. Matching may be of two types:

(1) Group Matching

(2) Individual Matching

⊡ Group Matching : Group Matching ensures that **cases and controls have the same proportion** of certain characteristics (*e.g.* age, sex, marital status). For example, if 25% of cases are married, then 25% of controls will also be selected as married.

Steps involved in Group Matching are :

  I. Select all cases first.

 II. Identify key characteristics.

III. Calculate proportions of key characteristics.

IV. Select controls to match those proportions.

⊡ Individual Matching : Individual Matching involves selecting one control for each case, matched on specific variables (*e.g.* age, sex, race). For example, a 45-year-old white female case is matched with a 45-year-old white female control. This creates matched case-control pairs, ensuring close similarity between each case and its control for key variables.

2.2.2 Selection of Multiple Controls

Multiple controls of the same type, such as two controls or three controls for each case, are used to *increase the power of the study*. Practically speaking, a noticeable increase in power is gained only up to a ratio of about 1 case to 4 controls.

2.3 Confounding Variables

Exposure of interest may be confounded by a factor that is associated with the exposure and the disease *i.e.* there is an independent risk factor for the disease. *Matching* can help us to control confounding variables by ensuring equal distribution of confounders among cases and controls. But we cannot later study the effect of the matched variable(s).

2.4 Effect Measure : Odds Ratio

After data collection, suppose we have the following contingency table.

Exposure Status	Outcome		Total
	Cases	Controls	
Exposed	a	b	$a+b$
Not exposed	c	d	$c+d$
Total	$a+c$	$b+d$	N

⚠ The primary statistical measure of risk in case-control study design is ***odds ratio*** defined as

$$\frac{\text{Odds of Exposure among Cases}}{\text{Odds of Exposure among Controls}}.$$

From the table above, Odds of Exposure among Cases $= \dfrac{\frac{a}{a+c}}{1 - \frac{a}{a+c}} = \dfrac{a}{c}$ and

Odds of Exposure among Controls $= \dfrac{\frac{b}{b+d}}{1 - \frac{b}{b+d}} = \dfrac{b}{d}.$

Thus, *odds ratio* $= \dfrac{a/c}{b/d} = \dfrac{ad}{bc}.$

2.5 Advantages

1. *Efficient for Rare Diseases* : Case-control studies are especially useful for investigating **rare conditions** or diseases with low incidence, as they start by identifying cases and then look back to examine exposures.

2. *Time and Cost Efficient* : Since data are collected retrospectively, these studies are generally **quicker and less expensive** compared to prospective designs like cohort studies.

3. *Requires Fewer Participants* : Smaller sample sizes are often sufficient to detect associations, making it a **practical option when resources are limited**.

4. *Can Study Multiple Risk Factors* : A single case-control study can assess **several exposures or risk factors** for a particular disease, offering a broad perspective on possible causes.

5. *No Long Follow-up Required* : Unlike cohort studies, there is no need to wait for outcomes to occur.

6. *Often no risk to subjects.*

2.6 Disadvantages

1. *Recall Bias* : Example - In a study of lung cancer (cases) versus healthy individuals (controls), cancer patients might over-report past smoking habits due to awareness of its link with cancer, while controls may under-report or forget minor smoking episodes.

2. *Selection Bias* : Suppose you want to study whether eating red meat is associated with colon cancer. Cases: You select patients who have colon cancer. Controls: You select patients from the hospital who are admitted for heart disease, diabetes etc. Problem: These hospital controls may not represent the general healthy population. People with heart disease or diabetes might have already reduced their red meat intake because of their condition. As a result, their diet doesn't reflect what a normal, healthy person's diet would be, and it could artificially increase the difference between cases and controls.

   If cases are selected from a single hospital, risk factors may be unique to that hospital, limiting generalizability.

3. *Susceptibility to Confounding* : Other unmeasured variables may distort the true exposure-outcome relationship. Example - A study shows alcohol consumption is associated with esophageal cancer. However, smoking - commonly associated with drinking - may be the actual cause, and if not adjusted for, the results may be misleading.