SOURCES OF BIAS IN EPIDEMIOLOGIC STUDIES

PART III: INFORMATION BIAS

SLIDES PREPARED BY AARON M. WENDELBOE, PHD
AND OTHER BSE FACULTY

1

Welcome to the third part of the series focused on bias. In this module, we will discuss information bias.

Biases in Study Design

- Learning objectives
  - ✓ Describe main types of information bias that can occur in conduct of observational studies
  - ✓ Explain impact of information bias on measure of association
  - ✓ Give examples of ways to assess and address information bias in the design and implementation of a study

2

Following the completion of this module, you will be able to:
Describe main types of information bias that can occur in conduct of observational studies
Explain impact of information bias on measure of association
Give examples of ways to assess and address information bias in the design and implementation of a study

Biases in Study Design

Most bias can be classified into 2 main types:

- Selection Bias
- Information Bias

3

Recall that there are two main types of bias:

Selection bias: all biases regarding how participants end up in the study
Information bias: how information is collected to assign exposure and disease status

This segment focuses on information bias.

Information Bias: Definition

- Systematic error in exposure assessment or outcome assessment
- Information biases result in misclassification of study subjects with respect to disease or exposure status
- Error in measurement of <u>categorical variables</u> is usually called <u>misclassification</u>
- Error in measurement of <u>continuous variables</u> called <u>measurement error</u>

4

We are now going to provide a few examples of types of information bias. As defined previously, information bias is a systematic error in assessing or measuring either the exposure or the outcome. Information bias reflects the manner in which data were collected. This often means that participants are misclassified in respect to either their exposure or outcome status.

Data are generally fall into two types, categorical variables and continuous variables.

When there are errors in categorical data it is usually referred to as misclassification.

When there are errors in continuous variables it is usually referred to as measurement error.

Misclassification can be:

- Non-differential
- Differential

5

We will discuss two different types of misclassification – non-differential and differential misclassification.

Non-differential Misclassification

- Classification error NOT depend on disease status or exposure status
- If error of exposure classification same for cases and controls (or disease classification is same in exposed and un-exposed), bias usually towards null
- Exceptions:
  - if there are mis-measured confounding factors
  - if > 2 categories of variable

6

Non-differential misclassification occurs when the systematic error occurs in a manner that does not depend on either exposure or outcome status.

When non-differential misclassification occurs, for example, if the misclassification of exposure is the same between cases and controls, it usually biases the estimate towards the null.

There are a few exceptions to this. They are 1) if there are mis-measured confounding variables and 2) if the variable that was mis-measured has more than two categories.

Example of Non-Differential Misclassification of Exposure in a Case-Control study

True Association

	Cases	Controls	OR
Exposed	250	100	4.0
Not Exposed	250	400	
Total	500	500	

Observed Association

	Cases	Controls	OR
Exposed	275	140	3.1
Not Exposed	225	360	
Total	500	500	

% unexposed cases misclassified as exposed = 10%
% unexposed controls misclassified as exposed = 10%

7

Consider this numeric example of non-differential misclassification of exposure in a case-control study.

In this case, there is non-differential exposure misclassification, and we see that 10% of unexposed CASES are classified as exposed and 10% of unexposed CONTROLS are classified as exposed. The misclassification percentage is the same between cases and controls. Even though same percent of cases and controls are misclassified, the results are still biased.

We see that the true association results in an odds ratio of 4.0; however, with 10% misclassification of exposed as unexposed among both cases and controls, the resulting odds ration is reduced to 3.1 and is biased towards the null hypothesis.

Example non-differential misclassification

- Study of heart attack and high-fat diet
- Food frequency questionnaires administered to both cases and controls
- Classification of diet categorized into high-fat and low-fat using cut-off value
- If same proportion of cases with high-fat diet are misclassified as low-fat as among controls, result in non-differential misclassification

8

As an example of non-differential misclassification, let's consider a study investigating the association between heart attacks and a high-fat diet. Assume that a food frequency questionnaire is administered to both cases and controls to measure consumption of high-fat foods. Based on the food consumption information, participants are classified into high-fat or low-fat diet groups based on a cut-point for defining a high-fat diet. If the same proportion of cases with high-fat diet are misclassified as low-fat as among controls, this results in non-differential misclassification.

Differential Misclassification

- Classification error DOES depend on disease status or exposure status
- Can lead to association when there is none or to an apparent lack of an association when one exists
- If error of exposure classification NOT same for cases and controls (or disease classification not same in exposed and unexposed), bias can go either toward or away from null depending on type of misclassification

9

Now let’s talk about differential misclassification. In this error, the classification error depends on the disease or exposure status. The direction of the error is unpredictable. It can lead you to conclude an association exists when it doesn’t, or it can lead you to conclude no association exists when it fact it does.

Example of Differential Misclassification in a Case-Control Study

- Mothers of children born with a congenital malformation tend to remember more mild infections during their pregnancies than mothers of normal infants, resulting in an over-estimate of true association between infection and congenital malformations

This would be <u>differential</u> misclassification of exposure in cases and controls

10

Here is an example of differential misclassification in a case-control study of congenital malformation and infection during pregnancy.

Because mothers whose babies born with congenital malformations are more likely to remember infections during their pregnancy more than mothers whose babies were born healthy, the observed association between infection and congenital malformations would be over-estimated.

This specific type of differential misclassification is also called recall bias – which we will discuss momentarily. In this case, there is differential misclassification of the exposure between the cases and controls.

Example of Differential Misclassification of Exposure in a Case-Control study

True Association

	Cases	Controls	OR
Exposed	250	100	4.0
Not Exposed	250	400	
Total	500	500	

Observed Association

	Cases	Controls	OR
Exposed	225	70	5.0
Not Exposed	275	430	
Total	500	500	

% exposed cases misclassified as unexposed = 10%
% exposed controls misclassified as unexposed = 30%

11

Let's now consider a numeric example. In this example, investigators are studying the association between exposure and case/control status. If 10% of exposed cases are misclassified as unexposed, and 30% of exposed controls are misclassified as unexposed, there is differential misclassification of the exposure because the percentage of exposed who are misclassified as unexposed differs between the cases and controls. In this case, the observed odds ratio will be 5.0 instead of 4.0 and the observed association results in an estimate that is biased away from the null.

Information Bias

Recall Bias

- Differences among cases and controls in accuracy or completeness of recall of exposure
- Results in OR away from null when cases more accurately remember exposure than controls or when cases over-report exposure history
- Avoid by validating exposure history from independent source or objective measures
- Use of “sick” (disease unrelated to the outcome or exposure of interest) controls may help equalize this bias

12

Recall bias is an important source of information bias, particularly in case control studies. This happens because cases tend to spend more time thinking about what could have happened to make them sick. They may have pulled out calendars and journals and talked to friends to think about what happened in the past that may have led or contributed to their disease.

When cases tend to over-estimate their exposure and/or controls tend to under-estimate their exposure history, this leads to estimating an effect estimate away from the null.

There are a couple ways to try to avoid recall bias. One is to use similar methods of data collection for controls as cases. For example, use calendars and journals to try to jog the memories of controls in a similar fashion as cases. You could rely on “hard” clinical outcomes that are documented in the medical record instead of relying on participant recall. Also, you might be able to use a control group that is sick, like the cases. The trick is, that their sickness can’t be related to the outcome or exposure of interest, otherwise you may introduce a new bias or confound your results.

Example of Recall Bias

- Case-control study of *in utero* x-ray exposure and childhood leukemia
- Cases were children with a first diagnosis of leukemia selected from a population registry and controls were of same age, sex, and geographic location and selected at random from local birth registries

13

Here is another example of recall bias. A case-control study of in utero x-ray exposure was hypothesized to cause childhood acute lymphoblastic leukemia. Cases were identified from a cancer registry and were selected from birth certificates. Controls were sampled to have a similar distribution of age, sex, and geographic location at the cases.

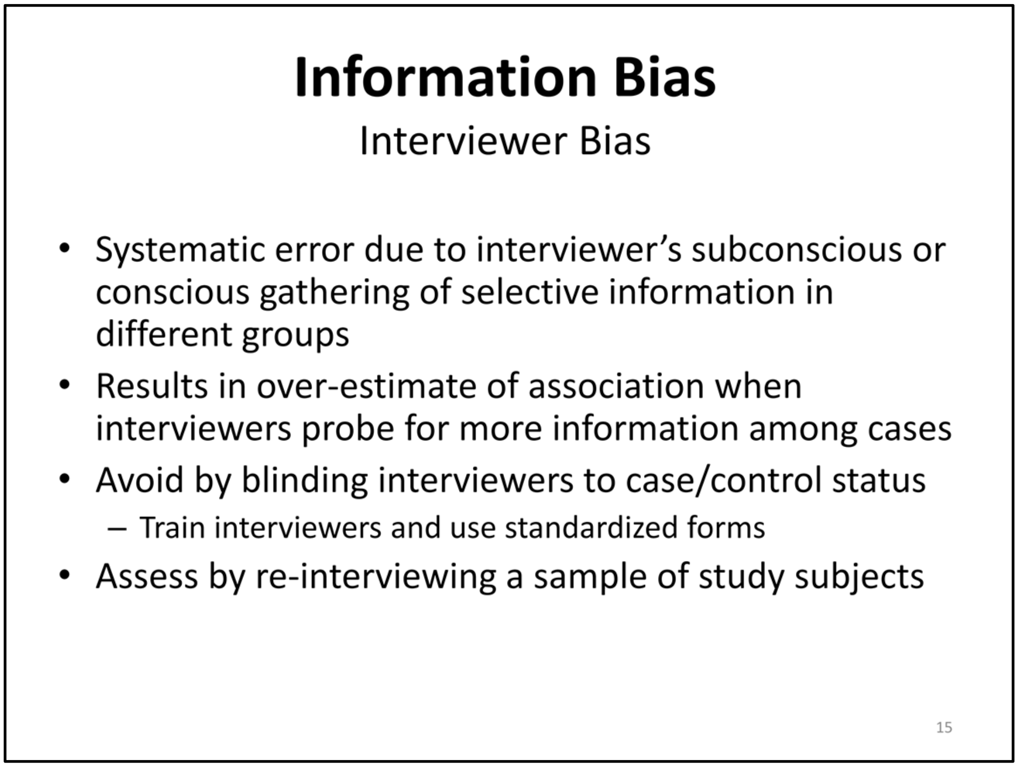
Example of Recall Bias cont….

- Information about exposure to pre-natal x-rays obtained by interviewing mothers of both cases and controls
- OR = 1.9
- To address issue of recall bias, subsequent study conducted in which exposure to x-rays ascertained through medical records
- Magnitude of effect was lessened after accounting for recall bias; however, still a statistically significant OR = 1.4

14

The mothers of the cases and controls were interviewed to collect information about in utero x-ray exposure. The study estimated an odds ratio of 1.9.

Because the investigators suspected recall bias may threaten the validity of their findings, a validation study was conducted to pull the medical records of both cases and controls to measure in utero exposure in a less biased way. The new OR was closer to the null, but still statistically significant at 1.4.



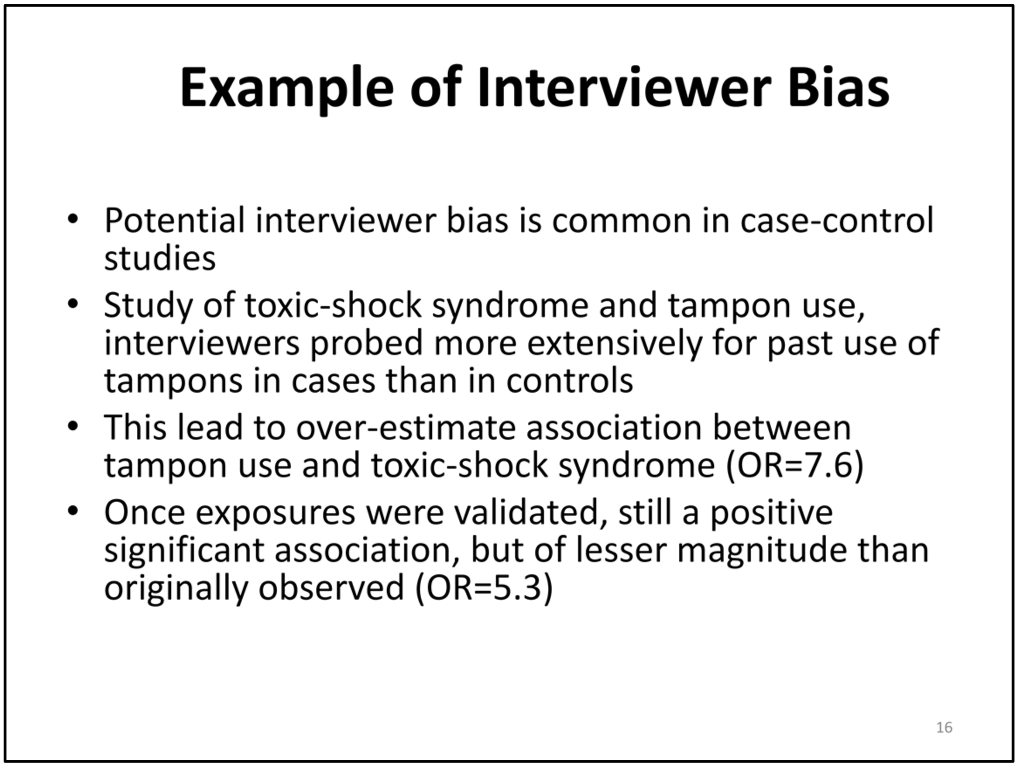
We now consider another source of information bias, interviewer bias.

Interviewer bias is a systematic error due to interviewer's subconscious or conscious gathering of selective information in different groups. Interviewer bias results in an over-estimate of the association between exposure and disease when interviewers probe for more information among cases, as an example.

Interviewer bias can be avoided by blinding interviewers to case/control status.

Investigators should train interviewers and use standardized forms to limit interviewer bias.

The extent of interviewer bias can be assessed by re-interviewing a sample of study subjects and comparing the responses between the interviewers, although one would need to consider the potential for the interview itself to "jog" the participant's memory and impact the recall on the second interview.



An example of interviewer bias can be seen in a study of toxic shock syndrome as a result of tampon use. Interviewers probed the cases more extensively than controls regarding their past use of tampons. The resulting odds ratio was 7.6.
Exposure histories were re-taken using a standard instrument for both cases and controls and a revised odds ratio of 5.3 was estimated.

Information Bias

Family Information Bias

- Occurs when cases more aware of their family history because occurrence of their disease has stimulated discussion or investigation of past disease history in family
- Results in increased (odds) estimate, since cases may be more aware of a positive family history than are controls
- Avoid by validating disease status of family members

17

The last bias we are going to describe in detail is family information bias. This occurs when there is a differential awareness of the disease status among family members and across generations. That is, a child suffering from a certain condition may be more aware of parents or grandparents who suffered from the same condition, whereas unaffected children may either be unaware or forget their parents or grandparents had that condition. When cases are providing the history, the result may be an increased effect measure estimate than if unaffected people were providing the information.

Family information bias can be avoided by collecting information about participants using documents such as medical records to validate information provided by cases. Instead of using self-report, you will instead utilize medical records as the source of information.

Example of Family Information Bias

- Demonstrated in study of rheumatoid arthritis and whether it clusters in families
- When investigators asked individuals with rheumatoid arthritis (cases and controls) whether their parents had arthritis, they obtained results which suggested disorder did "run in families"
- However, when investigators asked unaffected siblings of cases with arthritis whether or not their parents had arthritis, results were quite different

18

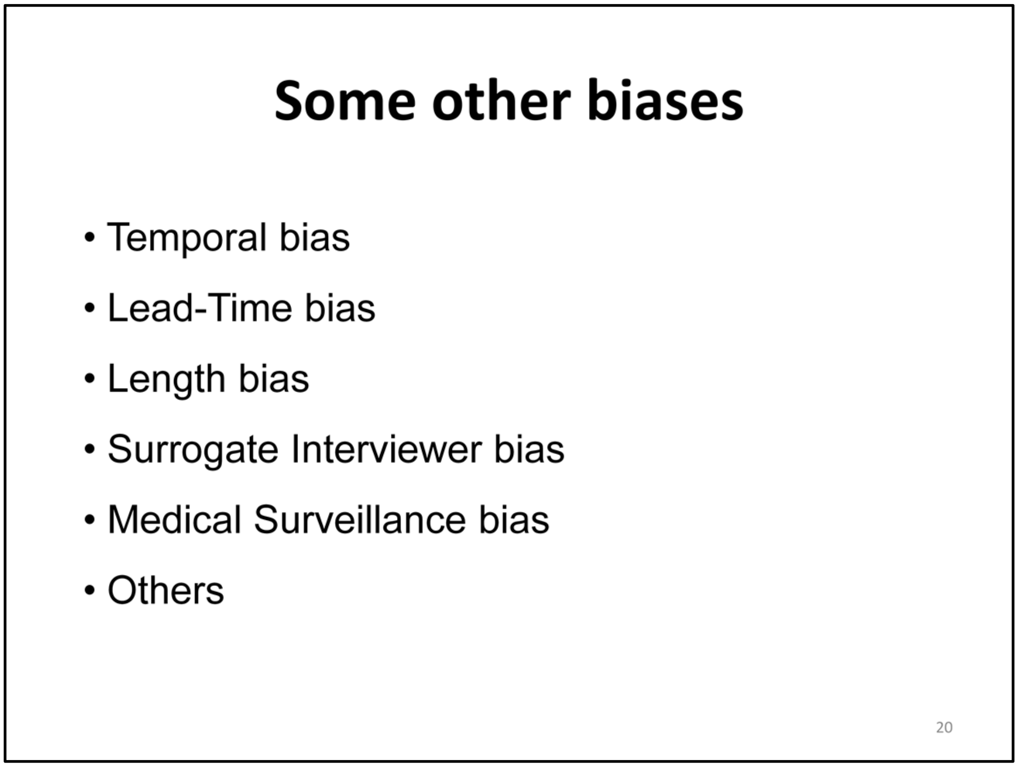
For example, there was a study conducted to determine if rheumatoid arthritis clusters in families. Investigators asked children who suffered from rheumatoid arthritis if their parents suffered from it also. They then interviewed siblings of cases the same question, about whether their parents suffered from rheumatoid arthritis. The results from the separate analyses provided different results.

Example of Family Information Bias cont

- Unaffected siblings reported parental history of rheumatoid arthritis much less frequently than their affected siblings
- This demonstrates that family history may vary drastically depending on whether individual providing information is a case or control

19

Unaffected siblings were more likely to deny that their parents had rheumatoid arthritis than affected siblings. This demonstrated the importance of knowing the health status of the person being interviewed when collecting information about family history of disease.



Other sources of information bias are listed on this page and we won't discuss them in detail.

Temporal Bias
In this situation, we don't know which came first, the exposure or the outcome. This may also be referred to as reverse causality.

Lead Time Bias
Lead Time - Amount of time diagnosis advanced as a result of screening
Lead Time Bias - Survival may falsely appear to be increased among screened group simply because the diagnosis was made earlier in the course of the disease

Length biased sampling:
Less aggressive forms of a disease are more likely to be picked up by screening because have a longer preclinical phase
Less aggressive forms of disease usually have better survival
Thus, screen detected cases appear to have better survival

Surrogate interviewer bias: when interviewing a parent of a child or next of kin for a deceased person. They may remember the deceased person better and report fewer risky behaviors than the person would report and that actually happened. Also, they may not know all of the deceased individuals behaviors.

Medical surveillance bias: cases of a chronic or life-threatening disease may have a more thorough physical exam than controls. Example- cases of childhood cancer will have a thorough physical exam, including imaging scans, at diagnosis and my have minor congenital anomalies detected more frequently than controls, even if they exist at the same frequency in both cases and controls. They weren't detected in controls without a thorough physical exam.

Summary

- Biases are a primary source of error in observational epidemiologic research
- Biases can be either conscious or unconscious
- Prevention: Appropriate data collection methods that are both valid and reliable
- Learn from other studies

21

Biases are an important source of error in observational studies. Multiple biases can be present in a single study. It is usually not a question of whether a given study is biased, it's a question of how much are the results of a given study biased.

Biases can result from both conscious and unconscious decisions of participants and investigators.

Some biases are hard to avoid, but by selecting the best study design and making efforts to validate data collection you can reduce the impact bias has on you study. Finally, it is helpful to read other studies and learn from them to determine how your study compares to others in the field.

This concludes our series on bias.