July 21 - 24, 2021 | Garki Hospital Abuja | Resource Person

Research Methodology Boot Camp

with Epi Info Training

Dr. Adamu Onu
MBBS, FWACP (FM)
MS Epidemiology & Biostatistics
PhD Public Health (Epidemiology)

Target Audience Clinical Researchers, Post-Part 1 Residents, and Others

Important Information

- Limited slots are available on a first come, first served basis
- Laptop running Windows 10 required
- Organized as morning lecture sessions and afternoon hands on coaching sessions

Highlights

- Research Methodology
- Research Design
- Data Management
- Sample Size Calculations
- Test Statistics
- Interpretation of Results
- Report Writing
- Hands-on training sessions
- Statistical consulting sessions

For further details contact
Email: epimetrix@gmail.com
Phone: +234 803 474 9930

epimetrix
health systems
data for health decisions

Epidemiological Statistics

> I hate definitions.
>
> — Benjamin Disraeli (1804 – 1881)

What is epidemiology?

Epidemiology is the study of how disease is distributed in populations and the factors that influence or determine this distribution.

- The premise underlying epidemiology is that disease, illness and ill health are not randomly distributed in the human population.

> … the study of the distribution and determinants of health-related states or events in specified populations and the application of this study to control of health problems.

The objectives of epidemiology

1. To identify the aetiology or cause of a disease and the relevant risk factors.
2. To determine the extent of disease found in the community
3. To study the natural history and prognosis of disease
4. To evaluate existing and newly developed preventive and therapeutic measures and modes of health care delivery
5. To provide the foundation for developing public health policy

Epidemiology and prevention

Types of prevention	Definition	Examples
Primary	Preventing the *initial development* of a disease	Immunization, reducing exposure to a risk factor
Secondary	Early detection of *existing disease* to reduce severity and complications	Screening for cancer
Tertiary	Reducing the *impact of the disease*	Rehabilitation for stroke

Epidemiology and clinical practice

- Critical also to clinical practice.
- The practice of clinical medicine relies heavily on population concepts.

Epidemiologic approach

Epidemiologic reasoning is a multistep approach

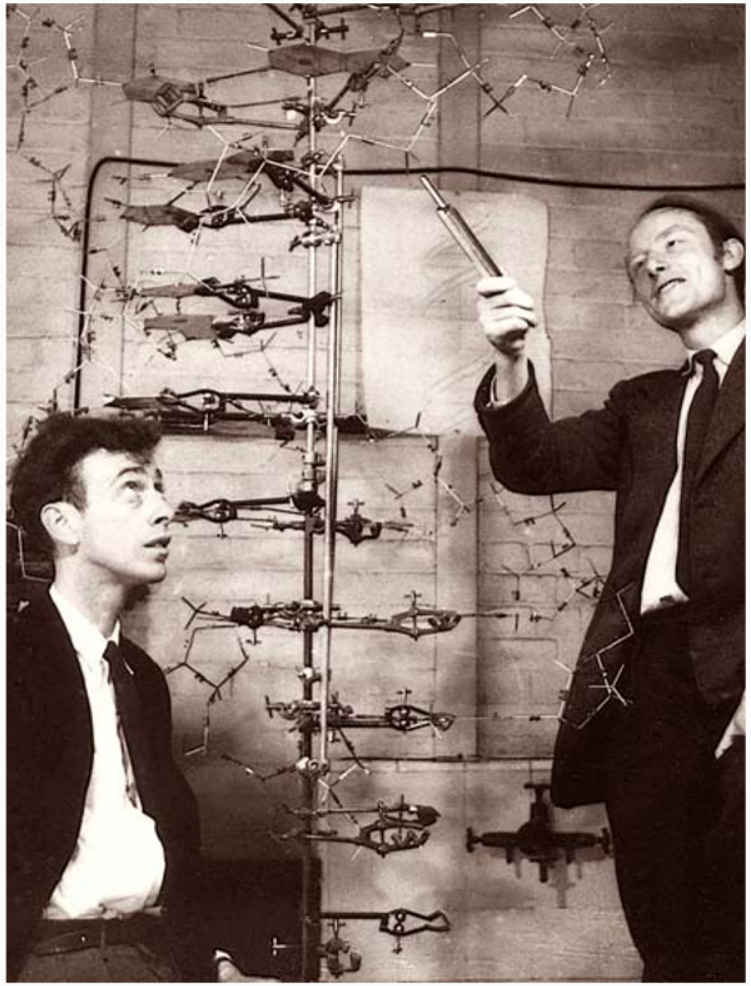
1. Determine whether an association exists between exposure to a factor, or personal characteristic and the development of the disease in question.
2. Derive appropriate inferences about a possible causal relationship from the patterns of association found.

Models

Understanding causal relations proceeds by a reiterative 3-stage process:

1. Observe phenomena.
2. Creating a model to describe or explain observations.
3. Using the model to predict future observations.

> It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.



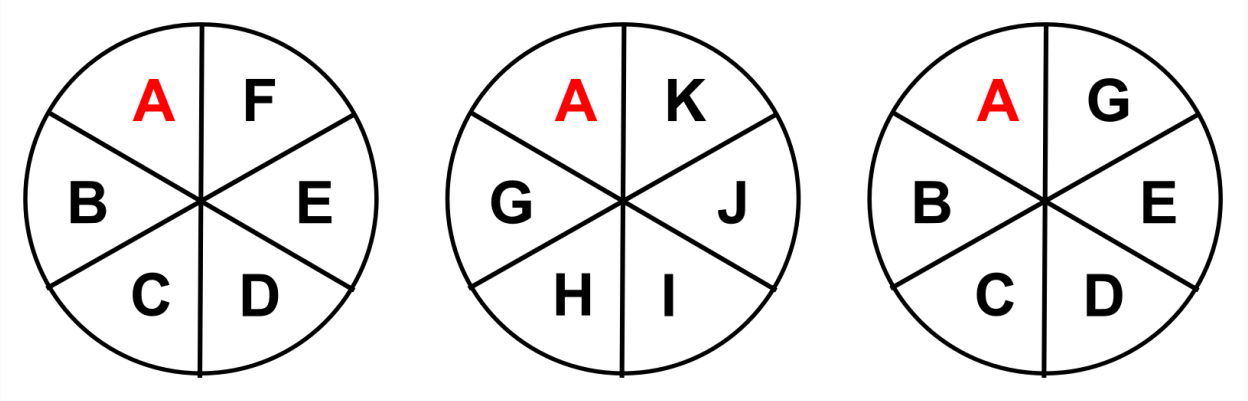
Models used to understand complex phenomena of disease causation

- Counterfactual paradigm
- Sufficient cause
- Induction and latent periods
- Directed acyclic graphs
- Study Designs
- Statistical probability

Counterfactual paradigm

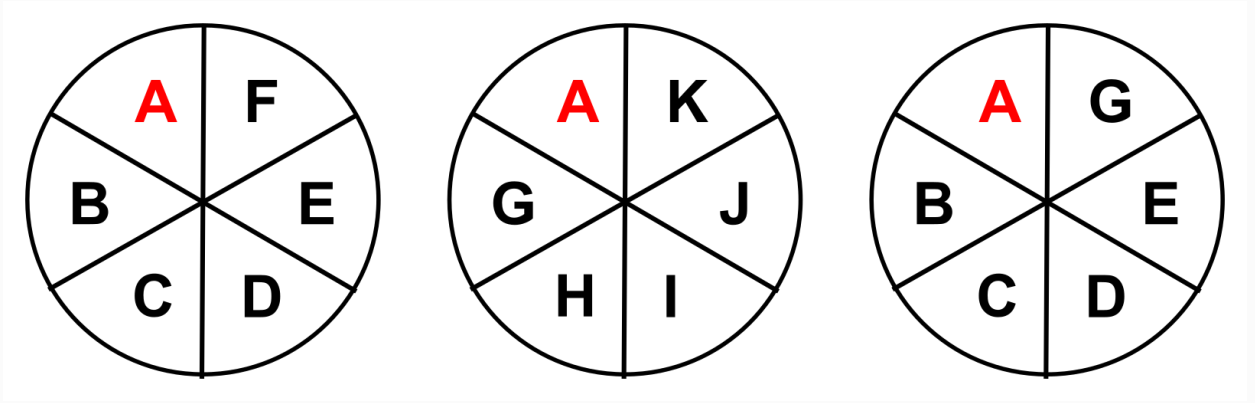
- The ideal comparison group would be people with themselves in both the exposed and unexposed state during the same time.
- This comparison is impossible.
- Every study evaluating causation is based on the counterfactual paradigm.
- Direct measurement of an effect is not possible – we must always depend on **substitution steps** when estimating effects
- The validity of an estimate will always depend on the validity of the substitutions.

Sufficient cause model



- **Sufficient causes** (causal mechanisms) are groups of **component causes** that result in disease.
- Sufficient causes are **constellations** of phenomena that correspond to individual characteristics.

Sufficient cause model

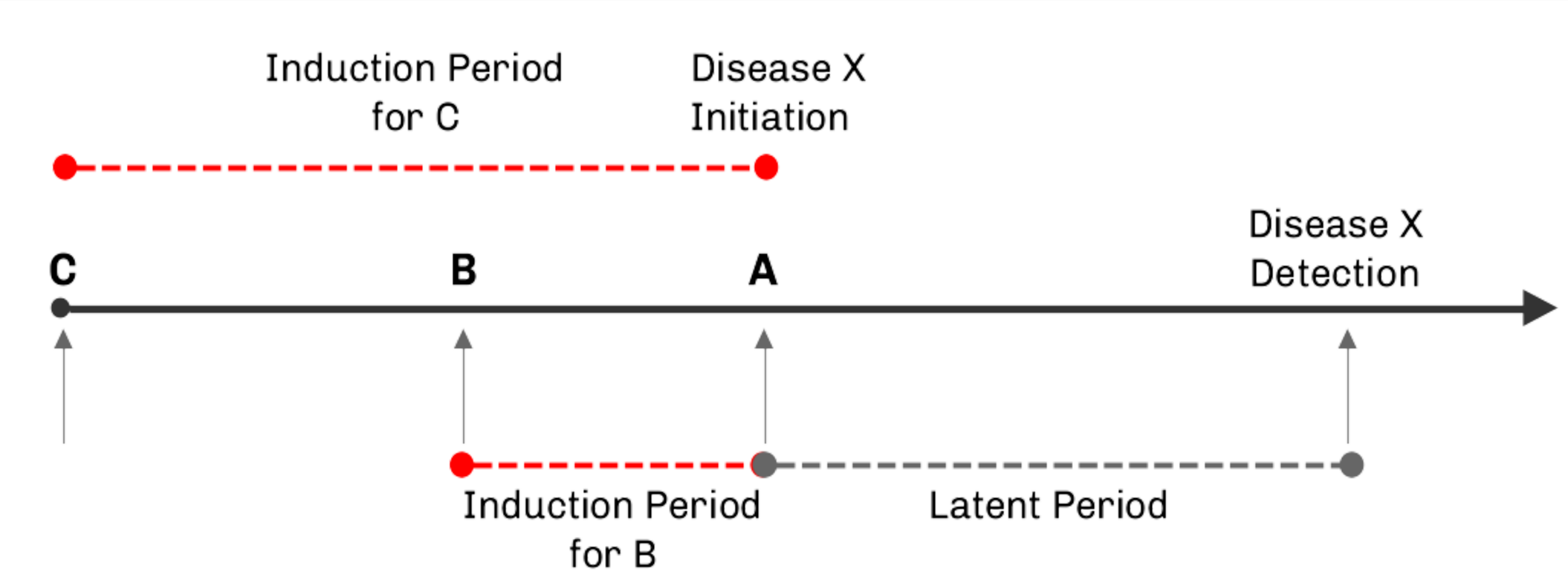


- Different sufficient causes correspond to different causal mechanisms of the disease.
- Each sufficient cause is restricted to the minimum number of **component** causes.
- **Component** causes that are members of every sufficient cause are **necessary** causes.

Timing of events

- If disease initiation begins when the sufficient cause (causal mechanism) is complete, then induction period is conceptualized only in relation to a specific component cause.
- Each component cause has a separate induction period and the last component cause's induction period is zero.
- Latent period is the time interval between disease initiation (completion of causal mechanism) and disease detection.

Aetiologic sequence for causal mechanism A + B + C



Prevention

- Identification of all the component causes of a sufficient cause is not necessary for prevention.
- Multiple component causes implies multiple points for preventive interventions.

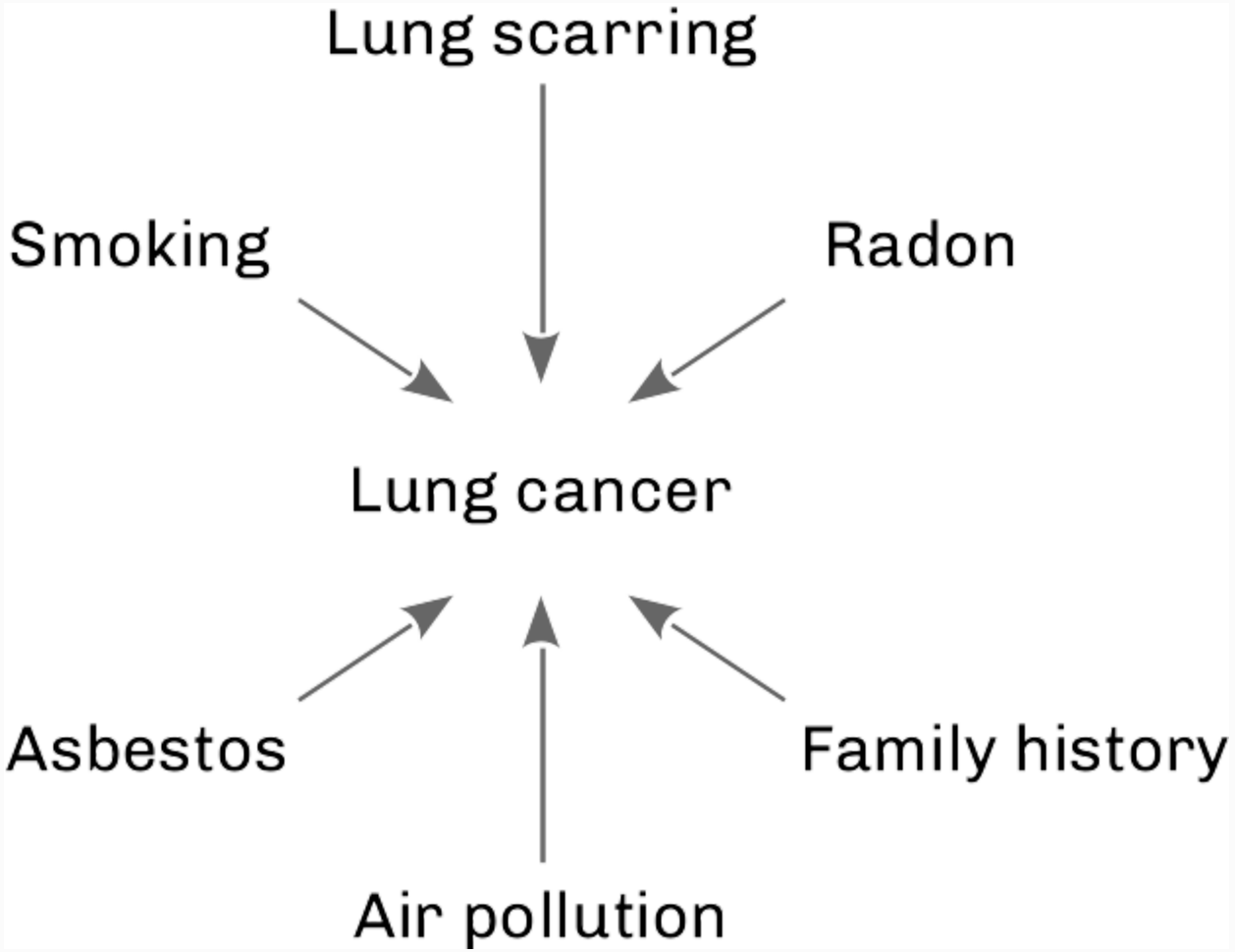
Conceptualizing multicausality for research purposes

- If several phenomena are involved in the development of a specific disease outcome, how do we conceptualize these phenomena in terms of causation for the specific disease outcome.
- Use of graphical models – path analysis, structural equation modeling, directed acyclic graphs – for representing a model of causal order.

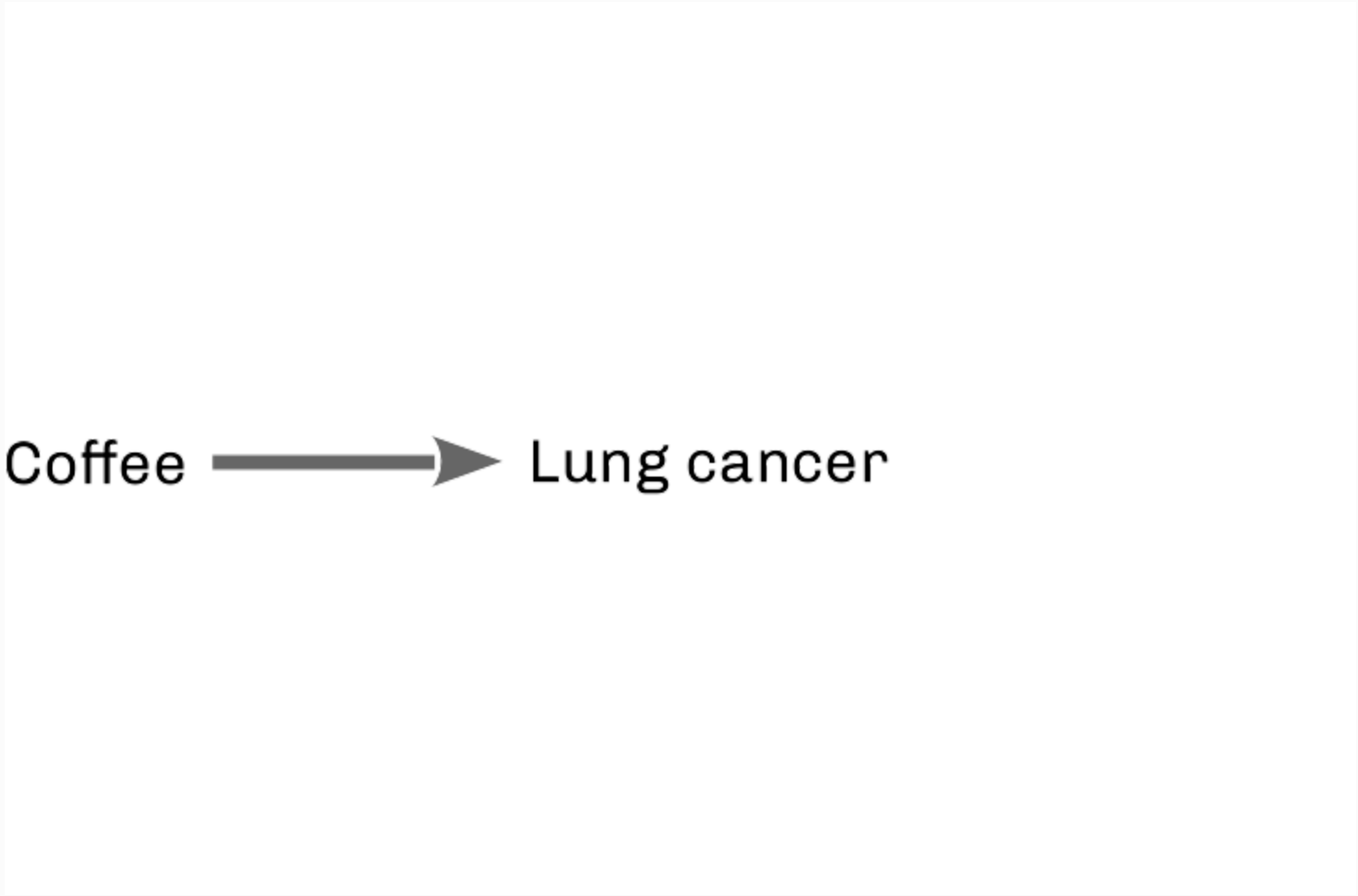
Directed acyclic graphs

- Constructed by abstracting causal assumptions from descriptions of hypothesized relations among potential study variables representing exposures, outcomes, and covariates.
- Compact graphical formulation.
- Helpful for recognizing and explaining complex relations, including effect modification and confounding.
- Graphs must be acyclic

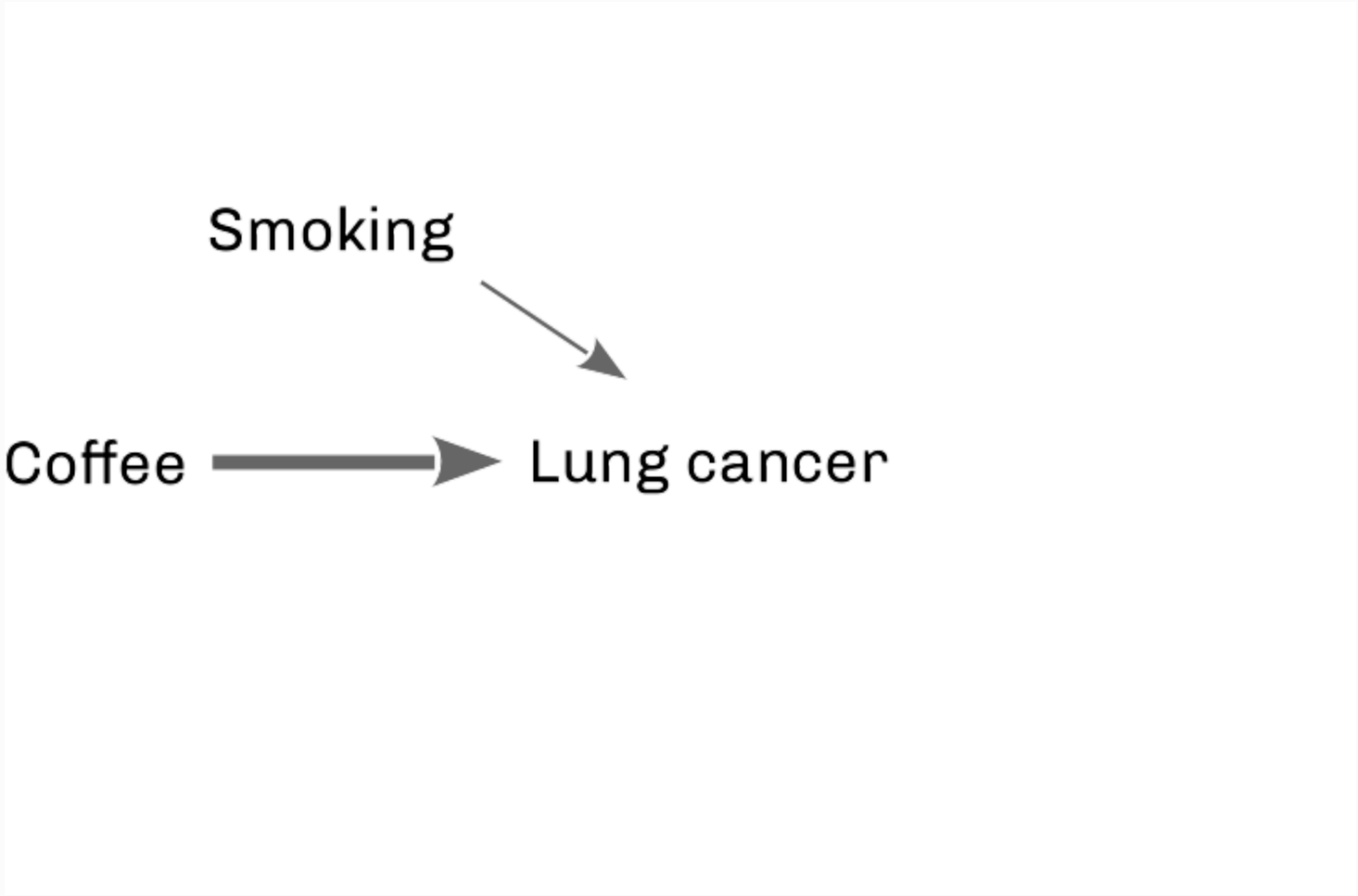
Graphic Model of Causation for Lung Cancer



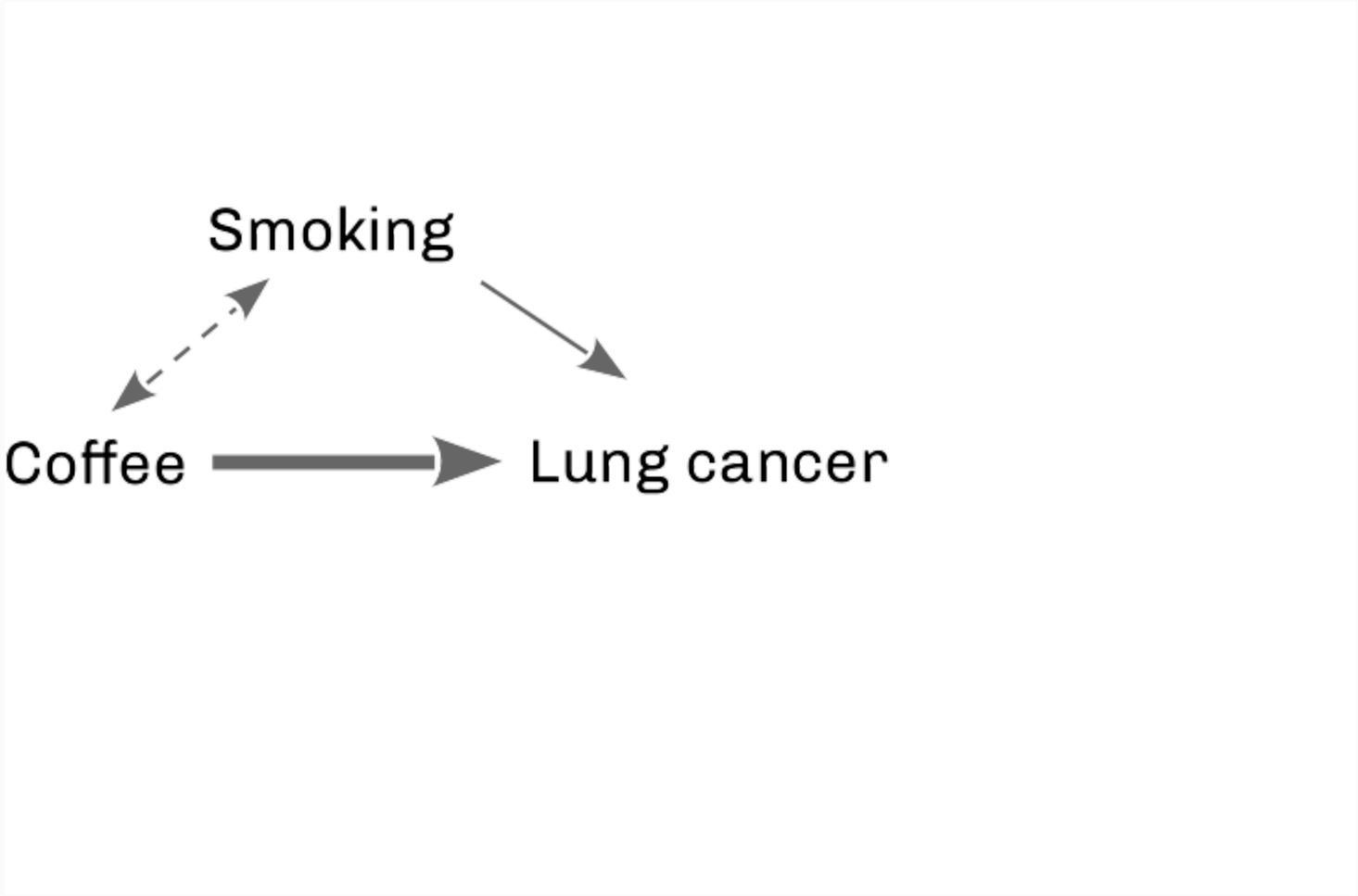
Graphic Model of Causation for Lung Cancer



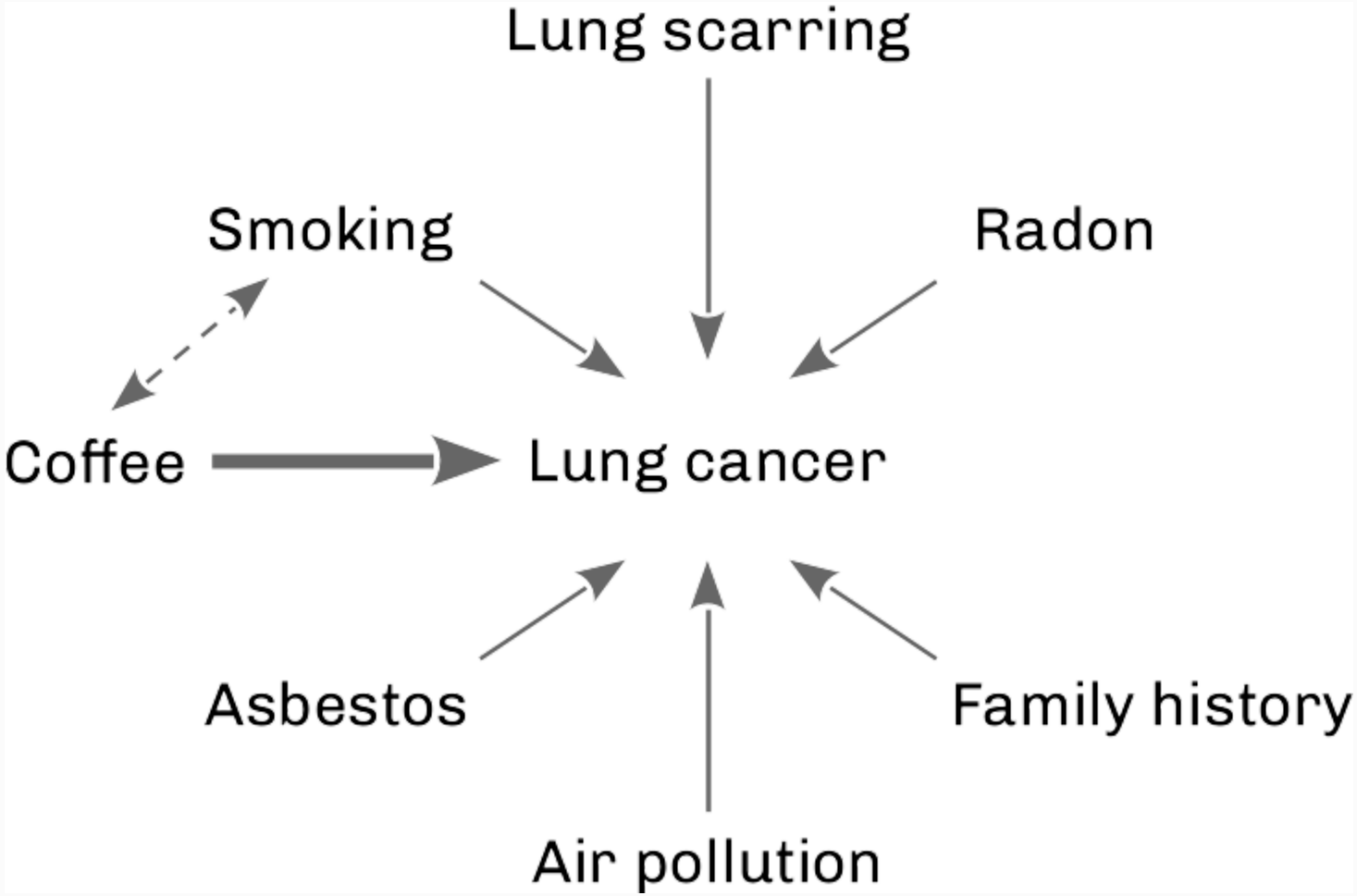
Graphic Model of Causation for Lung Cancer



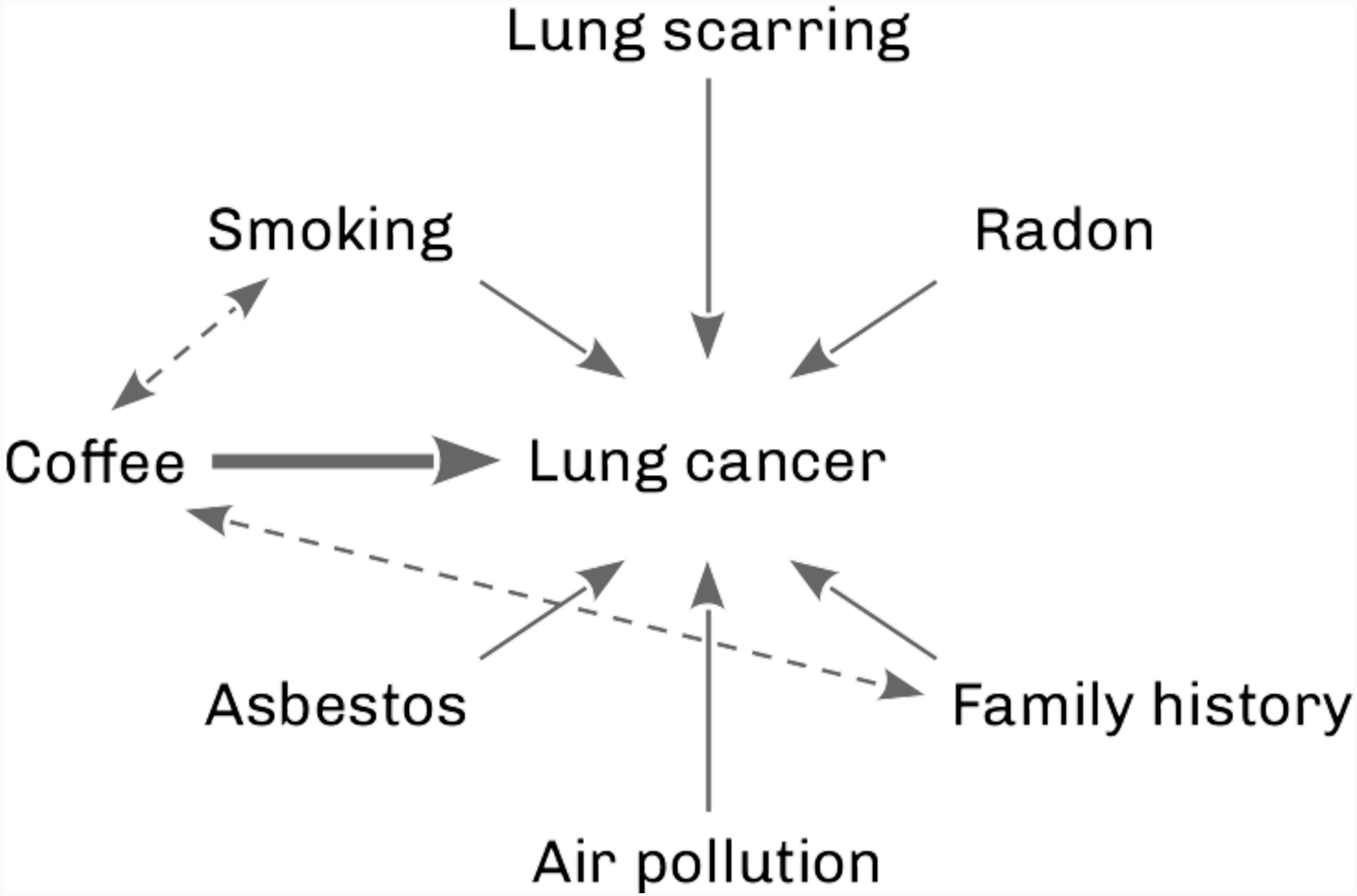
Graphic Model of Causation for Lung Cancer



Graphic Model of Causation for Lung Cancer



Graphic Model of Causation for Lung Cancer



Comparative Measures of disease occurence

- Measures of disease occurrence are compared either as ratios or differences to derive “measures of association” or “effect estimates” that quantify potential causal relations between exposure and disease.

- Choice of “effect estimate” to use in a study depends on:
  - Study objectives
  - What kind of outcome is being measured
  - Type of analyses
  - Conformation of data to statistical assumptions
  - How the investigator wants to convey the study's findings and interpretation.

Risk

- Widely used and readily understood.
- Interpreted as a probability.
- Must be interpreted over a known length of time.
- Average risk for a group is also referred to as the **incidence proportion** or **cumulative incidence**.

$$\text{Risk} = \frac{\text{No. or subjects developing disease during a time period}}{\text{No. of subjects followed for the time period}}$$

Risk

- Must consider induction and latent periods for the component cause being studied.
- Difficult to assess over long periods of time because of competing risks and losses to follow-up.
- Mortality studies have no competing risks.

$$\text{Risk} = \frac{\text{No. of subjects dying during a time period}}{\text{No. of subjects followed for the time period}}$$

Incidence Rate

- Used to address the problems of competing risks and follow-up losses.
- This is a true rate, not a proportion or probability, expressed as a change per unit time.
- Number of cases are the same, but the denominator is the person-time experience of the population being followed.

$$\text{Incidence Rate} = \frac{\text{No. of subjects developing disease during a time period}}{\text{Total time experienced for the subjects followed}}$$

Incidence Rate

- Person-time experience is the sum of the time that each person in the group being followed.
- Person-time experience can stop if disease occurs or can continue for recurrence diseases or outcomes.
- Time in the denominator should include every moment in which a person being followed is at risk for an event that would get tallied in the numerator.

Incidence Rate

- Mortality rate is an incidence rate where the event being measured is death.
- Death is an easily measured outcome, but may be caused for different reasons unrelated to the exposure being studied.
- Incidence rates treat one unit of time as equivalent to another.

Risk vs. Rate

Risk

- Range: 0 to 1
- Dimensionless (probability)
- Easily interpreted
- Whether disease occurred

Rate

- Range: 0 to ∞
- Units: Number/Time
- Requires calculation
- When disease occurs

$$\text{Risk} = \text{Incidence Rate} \times \text{Time}$$

- (Only an approximation: the probability of developing a disease cannot be obtained by simply multiplying average incidence rate by the observation period because the occurrence of disease removes them from the cohort.)

Methods of risk estimation

These methods estimate cumulative incidence or incidence proportion for disease outcome and has advantages, disadvantages, and specific assumptions.

- Simple risk – requires all subjects be followed for entire period of follow-up
- Density – requires prior incidence rate estimation
- Life table – does not require prior incidence rate estimation
- Product limit (Kaplan-Meier) – requires knowledge of follow-up time for each subject

Simple risk method

- This method is appropriate only when there are no or very few:
  - withdrawals from observation (e.g., death from competing cause or loss to follow-up) and,
  - changes that make the study subjects no longer at risk for the disease.

$$\text{Simple risk} = \frac{\text{Number of incident cases of disease observed during the follow-up period}}{\text{Number of persons at risk at the start of the follow-up period}}$$

- Validity of this method requires that all subjects be followed up.
- There are many inherent limitations.

Density method

$$\text{Risk} = 1 - e^{-\text{mean incidence rate} \times \text{observation period}}$$

- Exponentiation of relation between incidence rate and risk.
- *e* (Euler's constant) is the base of the natural logarithm.
- Incidence rate and observation period must be expressed in same units of time.
- Incidence rate must be the average over total time period.

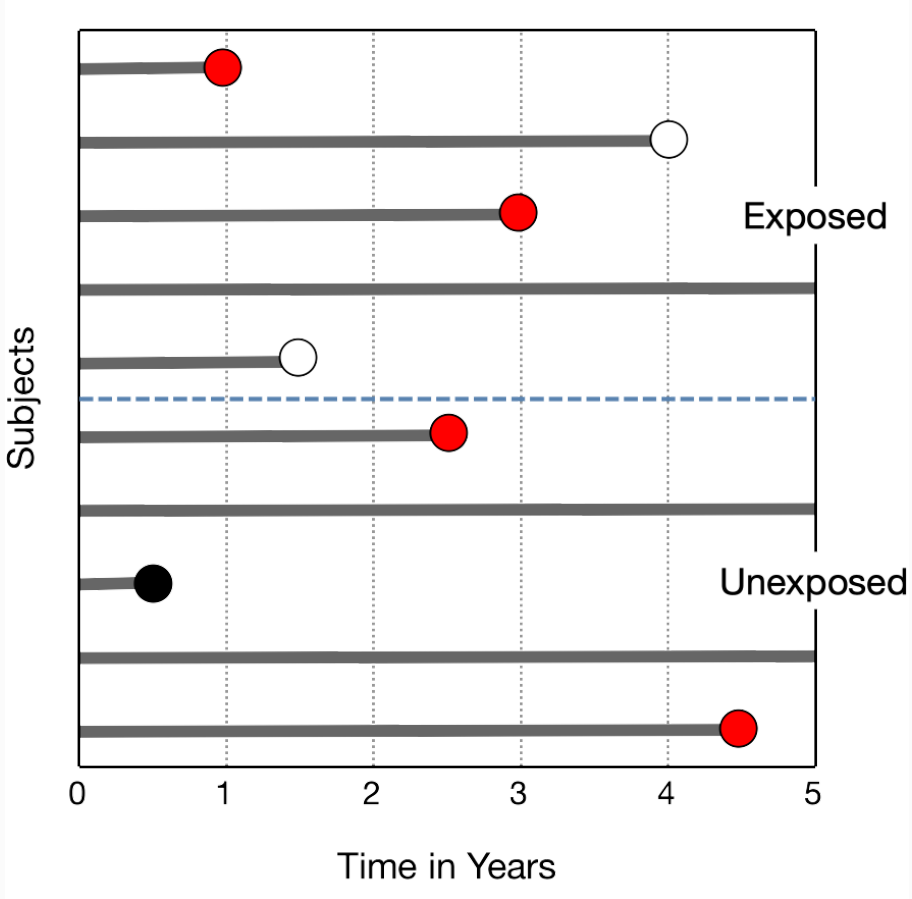
Density method

> Average incidence rate for a major coronary event in male smokers between ages of 40 and 64 is 13 per 1,000 men per year. What is the cumulative incidence over this 25 year period?

$$\text{Risk} = 1 - e^{-0.013 \text{ per year} \times 25 \text{ years}} = 0.277$$

- Not the same as 0.013 x 25 = 0.325
- If incidence rate changes, can calculate separately for subinterval of time periods

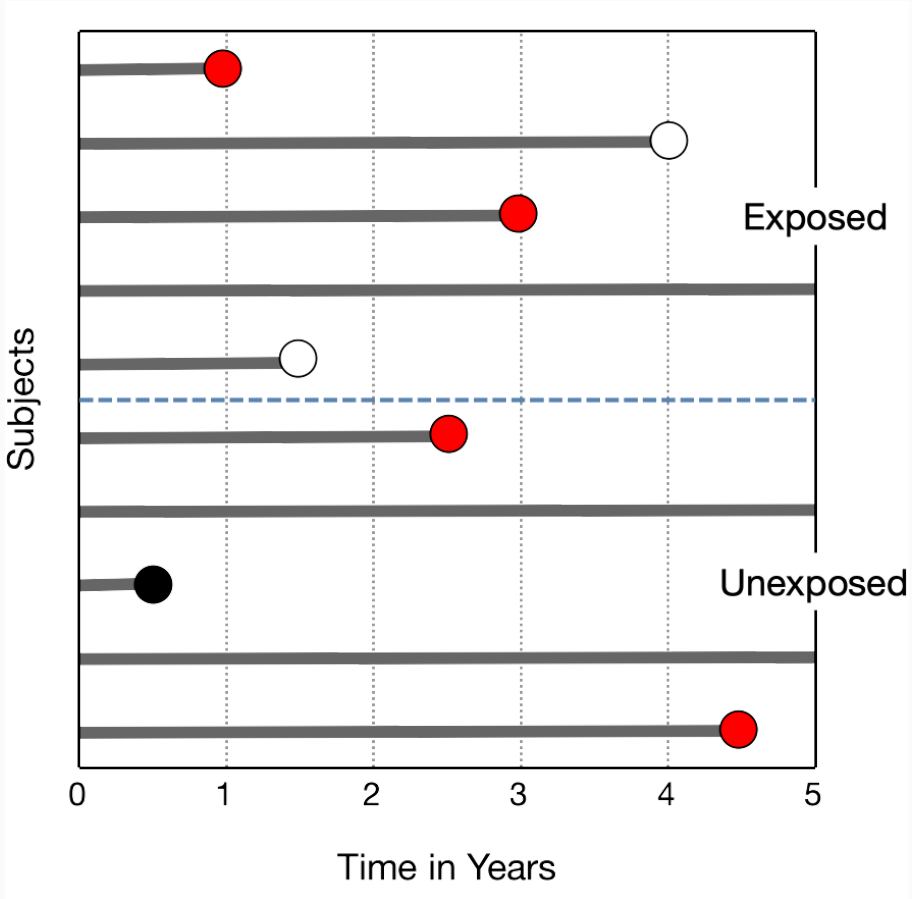
Rates and risks



- Red dots developed disease.
- White dots lost to follow-up and black dot died (i.e., censored)
- Simple risk estimation for entire cohort = 0.40 at five years.
- Combined experience at five years is 32 person-years.
- Average incidence rate is 4 cases / 32 person-years = 0.125 per year

Risk (cumulative incidence) = $1 - e^{-0.125 \text{ per year} \times 5 \text{ years}} = 0.465$

Comparison of rates and risks



- Simple risk estimation for both exposed and unexposed = 0.40 at five years (i.e., risk ratio = 1.0)
- Combined experience at five years is 14.5 person-years for exposed and 17.5 person-years for non-exposed.
- Incidence rates:
  - Exposed = 2 persons / 14.5 person-years (0.138/year)
  - Unexposed = 2 persons / 17.5 person years (0.114/year)

Comparison of rates and risks

- Simple risk estimation for both exposed and unexposed = 0.40 at five years (i.e., risk ratio = 1.0)
- Risk exposed = $1 - e^{-0.138 \text{ per year} \times 5 \text{ years}} = 0.498$
- Risk unexposed = $1 - e^{-0.114 \text{ per year} \times 5 \text{ years}} = 0.434$
- Incidence rate ratio = 0.138/0.114 = 1.21
- Cumulative incidence (risk) ratio = 0.498/0.434 = 1.15

Comparison of rates and risks

- Ratios measure relative effect of exposure.
- Rate and risk differences measure absolute effects and are call measures of impact.
- Risk difference for hypothetical cohort study = risk in exposed (0.498) – risk in unexposed (0.434) = 0.064.
- Synonymous with "**attributable risk**," which assumes a causal relation between exposure and outcome.

Life table method

- Cumulative incidence is estimated directly without having to estimate rates.
- Need to categorize follow-up period into intervals and count the numbers of new cases and withdrawals within each interval.
- Risk for getting disease is equal to one minus the probability of not getting the disease in the time period.
- Example - 10% of disease/year with 2 year follow-up.

$$\text{Risk} = 1 - (1 - 0.10)\,(1 - 0.10) = 1 - 0.81 = 0.19$$

Life table method

- Cumulative incidence 10% of disease/year for 3 years.

$$\text{Risk} = 1 - (1 - 0.10)(1 - 0.10)(1 - 0.10)$$
$$= 1 - (0.9)(0.9)(0.9)$$
$$= 1 - 0.729 = 0.271$$

- Risk for getting disease increases every year but never gets to one.
- The risk for disease can change from interval to interval.

Epidemiologic measures

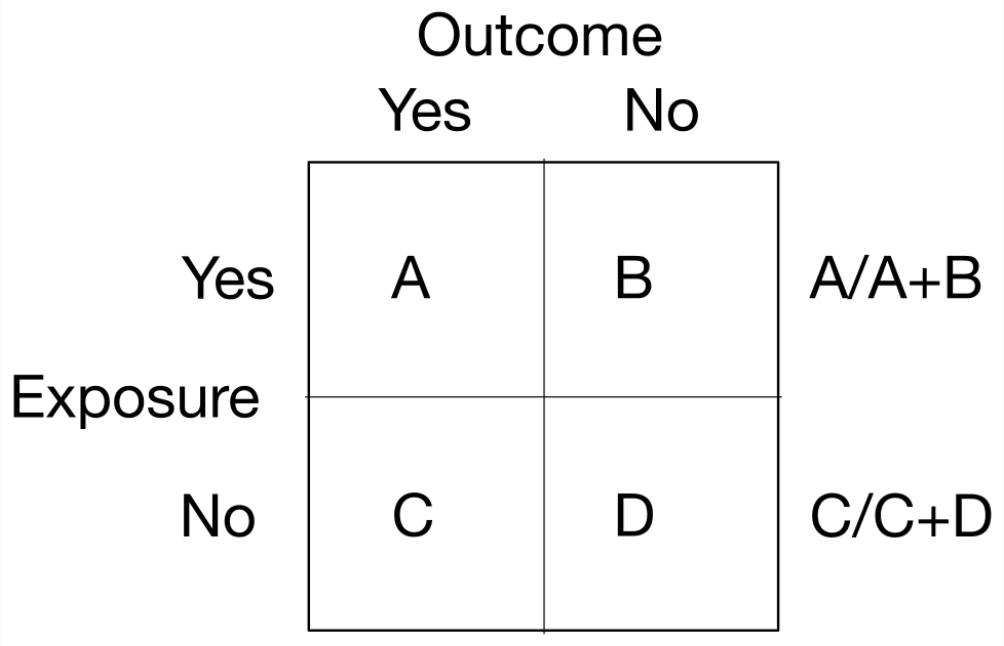
Measures of association

- Reflect the strength or magnitude of the statistical relation between exposure status and disease occurrence.

Measures of effect

- Certain measures of association involving disease incidence that reflect a causal parameter in a particular base population if the two exposure groups are comparable.

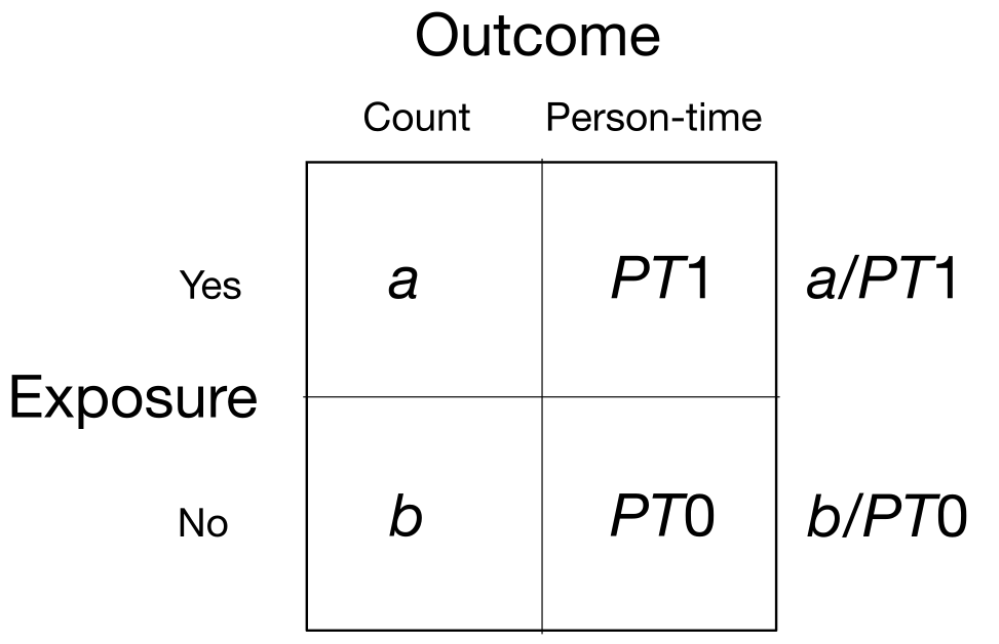
Closed cohort



Risk Ratio = (A/A+B) / (C/C+D)

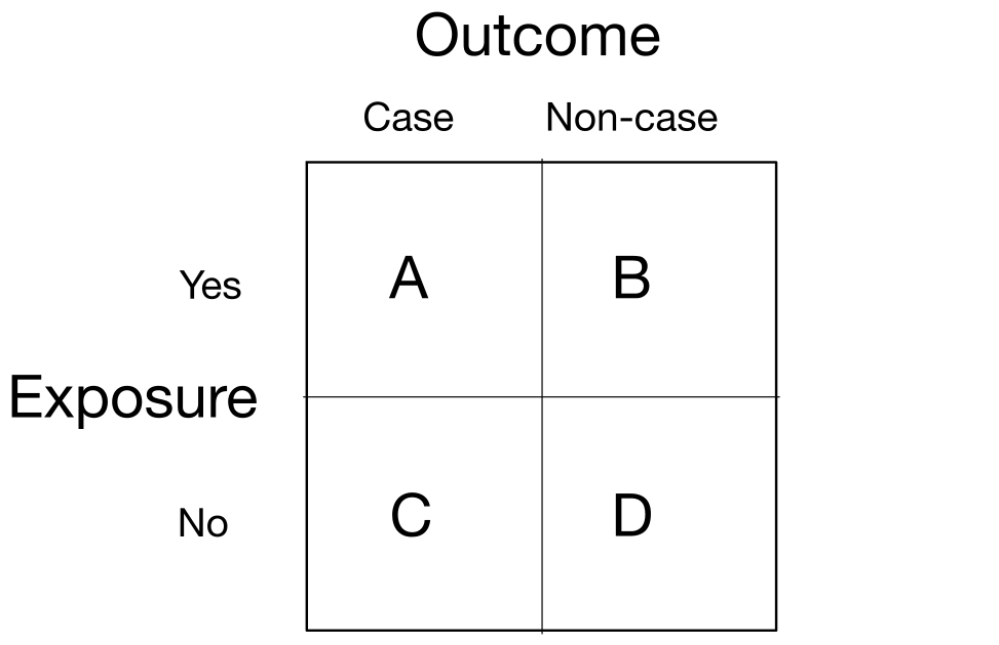
Risk Difference = (A/A+B) – (C/C+D)

Open cohort



- Rate Ratio = $(a/PT1) / (b/PT0)$
- Rate Difference = $(a/PT1) - (b/PT0)$

Case-Control



- Exposure Odds Ratio
  = (A/C) / (B/D)
  = Disease Odds Ratio
  = (A/B) / (C/D)
  = AD/BC

Odds ratio

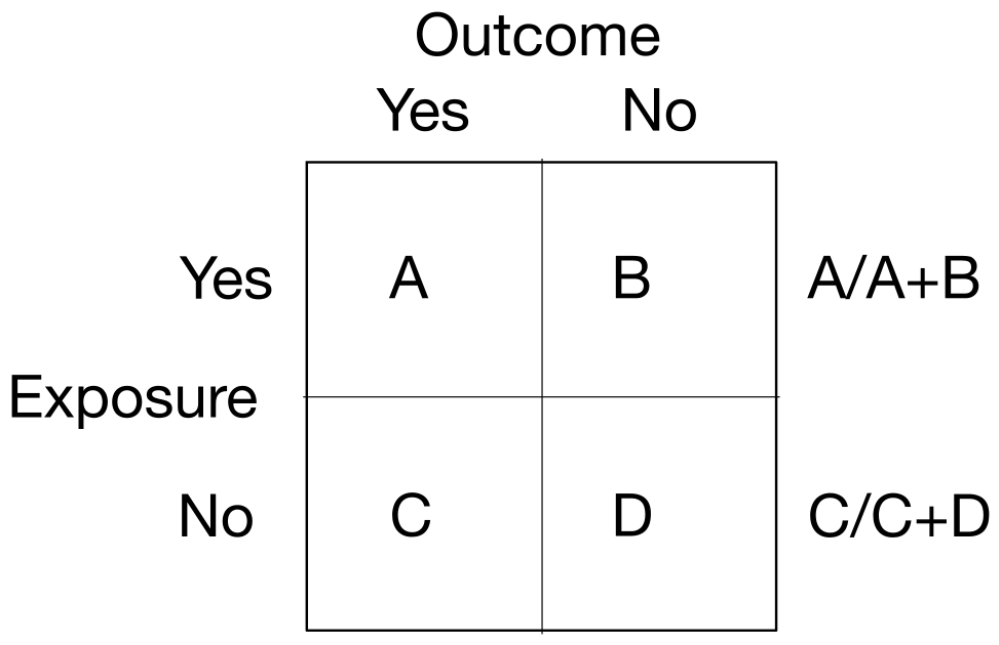
- Odds of having exposure among the cases = (A/A+C)/(C/A+C).
- Odds of having exposure among the controls = (B/B+D)/(D/B+D).

$$OR = \frac{(A/A+C)/(C/A+C)}{(B/B+D)/(D/B+D)}$$

$$= (A/C) \,/\, (B/D)$$

$$= AB/CD$$

Closed cohort



- Odds Ratio = (A/C) / (B/D)

Odds ratio

- Have specific mathematical advantages over other comparative ratios.
- Cannot estimate “odds difference.”
- Calculated for all case-control studies and for observational cohort and clinical trials with logistic regression analysis.
- **Overestimates** the true risk ratio, especially when the outcome is common.

Hypothetical closed cohort

Lung cancer

		Yes	No	
Treatment	A	50	950	50/1000
	B	20	980	20/1000

$$\text{RR} = \frac{50/1,000}{20/1,000} = 2.50 \text{ (95\% CI 1.50 - 4.17)}$$

$$\text{OR} = \frac{50/20}{950/980} = 2.58 \text{ (95\% CI 3.26 - 4.90)}$$

Hypothetical closed cohort

Lung cancer survival at 10 years

		Yes	No	
Treatment	A	500	500	500/1000
	B	200	800	200/1000

$$RR = \frac{500/1,000}{200/1,000} = 2.50 \text{ (95\% CI 2.18 - 2.87)}$$

$$OR = \frac{500/200}{500/800} = 4.00 \text{ (95\% CI 3.26 - 4.90)}$$

Mortality as outcome

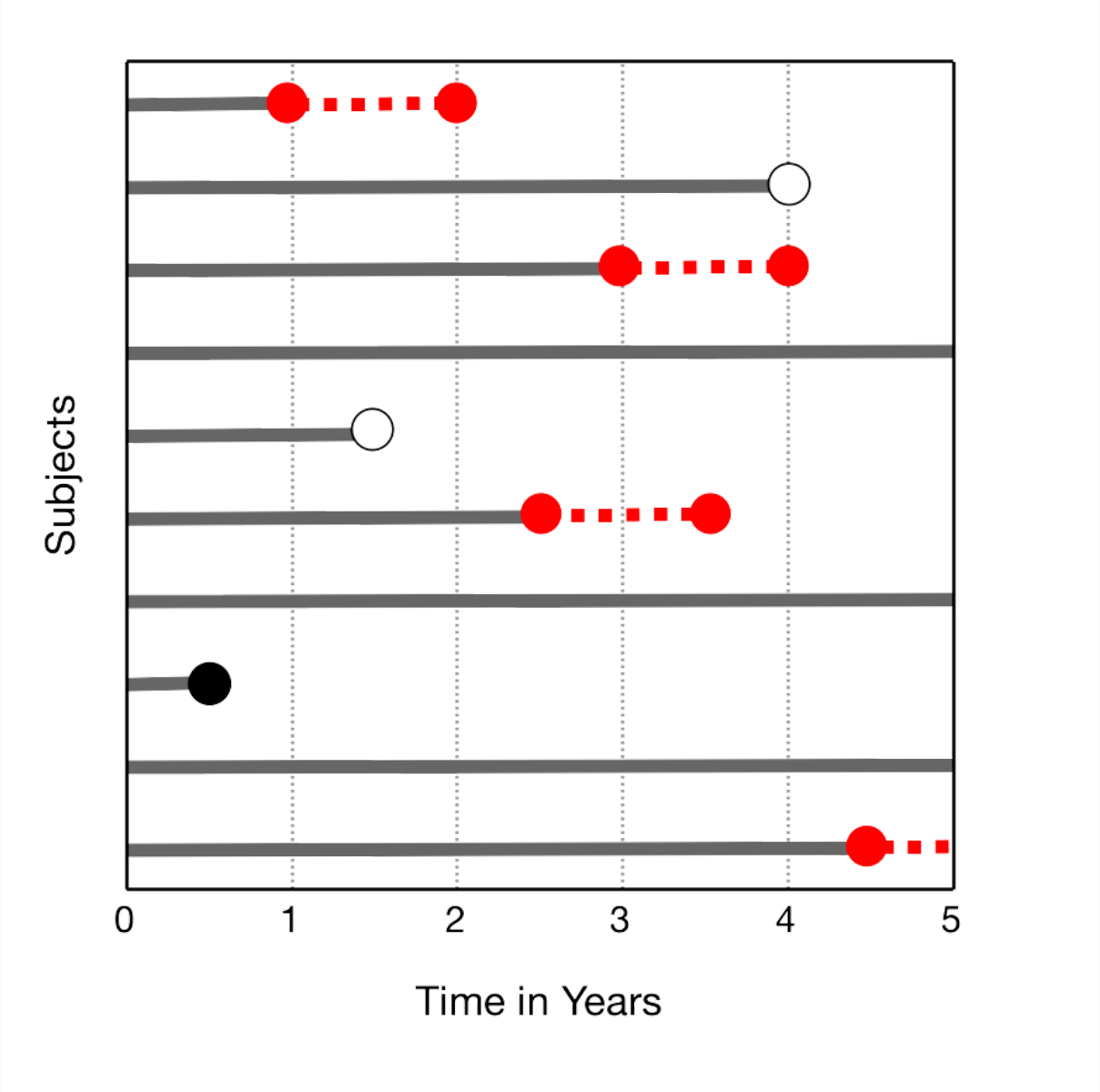
- Usefulness of mortality data as a measure of disease frequency depends upon:
  - High case fatality rate
  - Relatively easily diagnosed
  - Rapidly fatal
  - Accurately recorded on death certificate
  - A clinically important outcome
- Death is an easily measured outcome, but may be caused for different reasons unrelated to the exposure being studied.

Prevalence

- Measures disease status rather than onset.
- Proportion of people in a population that has the disease at one point in time.
- Affected by both incidence rate and duration of disease – greater incidence and longer duration results in higher prevalence.
- Short duration can result from rapid recovery or death.

Prevalence

- Factors affecting prevalence may actually affect duration not incidence.
- Not particularly useful for assessing causal association (major problem is temporal ambiguity).
- Very helpful for assessing disease burden, medical needs, and diseases of insidious onset.
- Steady state – incidence rate and disease duration are constant and no migration.


Subjects
0
1
2
3
4
5
Time in Years