SOURCES OF BIAS IN EPIDEMIOLOGIC STUDIES

PART I: OVERVIEW

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

1

Welcome to this series focused on sources of bias in epidemiologic studies. In this first module, I will provide a general overview of bias. In the second module, we will focus on selection bias and in the third, we will focus on information bias.

Biases in Study Design

- Learning objectives
  - ✓Discuss validity and precision
  - ✓Define “bias” in epidemiologic studies
    - ✓Describe how it threatens validity
  - ✓Explain impact of specific types of bias on measure of association
  - ✓Describe main types of bias that can occur in conduct of observational studies
  - ✓Give examples of ways to assess and address these biases in design of a study

2

By the end of this lecture, you should be able to achieve the following learning objectives.

You should be able to discuss validity and precision and differentiate between the two.
You should be able to define bias in epidemiologic studies and describe how, when present, it threatens the validity of the results.
You should be able to describe the main types of bias that can occur in observational studies and be able to explain the impact these biases have on the measures of association.
Finally, you should be able to provide examples of how to prevent or otherwise address biases in the design of observational studies.

Internal Validity vs. External Validity

- ***Internal Validity***: study provides unbiased estimate of what it claims to estimate
- ***External Validity***: results from study can be generalized to some other population

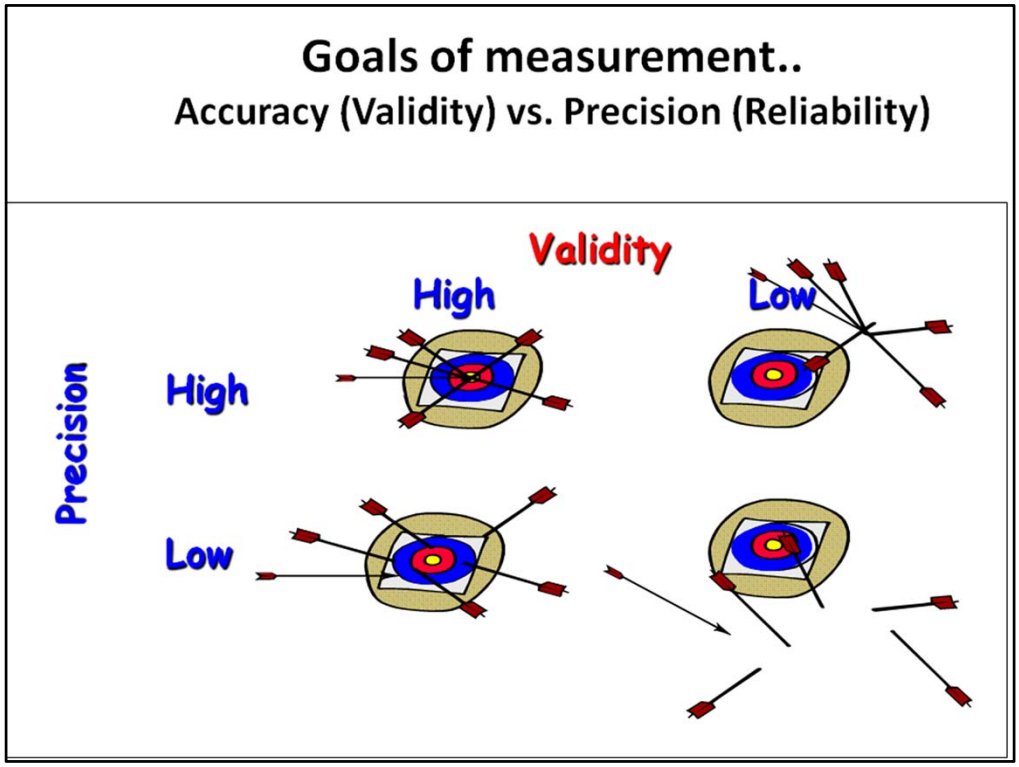
3

Let's now introduce some terminology.

Internal validity reflects the accuracy of the study and is achieved if the study provides an unbiased estimate of what it claims to estimate.

External validity reflects the generalizability of the study. External validity reflects the ability to extrapolate beyond the subjects in the study (to project or extend your findings from the sample population to a particular target population)

Example: a study restricted to white males may or may not be generalizable to all human males, but is not generalizable to females (all people).



Let's now consider two additional terms, validity and precision.

Validity relates to accuracy. Our primary goal is to generate accurate, unbiased, estimates of parameters such as the mean cholesterol in a particular patient population or the risk ratio summarizing the association between smoking and lung cancer.

Then, after accuracy is achieved, we want our estimates to be precise or reliable. Recall that precision is reflected by the width of our confidence interval.

When performing a study, we want our estimates to be both precise and valid.

Now, let's consider a diagram to relate the concepts of precision and validity.

As we just stated, validity is getting the right answer. It's like hitting the target or getting a bullseye. Precision can be thought of as getting the same answer repeatedly. It can also be thought of as how confident we are that the answer we got was a result of fact rather than chance.

Here we have four pictures of arrows being shot at a target as a way of trying to illustrate the relation between precision and validity. In the upper left hand target, we have a target that was hit repeatedly in the same spot. This is a demonstration of both high validity and high precision. In the lower left target, the target was hit every time, but not really in the same place. Thus, the validity is high, but not very precise.

In contrast, in the upper right hand target, the arrows all hit the same place repeatedly, indicating a high degree of precision, but none of them hit the target; thus resulting in poor validity. Finally, in the lower right target, none of the arrows hit the target and they didn't hit the same place, a picture of poor precision and poor validity.

In statistical and epidemiologic terms, if our point estimate (like the odds ratio) is accurate, then it has high validity. If the confidence interval is narrow, then our estimate is precise. A wide confidence interval shows that our estimate has poor precision. While we can regularly make a determination of the precision of our estimates, we rarely know how valid, or accurate, our estimates are.

We know a larger sample size WILL increase the precision of our estimate, reflected by narrower confidence intervals, but a larger sample size may not provide a more valid estimate, if for example, our estimation method or sampling method is not accurate and is biased a larger sample will not improve validity.

Biases in Epidemiologic Research

- Systematic error = bias
  - **Selection biases** – how study group chosen
  - **Information biases** – inaccuracy in measurement or classification of exposure, outcome, or covariates
    - results in measurement error/misclassification
- Any systematic error in design, conduct or analysis of a study that results in a mistaken estimate of an exposure's association with disease. i.e., produces a biased estimate of OR, RR, PRR etc.

5

Bias is defined as a systematic error. Bias is not necessarily intentional and it can occur at any stage of the study.

All observational studies are biased to some extent, but at what stage bias occurs, how can you minimize bias, and avoid bias if possible, are topics of this series.

One of the biggest reasons we can't get valid estimates from our observational studies is because of bias. That is, bias threatens validity. Bias is a systematic departure from the truth.

There are two general forms of bias. Selection bias and information bias.

Selection bias refers to how study participants are included in the study, for example, differences between those who did and did not participate in a study.

Information bias refers to systematically collecting poor information. For example, misclassification is a type of information bias that occurs when we classify participants as exposed/diseased when actually the participants are nonexposed/nondiseased.

Here is one example of selection bias in a study of the cost effectiveness of having Hospitalists manage hospitalized patients instead of patients of primary care physicians (PCPs). The hypothesis is that because Hospitalists have a more intimate knowledge of the hospital's services they can help patients receive all of the needed services in a more timely

and efficient manner than other physicians. Therefore, they can discharge patients earlier and thus save money while improving quality of care. A study found that patients of Hospitalists were more likely to be readmitted to the hospital within 30 days of initial discharge than patients being managed by their PCP and thus actually cost more than patients of PCPs. These results were affected by selection bias because the patients being managed by Hospitalists were sicker than the other patients. That is, they had more co-morbidities, like diabetes and heart disease, than the other patients.

Using the same study scenario, here is an example of information bias. Because Hospitalists were more familiar with the hospital's electronic medical record and billing system, they were able to correctly code for all of the services provided. On the other hand, PCPs entered their diagnostic and procedural codes in multiple places, so that it took longer for all of the services to be billed. Therefore, at the time the study was conducted, it appeared that patients managed by PCPs used less services and therefore were less expensive to manage than patients managed by Hospitalists; however, this difference was due in part to information bias.

Bias can be introduced in the design, conduct, or analysis stages of a study. The only way to address bias is to recognize it and then correct it to the extent possible.

Systematic vs. Random Error

- Bias = systematic departure from truth
- Random error = chance, sampling variability, sample size
  - -P values
  - -Confidence intervals

6

Bias is a systematic departure from the truth. In contrast, random error occurs due to chance variation and sampling variability, which may be higher when the sample size is small.

Note that we can't use p-values and CI's to assess for bias; these only tells us about random error.

We just provided two examples of how systematic differences between either the patients or the information collected on the patients resulted in error and consequently getting incorrect results. Bias does not affect the precision of the estimate.
In contrast, random error can threaten both the validity and precision of our estimates. Statistical methods can be used to address random error.

One important method of addressing random error is increasing the sample size. By having a larger sample, you are more likely to estimate the true population estimate and you are likely to be more precise about it, as reflected in your confidence interval.

In the previous two examples of bias, increasing the sample size would not have helped investigators get the right answer because their sampling or information sources were biased. All that a larger sample size would do is enable them to have a more precise estimate of the wrong answer.

What is bias in estimating a frequency?

- Wrong estimate of frequency of a characteristic in a population caused by:
  - Misrepresentation of population in study sample
    - Sampling methods are important
  - Error in measuring true state of individual characteristics
    - Validating measurement tools is important

7

Let's now consider how bias might impact our statistical estimates.

We will first consider measures of frequency, such as the number of adult females who smoke in the state of Oklahoma, and we may then calculate a prevalence proportion.

A biased frequency estimate might arise if we use sampling methods that are not reflective of the target population, for example, only sample women who live in a neighborhood that has a high level of poverty. A biased estimate might also arise due to our measurement method, for example, if we only ask about smoking cigars and don't include all tobacco products.

What is bias in estimating an association?

- Wrong estimate of measure of association between exposure and disease
  - Exposure or disease outcome NOT measured with a valid, standardized method (i.e., information bias)
  - Exposure status or disease status influence selection process (i.e., selection bias)

There is a THIRD variable associated with BOTH exposure and disease that can change results (confounding)

8

Besides counting frequencies of disease, epidemiologists also use odds ratios, rate ratios, and other measures of association to determine the relationship between an exposure and a disease.

The exposure, the outcome, or both can be measured systematically incorrectly and therefore result in bias. Thus, we often say the odds ratio or rate ratio is biased. This can arise from both selection and information bias. I will provide examples of these scenarios in this lecture.

There is another concept in epidemiology which contributes to getting the wrong (or an invalid) estimate. It is called confounding. Confounding is addressed in detail in a different lecture. However, it is important for you to understand that bias and confounding are different phenomena. Confounding is when the disease is associated with a factor other than the exposure under study. A confounder is a factor that is associated with both the exposure and the disease. An exaggerated example of confounding is concluding that wearing swimsuits causes sunburn. It's actually unprotected exposure to the sun which causes sunburn, but people wearing swimsuits are more likely be out in the sun than people not wearing swimsuits. People who wear swimsuits indoors all day, every day, without being exposed to the sun, are not going to get sunburned.

Estimating an association

- IMBALANCE between 2 groups (differential ascertainment)
  - In measurement/classification methods (information bias)
  - In sample selection methods (selection bias)

9

Another way of understanding bias is recognizing there is an imbalance between the two (or more) groups under study. Again, this imbalance can be the result of how participants were selected or recruited or how information about these participants was collected.

The null

- Ratio measures (e.g., rate ratio, odds ratio)
  - Null = 1.0
  - Positive associations >1.0
  - Negative (protective) associations >0.0 but <1.0
- Difference measures (e.g., rate difference)
  - Null = 0.0
  - Positive associations >0.0
  - Negative associations <0.0
- Not going to worry about statistical significance at this point

10

In the next four slides, we're going to talk about the direction in which bias can operate. But first let's review what the null means.

In measures of association, the null means there is no association between the factors under study, the exposure and outcome.

When using ratio measures, like the rate ratio, the null is 1.0. Any ratio measure > 1 is a positive association and any ratio measure between 0 and 1 is a negative (or protective) association.

For difference measures, like the rate difference or difference in means, the null is 0.0. Any difference measure > 0 is a positive association and any difference measure < 0 is a negative association.

We are not going to worry about statistical significance for now.

Direction of Bias

- ***Bias away from the null***
  - –Observed measure of association is farther from 1.0 than true value (bigger effect than true)
    - If true OR=3.0 and study finds OR=4.0 this result is biased away from null (1.0)
    - If true OR=0.76 and study finds OR=0.56 this result is biased away from null

11

For ratio measures, remember that a value of 1.0 indicates no association (i.e., the null value). Bias can occur in the positive (away from the null) or negative (towards the null) direction.

Another way of understanding bias is recognizing there is an imbalance between the two (or more) groups under study. Again, this imbalance can be the results of how participants were selected or recruited or how information about these participants was collected.

When bias occurs away from the null in ratio measures, it means it is farther from 1.0 than the true value. It can go away from 1.0 in a positive or negative direction. In order to understand this concept, it is helpful to realize there are two measures to consider. There is the true measure and there is the measure that our study estimated.

Let's say the true odds ratio summarizing the relationship between an exposure and disease is 3.0. If we conduct a study and estimate an odds ratio of 4.0, then our estimate is biased away from the null.

Now let's look at the other side of the null. Let's say the true odds ratio is 0.76 and the study estimates an odds ratio of 0.56. This odds ratio is again biased away from the null.

If we were calculating difference measures, the null would be 0.0 and away from the null would be farther from 0.0 on either side.

Direction of Bias

- ***Bias towards the null***
  - –Observed measure of association is closer to 1.0 than true value (smaller effect than true)
    - If true OR=3.0 and study finds OR=2.5 this result is biased toward null (1.0)
    - If true OR=0.76 and study finds OR=0.88 this result is biased toward null

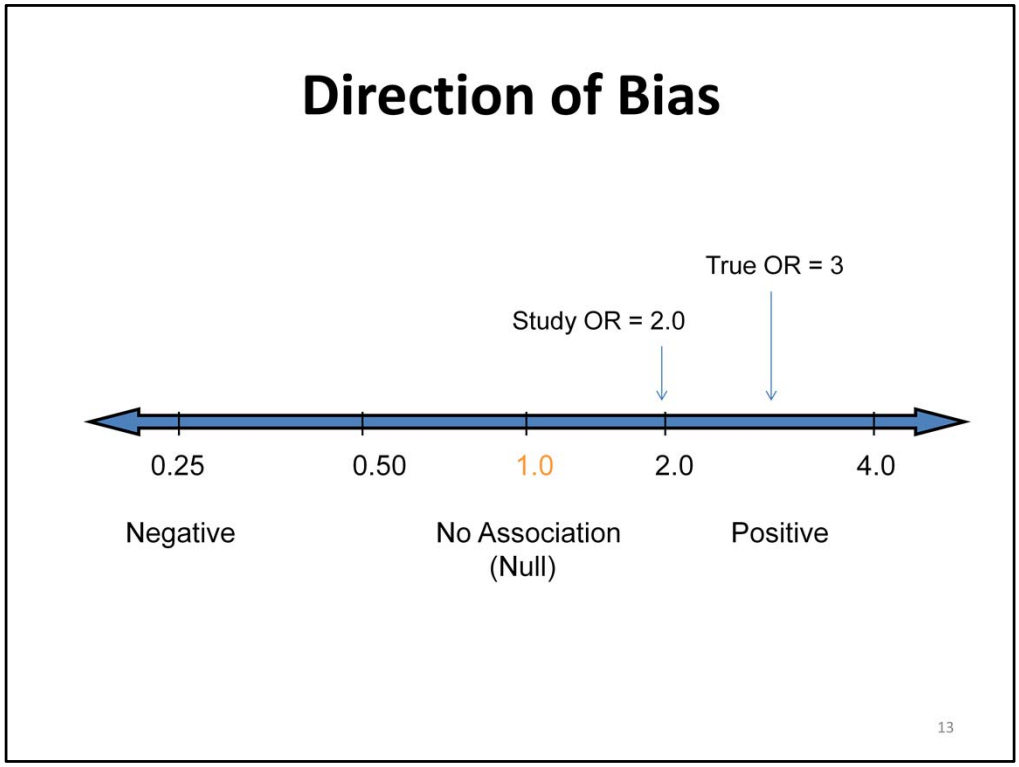
12

When bias occurs towards the null in ratio measures, it means it is closer to 1.0 than the true value.

Let's say the true odds ratio summarizing the relationship between an exposure and disease is 3.0. If we conduct a study and estimate an odds ratio of 2.5, then our estimate is biased towards the null.

Now let's look at the other side of the null. Let's say the true odds ratio is 0.76 and the study estimates an odds ratio of 0.88. This odds ratio is again biased towards the null.

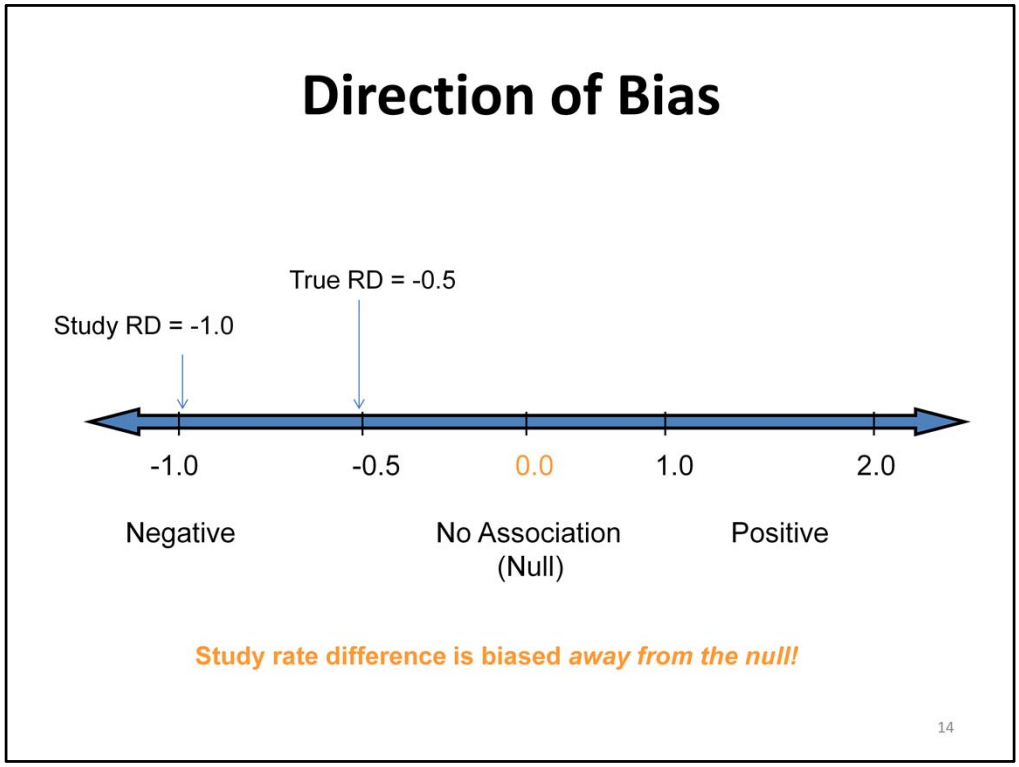
If we were calculating difference measures, the null would be 0.0 and towards the null would be closer to 0.0 on either side.



Now, let's consider a diagram to visually summarize the direction of bias.

We continue to work with ratio measures so that the null is 1.0.

We can use this continuum to identify the true odds ratio and the study odds ratio. We write down the true odds ratio of 3.0. [click animation] Then we record the observed, or study OR. [click animation]. As you can see, the study OR is biased towards the null because it is closer to the null value of 1.



Now let's consider a difference measure where the null value is 0.

The true rate difference is -0.5. The observed, or study rate difference is -1.0. Is this rate difference biased away from the null or towards the null? I am going to wait 10 seconds while you think about this.

In this case, we would conclude that the rate difference is biased away from the null.

Which direction is worse?

- Do you want your effect biased
  - – Towards the null

  or

  - – Away from the null

15

As an investigator, in which direction would you rather have your estimate biased?

The answer is context specific, but most the time it is preferable to have your estimate biased toward the null. If you are trying to show that an exposure causes a disease and you are worried about your effect estimate being biased, it is easier to report something like "We estimated the odds ratio to be 2.0. This estimate is probably biased towards the null because of factor x. Thus, the true association is probably even stronger than what we reported."

When an effect estimate is biased away from the null, it's difficult to demonstrate the public health importance of your study. Suppose your report went something like this "We estimated the odds ratio to be 3.0. This estimate is probably biased away from the null because of factor y. Thus, we don't really know how important a role the exposure plays in causing our disease of interest."

Biases in Study Design

Most bias can be classified into 2 main types:

- Selection Bias
- Information Bias

16

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

The second part of this series focuses on selection bias and the third part of this series focuses on information bias.

Summary

- Biases are systematic, non-random errors and can occur at any stage of research process
  - assembling subjects, collecting data, analyzing results
- Bias results in invalid (not 'real') findings and can produce either an under-estimate or over-estimate of true association, depending on type of bias
- Bias includes selection biases, information biases

17

To summarize our discussion, we have learned the following:

Biases are systematic, non-random errors and can occur at any stage of research process
assembling subjects, collecting data, analyzing results

Bias results in invalid (not 'real') findings and can produce either an under-estimate or over-estimate of true association, depending on type of bias

Bias includes selection biases, information biases