Berkeley School of Public Health

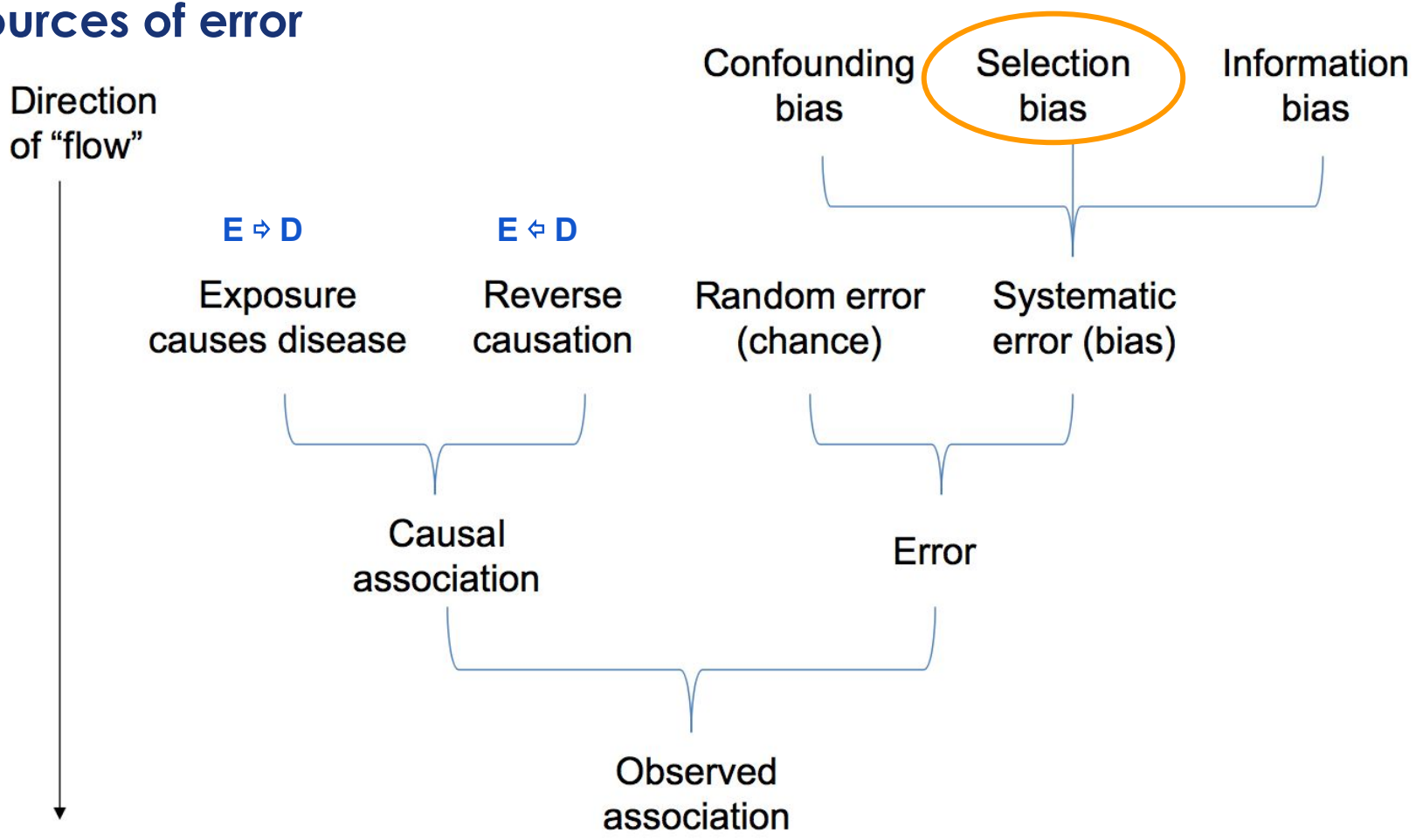
Selection bias

PHW250B

Outline

- Overview of sources of error
- Definition of selection bias
- Types of selection bias
  - Self-selection bias
  - Healthy worker effect
  - Differential loss to follow-up
  - Berkson's bias
- Quantitative assessment of selection bias
- Distinguishing selection bias from confounding
- Preventing and correcting for selection bias

Sources of error



Overview of selection bias

- **Selection bias:** “distortions that result from procedures used to select subjects and from factors that influence study participation.”
- The relationship between exposure and disease is different for those who participate and for those who should theoretically have been eligible for the study.
- When selection bias is present, associations represent a mix of:
  - forces that determine participation and
  - forces that determine disease.

Rothman et al., *Modern Epidemiology*, 3rd Ed.

Self-selection bias

- This bias occurs when people are allowed to self-select into a study and their desire to participate in the study is associated with the exposure and outcome.
- **Example**: In a study of leukemia among troops who worked at the Smoky Atomic Test in Nevada, 18% of troops with known leukemia status asked investigators about participation in the study (instead of being contact by investigators).
- Among troops with known leukemia status,
  - 22% of self-selecting troops had leukemia
  - 5% of non self-selecting troops had leukemia
- This suggests that self-selection bias was present.

Rothman et al., *Modern Epidemiology*, 3rd Ed.

Healthy worker effect

- The healthy worker effect occurs when a study focused on the health of workers compares health outcomes among workers to the general population and the health of workers is better than that of the general population.
- The general population includes people who are not able to work because of illness disability, retirement, etc.
- Traditionally this effect is considered to be a type of self-selection bias, but in the modern view it is classified as confounding (more on this later).

Rothman et al., *Modern Epidemiology*, 3rd Ed.

Differential loss to follow-up

- Occurs when exposed vs. unexposed follow-up rates differ or disease vs. non diseased follow-up rates differ.
- Most common in cohort studies.
- Can think of as “backwards” selection bias since it occurs during and at the end of the study rather than during enrollment.

Rothman et al., *Modern Epidemiology*, 3rd Ed.

Berkson's bias (or "Berksonian bias")

- First described by Berkson in 1946.
- Occurs when both exposure and disease are associated with the risk of hospitalization. If so, a false association is induced between exposure and disease.
- Bias occurs because people are more likely to be hospitalized for two conditions than one.
- **Example**: a hospital-based study of the effect of hypertension on skin cancer.
  - People with hypertension are more likely to be hospitalized (regardless of skin cancer status)
  - People with skin cancer are more likely to be hospitalized (regardless of hypertension status)
  - The proportion of people in the study with neither exposure nor disease will be smaller than in the general population

Rothman et al., *Modern Epidemiology*, 3rd Ed.

Quantitative assessment of selection bias

General population		
	Disease	No disease
Exposure	A	B
No exposure	C	D

Study population		
	Disease	No disease
Exposure	a	b
No exposure	c	d

- To assess possible selection bias, we can assess how well each element of the 2x2 table for the study population reflects each element for the 2x2 table for the general population.
- If the "flow" from the general population to the study population is equal for A, B, C, and D, there is no bias.

Quantitative assessment of selection bias

General population		
	Disease	No disease
Exposure	A	B
No exposure	C	D

Study population		
	Disease	No disease
Exposure	a	b
No exposure	c	d

- If the “flow” from the general population to the study population is different between A, B, C, and D, bias occurs. (shown by thicker arrow)
- This is because (A/C)/(B/D) ≠ (a/c)/(b/d)

  (other measures of association can be assessed as well)

Szklo & Nieto 3rd Ed.

Quantitative assessment of selection bias

Example of case-control study with no selection bias

General population

	Disease	No disease
Exposure	500	1800
No exposure	500	7200

OR = (500/500) / (1800/7200) = 4.0

Study population

	Disease	No disease
Exposure	250	180
No exposure	250	720

OR = (250/250) / (180/720) = 4.0

No selection bias because (A/C)/(B/D) = (a/c)/(b/d)
→ Flow is equal for A, B, C, and D

Quantitative assessment of selection bias

Example of case-control study with selection bias

General population		
	Disease	No disease
Exposure	500	1800
No exposure	500	7200

OR = (500/500) / (1800/7200) = 4.0

Study population		
	Disease	No disease
Exposure	300 ⇧	180
No exposure	200 ⇩	720

OR = (300/200) / (180/720) = 6.0

Selection bias is present because (A/C)/(B/D) ≠ (a/c)/(b/d)
→ Bias away from the null
→ A is overrepresented in the study and B is underrepresented

Quantitative assessment of selection bias

Example of case-control study with "compensating" selection bias

General population

	Disease	No disease
Exposure	500	1800
No exposure	500	7200

OR = (500/500) / (1800/7200) = 4.0

Study population

	Disease	No disease
Exposure	300 ⇧	245 ⇧
No exposure	200 ⇩	655 ⇩

OR = (300/200) / (245/655) = 4.0

Selection bias is present but cancels itself out, so (A/C)/(B/D) = (a/c)/(b/d)
→ A & C are overrepresented and B & D are underrepresented in the study

Distinguishing confounding and selection bias

- Depending on an epidemiologist's particular definition of confounding and selection bias, these concepts may overlap.
- **Confounding** results from differential selection that occurs before exposure and disease leads to and can be controlled for in the analysis.
- **Selection bias** arises from selection affected by the exposure under study (or both exposure and outcome in the case of Berkson's bias) and in most cases cannot be corrected for in the analysis.
- Directed acyclic graphs can be used to distinguish between these two sources of error (more on that later).

Preventing selection bias

- Choose study subjects from defined reference populations
  - Case-control studies
    - Define controls using a primary study base
    - Be very careful about control selection in hospital-based case-control studies.
- Try to minimize loss to follow-up as much as possible
- Avoid enrolling participants based on self-selection
- Attempting to compensate for selection bias is generally not recommended

Correcting for selection bias

- Methods exist to correct for selection bias due to differential loss to follow-up.
- Statistical methods that impute outcomes for people with missing exposure/outcome data, but these approaches may rely on strong assumptions
- Sensitivity analyses can be done that compare results over a range of different assumptions and imputation approaches.

Summary of key points

- Selection bias can occur in any type of epidemiologic study.
- Good epidemiologic design practices can minimize selection bias in most studies.
- Case-control studies are particularly prone to selection bias and warrant a careful assessment of selection bias in the design phase.

Berkeley School of Public Health

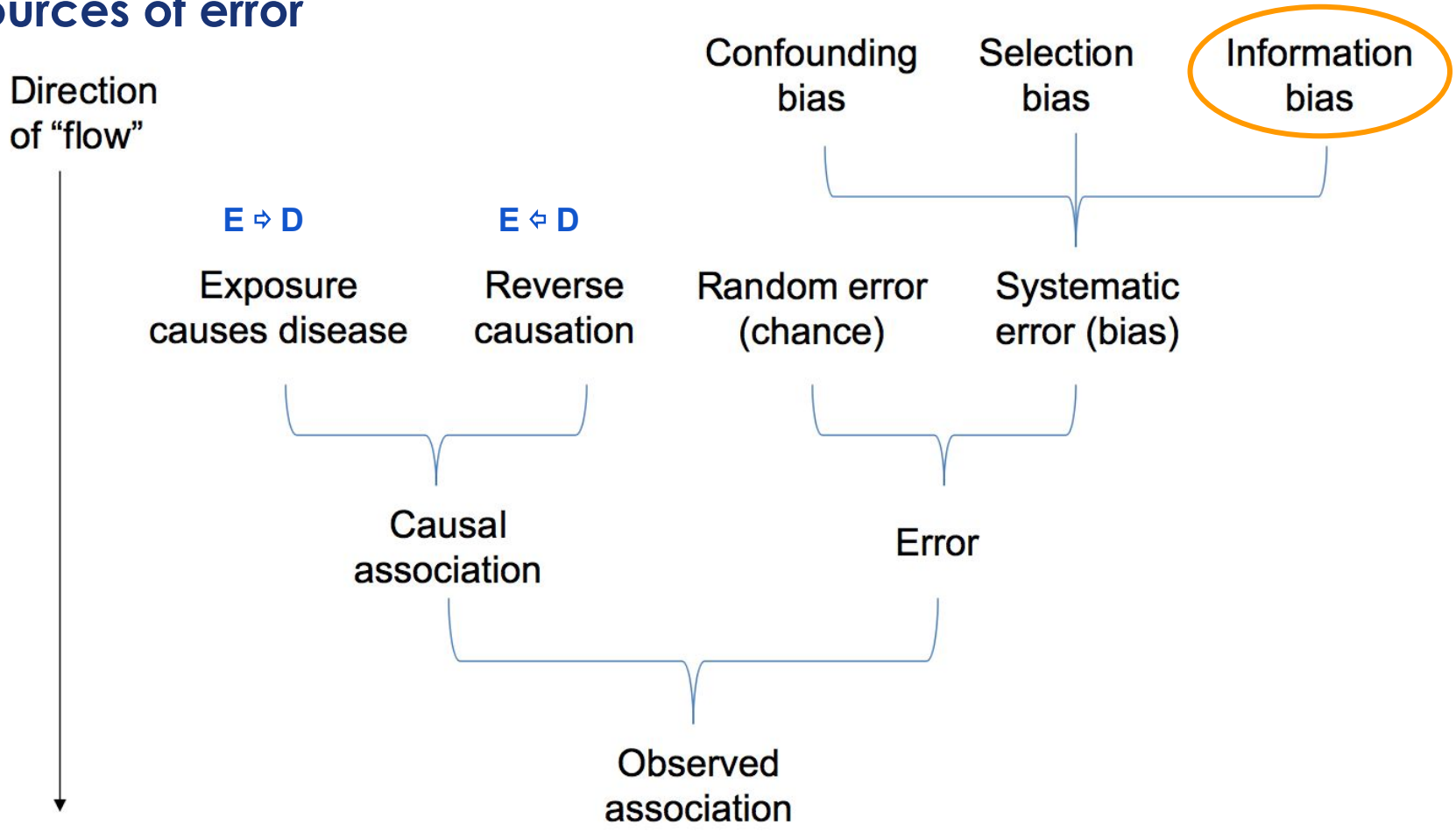
Information bias

PHW250B

Outline

- Definition of information bias
- Types of information bias
  - Recall/respondent bias
  - Interviewer/observer bias
- Result of information bias: misclassification
  - Non-differential misclassification
  - Differential misclassification
- Quantitative assessment of misclassification

Sources of error



Definition of information bias

- **Information bias:** bias that results from imperfect definitions of study variables or flawed data collection procedures
- We can divide this into two categories:
- **Exposure identification bias:** due to problems in the collection of exposure data or definition of exposure
  - Primarily affects cohort studies and case-control studies in which exposure status is assessed after disease status
- **Outcome identification bias:** results from differential or nondifferential misclassification of disease status
  - Occurs in any type of epidemiologic study, especially if self-reported outcomes are used
  - More common in case-control and cohort studies

Recall/respondent bias

- Results from inaccurate recall of past exposure or disease.
- Particularly a concern in case-control studies when cases and controls know their disease status since cases and controls may differentially recall exposure status based on their disease experience.
- Can occur in any study design (e.g., a trial with self-reported outcomes)
- More likely to occur with long recall periods

Szklo & Nieto 3rd Ed.

Recall/respondent bias

- **How to minimize:**
  - Validate responses from study subjects
    - E.g., compare reported exposure to medical charts
  - Use objective exposure markers
    - E.g., in a study of circumcision, use physical examination instead of self-report to assess circumcision status
  - Use diseased controls in case-control studies
    - Can ensure that the amount of “rumination” about causes of disease is similar between cases and controls

Interviewer/observer bias

- Occurs when interviewers are not masked:
  - with regard to disease status of study participants when measuring exposure status (e.g., in a case-control study)

    or

  - with regard to exposure status of study participants when measuring disease status (e.g., in a cohort study or trial)
- Can occur if interviews ask or clarify questions in a different way depending on exposure / disease status
- E.g., in a trial if the interviewer knows the intervention arm a study participant is in, they may consciously or unconsciously ask questions in a way that bias results towards a beneficial effect

Interviewer/observer bias

- **How to minimize:**
  - Use rigorous survey design with a specific protocol to avoid probing by the interviewer that may introduce bias
  - Mask interviewers to disease/exposure status
  - Conduct a reliability or validity substudy to compare interviewer's assessment to a gold standard
  - Compare interviewer's assessment and conduct periodic retraining when systematic differences in exposure/outcome classification occur

The result of information bias: misclassification

- Misclassification of exposure / disease:
  - A participant is classified as exposed but is truly unexposed
  - A participant is classified as unexposed but is truly exposed
  - (same for disease status)
- Due to information bias
- Misclassification can also be due to measurement error
  - e.g., a diagnostic test is used that has imperfect sensitivity or specificity

Types of misclassification

- **Non-differential** misclassification of exposure
  - Exposure misclassification does not depend on disease status
- **Non-differential** misclassification of disease
  - Disease misclassification does not depend on exposure status
- **Differential** misclassification of exposure
  - Exposure misclassification depends on disease status
- **Differential** misclassification of disease
  - Disease misclassification depends on exposure status

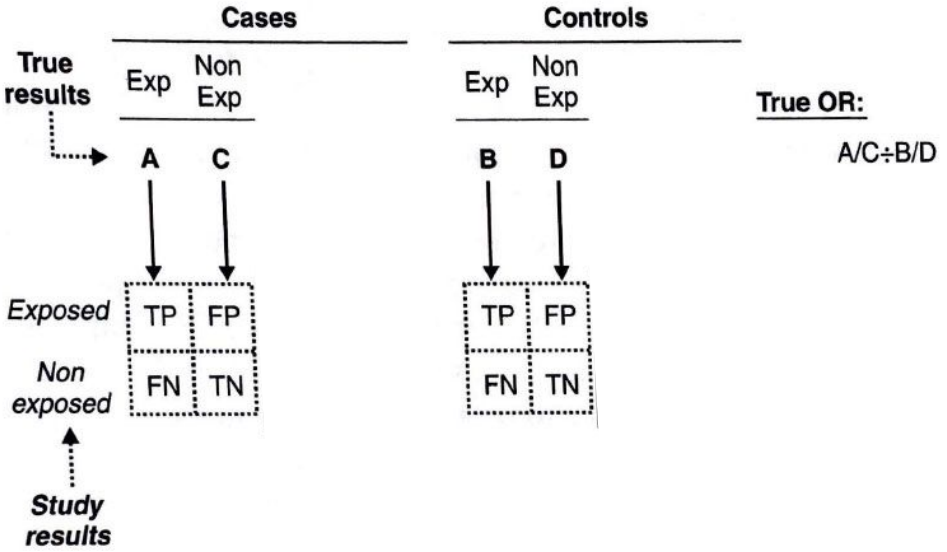
How to assess misclassification

- Often misclassification can occur in multiple forms concurrently
  - E.g., exposed individuals are classified as unexposed AND unexposed individuals are classified as exposed
- As a result, it is useful to use the concepts of sensitivity and specificity when assessing potential misclassification
  - **Sensitivity of exposure classification**: probability that an exposed person is classified as exposed
    - Probability of a true positive
  - **Specificity of exposure classification**: probability than an unexposed person is classified as unexposed
    - Probability of a true negative
  - Analogous definitions for disease classification

Applying sensitivity & specificity to misclassification

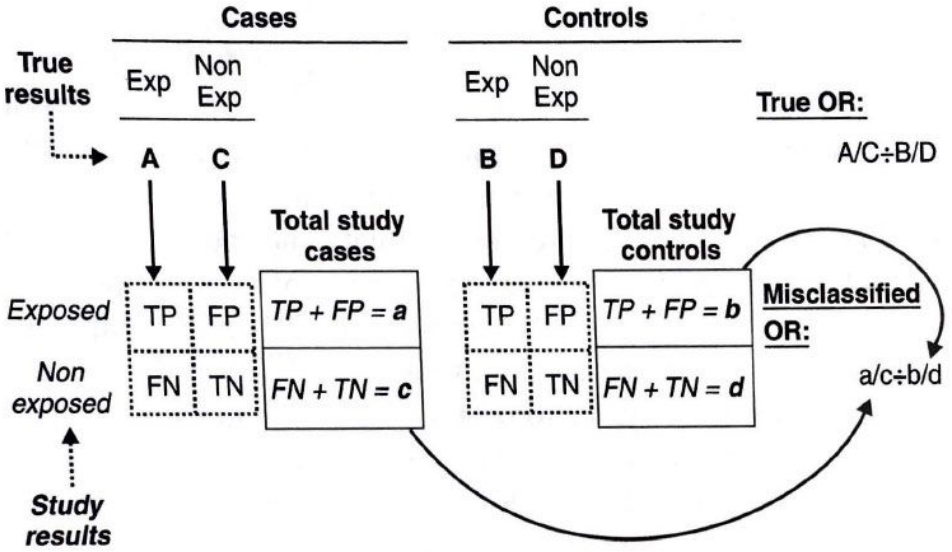
True results	Cases		Controls		
	Exp	Non Exp	Exp	Non Exp	True OR:
→	A	C	B	D	A/C÷B/D

Applying sensitivity & specificity to misclassification



EXP = exposure; TP = true positive; FP = false positive; FN = false negative; TN = true negative

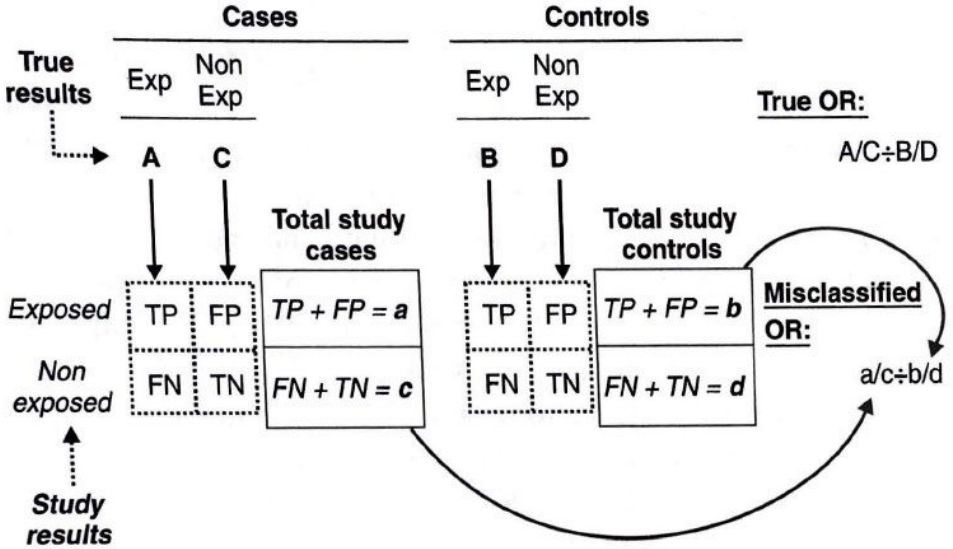
Applying sensitivity & specificity to misclassification



EXP = exposure; TP = true positive; FP = false positive; FN = false negative; TN = true negative

Sensitivity = TP / (TP+FN)
Specificity = TN / (TN+FP)

Applying sensitivity & specificity to misclassification

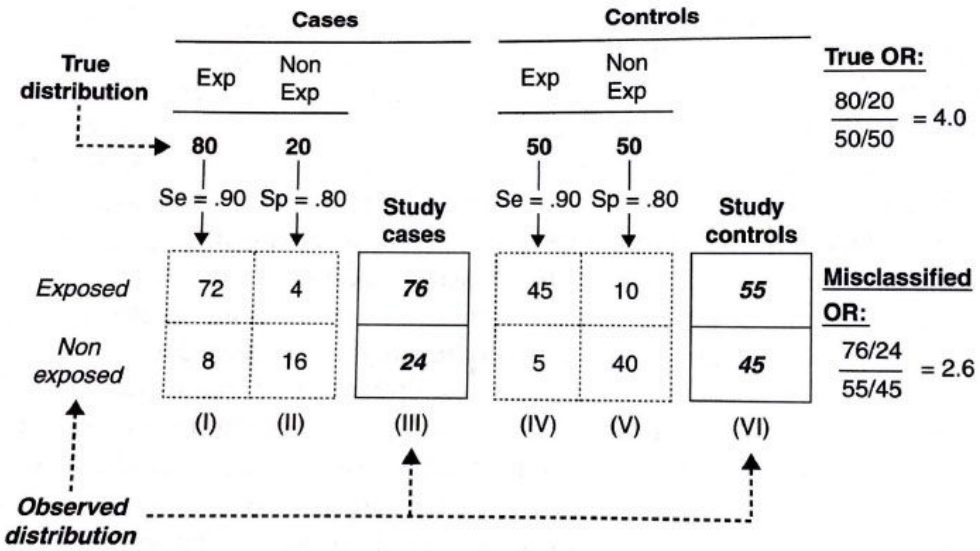


EXP = exposure; TP = true positive; FP = false positive; FN = false negative; TN = true negative

Sensitivity = TP / (TP+FN)
Specificity = TN / (TN+FP)

- If Sensitivity = 1 and Specificity = 1, cell counts (a, b, c, d) in the study are equal to those in the true distribution, and there is no misclassification.

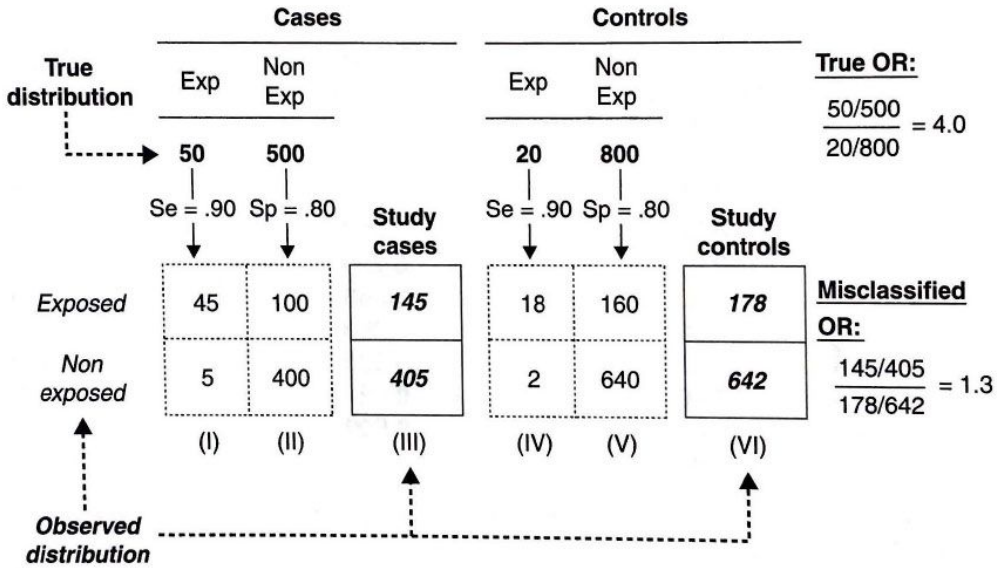
Example of non-differential misclassification



Exp: Exposed; Non exp: Non exposed; Se: Sensitivity: Sp: Specificity; OR: Odds ratio

- In the example, there is misclassification of the exposure in both directions since Se = 0.90 and Sp = 0.8
  - 10% of people classified as unexposed are truly exposed
  - 20% of people classified as exposed are truly unexposed
- Non-differential because Se and Sp are the same for cases and controls
- **Result: bias towards the null** (Study OR = 2.6 < True OR = 4.0)

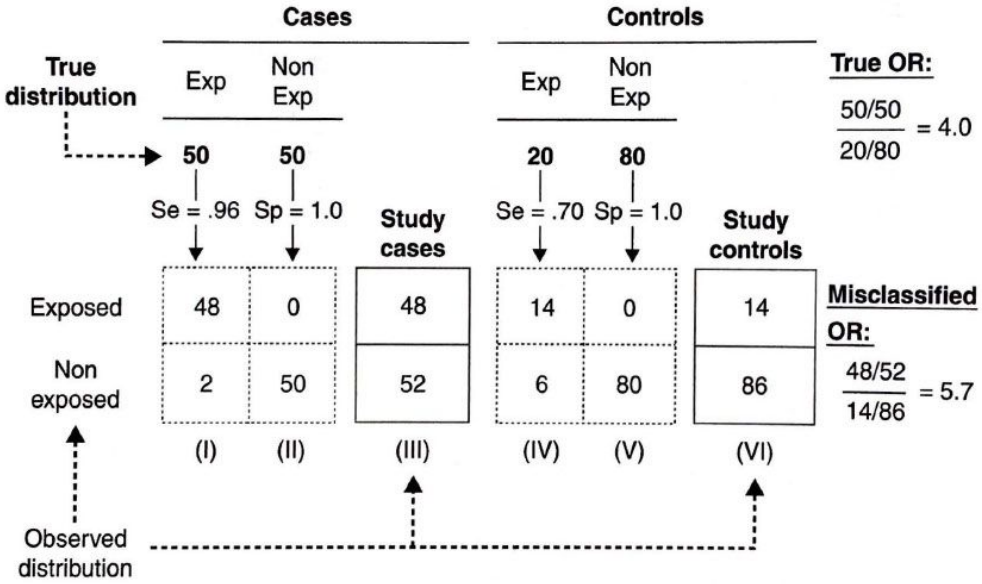
Example of non-differential misclassification with low prevalence of exposure



Exp: Exposed; Non exp: Non exposed; Se: Sensitivity: Sp: Specificity; OR: Odds ratio

- Same as previous example but the exposure is less common (26% of cases and 22% of controls are exposed)
- Non-differential because Se and Sp are the same for cases and controls
- **Result: stronger bias towards the null**
  (Study OR = 1.3 < True OR = 4.0)

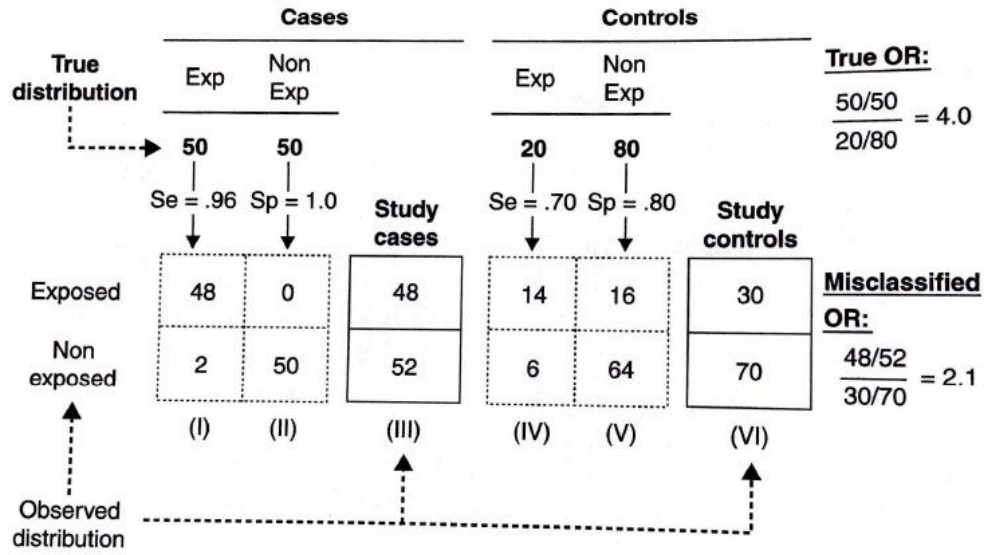
Example of differential misclassification with imperfect sensitivity and perfect specificity



Exp: Exposed; Non exp: Non exposed; Se: Sensitivity; Sp: Specificity; OR: Odds ratio

- In the example, there is misclassification of the exposure in one direction since Se<1 and Sp=1
  - 4% of cases and 30% of controls classified as unexposed are truly exposed
- Differential because Se differ between cases and controls
- **Result: bias away from the null** (Study OR = 5.7 > True OR = 4.0)

Example of differential misclassification with imperfect sensitivity and specificity



Exp: Exposed; Non exp: Non exposed; Se: Sensitivity; Sp: Specificity; OR: Odds ratio

- In the example, there is misclassification of the exposure in both directions since neither Se or Sp equal 1
- Differential because Se and Sp differ between cases and controls
- **Result: bias towards the null** (Study OR = 2.1 < True OR = 4.0)

Misclassification of a confounder

- In addition to confounders and exposures, confounding variables can be misclassified.
  - E.g., a participant's education level might be misclassified
- Non differential misclassification of a confounder results in imperfect adjustment when the variable is controlled for in an analysis or matched on in the design phase.
- Thus, when there is misclassification of a confounder, there can be residual confounding of the measure of association between exposure and disease.

Summary of key points

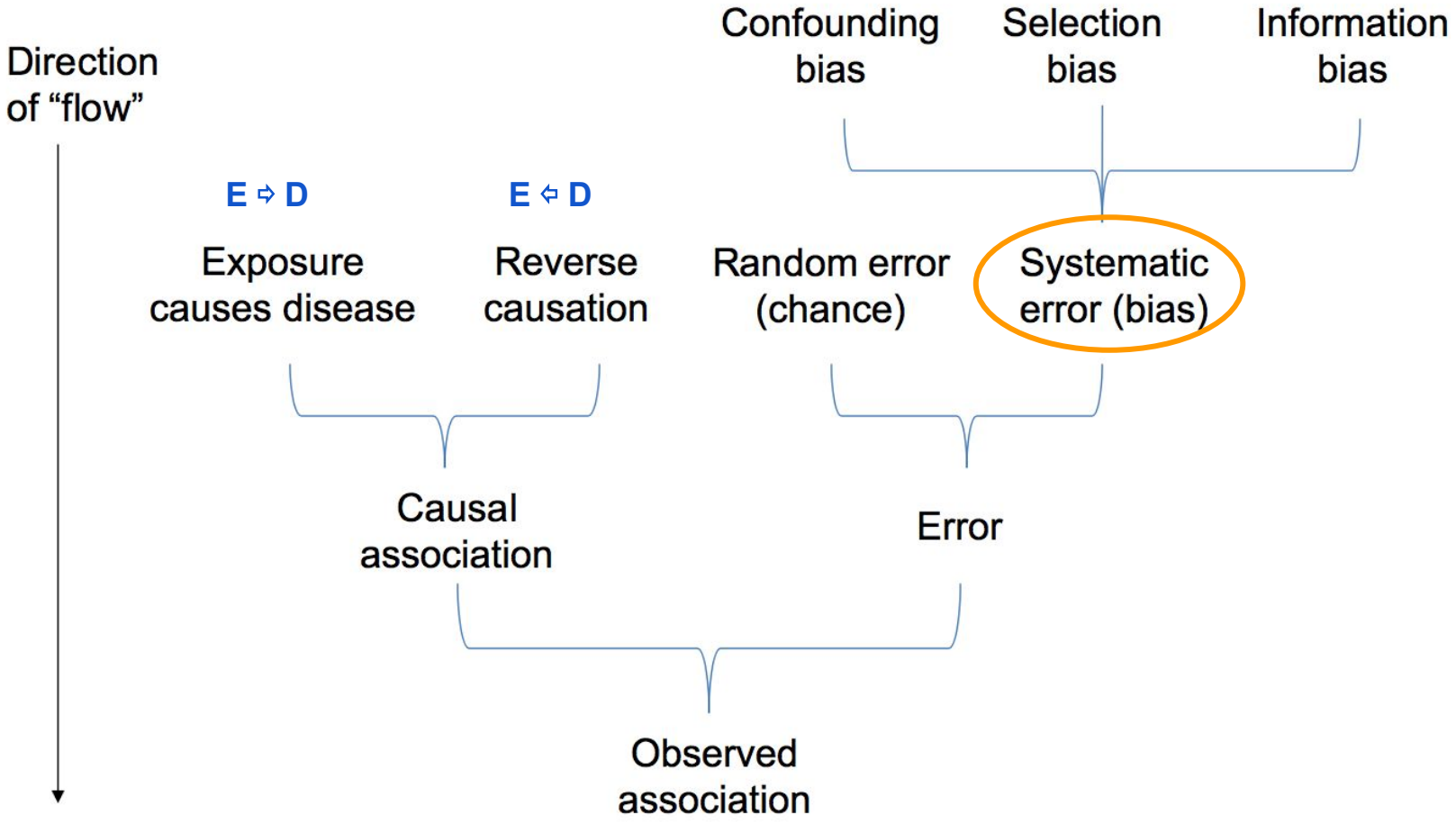
- Information bias can result from the way exposure, confounders, or outcomes are measured in a study.
- Information bias can occur in any type of epidemiologic study.
- Careful study design and detailed training and study protocols can help reduce information bias.
- Misclassification is the result of information bias.
- Non-differential misclassification biases results towards the null.
- Differential misclassification biases results in an unknown direction.

Berkeley School of Public Health

Generalizability

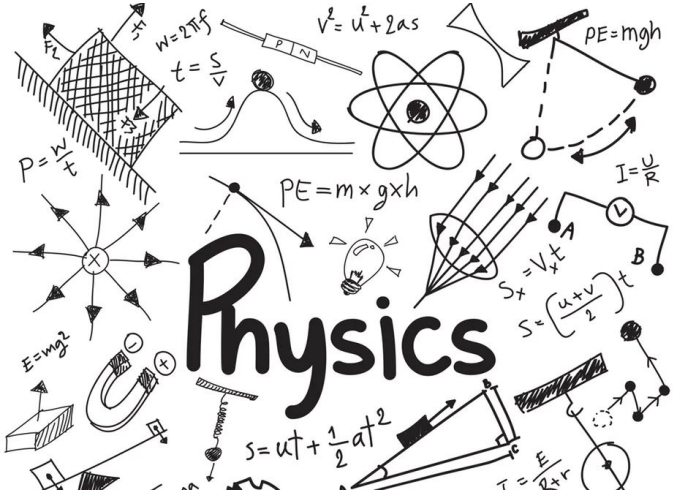
PHW250B

Sources of error



Generalizability in scientific disciplines

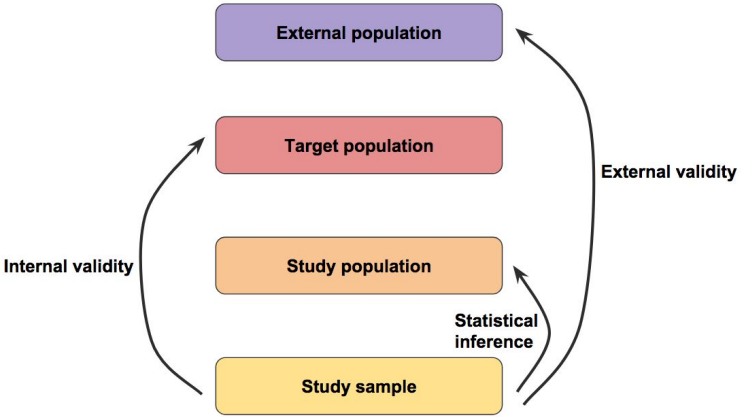
- In some fields, like physics, it's reasonable to assume that the laws of nature have universal applicability (ie, high generalizability)
- However, in biomedical science, we often restrict our inferences to specific populations.
- This stems from the assumption that biological effects can differ across populations



Rothman et al., *Modern Epidemiology*, 3rd Ed.

Generalizability in epidemiology

- In epidemiology, many studies enroll a study population with the goal of making inferences about a target population from which that study population is sampled.
- However, the focus on selecting study populations that represent target populations can detract from making valid causal inferences that may be true across populations.
- In biology, experiments are often done among animals with characteristics that will maximize the internal validity of the experiment with little concern about external validity.
- In epidemiology there may be a trade off between external and internal validity— enrolling a population ideal for internal validity often reduces generalizability of a study.



Rothman et al., *Modern Epidemiology*, 3rd Ed.

Internal vs. external validity

- Epidemiologic studies that prioritize external validity may make it more difficult to control for confounders that vary in the population and difficult to ensure uniformly high levels of cooperation and accurate measurements.
- To maximize validity, it is ideal to select study groups that have similar levels of confounders, who are very cooperative, and for whom accurate measurement can be done.
- **Examples**: British Physicians' Study and Nurses' Health Study both were large, impactful studies in non-representative populations that allowed for high internal validity.

Rothman et al., *Modern Epidemiology*, 3rd Ed.

Summary of key points

- Generally, for a given **etiologic research question** about whether an exposure causes a disease it is best to first conduct studies that maximize internal validity to establish the direction and size of effects.
- Then subsequent studies can be done to assess whether these effects hold true in other populations.
- Some studies, such as **impact evaluations** evaluating the impact of a specific program conducted in a specific population at a specific time, do not aim to study etiology but rather are focused on inferences about that population.
    - In this case, it is often important to design studies to balance both internal and external validity.

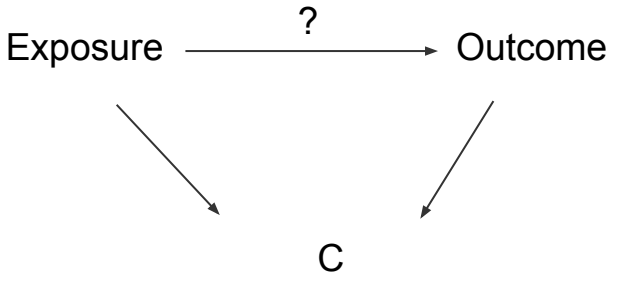
Berkeley School of Public Health

Diagnosing bias using DAGs

PHW250B

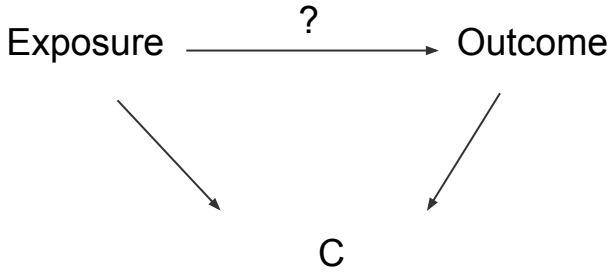
Directed acyclic graphs (DAGs) and bias

- DAGs are a powerful tool in the modern epidemiology toolkit for diagnosing bias.
- They allow us to diagnose bias using our beliefs about the causal relationships between variables in our study.
- They can be used during the design or analysis phase to detect possible confounding.

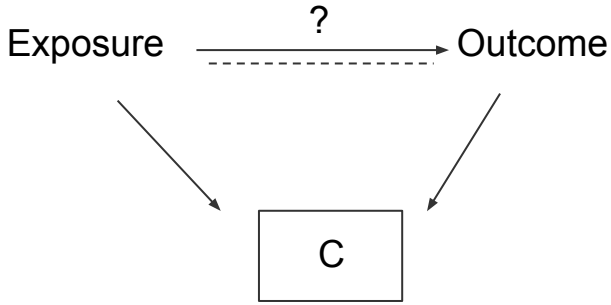


Colliders

- When diagnosing bias, we will look for a node within our DAG called a collider.
- A **collider** is a DAG configuration with a variable that has two directed paths into it.
- In this DAG, C is a collider.
- Conditioning on a collider is indicated by drawing a box around that variable.
  - Conditioning means we **stratify** for that variable or **adjust** for that variable.
  - In other words, we fix the levels of that variable in our design or analysis.
- Conditioning on a collider induces a statistical association between its parents.



Conditioning on a collider

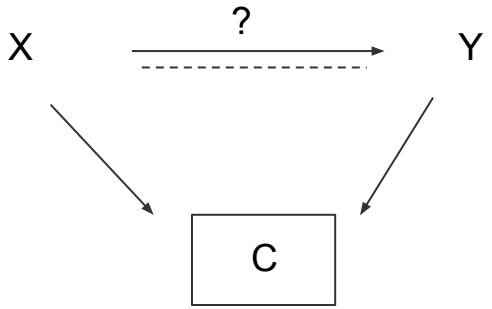


Why does conditioning on a collider induce an association between its parents?

Example:

- C = X + Y
- If you know X = 3 you don't know the value of Y because X and Y are independent.
- If you know C = 10 and X = 3 then you know the value of Y.
- This is because X and Y are dependent **given that C = 10**.
- i.e., when we set C to a fixed value, knowing the value of X allows us to know the value of Y and vice versa.

Conditioning on a collider

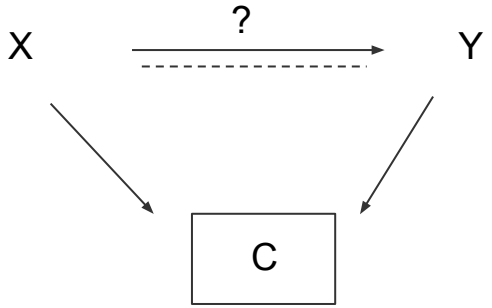


What we mean by "condition"

If C is hospitalization, conditioning on C could mean:

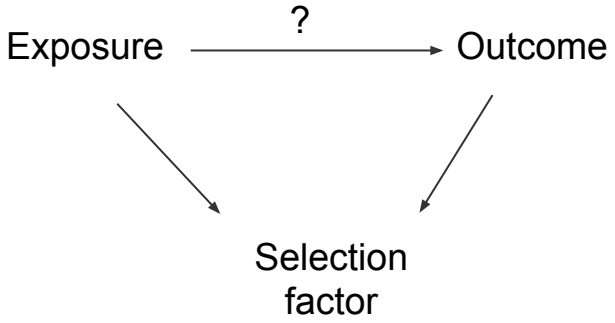
- **Restriction**: We restrict the value of C to a single number.
  - We restrict to only people who were hospitalized.
- **Stratification/ Adjustment**: We estimate the exposure-outcome association within each level of C.
  - We estimate the association separately among the hospitalized and among the not hospitalized.

Conditioning on a collider

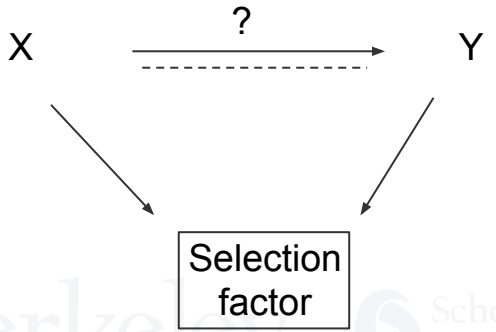


Diagnose selection bias with a DAG

- The key relationship to look for when diagnosing bias:
  - Exposure causes a selection factor
  - Outcome causes a selection factor
- In other words, check whether there is a selection factor that is a collider of the exposure outcome relationship.
- Selecting on a factor that is a collider is the same as conditioning on a collider (by restricting), which induces a false exposure-outcome association.
  - You only have data on people with the selection characteristic.

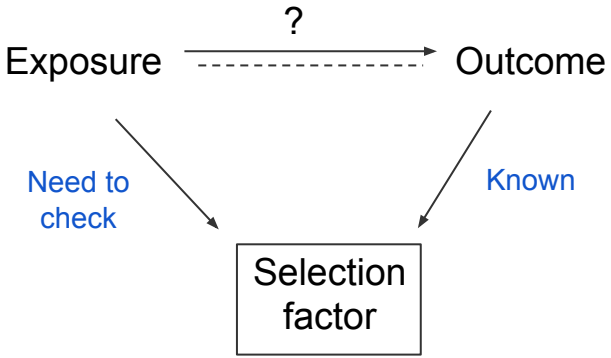


Conditioning on a collider

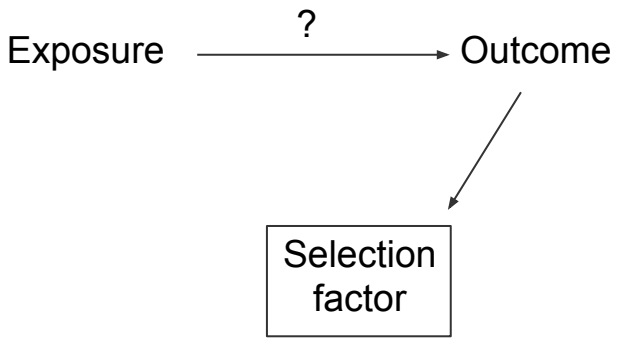


Selection bias in a case-control study

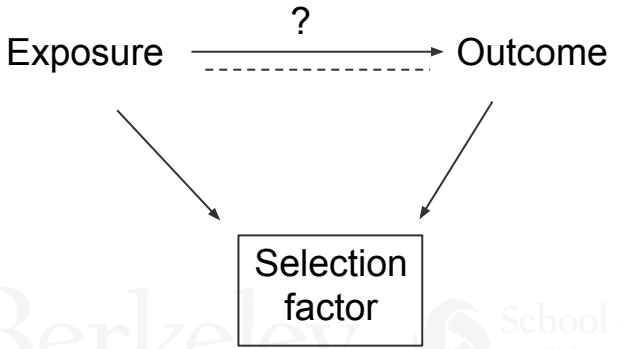
- In a case-control study, we know whether we selected cases by restricting to a certain population subgroup (e.g., hospital patients)
- We don't know whether the exposure causes the selection factor so we need to assess this.



No selection bias

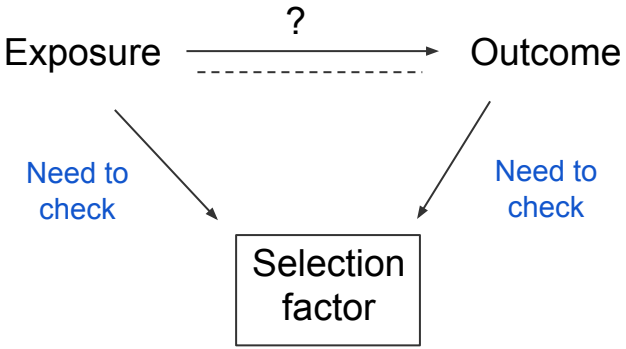


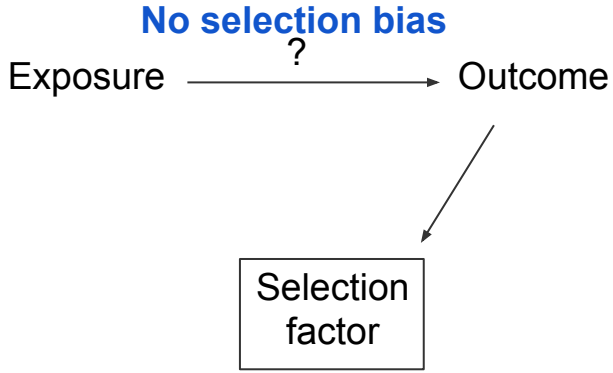
Selection bias

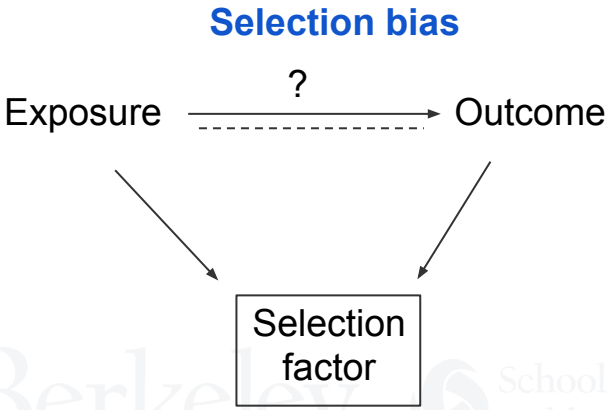


Selection bias in a cohort study

- In a cohort study, we need to assess both whether the exposure and the outcome cause a selection factor.

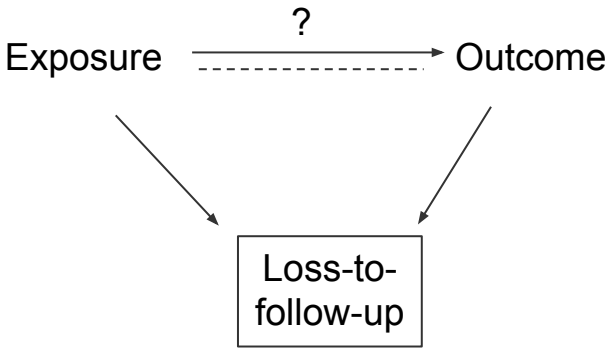






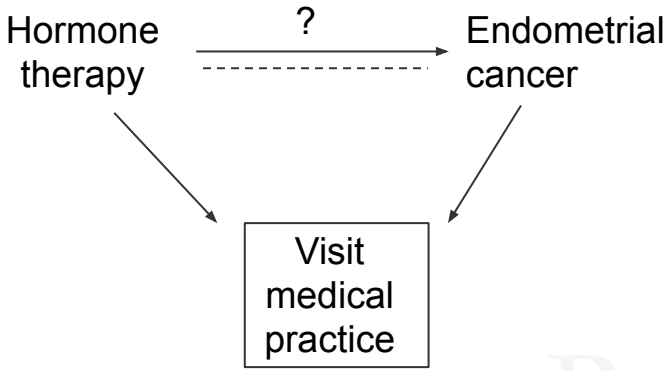
Loss to follow-up and selection bias

- When both the exposure and outcome cause loss to follow-up, selection bias occurs.
- This may happen if the outcome makes people too sick to participate in the study and the exposure also leads to a different disease that causes people to drop out of the study due to illness.



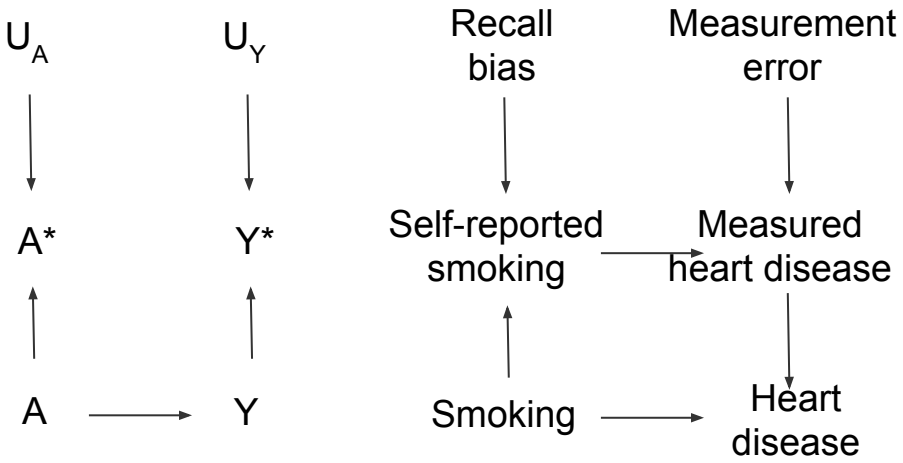
Example

- A case-control study of postmenopausal hormone therapy and endometrial cancer
- Cases and controls enrolled from same medical practice
- Both women with endometrial cancer and those with postmenopausal hormone therapy are more likely to have bleeding and seek medical care.
- Visiting the medical practice is a collider that is conditioned on, so selection bias is present.



Characterizing imperfect measurement in a DAG

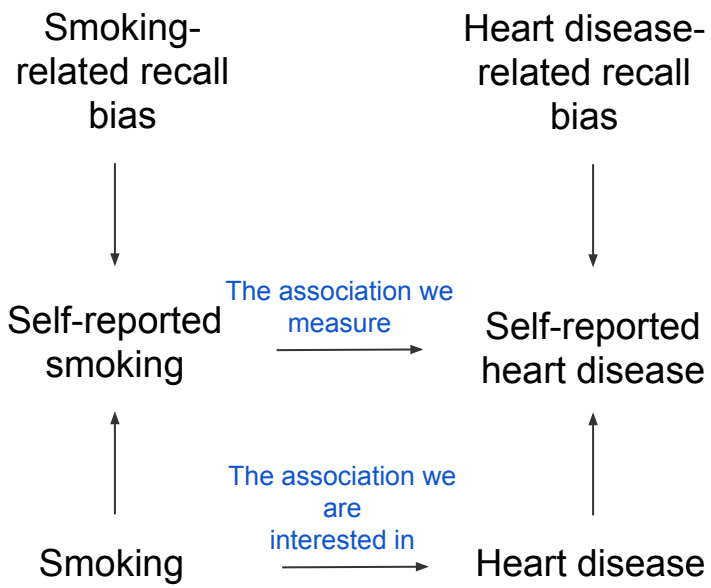
- A = true exposure
- A* = measured exposure
- U_A = all factors other than A that determine the value of A*
- Y = true outcome
- Y* = measured outcome
- U_Y = all factors other than Y that determine the value of Y*
- The difference between the A-Y and A*-Y* associations is information bias



Hernan & Cole, 2009

Example of nondifferential misclassification

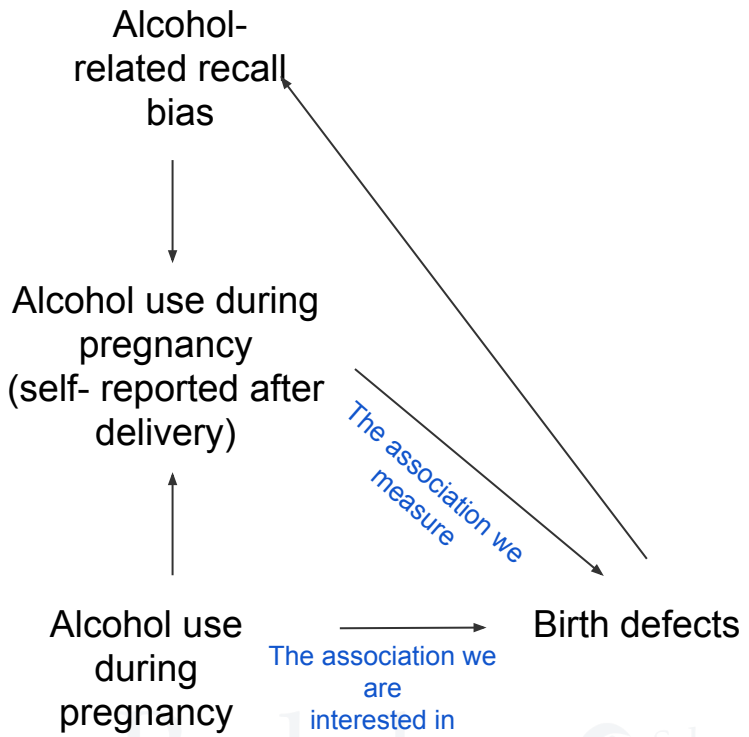
- This DAG shows nondifferential measurement errors.
- The error for the exposure (smoking) is independent of the true value of the outcome (heart disease)
- The error for the outcome is independent of the true value of the exposure



Hernan & Cole, 2009

Example of differential misclassification

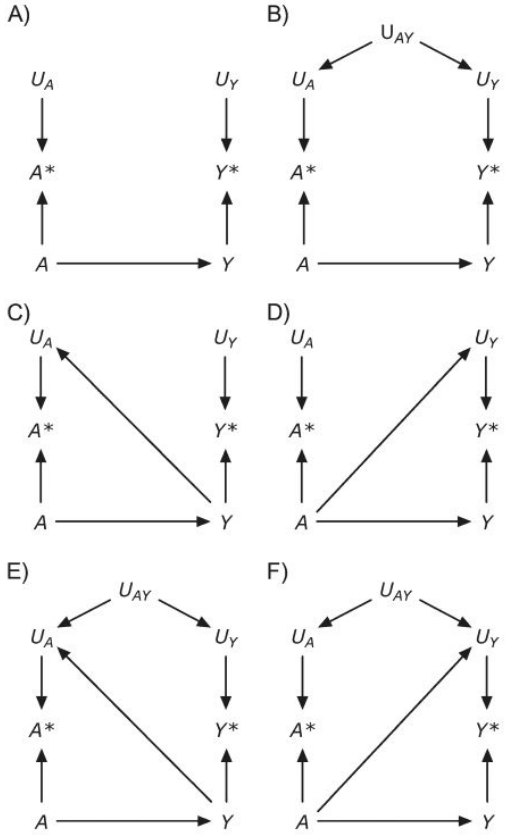
- This DAG shows differential measurement errors.
- The true value of the outcome affects the measurement of the exposure.
- Once a child is born, their birth defects status is likely to influence recall of alcohol use during pregnancy.



Hernan & Cole, 2009

Diagnosing information bias using DAGs

Many possible DAG configurations exist that could reflect different forms of information bias.



Hernan & Cole, 2009

Summary of key points

- DAGs are a powerful tool to help diagnose selection and information bias in a particular study.
- To diagnose selection bias, identify whether a selection factor is a collider of the exposure-outcome that is conditioned on in a DAG.
- There are many DAG configurations that depict information bias (this is a topic for more advanced Epidemiology courses).
- DAGs are ideally used during the study design phase to help minimize potential bias.