

**Health Research Fundamentals**  
**Dr. Tarun Bhatnagar**  
**ICMR-National Institute of Epidemiology, Chennai**

**Lecture – 08**  
**Validity of Epidemiological Studies**

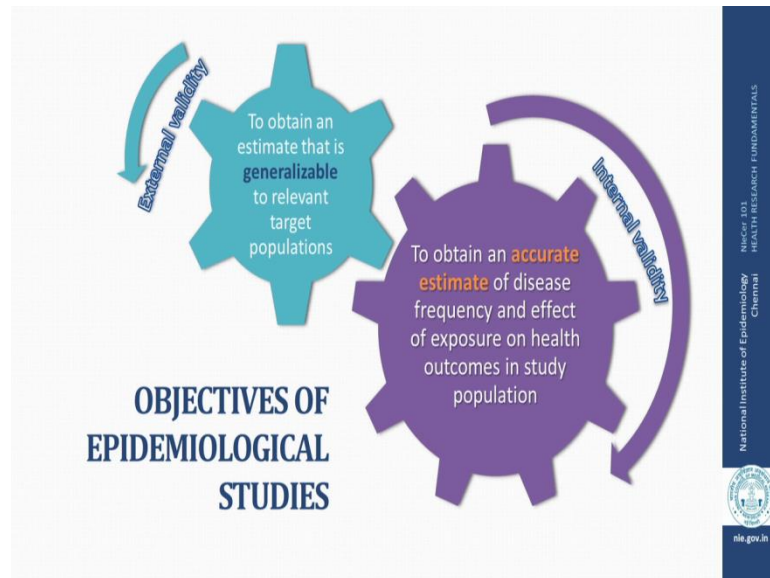
Hello, welcome to this session of Health Research Fundamentals. Today, we are going to talk about Validity in Epidemiological Studies.

(Refer Slide Time: 00:17)



What if you come across one day, a headline in the newspaper that a study says, that coffee drinking doubles the risk of heart attack? What is going to be your reaction? In order to further go into indepth into the study, we will actually need to look at how the study was done? And how valid are the results of the study?

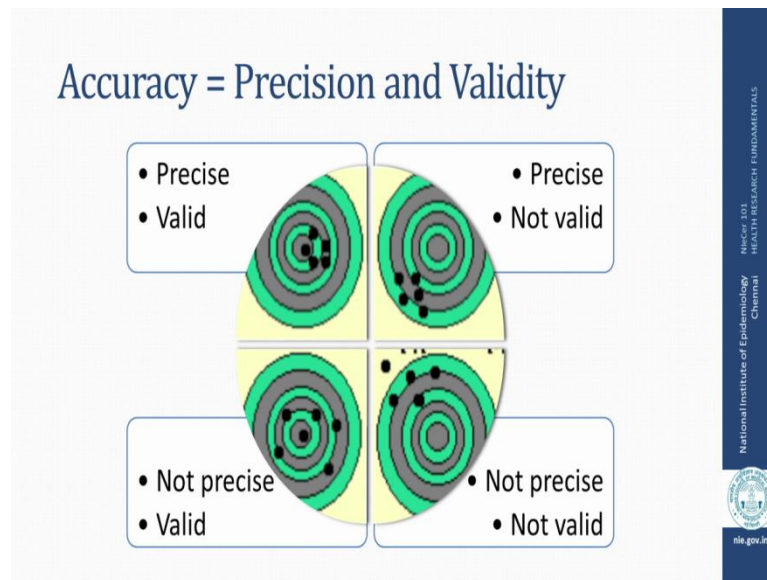
(Refer Slide Time: 00:42)



If you look at any epidemiological study, the basic goal of any epidemiological study is twofold; one is to obtain an accurate estimate of whatever is being studied, whether that is the frequency of a disease or the effect of an exposure on a health outcome. And all of this, we study in a certain sample of the population. Now, this aspect of any epidemiological study is known as internal validity of the study. How valid are the method that are being used to either estimate the frequency or determine the effect of an exposure?

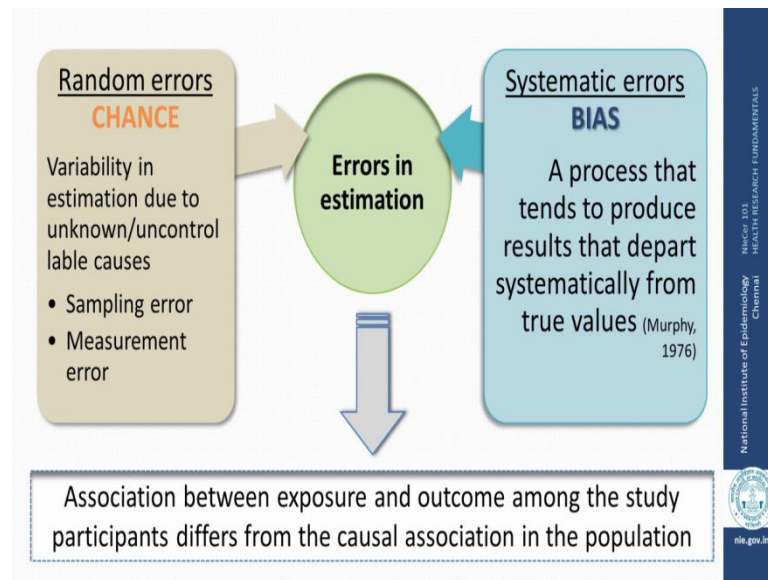
Now, in the long run what we would want is we would want that this estimate is generalizable to the relevant target populations among which the study is being done. Now, this aspect of any epidemiological study is known as external validity, so that the results of the study can be extrapolated to the whole population.

(Refer Slide Time: 01:40)



When we talk of accuracy of the estimate, what accuracy actually means, it consists of two things, precision and validity. If we look at say, a bull's eye and we want to hit the mark, what we would want to be, is to be precise and as well as valid, so that we try to hit the bulls mark as many as times as possible. So, similarly every epidemiological study can have results, which are both precise and valid, which is what we would actually want in every study. **However**, there could be studies, where the results may be precise which means that every time the study has been done you get the similar results, but it may be that the methodology was not correct and so they are not valid. It may happen that the results are not precise, but sometimes they may be valid or in the worst-case scenario, either result may be neither precise nor valid. So, when we are looking at any epidemiological study, we need to be wary of both precision as well as validity.

(Refer Slide Time: 02:45)

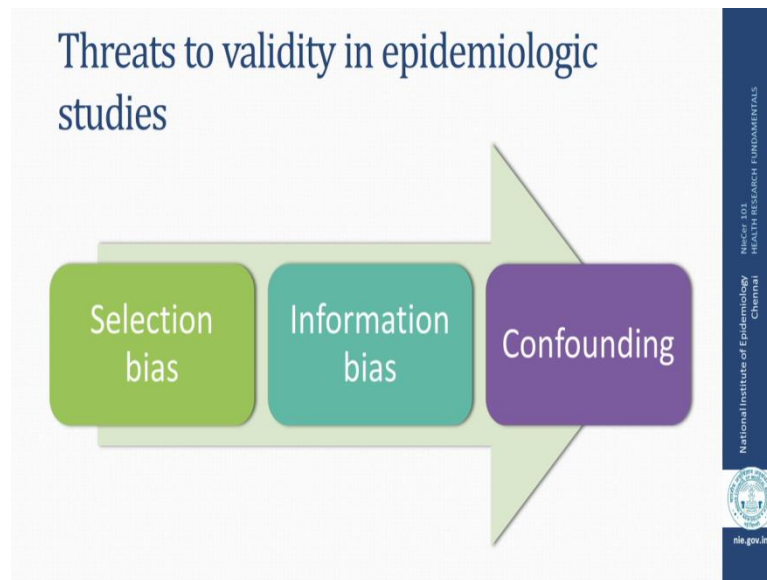


As I told you in epidemiological studies all that we are doing is, basically estimating. We are estimating, either the frequency of a disease or a health outcome or we are estimating what is the effect of an exposure on an outcome?

Now, when we are doing these estimations there are bound to be errors that may happen in our studies. There are two kinds of errors that we come across when we are doing epidemiological studies. One are called random errors or errors that happen due to chance, which is basically the variability because of any unknown or uncontrollable causes such as errors in sampling or errors in doing measurements. However, the more problematic error that we may face in any study, are called systematic errors or biases. These are the errors that are basically a threat to validity of any epidemiological study.

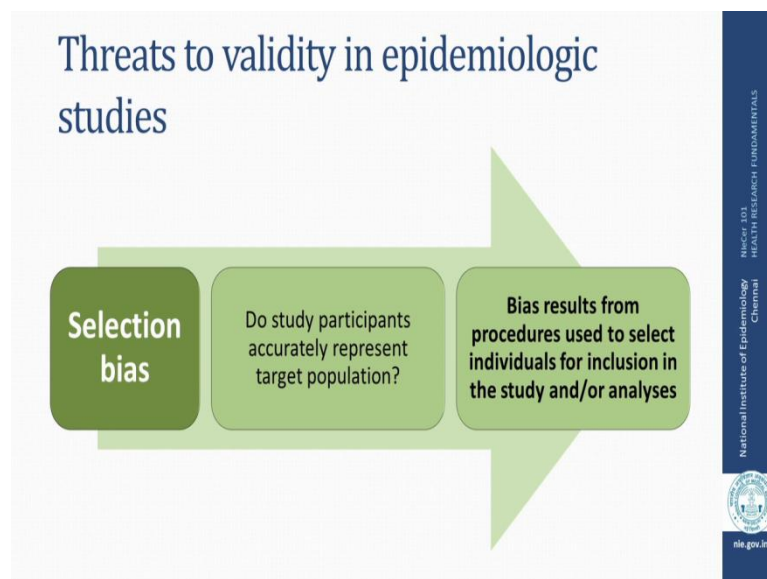
And, how do these errors happen in any study? Basically, the way we do the study, the methodology that we use to do an epidemiological study and if it is done in a certain way that tends to produce results, that are not the true results, then that leads to errors which are called biases. Ultimately, what we would see is that, either the estimates or the associations that we are trying to assess between the exposure and the outcome in the study sample may defer from the true causal associations between the same exposure and outcome that may be there in the source population.

(Refer Slide Time: 04:16)



So, let us look at the various kinds of biases or Threats to validity in epidemiological studies. There are essentially three kinds of biases that may occur in epidemiological study. These are called as selection bias, information bias and confounding.

(Refer Slide Time: 04:35)



So, let us go through one by one. Coming to selection bias, selection bias happens when we use procedures to select populations. Remember that, in epidemiological study we are sampling a certain number of individuals to participate in the study. The way, in which we select the study participants, are we sure that these study participants really

accurately represent the target population? And if there is any issue in which, the way we select these people that results into what we call as selection bias.

(Refer Slide Time: 05:08)

**Selection bias in epidemiological studies**

- **Surveillance** - Systematic notification of exposed cases
- **Screening / diagnosis** - Systematic case search among exposed
- **Admission to health care facilities** - Systematic admission of:
  - Case-patients exposed / unexposed
  - Control-subjects exposed / unexposed
- **Selective survival** - Systematic inclusion of cases who survived and who may be more or less exposed
- **Non response / loss to follow up** - Systematic inclusion of subjects more likely to participate who may be:
  - More or less exposed
  - More or less at risk

National Institute of Epidemiology  
Chemical  
NIC-101  
HEALTH RESEARCH - FUNDAMENTALS  
nitg.gov.in

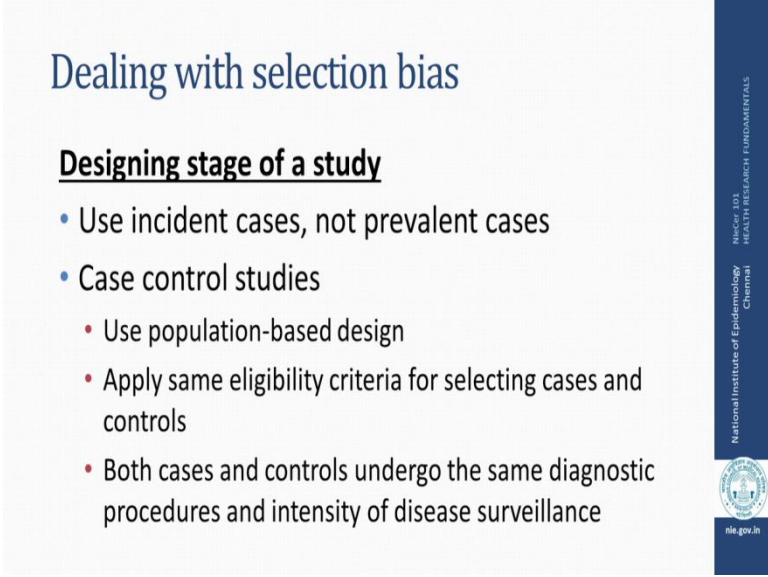
Now, how do all these things happen in epidemiological studies? Remember that, we are selecting our cases and controls and these may happen through either, we are using a surveillance mechanisms from which there is a systematic notification of cases and if we are taking more of exposed cases from this surveillance mechanisms that could be one way in which selection bias could occur. We will be screening and doing diagnosis more systematically among those, who are exposed. If we know their exposure history before hand and then that can artificially create biases. Again, selection biases can occur in, if we select our cases and controls from health care facilities...hospitals and where if, it is likely that more of the case patients who are exposed are admitted or the other way round that can lead to selection bias.

Another common way in which selection bias occurs is, when we select those cases who are alive, the cases of the disease who are dead would not be part of our studies and it may be that, the reason why these cases are alive, may have to do with the exposure status and hence, selective selection of survived patients can actually lead to selection bias.

In cohort studies, selection bias usually occurs when there is a lost to followup. Remember that, we have to follow up people over a period of time in cohort studies and

if it is likely that people, who are less exposed or more exposed, are more likely to be lost or people who are at more risk or at lesser risk are more likely to be lost to follow up, that eventually can lead to results that are biased and that would be attributable to selection bias.

(Refer Slide Time: 07:01)



The slide is titled "Dealing with selection bias" in a blue serif font. Below the title, the section "Designing stage of a study" is underlined in bold. It contains a bulleted list of four points: "Use incident cases, not prevalent cases", "Case control studies" (with three sub-bullets: "Use population-based design", "Apply same eligibility criteria for selecting cases and controls", and "Both cases and controls undergo the same diagnostic procedures and intensity of disease surveillance"), and "Both cases and controls undergo the same diagnostic procedures and intensity of disease surveillance". On the right side of the slide, there is a vertical blue bar with white text: "NICOR 101 HEALTH RESEARCH FUNDAMENTALS Chennai" and the logo of the National Institute of Epidemiology, Chennai, with the website "nie.gov.in" at the bottom.

How do we deal with selection bias? We can deal with selection bias at any stage of our study. Ideally, we would want to make sure that the way in which we design the study is free from selection bias. So, one way would be to use incident cases and not prevalent cases because prevalent cases have the issue of survival bias. Especially, case control studies are more prone to this, to selection bias and various ways in which, to deal with selection bias in case control studies is to use population based design rather than hospital based design, such that the cases and controls are actually selected from the community or the population and not from few or a particular health care facilities.

We need to make sure that we apply the same eligible criteria, when we are selecting cases and controls and we are not leaning towards a particular exposure among the cases and controls. Again, both the cases and controls should undergo the same diagnostic procedures and the same intensity of surveillance in order to identify them as cases and controls. So, that we are not biased at the time of their selection.




(Refer Slide Time: 08:12)

## Dealing with selection bias

### Data collection stage of the study

- Minimize nonresponse, nonparticipation and loss to follow-up (Cohort studies)
- Keep a record of all losses and collect baseline data on them
- Make sure that diagnosis of disease is not affected by exposure status (blinding)



National Institute of Epidemiology  
Chennai  
NICCR-101  
HEALTH RESEARCH FUNDAMENTALS  
nie.gov.in

Now, at the time of data collection what we need to ensure is to minimize non-response, to minimize non-participation and make sure that we do not lose many people, especially in cohort studies over a long follow-up period. Even if, we should anticipate actually that we may lose people and so it would be a good idea to actually keep a record of all these losses, at least some basic socio demographic characteristics of these people. So, that later on, at the analysis stage, we can actually compare people who were lost to follow up versus those who remain in the study and see if there were any major differences in these two populations, which could lead to selection bias.

We also need to make sure at the time of data collection that the diagnosis of disease is not affected by the exposure status, which means at the time of selecting, who the cases and controls are, the person who is selecting the cases and controls should not be aware of what the exposures status of this population, of the cases and controls are and this one way in which we do, this is called blinding.




(Refer Slide Time: 09:19)

## Dealing with selection bias

### Analysis stage of study

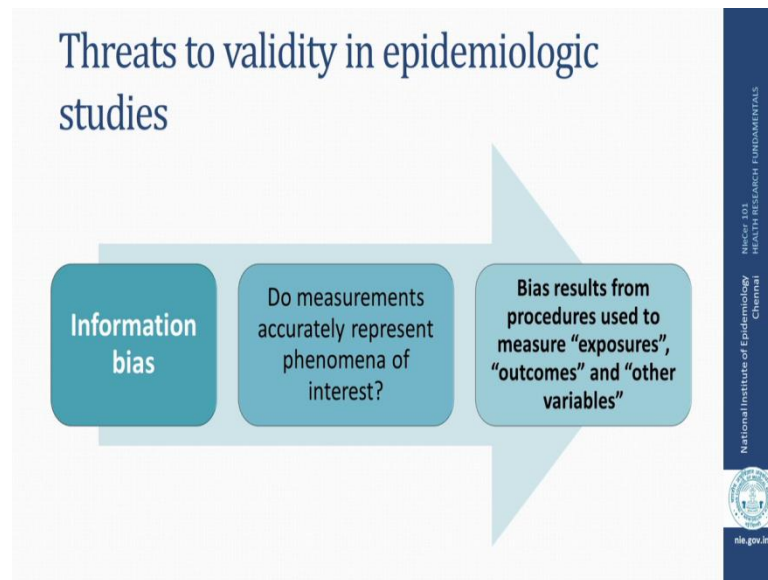
- Compare non-responders/dropouts with responders/non-dropouts with respect to baseline variables
  - Large differences strongly suggest selection bias
  - Small differences do not rule out selection bias
- Use study results and external information to deduce the direction of biases and assess magnitude of biases
  - Do sensitivity analysis



National Institute of Epidemiology  
Chennai  
NIC-101  
HEALTH RESEARCH FUNDAMENTALS  
nik.gov.in

Now, even at the analysis stage, what we can do is, as I told you before, we can compare those who responded or those who did not respond. Those who are dropouts compared to those who are left in the study with respect to the base line variables and see if there are any large or small difference between these two groups. If we find large differences, it is suggestive of selection bias; however, small differences do not rule out selection bias. So, we need to be wary of that. Again, another way to assess whether there may be a selection bias may have occurred in our study, is to do what we call as sensitivity analysis, in which we try to do an analysis assuming how much bias could have happened, and what direction it could have gone and try to see how it affects the study results. If the study results are affected in a major way then we can assume that, yes selection bias has occurred.

(Refer Slide Time: 10:19)




Moving on to the next threat to validity and that is called information bias. Information bias is essentially a bias that can occur when we are measuring the characteristics of study participants. Now, what do we measure, we measure exposures, we measure outcomes and we measure other variables which may influence the exposures and the outcomes, which are called as third factors or confounders or modifiers. What we need to make sure that the measurements that we are doing accurately, represent what it actually is, the level of exposure is accurately measured whether there is an outcome present or absent is accurately measured and other variables, such as socio demographic, age, gender, education, income all those variables are also appropriately measured.

(Refer Slide Time: 11:11)

## Information bias in epidemiological studies

- **Case control study**
  - Collection of information leaning towards specific exposure status
  - Recall - Cases may recall exposure more than controls
  - Better exposure data available on cases compared to controls
- **Cohort study**
  - Collection of information leaning towards specific outcome status
  - Better outcome data available on exposed compared to unexposed
- **Investigator** - Systematic collection of information supporting expected conclusions (Unconsciously or Consciously)
  - May be examined in the analysis
- **Prevarication** - Systematic distortion of the truth by subjects



National Institute of Epidemiology  
National Institute of Health Research - Fundamentals  
Chennai  
nie.gov.in

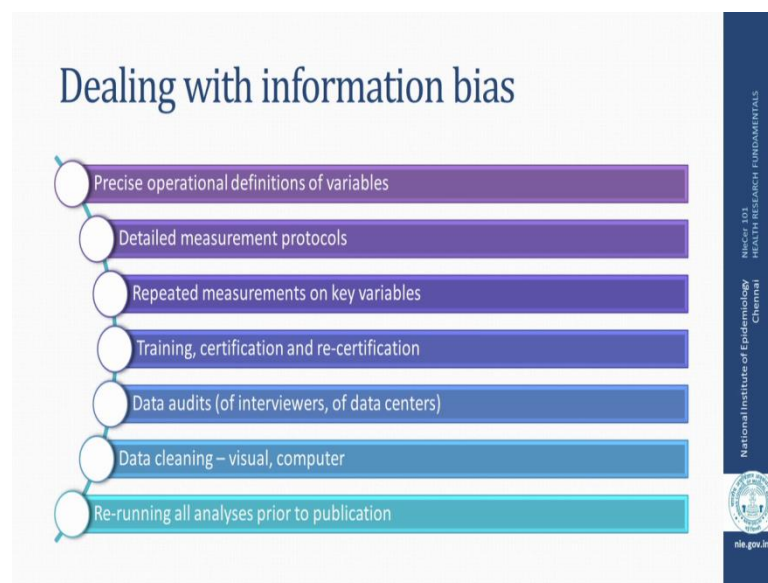
Now, how does this happen? In case control studies, information bias can happen if we are collecting exposure information which is leaning towards a particular exposure status, if we are trying to collect more of people, who are exposed compared to the unexposed or the other way round. This can lead to information bias. One of the very common ways in which information bias occurs in case control studies is through the process of recall. Remember, that we have cases and controls and we are trying to, we are asking them to recall the past history of exposures and it may be, it may so happen that those people who are diseased or who have a certain health event may be more likely to recall certain exposures compared to those people who are healthy and this is what we call as recall bias.

It may also be possible that better exposure data is available on cases compared to the controls and that again can lead to information bias. In cohort studies, information bias can happen, if we collect information leaning towards a specific outcome status. If we follow the exposed population much more rigorously compared to the unexposed population, that is something that can lead to information bias in cohort studies. It may also be possible that better outcome data is available among the exposed and then again compared to the unexposed, which can again produce information bias in the study.

Information bias can be introduced in a study both, either by what the investigator does in the way in which the investigators collect the information about the cases, about the

controls, about the exposure, about whether they get the disease or not get the disease and if there is a systematic way in which this is being done irregularly that can lead to information bias and last, but not the least, of course, remember that in general, in observational epidemiological studies, we are dependent on what **our** study participants tell us and if there is any systematic distortion of the true facts by the study participants, that is anyways going to lead to information bias.

(Refer Slide Time: 13:34)



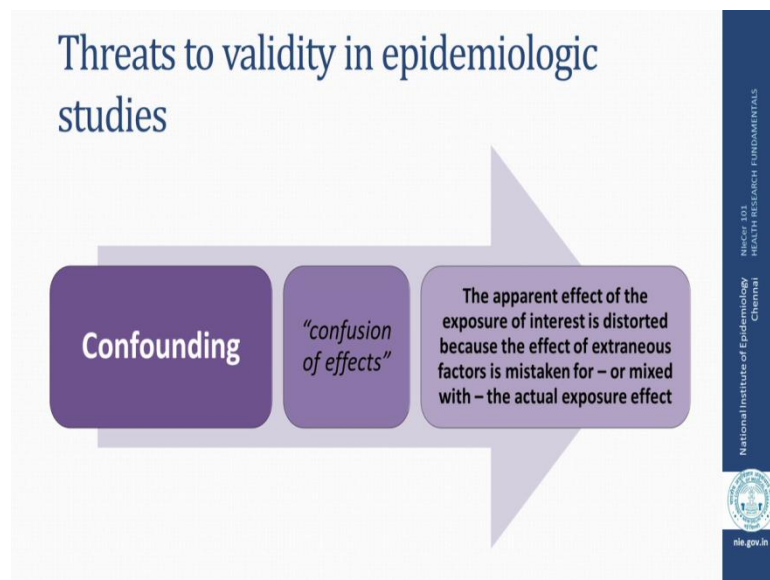
Now, how do we deal with information bias? First of all because we are measuring the exposure variable, the outcome variable and the other variables, we need to set up precise operational definitions of what we are going to measure and how much is it going to be. We need to have detailed measurement protocols in the way we are going to measure each of these variables.

Sometimes, it is also good way to do repeated measurements on key variables say, for example, blood pressure and we know that blood pressure can vary from time to time. So, we may take more than one readings of blood pressure and then take an average of that reading, in order to say, what the actual blood pressure of that individual is at that particular point of time. It is very important, that the investigators are trained and certified in the way in which they follow the study protocol and all the methodology that needs to be done to collect information. There we can do data audits, both of the interviewers and of course, of the data management centers where the data is stored to, in

order to make sure that the way in which the data is collected, the data is retrieved, the data is stored is done correctly and there is no information bias happening because of the same.

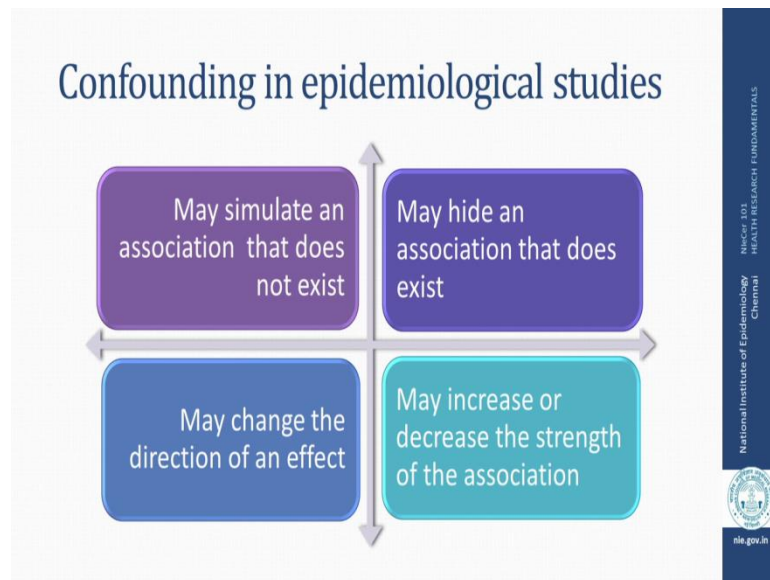
Once, the data is collected, we need to make sure that the data is clean. We need to go through the data both visually as well as through computer programs, softwares and make sure that we getting clean data. It is also good practice to actually rerun all your analysis before you are trying to do give say, send your paper for publication, just to make sure that there is no possibility of any information bias occurring because of the way the analysis was done.

(Refer Slide Time: 15:22)



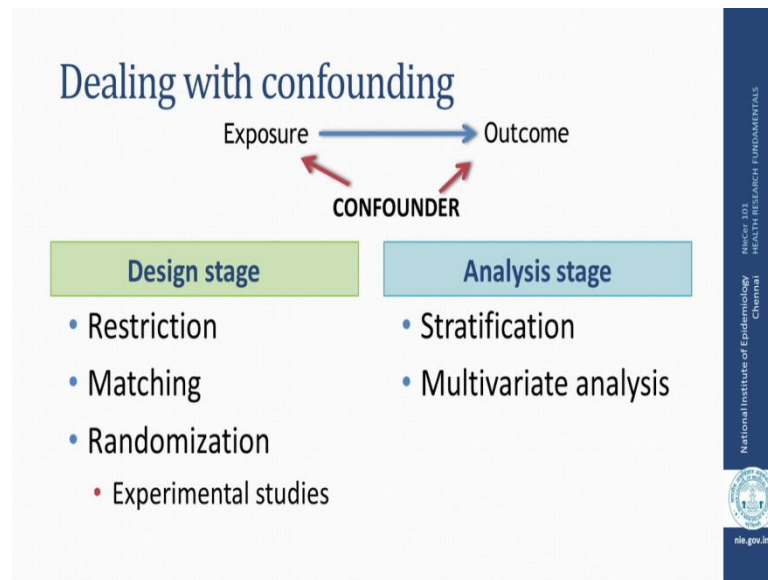
Now, we are going to look at the next threat to validity, which is called confounding. Confounding comes from a French word, which actually means confusion of effects. Now, what effects are we talking about here?, remember what we are doing in epidemiological study is looking at the effect of an exposure on the outcome, whether if you are more exposed are you more likely to get the disease or vice versa. Now, what we want to know is the effect of this exposure on a particular outcome. This effect can be confused with the effect of a third factor, which can have an influence both on the outcome as well as the exposure and this is what leads to the phenomenon of what is called confounding.

(Refer Slide Time: 16:07)



Now, what does confounding do? Actually, confounding is probably the most, the biggest threat to validity in any epidemiological study because confounding can actually simulate, can show you an association even when it does not exist. Confounding may hide associations that is actually there or confounding may actually increase or decrease the strength of the association. So, you may say that an exposure is more associated with the outcome or less associated with the outcome than what it actually is, and in the worst case scenario, confounding can actually change the direction of an effect. If an exposure, say, causes an outcome, because of confounding you may see that the exposure is preventing the occurrence of that outcome and that is the most dangerous threat to validity in any epidemiological study.

(Refer Slide Time: 17:06)



So, how does confounding happen? So, diagrammatically what we represent that confounder is a third factor, is a variable, which influences both the exposure and the outcome and when we are trying to determine what is the association between the exposure and outcome, this association is influenced by this third factor. Now, we can deal with confounding, both at the design stage and at the analysis stage. It is always better to deal with it at the design stage than to take care of it at the analysis stage. So, at the design stage, we can do several things.

One, we can do what is called restriction; we can restrict our study participants to only those people who are in one stratum of the confounders, so that the confounders cannot play a role in the association between exposure and outcome.

Secondly, we can do, what is called matching. If we already know what the potential confounders could be in a particular study we can match our cases and controls on those particular confounders and which will negate the effect of the confounders and then the association that we see between the exposure and outcomes would be without the influence of the confounders. Of course, remember that, if you do matching you have to do what is called matched analysis.

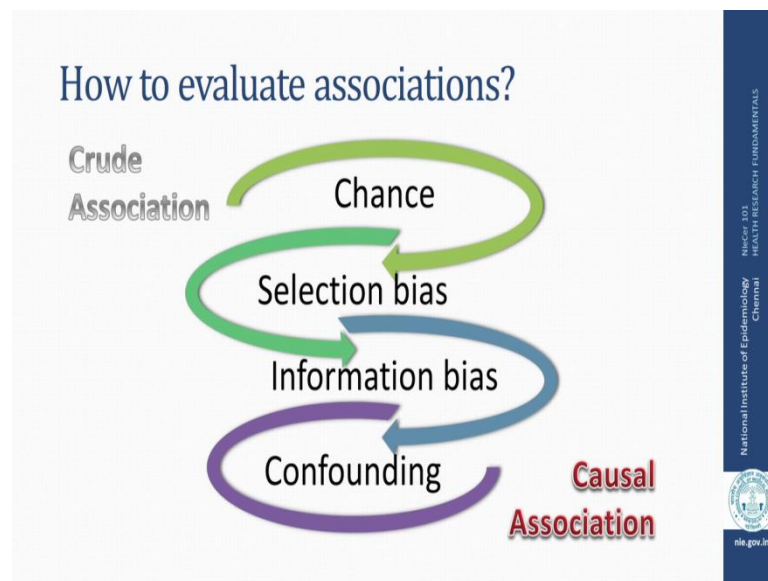
In experimental studies, we do what is called randomization and that is something that actually automatically takes care of the confounders and makes sure that the two arms in a randomized trials are similar in all ways in terms of the confounding variable. Now, at



the time of analysis, what we need to do in every study, is to actually first test whether there is any confounding or whether there are variables, which could be acting as confounders, which need to be taken care of at the time of analysis and this is where we do what is called stratified analysis. And we stratify our data in various stratum of the confounder, and then try to find associations and which helps us to identify whether there is confounding or not.

Now, in order to take care of these confounders, we can do, what is called a multivariate analysis, where in we do, we use regression techniques, whether it is logistic regression, linear regression or other advanced methodologies in order to take into account the effect of confounding and then the associations that we get between the exposure and outcome are without the influence of the confounder or as we say adjusted for the confounders.

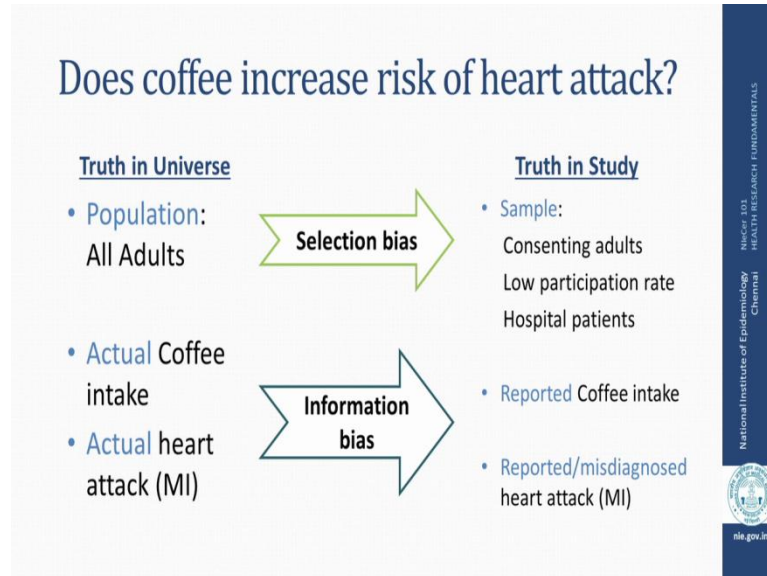
(Refer Slide Time: 19:49)



So, how do you evaluate associations? Whenever you see a study, whenever you see a risk ratio or an odds ratio, what you see is a crude association. Now, how do we make sure that this crude association is actually the true or the causal association, that is the true relationship between the exposure and the outcome? What we need to make sure is, we need to go through this spiral. We need to make sure that it is not because of chance. We need to ensure that there is no selection bias. We need to check if there could be any information bias. We need to understand if there could be confounding and if that

confounding has been taken care of. Only after going through this process, we would be able to say that whether the crude association is actually the casual association or not.

(Refer Slide Time: 20:46)



So, coming back to our problem, does coffee really increase the risk of heart attack? Well, let us analysis this. What we wanted to do is to look at all population, all adults in the population, who are drinking coffee. Now, in the study what we get is a sample of people who agreed to take part in the study. Now, these people could be people who are more likely to drink coffee or less likely to drink coffee. These people, may be hospitalized patients, and if you are doing a study in a hospital and it maybe that these patients are hospitalized for say, gastric ulcer and that is because of coffee drinking. So, the way in which we select these participants can actually lead to a bias and that is what is called selection bias.

Now, what are the exposures that we are trying to assess here, is the coffee, actual coffee intake of the study participants and what we get from the study participants is actually what they report? Are they reporting the true coffee intake? Do they actually remember how many cups of coffee they have had in the past? What is the average number of coffee they drink? Whether they drink coffee with milk, without milk? What is the strength of the coffee? All of these issues can actually influence whether the coffee intake that we are measuring is actually the true coffee intake and that can lead to information bias.

Again, remember that we are also trying to see whether the people really had a heart attack or not and it is possible that they maybe have may have been a misdiagnosis of a heart attack. There could be other, the chest pain that the study participants may report as heart attack, may actually have been may be due to other causes and that is reported as heart attack. So, actually what we may be seeing is not heart attack, but some other causes for chest pain and that is again the study results would then be influenced by information bias.

(Refer Slide Time: 22:50)

### Does coffee increase risk of heart attack?

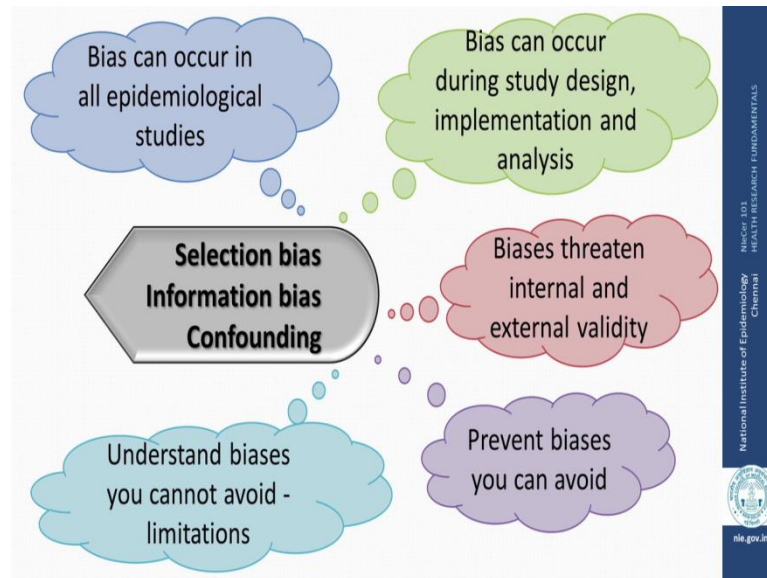
- Was the association between coffee and MI due to **CONFOUNDING** by smoking?
  - “A confounder is associated with both the exposure (coffee) and the outcome (MI).”
  - Smoking in
    - coffee (+) 86%
    - coffee (-) 27%
    - MI (+) 80%
    - MI (-) 40%

```
graph LR; Coffee -- "?" --> MI; Smoking --> Coffee; Smoking --> MI;
```

NIMRCH 101  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai  
nitg.gov.in

Then of course, there is confounding. Could it be that this association that we saw between coffee and heart attack, what we call myocardial infarction in medical terminology, could it be confounded by smoking? Is it possible that, we know that those who are smokers are more likely to, is a known risk factor for heart attack. It is also known that those who are smokers are more likely to be coffee drinkers and it is possible that because we may see more of smokers among the coffee drinkers and more of smokers among those who had an heart attack the association that we are seeing between coffee and heart attack is not due to the actual coffee, but it is actually because of the effect of smoking on heart attack. So, the result of this association between coffee and MI could have just been confounded by the effect of smoking.

(Refer Slide Time: 23:49)



What we need to understand is that, there are various threats to validity in any epidemiological study and these biases can occur in all epidemiological studies, more so in observational studies, such as case control and cohort study and less so, in randomized trials. Biases can occur during all stages of the study, when we are designing the study, if the study is not designed appropriately, if the study is not conducted appropriately or if the analysis is not done appropriately, all of which can lead to one or the other biases. And we know that biases threaten both the internal and external validity.

Remember that, the study which has no internal validity cannot be generalized and so it does not have any external validity. So, we need to keep in mind is that, when we are designing a research study, we need to be thinking of all the possible ways in which these various biases could creep-in into our study and design it appropriately and try to prevent as many biases as possible at the time of designing and implementing the study.

However, we should also remember that, there could be some biases which cannot be avoided. What we need to understand is at the time of analysis of the results, we need to be aware of what these biases could have been and state these biases in the form of limitations of the study. So, it is critical that whenever we look at the results of any epidemiological study, we need to be wary of what possible threats could be to the validity of these studies and make sure that the investigators have taken care of these various threats.

Thank you.