

Health Research Fundamentals
Dr. Manickam Ponnaiah
ICMR-National Institute of Epidemiology, Chennai

Lecture – 18
Overview of data analysis

Hello friends, welcome to this session, the course Health Research Fundamentals.

(Refer Slide Time: 00:13)

Key objectives of data analysis

- Plan the analysis
- Programme the crude analysis
- Deal with chance, biases and third factors
- Assess causality
- Measure clinical/public health impact

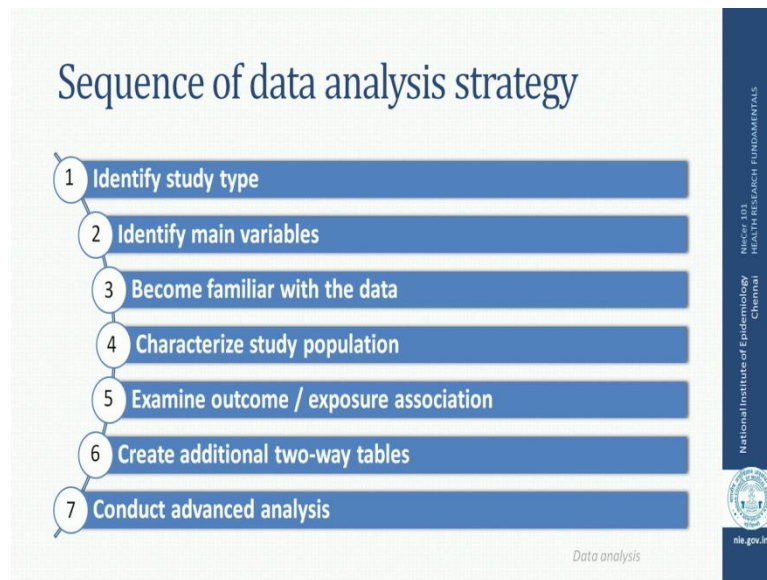
Data analysis

NICE: 101
HEALTH RESEARCH FUNDAMENTALS
National Institute of Epidemiology
Chennai

nit.gov.in

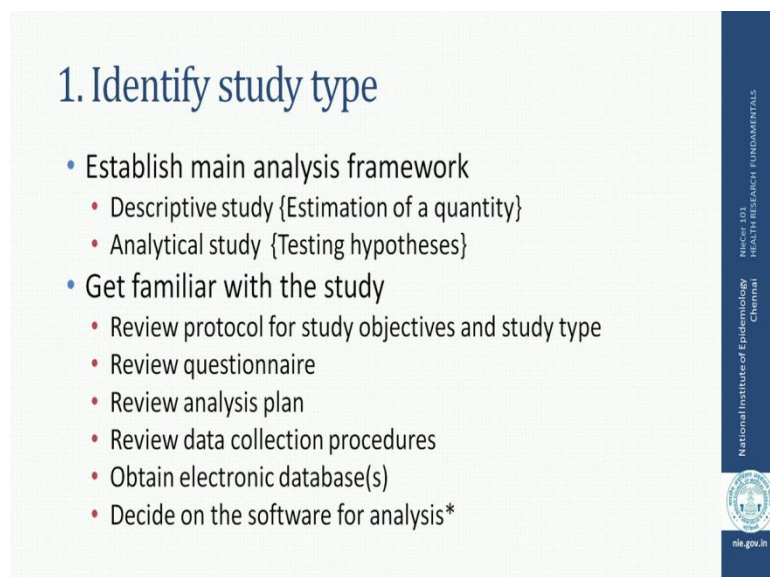
I am going to look at Data Analysis. What are the key objectives of data analysis? The objectives are, to plan an analysis, program analysis; deal with chance, biases and third factors, to assess causality that is what Dr. Sanjay Mehendale talked in the first session. The essence of research is to link an exposure to an outcome and finally, to measure the impact it has in clinical or in domain of science.

(Refer Slide Time: 00:50)



The data analysis strategy has its sequence and I am **going** to talk about these 7 steps. The first step is identifying the study type, which is essential before you venture into data analysis. The second is in terms of identifying main variables. The third, becoming familiar with the data. Fourth, characterize the study population and fifth, examining the association between exposure and outcome based on the study type and sixth, in terms of a creating additional tables and finally, to conduct an advance analysis.

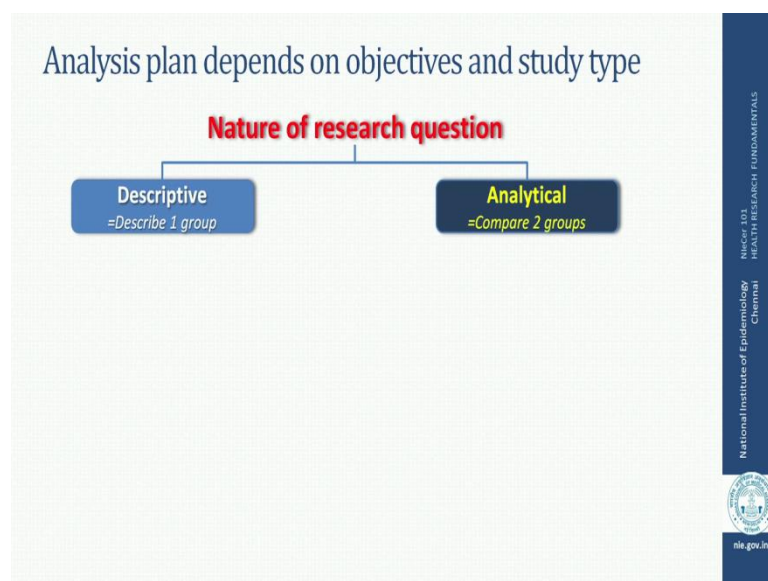
(Refer Slide Time: 01:30)



Identifying a study type is the first and foremost step before you enter into data analysis, because it establishes the main frame work. I am going to repeat again that you need to know whether you are dealing with the descriptive question or descriptive study or you are dealing with an analytical study. If a descriptive study is involved, you need to measure a quantity and estimate appropriate indicator, which I am sure you have seen in the measurements lecture and if it is an analytical study, you need to test a hypothesis using statistical test.

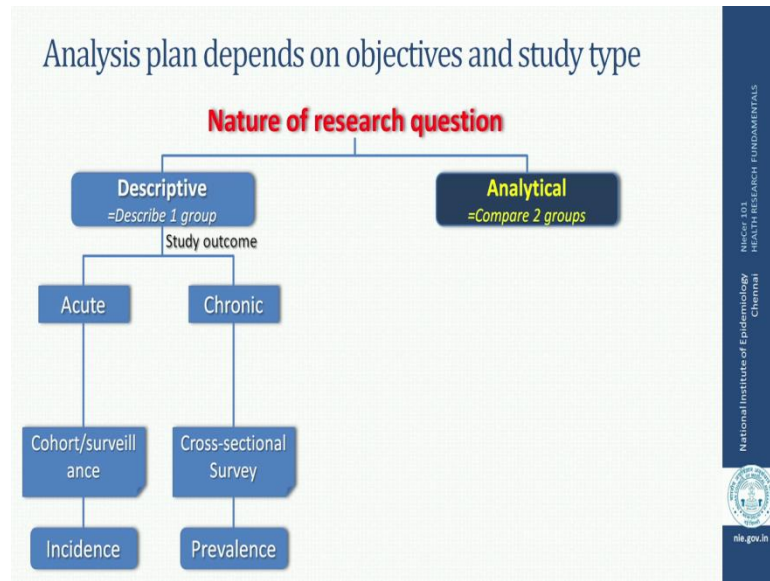
For that clarity, you need to be familiar with the study, you need to look at the protocol especially for study objectives and study type, you need to review the entire questionnaire, you need to look at the analysis plan, you need to look at the data collection procedures, you need to obtain the electronic data base and you may have to decide on software for analysis about which we will see at the end.

(Refer Slide Time: 02:28)



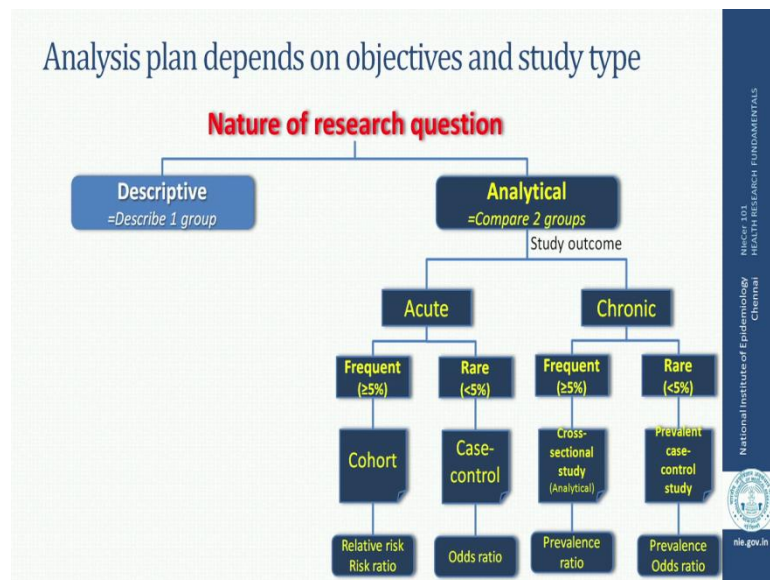
This analysis depends on the statement of objectives and study type. Therefore, the nature of research question has to be very clear, whether it is descriptive involving, describing one group measuring a quantity or is this analytical involving some intervention comparing two groups and hypothesis.

(Refer Slide Time: 02:52).



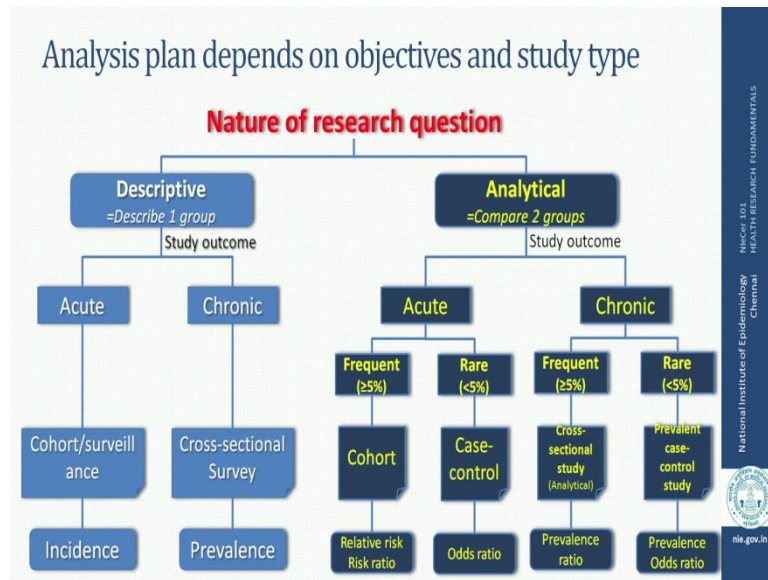
If it is descriptive, you need to ask this question whether it is acute or chronic outcome and then accordingly, you will zoom into a study design and then measure either incidence or prevalence.

(Refer Slide Time: 03:04).



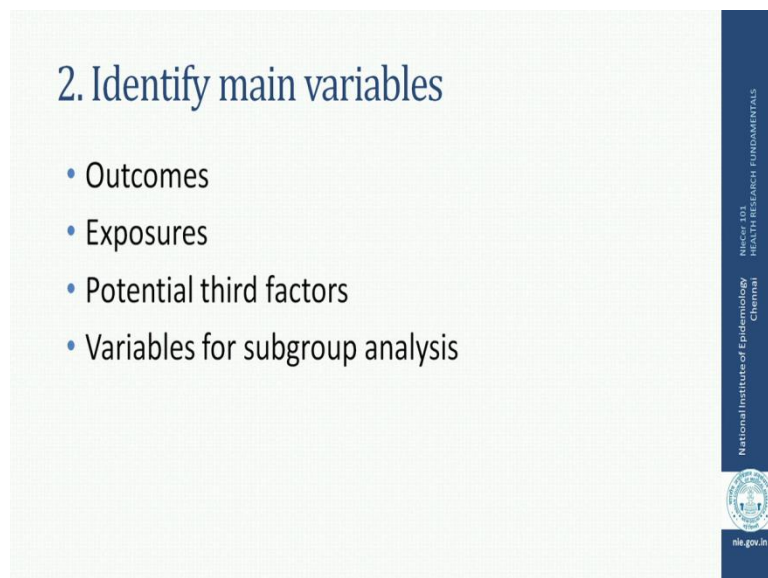
If it is analytical and again you will ask question what is the study outcome? You may decide on acute or chronic outcome and then choose an appropriate measure of indicator, it could be relative risk or risk ratio, if a cohort study or it is odds ratio based on case-control study or prevalence ratio or prevalence odds ratio.

(Refer Slide Time: 03:29)



So therefore, your analysis plan depends on your review of the study with reference to; what is the main frame work? What is the research question that they are trying to answer?

(Refer Slide Time: 03:37)



The second thing you need to identify the main variables. What are the outcomes? What are exposures? What are the potential third factors? And what is the variables that be need to consider for subgroup analysis?

(Refer Slide Time: 03:49)

3. Become familiar with the data

- Perform
 - Frequency distribution
 - Examine frequency of all the variables
 - Descriptive statistics
 - All the variables describing the study population
- Review number of observations by status in the database
 - Look for duplicates
 - Look for missing observations
- Check ranges and legal values
- Check consistency

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

We need to be familiar, intimate with the data. How do you do that? You need to perform a frequency distribution of all the variables. Look at the frequency of all the variables in your data set and then you look at descriptive statistics which helps you to describe the study population. This will give you a fairly good idea of what this data set is all about. And secondly, you review the observations by its status in the database, are they duplicates, cross check, looks for missing observations, check the ranges and legal values against the specified in the data dictionary, check for consistency in the pattern of the data. So this is the crucial step, the third step which you need to spend a sufficient time.

(Refer Slide Time: 04:36).

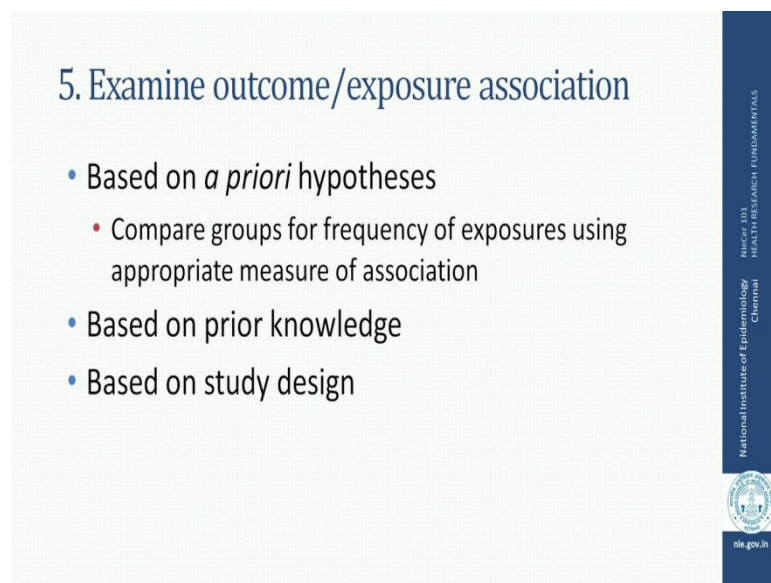
4. Characterize study population

- Baseline characteristics
 - Distribution of study participants by socio-demographic- economic variables
 - e.g., Age, gender, income
- Frequency of clinical features/ health problems
- In analytical study → for compared groups

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

Fourth, you need to characterize the study population with reference to the basic characteristics of the study population. These could be in terms of you know socio-demographic and economic variables by age, gender, what income groups and things like that. And then if it involves analytical study you look at in the two groups or three groups, the comparison groups these variables. You may want look at frequency of clinical features as well in characterizing study population.

(Refer Slide Time: 05:10)



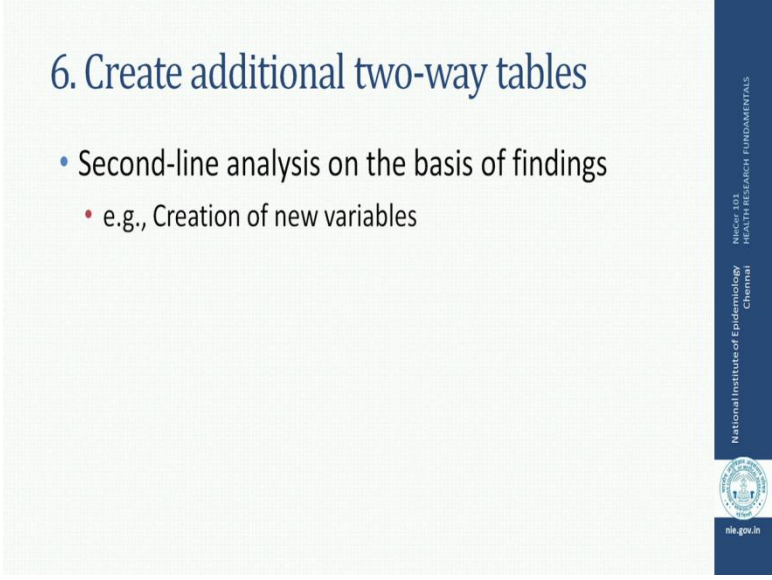
5. Examine outcome/exposure association

- Based on *a priori* hypotheses
 - Compare groups for frequency of exposures using appropriate measure of association
- Based on prior knowledge
- Based on study design

National Institute of Epidemiology
Chennai
NIER 301
HEALTH RESEARCH FUNDAMENTALS
nie.gov.in

Fifth, is in terms of examining the outcome and exposure association, this is the most interesting part. This is based on a priori hypotheses or hypothesis, where you compare the groups for frequency of exposures using appropriate measure of association which we talked about earlier. This has to be based on prior knowledge; it has to be based on study design because as you of seen earlier, each of the study design has its own measurement of association for the specific exposure and outcome. Therefore, this is a very critical step. And then you may also apply the principles that you know you might have learnt from Dr. Tarun, when he talked about the biases and confounding that can complicate this association apart from the biological and knowledge that we have about particular exposure and outcome.

(Refer Slide Time: 06:07)



6. Create additional two-way tables

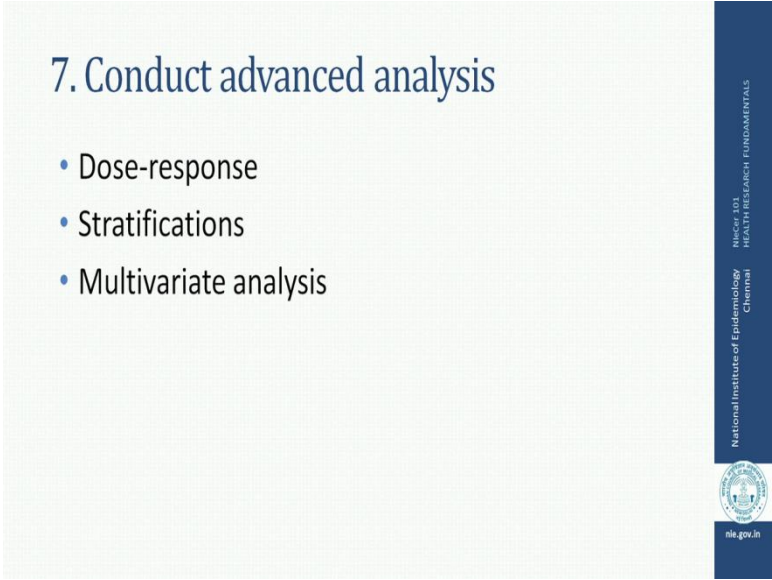
- Second-line analysis on the basis of findings
 - e.g., Creation of new variables

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

This slide features a light green background with a dark blue vertical sidebar on the right. The sidebar contains the text 'NICOR 101 HEALTH RESEARCH FUNDAMENTALS', the logo of the National Institute of Epidemiology, Chennai, and the website 'nie.gov.in'. The main content area has a title '6. Create additional two-way tables' and a single bullet point: '• Second-line analysis on the basis of findings', which includes a sub-bullet '• e.g., Creation of new variables'.

The sixth step is in terms of creating additional two-way table for analyzing new variables that you may detect on the basis of findings that have already generated.

(Refer Slide Time: 06:20)



7. Conduct advanced analysis

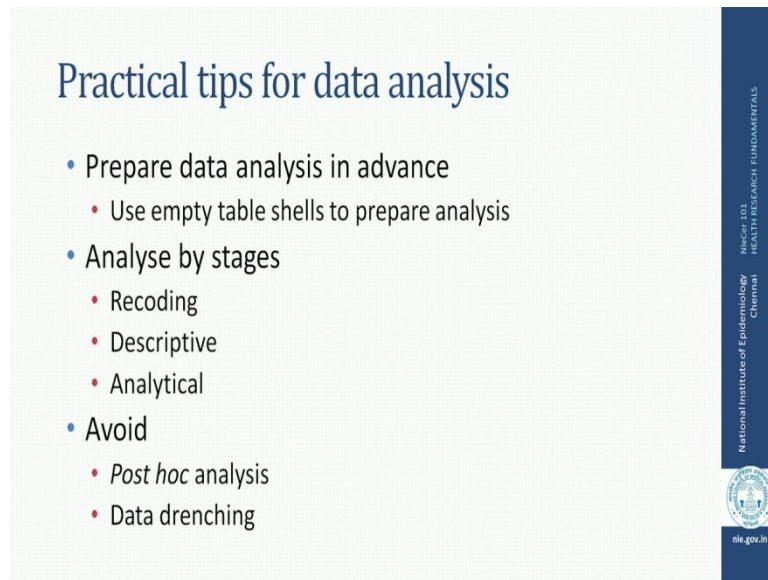
- Dose-response
- Stratifications
- Multivariate analysis

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

This slide features a light green background with a dark blue vertical sidebar on the right. The sidebar contains the text 'NICOR 101 HEALTH RESEARCH FUNDAMENTALS', the logo of the National Institute of Epidemiology, Chennai, and the website 'nie.gov.in'. The main content area has a title '7. Conduct advanced analysis' and three bullet points: '• Dose-response', '• Stratifications', and '• Multivariate analysis'.

The seventh step could be in terms of a dose-response, stratifications and multivariate modeling analysis.

(Refer Slide Time: 06:34)



Practical tips for data analysis

- Prepare data analysis in advance
 - Use empty table shells to prepare analysis
- Analyse by stages
 - Recoding
 - Descriptive
 - Analytical
- Avoid
 - *Post hoc* analysis
 - Data dredging

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

Therefore, these are the steps that you will conclude by analyzing the data for generating information that you think will be useful to improve the health of the study population. Some practical tips that you may want to keep in your notes in terms of a prior plan of analysis. It is very useful, it is has to be done. The data analysis has to be done; plan has to be prepared well in advance.

We will recommend you to use empty tables to prepare analysis. In your reading material you will see such empty tables for each of the study designs. We have prepared an empty table shell, so for your study you can prepare a plan of action for analysis, this is how my table will look like and then at the time of analysis your attempt is to fill up that empty shells. The second step is in terms of analyzing by stages go by stages. You first do the recoding that is necessary, you may create new variables, you may dichotomize, you may look at the descriptive information, you may decide to change the way you cut the two groups, you may decide on three or four groups for measuring those response based on your understanding from the data, you may do the next step of descriptive analysis and then finally, analytical analysis.

So this has to be a sequential process, it should not jump. People have the tendency to jump the steps; I think it has to be in the measured manner one by one. And finally, please avoid any analysis that is driven by the data that you analyze without any plan, this is called Post Hoc analysis. You did not have a plan, you finally find something and

then you tried to make news out of it. Last but very important, you do not look squeeze of the data because you want something.

(Refer Slide Time: 08:37)

Initial stages of the analysis:
e.g., Effect of brisk walking on fasting blood sugar levels in diabetics

- Recoding stage
 - Create outcome data
 - Recode key variables e.g., age-groups, income
- Descriptive stage
 - Calculate frequency of outcome


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

As an example, we talked about exercise and diabetics you remember. In the initial stages of the analysis, if this question has to be answered through data analysis, we will recode the data, we will recode in terms of creating an outcome data. Here, we talked about blood sugar levels so we have to group the outcome into positive outcome, yes, reduce blood sugar level, negative outcome, no, not reduced, blood sugar is not reduced, so that is the outcome. Then we may recode some key variables such as, age-groups can be cutoff, income level can be used to cutoff as you know ok income, not so **ok** income, or below poverty line or above poverty line and things like that. This recoding we may do for number of variables including for examples, in terms of exercise, moderate, heavy, mild or no exercise, we can create groups during this recoding stage. And then in the descriptive stage we calculate the frequency of the outcome by each of these groups.

(Refer Slide Time: 09:42)

Analytical stage of the analysis:
e.g., Effect of brisk walking on fasting blood sugar levels in diabetics

- Univariate analysis
 - Frequency of outcome by age, gender and income
 - Frequency of outcome by income categories (potentially examine dose-response effect)
- Stratified analysis
 - Frequency of outcome by income, stratified for age, gender and income
- Multivariate analysis
 - Logistic regression model




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

Analytical stage, we do it in three steps. One is Univariate analysis, where we look at one upon the other. We look at the outcome which is in terms of reduction of blood sugar level by age, gender or income or such similar variables that are collected in the data base. Then, if you want to do a stratified analysis or dose response effect you can examine the outcome by categories of a particular variable. For example, if there are income categories quartiles or levels of income, low income group, middle income group, high income group, what is the frequency of outcome? You can examine. And then you may do a stratified analysis, for example if you want to look at the exposure that is exercise and its relationship with reduction in blood sugar level among income groups stratified by age, gender and income. You can look at all of this in the next step. And finally, a model, logistic regression model would tell us whether exercise can predict reduction in fasting blood sugar level in the diabetes.


(Refer Slide Time: 10:53)

Software for data management and analysis

- ✘ Avoid spreadsheets for data management /analysis of any type /size
- ✓ Use software with data management & analysis tools
- ✓ e.g., EpiInfo*



*EpiInfo 3.5.4 or EpiInfo 7.1.5; www.cdc.gov/epiinfo

NICRI 101
HEALTH RESEARCH FUNDAMENTALS
National Institute of Epidemiology
Chennai

nik.gov.in

Finally, for beginners or people who have no exposure handling databases, you may wonder what data base to use. The usual trap that we get into is using spreadsheets for data management and analysis of any type in size. We **think** that it is a small data, let use it in we know as spreadsheet but I think, we recommend that you better avoid that temptation because spreadsheet are not meant to be data management or data analysis tools. It is preferable to use softwares that can give you both the capabilities data management and as well as data analysis.

For example, we are just suggesting that this is one of the softwares that we find where we using free software call EpiInfo, which has the capability to for example, create a data collection instrument format, enter the data, analyze using the even you know for some the latest visual analysis and then it has the capability to map the information if that is part of your analysis. And finally, it can also be used for a fairly a good amount of statistical analysis. So, such softwares may be very helpful. We wish you good luck with your database a management as well as data analysis because that leads to report generation.

Thank you very much.