

# HEALTH RESEARCH FUNDAMENTALS



Prof. Sanjay Mehendale | Prof. Manoj V. Murhekar,

Prof. R. Ramakrishnan | Prof. Tarun Bhatnagar,

Prof. Prabhdeep Kaur | Prof. P. Manickam

Prof. P. GaneshKumar

Multidisciplinary  
National Institute of Epidemiology

**NIE**

NATIONAL INSTITUTE OF  
EPIDEMIOLOGY

# **INDEX**

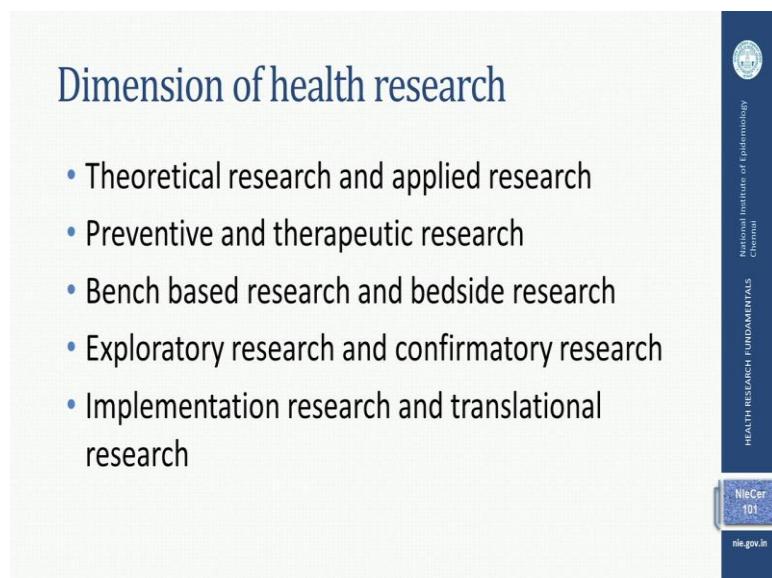
<b>S.No</b>	<b>Topic</b>	<b>Page No.</b>
	<b><i>Week 1</i></b>	
1	Introduction to Health Research	01
2	Formulating research question, hypothesis and objectives	18
3	Literature review	40
	<b><i>Week 2</i></b>	
4	Measurement of disease frequency	58
5	Descriptive study designs	74
6	Analytical study designs	87
	<b><i>Week 3</i></b>	
7	Experimental study designs: Clinical trials	110
8	Validity of epidemiological studies	124
9	Qualitative research methods: An overview	144
	<b><i>Week 4</i></b>	
10	Measurement of study variables	159
11	Sampling methods	175
12	Calculating sample size and power	195
	<b><i>Week 5</i></b>	
13	Selection of study population	218
14	Study plan and project management	234
15	Designing data collection tools	250
	<b><i>Week 6</i></b>	
16	Principles of data collection	270
17	Data management	284
18	Overview of data analysis	300
	<b><i>Week 7</i></b>	
19	Ethical framework for health research	314
20	Conducting clinical trials	331
	<b><i>Week 8</i></b>	
21	Preparing a concept paper for research projects	342
22	Elements of a protocol for research studies	360

**Health Research Fundamentals**  
**Dr. Sanjay Mehendale**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**Lecture - 01**  
**Introduction to health research**

Hello, in this online course on Health Research Fundamentals being offered by National Institute of Epidemiology and Indian Council of Medical Research.

(Refer Slide Time: 00:19)



**Dimension of health research**

- Theoretical research and applied research
- Preventive and therapeutic research
- Bench based research and bedside research
- Exploratory research and confirmatory research
- Implementation research and translational research

National Institute of Epidemiology  
Chennai

HEALTH RESEARCH FUNDAMENTALS

NieCer 101

nie.gov.in

Today, I am going to give you Introduction to Health Research. Primarily, we have to understand that there are many dimensions to health research. Health research is conducted for theoretical purposes and when we have some good evidence created out of that we go for applied research. There are also prevention and therapeutic angles to doing health research as well, in the sense some of the preventive technologies as they emerge, they have to be tested out through research methods as well as newer therapeutic options become available they need to be tested out through research methods.

Essentially, we have many a times heard of and we have seen in the movies the bench best research, where we have seen scientist in white coats working in the laboratories

trying to do something with animals like mice and monkeys and so on and so forth, but what they do there has no significance until the findings or the learning's that they have got in that particular research are carried further and they become available to the bed side, that is they are they can be applied to the human beings either for elevating diseases or for preventing diseases. And this is a process which one has to understand and one has to really follow correctly.

Sometimes the research is of exploratory in nature, where in the sense we do not know much about it right in the beginning, but what we try to do is employ various methods and try to figure out if we can get some clues for further research. Sometimes, it can be confirmatory, in the sense some of the clues that have been obtained by people can they be say sort of strengthen or can we get additional information which could be of practical significance that is called as Confirmatory Research.

There are other terminologies which are also employed; one which has been recently employed in a big way is Implementation Research and Translational Research. Many of the government programs get implemented and what is important to understand at this particular point of time is how these programs are functioning well and it is also therefore important to make mid course corrections, if they are necessary or to decide in which areas you need specific angles of the programs to be strengthened and so on. Transnational research basically talks about the earlier concept of bench to bed site that I talked about. This is a process of development of technologies for human benefit and human welfare.

(Refer Slide Time: 02:47)

## Fundamental principles to be followed

- Planning stage is very critical – it is important to spend enough time and involve the right people in planning the study
- Team work is critical
- Three levels of review are essential
  - Scientific review: novelty, rationality, justification
  - Ethics review: human subjects protection
  - Regulatory review: foreign funding, sample shipment, intellectual property, exchange of visitors

National Institute of Epidemiology  
Chennai

HEALTH RESEARCH FUNDAMENTALS

NIECer 101

nie.gov.in

There are some important principles that we have to understand when we conduct health research, but one thing which is really critical is that the planning stage in research is absolutely critical, why? Because, if we do not spend enough time and if we do not involve the right kind of people in planning the research study, it is very likely that some of the issues that we could have earlier prevented or they get missed which we could have prevented in the real sense. Hence, in my opinion team work is really critical in research, very rarely solo research succeeds. But for any kind of research to happen it has to undergo several layers or levels of reviews.

Scientific reviews are important because they look at, what is the novelty in the concept that is being looked at? Is there rationality behind doing that? And what is the justification for that? And in that particular context the rational for doing that particular study in a specified country is also critically important. Whereas, the ethics review primarily focuses towards finding out, whether the human subjects protection issues have been adequately taken care of or not. What is important is research definitely means development, research definitely means advancements, but it is cannot be at the cost of human subjects. People who are participating in the research are described as human subjects or human participants and we should do nothing that can really harm them in the long run or in the short run. So, ethics review ensures that this does not happen.

There are certain in country procedures or in country reviews called as Regulatory Reviews and they are basically there to decide about, what kind of foreign funding is being received for that particular project? Are there any sample shipments that are going to happen? Because there is a lot of intellectual property also which is attached to this data sharing which happens, we always should be protective of our own intellectual properties and so the regulatory authorities in our country do take care of this and they ensure that the our intellectual property is properly protected. Some projects do also involve exchange of visitors and several countries have their own rules and regulations regarding the visitors, who are coming and going. These are mandatory aspects and it is important that the regulatory committees do review that aspect as well.

(Refer Slide Time: 05:27)

## Process of health research

- Ensure that data is collected systematically
- Draw meaningful conclusions
- Make appropriate decisions
- Take appropriate actions for prevention and control of diseases, conditions: Evidence based actions
- These should help in reduction of suffering and ultimately improve health and well-being of the community

National Institute of Epidemiology  
Chennai  
HEALTH RESEARCH FUNDAMENTALS

NIECer 101  
nie.gov.in

One thing I would like to stress, the health research or for that matter any research is a process and it contains multiple components and each and every component in this process is of critical importance. It all starts with collecting data. We have to collect data with a specific purpose and for that data to be of high quality, the data collecting instruments also have to be appropriately designed. If the data quality is good then we can draw meaningful conclusions based on that and then we can make appropriate decisions.

What is important is once these decisions are made, the policy planners and the program managers of the country they decide, whether this is the right time to take these particular learning's from research to appropriate actions which can be employed at an individual level or at a mass level. This is what we call as Evidence based action. Primarily, all this is done so that we ensure that there is a reduction in suffering and ultimately improvement in health and well being of the people or the community.

(Refer Slide Time: 06:40)

Breadth and depth of inquiry in health research

- Human host: healthy, susceptible, with disease, dead
- Surrounding environment and society: climatic factors, housing, vectors, animals, socio-cultural practices, family structure
- Health care infrastructure and delivery

National Institute of Epidemiology  
Chennai

HEALTH RESEARCH FUNDAMENTALS

NIECer 101

nie.gov.in

But when we talk about the kind of information that will collect in research, it can have wide range of breadth and depth as well. There is some information which is required to be collected with respect to the human host for example, and they can be sometimes healthy, they can be sometimes susceptible to a particular condition or they can be suffering from a particular disease, some of them may have died because of the that particular disease also. We have to figure out, how we can get the required information from these kinds of subjects from these various kinds of host that I just talked about.

The disease cannot occur in an individual unless and until there are many factors which come together and that is why we in the modern times always say that there is a multi factorial origin for occurrence of any particular disease and we all know, the environment which is surrounding us and the society in which we live, plays a very

significant role in occurrence of diseases. There could be factors like climatic factors, there could be housing factors, housing related factors, the vector around us, the animals around us, various socio-cultural practices that we follow, the family structure with which we live, all these factors can effect not only the health of individuals which could be physical or mental as well but they can lead to a occurrence of a disease as well.

In addition to that one more dimension that comes into the picture and which is really important from the context of health is the Health care infrastructure. Sometimes, people find it very difficult to access health. The access to health is a critical component because many of these studies have shown that if people are not able to access health in time, the kind of complications that arise and the death rates that result from that are unusually high in nature. So, there could be multiple angles which are require to be collected, we call all those like study variables. We have to collect correct information on those study variables when we carry out health research.

(Refer Slide Time: 08:52)

## Broad scope of health research

- Getting additional or new information
  - Are more of diphtheria and pertussis reported among adults in recent times?
  - What are the differences in full genome structure of HBV and HEV?
- Verifying and confirming available information
  - Are etiologies of pediatric pneumonia different in the children aged 5 or less in developed and resource limited countries?
  - Have the incidence and complications of diabetes changed with increasing consumption of pre-cooked and packaged food?
- Explaining cause and effect relationship
  - Does presence of a particular co-receptor [cause] on CD4 cells protect against HIV infection [effect]?
  - Are breast cancers [effect] more common in breast implant [cause] recipients?

Primarily, when we talk about the scope of health research or the objectives of health research they could be many. For example, most of the times when we think of research, we think of something like creating new information but the objective could be creating totally new information but we sometimes also get additional information on something

which already exists. Say for example, are more of diphtheria and pertussis cases reported among adults in the recent times? We used to know that these were the diseases of childhood earlier but now, it has been seen that some rare instances the cases in adults are also getting reported, so we need to figure out why this is happening? And where is it happening?

Similarly, in terms of getting new information somebody might want to do research to find out the full genome structure of hepatitis B virus and hepatitis E virus. It could have lot of implications in terms of understanding what kind of pathogenic impact it would have in human beings, also it could have significant decision making with respect to development of vaccine against those viruses.

Another objective that we also can pursue through health research is to verify and confirm available information. Here is where, most of the research that happens in our country is now a days happening. For example, are etiologies of pediatric pneumonia, you all know, pneumonia is a very serious disease in childhood particularly in children below 5 years of age. So, are the etiologies different in the developed countries and a resource limited countries like India. If one wants to figure this out, this could fit under this particular objective. Also have the incidence and complications of diabetes changed with increased consumption of pre-cooked or packaged food. We have seen this to change, the eating habits of people have changed but has it got any relation with the incidence and complications we might want to study through health research.

Many a times health research is also focused on finding out the cause and effect relationship and this could be applied if you think about it in multiple situations. For example, there is a presence of a specific receptor on a type of white blood cells which are called CD4 cells in the human body and they are believed to protect against HIV infection. But whether they project against a particular type of HIV infection, a particular type of HIV infection or they have a generalized effect, this could be evaluated under a cause and effect relationship related research. You must also have heard that in the recent times breast implantation operations are undertaken, but do this breast implant operations or do these recipients are they more likely to develop breast cancers, this is something like a cause and effect relationship, where we can think of breast implant is a cause and

breast cancers is a effect.

(Refer Slide Time: 12:00)

**Broad scope of health research**

- Testing new drugs, vaccines, tools or interventions for prevention, treatment and control of a disease
  - Can INH prophylaxis delay the onset of tuberculosis in HIV infected persons?
  - Will introduction of smokeless stoves result in reduction of respiratory morbidity and mortality in rural areas?
- Evaluating ongoing programs and assessing feasibility of new programs
  - Is injectable iron sucrose a better alternative to deal with pregnancy related anemia than oral iron?
  - Will the Integrated Disease Surveillance Program be able to predict the epidemics of influenza in India?

National Institute of Epidemiology  
Chennai

HEALTH RESEARCH FUNDAMENTALS

NICer 101

nie.gov.in

One more important area in health research is that testing of new interventions like says, new drugs, new vaccines new tools or newer prevention treatment technologies, etcetera. For example, it has been said that tuberculosis is one of the most common graded complications or diseases to happen in HIV infected individuals and mostly this is the disease which eventually kills them. So can we introduce INH profile access? INH is a potent drug which is used against tuberculosis, so can it be introduced among HIV infected individuals, so that they can be protected from getting tuberculosis. This is an important intervention which could be tried out and to reduce mortality associated with HIV because of prevention of tuberculosis in them.

Or for example, we are nowadays talking a lot about indoor air pollution particularly in the rural areas. So, if we decide to introduce smokeless stoves in those areas, would there lead to reduction in mortality and morbidity in the rural areas. This could again be a part of a intervention evaluation. Sometimes, the government has various on going programs, Public Health Programs which are vertically in nature, which are operated by the federal government or the central government in our country. All of you know that the general health services are provided in our country by the state level health services but the

national level control programs are operated and funded by the central government.

So for example, we have seen that there has been large number of women noted with anemia and this is one reason, why many complications occur in pregnant women. There has been this oral iron supplementation has been the main stage for treatment of this particular condition for many years, but nowadays iron sucrose inject able iron has become available and it is believed that it probably can improve or tackle anemia during pregnancy much better than oral iron. So, evaluating that as been given in the program could be an important thing that we can think about under this part of research. Also the government has introduced a program for several years now called as Integrated Disease Surveillance Project or Integrated Disease Surveillance Programs now. So can you predict the epidemics of newer and emerging diseases like H1N1 influenza that has been affecting us recently.

(Refer Slide Time: 14:40)

## Making the right choice of study design

- Qualitative studies or Quantitative studies
- Observational studies or Experimental studies
- Retrospective studies or Prospective studies

National Institute of Epidemiology  
HEALTH RESEARCH : FUNDAMENTALS  
NIeCer 101  
nie.gov.in

So, these are some important objectives of research that we have to keep in mind but one important thing is, if you want to do sound research, it is really critical that we make the absolutely a right choice about the study design over here because it some kind of a miss judgment here can lead to a futile kind of a research or which does not help you to understand it. This can be broadly say described in terms of some enquirers or some kind

of health research can be qualitative in nature, which requires person to person interviews or discussion in focus groups or just say free listing or observations and things like that. It mostly is observational in nature and it is open ended and people do probe at times to find out this information. Whereas quantitative studies are mostly based on structured questionnaires, that is where people deal with questionnaires, which are previously thought of and only the questions with very specified options of answers that can be made available are used or employed in the study.

One more important differentiation that is made is the observational study and experimental studies. In epidemiological terms, observational studies are those wherein the investigators do not change the environment in which the study participants are living and the main thing that distinguishes the experimental studies is, this is where the participants are exposed to some kind of an intervention at the will of the investigators. So, some kind of a change is made and that is how this kind of studies becomes different kinds of studies. They also described as retrospective studies or prospective studies.

Retrospective studies are classically where the information on the outcome that we are trying to study in a particular health research is already available. But in a prospective study we only start the study with people who are at risk or who are susceptible to a particular disease also have a comparator arm or control group, comparison group and then follow these two groups to find out how many people in either of these two groups develop a particular disease, which they were free off at the beginning of the study. So outcome happens sometimes in the future and it is called as a Prospective study design.

(Refer Slide Time: 17:06)

The slide has a blue header bar with the text 'Some critical considerations in planning phase'. On the right side, there is vertical text: 'National Institute of Epidemiology', 'Chennai', 'HEALTH RESEARCH FUNDAMENTALS', 'NIECer', '101', and 'nie.gov.in'. The main content is a bulleted list:

- There should be adequate justification to conduct the research study
- The research question should have clarity and focus
- Case definitions of study variables and outcomes should be standard and unambiguous
- Sample and sample size:
  - Should be representative of the population [External validity or generalizability]
  - Should be adequate [enough power to draw meaningful inferences]

As I mentioned earlier, planning is a very critical stage in research and some important considerations in this particular phase, which have to be kept in mind include, there should be adequate justification for conducting that particular research. Always please remember research involves investment. Investment in terms of money, investment in terms of manpower and hence, it is absolutely important that there should be adequate rational or justification for conducting any research study. For this to happen, the original research question should have total clarity and focus only then the study becomes a good study.

Another important point to be considered is that, the case definitions used for various study variables and outcomes should be standard and unambiguous. For example, if you are looking at cancer service then what we call as carcinoma in C2 and what we call as invasive cancer has to be known to everybody, who is involved in that particular research. For example, if you are collecting information on oral contraceptive pills use, what is considered to be as those women who are using oral contraceptives pills versus those, who are not, should be adequately clarified.

Another critical aspect that has to be kept in mind is the sample and a sample size. It has two dimensions to that, sample has to be qualitatively representative of the population in

which the study is being conducted because here this leads to what we call as external validity or generalizability. In simple terms, it means that we are able to just generalize the findings of our study which are based on a particular sample adequately and comfortably to the whole population from which the sample was drawn then we have served our purpose.

But this can also has to be complimented with adequate sample size as well because this is, we should have enough power to draw meaningful inferences and hence sample size should be appropriately decided for any kind of a research study.

(Refer Slide Time: 19:16)

Research can never be free of errors, but errors can be predicted and minimized

- Random error representing wrong result due to chance: unknown sources of variation that can distort findings
  - Can be minimized by increasing sample size and increasing precision
- Systematic error signifying wrong result due to bias - mostly due to variation that would distort the results in one direction .. Either over-estimation or under-estimation
  - Can be minimized by improving study design

National Institute of Epidemiology  
Chennai

HEALTH RESEARCH FUNDAMENTALS

NICer 101

nicer.gov.in

We always have to remember research can never be free from errors, but if you predict them well then they can be minimise through appropriate study design, this is an important aspect. Sometimes errors happen because no 2 human beings are equal, one person has a different haemoglobin level than the other, one person has a different height and a different weight than the other and hence, we look at many such people together there would be some kind of a variation which would naturally be happening in them and this is what we call as a random error, which anyway is going to be there in any kind of a research study.

How we minimize it is by taking a large sample size. If we take a large sample size this intra individual variation within a particular sample gets minimise to a large extent. But sometimes, there could be another type of error that could occur and the results could be due to something called as bias. And this is a variation due to some kind of distortion by some kind of a faulty procedure, it might be a measurement error, it is may be due to the kind of where the information is collected, it might be due to the process in which the participants are enrolled in the research study. So, one has to remember this kind of an error can only be minimised by improving the study design and taking care of that very adequately.

(Refer Slide Time: 20:41)

Challenges in designing and implementation of research studies

In a scenario when we desire to study the relationship between a variable and an outcome

- Confounders: Affect both the study variable as well as the outcome
  - Effect can be minimized by proper study design and through stratified analysis
- Effect modifiers: Can alter or distort the true relationship between the study variable and the outcome by independently affecting the outcome
  - Good to be aware of them through adequate literature review and not to include them in the study

National Institute of Epidemiology  
Chennai

HEALTH RESEARCH FUNDAMENTALS

NieCer 101

nie.gov.in

There will always be challenges in designing and implementing any kind of a research study. But the important thing that we have to understand is how we predict those, and one way to tackle this very appropriately is to do a thorough literature research before we do any kind of a study because it gives us an idea about two important say entities which can affect the result, they are called as Confounders and Effect modifiers.

Well, confounders are the entities which affect both the study variable as well as the study outcome and invariably, if you just leave out all those people having this confounding characteristics, it is quite possible that you may land up with a situation

where you do not have enough people to do research at all and in this particular scenario therefore, it is important that we understand which are the confounders. Collect the right kind of information on all the compounding factors because statistical analysis can take care of this confounding, which happens in a research setting. Effect modifiers are little difficult to deal with because they can alter or distort the true relationship between the study variable that we are looking at and outcome by independently effecting the outcome itself. So theoretically, it is a good idea to be aware of them and to also understand the kind of effect they are likely to make on the outcome variable, also one strategy that could be tried is to not to include people with effect modifiers in the study.

(Refer Slide Time: 22:16)

## Study methods and measurements: Major issues

- Pilot study
- Study participants: Inclusion and exclusion criteria, recruitment targets and strategies
- Data collection instruments
- Measurements tools and assay
- Plan for statistical analysis
- Quality control and assurance at all levels



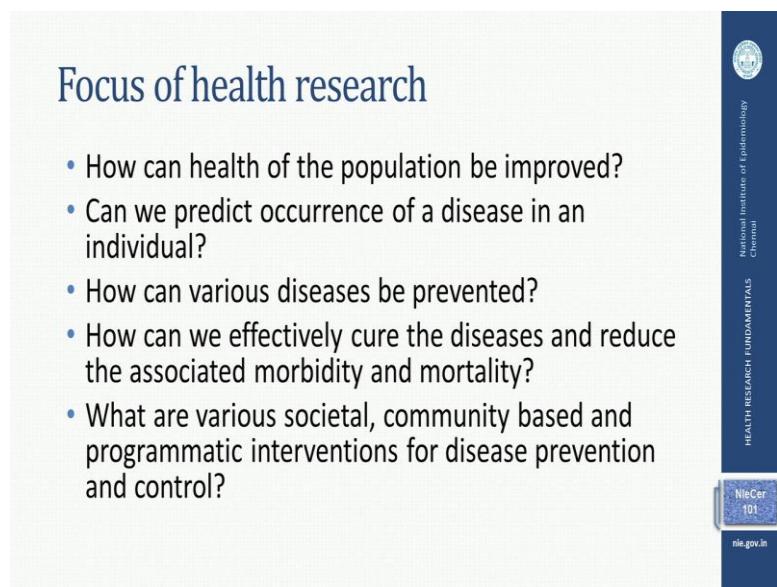
When we do any kind of a study, it can have multiple stages. It might be a good idea to do a pilot study. Pilot study just helps us to carry out all, what we want to do in the study in a say sort a few individuals this kind of becomes a rehearsal. It helps us to understand what kind of difficulties we are likely to face in collecting the information, carrying out certain procedures, doing the interviews, doing the inform consents, peoples understanding about informed consent forms and so on which can be changed.

There have to be certain issues related to study participants that need to be tackled very well, which are the inclusion criteria? And which are the exclusion criteria? One has to

be very specific about all of them. Why? Because, if we are very specific about inclusion and exclusion criteria, enrolment of people who are not eligible to participate in the study can be substantially minimized. Also every research study has to move with very specific research targets and it should follow certain strategies, like in some studies we might want to do the recruitments by going out in the community, whereas in some we might just want to do it based in the healthcare facilities, this decision has to be taken up front. Data collection instruments are really critical because if there are any mistakes the wave they have been designed we cannot make any kind of a change at a later stage. So, lot of work has to be done before hand in a drawing or in deciding the right kind of data collection tools.

Measurements, here I am referring to all those measurable items which are either using the laboratory methods or asses. They have to be properly standardized. They should have proper internal and external controls, here positive and negative controls for quality control. This lab also should probably be a part of an external quality assurance program because all this ensures quality control and assurance at all levels. It is also important that the plan for statistical analysis for any kind of a study has to be decided right up front, right in the beginning because that gives you a very clear idea about how we are going to collect this particular information and how the results of this particular study are going to look like eventually.

(Refer Slide Time: 24:36)



The slide has a light blue background with a dark blue vertical bar on the right side containing text and logos. At the top, the title 'Focus of health research' is displayed in a large, bold, dark blue font. Below the title is a bulleted list of six questions, also in a dark blue font. The dark blue vertical bar on the right contains the text 'HEALTH RESEARCH FUNDAMENTALS' at the top, followed by the National Institute of Epidemiology logo, the text 'National Institute of Epidemiology' and 'Chennai' stacked vertically, and the text 'NIECer 101' and 'nie.gov.in' at the bottom.

- How can health of the population be improved?
- Can we predict occurrence of a disease in an individual?
- How can various diseases be prevented?
- How can we effectively cure the diseases and reduce the associated morbidity and mortality?
- What are various societal, community based and programmatic interventions for disease prevention and control?

So when we are doing health research, we have to have some kind of an orientation towards either promoting behaviour or promoting health then preventing disease and also preventing mortality and then maybe. How can we do that? How can health of the population be improved? This is something like a direction that we have to think about. How can we predict occurrence of a disease in an individual? How can various diseases be prevented? How can we effectively cure the diseases and reduce the cost and also the associated morbidity and mortality? What are the various societal community based and programmatic interventions for disease prevention and control? These are the kinds of directions which health research tests and follows.

(Refer Slide Time: 25:24)

Health research aims at finding answers or practical solutions at individual and community levels

- At individual level
  - Promote healthy behavior, prevention at individual level, early diagnosis, adequate and appropriate treatment, rehabilitation
- At community level
  - Improve community behavior and practices, prevention and control programs, support to affected people, stigma reduction
- Healthy individuals build healthy nations!

National Institute of Epidemiology  
Chennai

HEALTH RESEARCH FUNDAMENTALS

NieCer 101

nie.gov.in

But basically what we try to do is to find answers or practical solutions at individual and community levels. For example, at community levels we have to see how healthy behaviour can be promoted? What can be the personal prevention achievable at an individual level? How can early diagnosis be achieved at an individual level and adequate an appropriate treatment be instituted? What kind of rehabilitation done at an individual level? But when we think of the community level we have to think of improving the community behaviour and practices, prevention and control programs as supporting the effected people or stigma reduction. All in all healthy individuals only can build healthy nations.

Thank you very much.

**Health Research Fundamentals**  
**Dr. Manickam Ponnaiah**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**Lecture - 02**  
**Formulating research question,**  
**hypothesis and objectives**

Hello, welcome to today's session of the Health Research Fundamentals course. In the introductory talk, Dr. Sanjay Mehendale talked about, the scope and focus of health research. He mentioned that the goal of research is to establish facts or principles through careful and systematic investigation in a particular area. Ultimately, the result of such investigation is to improve the health of the population. The first step in such research is formulating research question.

(Refer Slide Time: 00:51)

**Key areas**

- Spell out research question
- State research hypothesis
- Formulate objectives

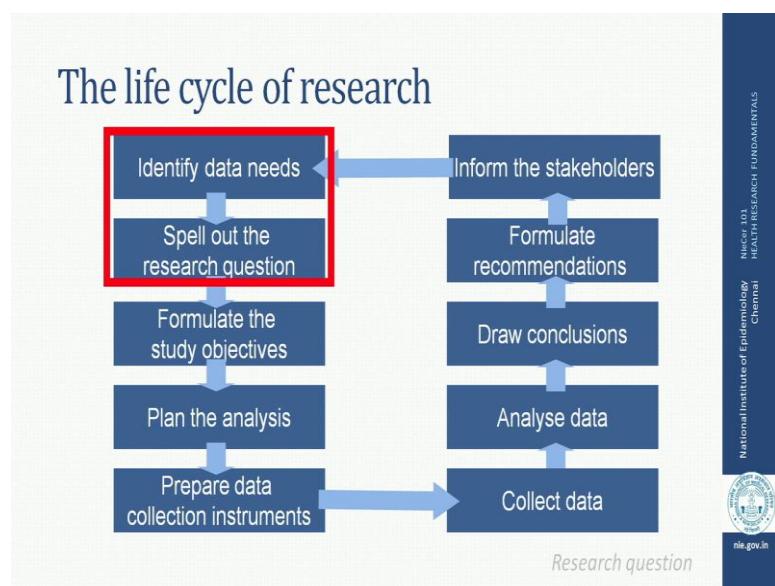
National Institute of Epidemiology  
Chennai



nie.gov.in

Today, we are going to cover 3 areas - spelling out research question, stating research hypothesis and formulating study objectives.

(Refer Slide Time: 01:07)



Any research has a life cycle. It begins with an uncertainty or needs in a particular area and that needs is translated into research question. Subsequently as study objectives and a plan of analysis is formulated to guide framing data collection instruments. Using these tools, data is collected and then subsequently is analyzed as per the plan and as per the objectives conclusions are drawn and recommendations are formulated. Ultimately, this particular recommendation is shared with the stakeholders for whom it matters. This process ends with another uncertainty that may begin the cycle once again. Therefore, the first two steps clearly indicate that we need to start with a good research question. Therefore, we are going to see what is research question?

(Refer Slide Time: 02:02)

## What is research question?

- ‘Uncertainty’ about something in the population that the investigator wants to resolve by making measurements in the study population
- Uncertainty = ‘data needs’
- Clear question facilitates to
  - Choose the most optimal design
  - Identify who should be included, what the outcomes should be, and when the outcomes need to be measured

Research question

NIECR IRI  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

Research question is an uncertainty about something in the population that the investigator wants to resolve by making measurements in the study population. The uncertainty is referred to as data needs in the life cycle. Therefore, we need a clear question to facilitate choosing an optimal study design and identify whom to be included, what are the outcomes that we need to measure and when these outcomes to be measured.

(Refer Slide Time: 02:34)

## Refining 'ideas' into research questions

- Begins with general uncertainty about a health issue
- Narrows down to a concrete, researchable issue

Research question is all about refining your ideas into systematic process of framing a question. It begins with the general uncertainty about a health issue in the context of health research and then it is narrowed down into a concrete researchable issue.

(Refer Slide Time: 02:54)

## Translating uncertainty to research question

- Frames problem in specific terms (clinical/public health/...)
- Focuses on one issue
- Is written in everyday language
- Can use more than one operational verb, if needed
- Should link the question to the potential action that would be taken once the question is answered
- *Is stated as a question!*

While translating uncertainty to research question, one frames the problem in specific

terms. In health research, it could be in clinical or public health terms. We need to focus on only one issue at a time and it is written in everyday language. So, that everybody understands what the question is. You may choose to use more than one operational work, if needed. It should link the question, if answered what action will be taken and it is stated as a question that is why it is called research question.

(Refer Slide Time: 03:29)

Research question sets out

- ✓ What the investigator wants to know
- ✗ NOT
- ✗ What the investigator might *do* or
- ✗ What the results of the study might ultimately *contribute* to that particular field of science

NINCH-IHL  
HEALTH RESEARCH FUNDAMENTALS

National Institute of Epidemiology  
Chennai

nie.gov.in

Research question

A research question sets out clearly what the investigator wants to know and definitely not what he or she may do or what this results will ultimately contribute to. It does not just include these two aspects; it should spell out what the investigator wants to know.

(Refer Slide Time: 03:51)

## Sources of research questions

- 1. Mastering the published literature**
  - Continue review of work of others in the area of interest
- 2. Being alert to new ideas and techniques**
  - Attending research meetings / conferences
  - Having a skeptical attitude about prevailing beliefs
  - Applying new technologies to old issues
- 3. Keeping the imagination roaming**
  - Careful observation; teaching, tenacity
- 4. Choosing a guide/mentor**

*Research question*

SB Hulley et al. Designing Clinical Research, 3<sup>rd</sup> ed. Lippincott Williams & Wilkins 2007

NATIONAL INSTITUTE OF EPIDEMIOLOGY  
CHENNAI  
nie.gov.in

There are many sources from which the resource questions or ideas can arise from. I am spelling out here four such sources. First is definitely a scholarship in the area of research interest, an up-to-date information from literature will help in generating research questions. The second is being allowed to new ideas and techniques, how does that happen? It can happen through attending research meetings or conferences, where the latest findings are shared and there may be a good discussion in peer group during the conference or meeting. Having a good attitude, skeptical attitude in particular, about the prevailing beliefs and last, but very important, applying new technologies to old issues.

The third aspect could come from careful observation in your clinic or in your basic science work, on your sphere of life. Teachers get enormous opportunities, while they are preparing for teaching and teaching and interacting with students and then finally, tenacity to go to the bottom of the things. Last, but very important, if you have a good guide or a mentor, he can help you in identifying and framing research questions.

(Refer Slide Time: 05:05)

Two categories of research questions

**1. Descriptive questions**

- Involve observations to measure quantity
- No comparison groups / interventions

**2. Analytical questions**

- Involve comparisons / interventions to test a hypothesis

Research question

NIECR IDI  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai  
nie.gov.in

There are two categories of research questions that we need to remember. There are only two categories let me put it that way. One is a descriptive question another is analytical question. So, your research question should fall into either of these categories. A descriptive question is one which involves observations to measure a quantity. When I say quantity, it could be height, it could be knowledge, it could be to what extent the problem is present in a community, it could be to what extent something is present in a given context. There are no comparison groups, no interventions in a question that is supposed to be called descriptive question.

Whereas analytical question involves comparison groups or it could be involving an intervention or experimental to test a specific hypothesis, so therefore, your research question when you are framing you need to find out, which of these two categories your research question may fall into. This has implications later, while we discuss about the statement of objectives and choosing study designs.

(Refer Slide Time: 06:20)



I am just going to give you some 6 steps as to how one can conceive a research question. One, a review of state-of-art information from the literature, second, you raise a question and the third you decide whether it is worth investigating through a peer review. Forth, by defining a measurable exposures and outcomes and fifth sharpening the initial question based on the above and defining the question by specifying details based on all of the steps that were explained earlier.

(Refer Slide Time: 06:54)

## Steps in conceiving a research question

e.g., Should diabetics do exercise daily?

### 1. Review of state-of-art information

- Exercise reduces blood sugar, body fat
- Exercise improves protection against developing diabetes related complications

Research question

NIECE I01  
National Institute of Epidemiology  
Chennai

nie.gov.in

Suppose, let us take an example, should diabetics do exercise daily? That is a very common sensitive question; one reviews a literature as to what is the effect of exercise on human body to begin with. Literature shows exercise reduces blood sugar level and body fat, exercise improves protection against developing complications due to diabetes. So, definitely it is worth investigating.

(Refer Slide Time: 07:24)

## Steps in conceiving a research question

### 2. Raise a question

- Can exercise help control blood sugar level?

Rather vague; Need to define

- 'exercise' & 'blood sugar level'

Research question

NIECE I01  
National Institute of Epidemiology  
Chennai

nie.gov.in

Then, let us raise a question based on our review of literature. Can exercise help control blood sugar level? Sounds better than the earlier question, but definitely this is vague. We need to define it especially, what do you mean by exercise? What do you mean by blood sugar level?

(Refer Slide Time: 07:42)

Steps in conceiving a research question

### 3. Decide worth investigating by peer-review

- What is the level of reduction in blood sugar?
  - Fasting or random or post-prandial <i.e., after food>
- What are optimal type, frequency, intensity and duration of exercise?
- What are the risks? What are the other benefits?

Research question

NATIONAL INSTITUTE OF ENVIRONMENTAL SCIENCE AND TECHNOLOGY  
NATIONAL INSTITUTE OF EPIDEMIOLOGY  
CHENNAI  
niehs.gov.in

Then we go in to the literature and talk to people, peer group and talk to investigators who was special expertise in this areas, what do you mean by blood sugar? What is the level of reduction? Is it blood sugar level after you do not eat at all or any time during the day, that is called random or after a meal that is called postprandial. So, which type of blood sugar gets reduced? Which of these three types of blood sugar gets reduced? Regarding exercise, a look at the literature suggests more questions, what is optimal type? What is the frequency? What is the intensity? What is the duration of such exercise? And are there any risks for a diabetic to be engaged in exercise? Or there other benefits other than a possible reduction in blood sugar?

(Refer Slide Time: 08:41)

## Steps in conceiving a research question

### 4. Define measurable exposures & outcomes

- **Exposure:** Exercise
  - *Pre-determined physical activity comprising of any body movement produced by skeletal muscle, resulting in an increase in energy expenditure*
  - *At least one session of 60 minutes every day for one year*
  - Could be specific: *walking, jogging or cycling or aerobic...*
- **Outcome:** Fasting blood sugar level

Research question

NIEHS IRI  
National Institute of Epidemiology  
Chennai  
  
nie.gov.in

The fourth step is to now, tighten up the framing of exposure and outcome. For example, here exercise could be defined as a predetermined physical activity comprising of a body movement produced by a muscles resulting in increased energy expenditure. At least one session of 60 minutes every day for one year, you can be specific about this exercise or physical activity by specifying whether it is walking, jogging, cycling or aerobic or even dancing. Finally outcome, now we have specified, it is fasting blood sugar level which means after the dinner anybody goes into sleep and the next day morning the stomach is empty after 8 hours. At that time, if you measure blood sugar that is called fasting blood sugar and we are defining that as our outcome.

(Refer Slide Time: 09:37)

## Steps in conceiving a research question

### 5. Sharpen the initial question

- Among diabetics, does physical activity for one hour daily help in reducing fasting blood sugar level?

Research question

NICER IODI  
National Institute of Epidemiology  
Chennai  
nie.gov.in

The fifth step therefore, is to sharpen the initial question with these steps in which we have progressed further. Among diabetics, does physical activity for one hour daily help in reducing fasting blood sugar. You can still refine it by specifying further details.

(Refer Slide Time: 09:59)

## Steps in conceiving a research question

### 6. Refine the question by specifying details

(Study population, operational definitions of variables and study design)

- What is extent of walking practiced by diabetics (type 2 diabetes) regularly? [Descriptive question]
- In order to improve management of type 2 diabetes, we wish to know whether brisk walking by diabetics for atleast one hour daily reduce fasting blood sugar level as compared to those who do not?  
[Analytical question]

Research question

NICER IODI  
National Institute of Epidemiology  
Chennai  
nie.gov.in

I have just given an example of such specifications in the form of descriptive question as

well as analytical question. When you are specifying the details, you need to specify study population, operational definitions of variables which are exposed and outcome and also if possible study design. An example of descriptive question in the example that we are discussing is that, what is the extent of walking practiced by diabetics that is type 2 diabetes regularly? Or an analytical question could be, in order improve management of type 2 diabetes, we wish to know whether brisk walking by diabetics for at least one hour daily reduce fasting blood sugar as compared to those who do not?

(Refer Slide Time: 10:47)

Good research question should pass  
the 'so what?' test

- Feasible
- Interesting
- Novel
- Ethical
- Relevant

Research question

SB Hulley et al. Designing Clinical Research, 3<sup>rd</sup> ed. Lippincott Williams & Wilkins 2007

NIEHSI  
National Institute of Epidemiology  
Chennai  
nie.gov.in

After framing a good research question, we need to test this question to a test called 'so what?' This test comprises of five elements, it is called FINER as an acronym. Is this research question feasible to answer? Is this interesting to answer? Is it novel? Is it ethical to do studies around this research question? Is it relevant?

(Refer Slide Time: 11:12)

Good research question should pass  
the 'so what?' test

- **F**easible
  - Adequate number of participants, technical expertise & resources
- **I**nteresting
- **N**ovel
  - Confirms, refutes or extends previous findings
  - Provides new information
- **E**thical
  - Amenable to a study that ethics committee will approve
- **R**elevant
  - Advance scientific knowledge, improve practice, influence policy

*Research question*

NIEHS IRI  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai  
  
nie.gov.in

So, basically feasibility means whether we will get adequate number of study participants. Do we have technical expertise to do this study? Do you have resources both material and manpower to do this? Is it interesting? Does it really enthuse people to engage in this particular research? Is it worth doing it? In terms of novelty, does it confirms, refutes or extends the previous findings? Or does it provide you information? That is the question that we are interested in answering and the forth test is ethical angle. Does this research based on the research question is allowable under the ethical norms? Will an ethics committee pass this research based on the question? Finally, is it relevant in terms of advancing science, advancing practice and also influencing policy?

(Refer Slide Time: 12:10)

**Statement of research hypothesis**

- A specific version of research question
  - Summarizes main elements of study
  - Establishes basis for test(s) of statistical significance
    - Main elements: *Sample, Exposures and Outcomes*
- Stated for analytical questions with comparison groups
  - For research questions with terms: *greater or less than, causes, leads to, compared with, more likely than, associated with, related to, similar to or correlated with*
- Purely descriptive questions DO NOT require hypothesis

Hypothesis

NIEHS IRI  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Environmental Health Sciences  
Chennai  
nie.gov.in

Now, let us look at hypothesis, what is research hypothesis? It is nothing, but a specific version of the research question that summarizes the main elements of the study that establishes the basis for statistical test of significance. So, it is stated for statistical purposes. The main elements that I mention include the sample, the exposures and outcomes. A hypothesis is stated only for analytical questions with comparison groups. Remember, we talked about 2 types of research questions, descriptive questions analytical questions. So, only the second type of research questions involving analytical aspect needs statement of hypothesis.

If you have any doubts about, what is your analytical question? You check in your research questions, if that contains terms such as greater or less than, causes, leads to, compared with, more likely than, associated with, related to, similar to or correlated with. If these terms are contained in your research question, this is an analytical question that needs a statement of hypothesis. Purely descriptive questions do not require a statement of hypothesis.

(Refer Slide Time: 13:37)

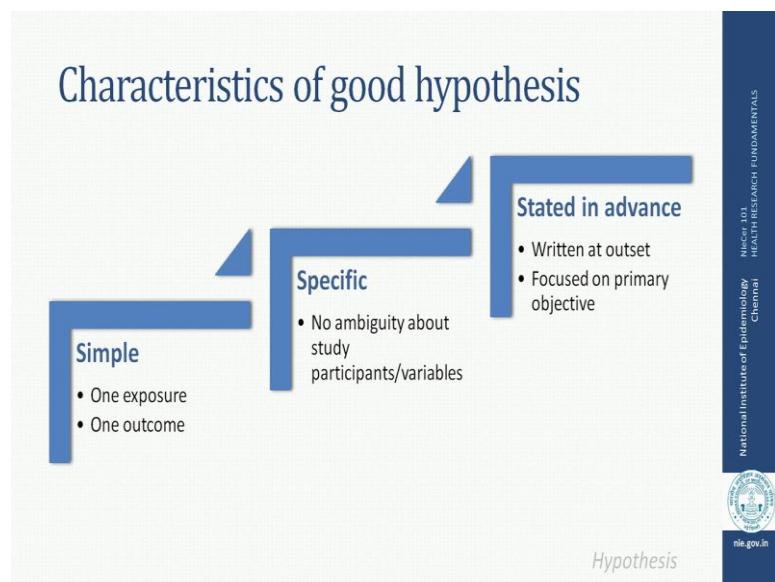
## Example of research hypothesis

*Among diabetics (type 2 diabetes) from the study area, who do brisk walking for atleast one hour daily results in average reduction of 10 mg% of fasting blood sugar level as compared to those who do not*

Hypothesis

An example of research hypothesis on the basis of example that we are discussing on diabetes and exercise, among diabetics type 2 diabetes, which means they do not take insulin injection to control the blood sugar level. From an area, who do brisk walking for at least one hour daily results in an average reduction of 10 milligram percentage of fasting blood sugar level as compared to those who do not, this specification of what much level? What study area? What group of study participants helps investigators in tightening the rope in terms of statement of a null hypothesis, statement of alternative hypothesis about which we will see in the sample size section and calculating sample size? So, research hypothesis helps in statement of specifying certain details in the context of statistical test and sample size.

(Refer Slide Time: 14:42)



What is a good hypothesis? A good hypothesis should be simple, should be specific and should be stated in advance. What do I mean by simple? It should be one exposure, one outcome. What do I mean by specific? There should be no ambiguity about the study variables or study participants. It should be stated in advance, *a priori* that is a terminology used. It should be stated in advance, it should not be discovered at a later part of the study and the hypothesis is focused around the primary objective.

Now, let us come back to the life cycle of research. We have now talked about research question and how research question is translated in to research hypothesis for analytical questions. Now, let us see how research question is converted into statement of objectives.

(Refer Slide Time: 15:32)

## Translating research questions to objectives

- Frame in scientific/epidemiological terms
- Take the question in a few limited axis
- Write in scientific/epidemiological language
- Make use of no more than one verb for each
- Sort as primary and secondary
- Be clear about the type of question:
  - Descriptive questions {Measuring a quantity}
  - Analytical/experimental questions {Testing a hypothesis}

*Objectives*

NICER IRI  
National Institute of Epidemiology  
Chennai  
  
nie.gov.in

When you translate research question into objective as compared to research question, a statement of objective is stated in scientific and epidemiological terms. It takes the research question in only few limited axis. It is written in scientific and epidemiological language. It should ideally make use of no more than an operational web for each of those questions. It is ideal to sort them out as a primary objective and secondary objective and the statement of objective should be very clear about what research question it is trying to answer. Is it descriptive question or an analytical or experimental question? So, it should clearly spell out, whether we are answering descriptive or analytical question.

(Refer Slide Time: 16:24)

## Objectives for descriptive vs. analytical studies

- **Descriptive:** Estimating a quantity
  - Use the verb “Estimate”
    - E.g., Estimate prevalence of physical activity
- **Analytical:** Testing a hypothesis
  - Use the verb “Determine”
    - E.g., Determine whether exercise reduces blood sugar level



An objective is again based on this understanding; whether it is descriptive analytical is stated as a descriptive objective or analytical objective. We recommend that the statement objective contains scientific and epidemiological terms. We also recommend that it uses the terms that denote the descriptive nature of the study and analytical nature of study. For instance, for a descriptive study in which, we expect that you estimate a quantity by observations, the verb estimate is preferred.

For example, to estimate prevalence of physical activity in diabetics, that is a descriptive objective. For analytical, you use the terms that are more powerful denotes this comparison, connotation. Testing hypothesis has to be denoted in the verb for example, the verb, determine may be preferred for example, in the diabetes and exercise related studies that we are discussing. Determine whether exercise reduces blood sugar level? As you can see, the verb is also equally important and therefore, we do not recommend use of the word study, to study in the statement objectives at all. To study is a very poor statement of objective.

(Refer Slide Time: 17:56)

## The research question

- In order to improve management of type 2 diabetes, we wish to know whether brisk walking by diabetics for atleast one hour daily reduces fasting blood sugar level as compared to those who do not?

## Primary objective

- Determine the effect of brisk walking for atleast one hour daily on fasting blood sugar level of patients with type 2 diabetes compared those who do not

Objectives

Now, let us go back to our example, the research question that we stated was, in order to improve management of type 2 diabetes, we wish to know whether brisk walking by diabetics for at least one hour daily reduces fasting blood sugar level as compared to those who do not? We can translate this into a primary objective to determine the effect of brisk walking for at least one hour daily on fasting blood sugar level of patients with type 2 diabetes compared to those who do not. So, determine is a very useful word especially when you have an analytical objective.

(Refer Slide Time: 18:33)

## Good and bad examples of study objectives

- Determine importance of sedentary lifestyle among diabetics
  - ✓ Estimate prevalence of physical activity among diabetics
- Assess physical activity and diabetic complications
  - ✓ Estimate effect of physical activity on the rate of diabetic complications
- Evaluate depression and diabetes
  - ✓ Determine whether depression is more common among diabetics as compared to healthy individuals

*Objectives*

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

Let me show you some of the good and bad examples of study objectives. Determine importance of sedentary lifestyle among diabetics. The word determines is not particularly suited for this particular statement because it seems to be descriptive studies. Therefore, estimate prevalence of physical activity among diabetics is ideally suited for this statement of objective.

Assess physical activity and diabetic complications. I think they are trying to do again a descriptive study. Therefore, estimating the effect of physical activity on the rate of diabetic complications in a group of diabetes is what they are doing. So, the word estimate is preferable. Evaluate depression and diabetes, it may be preferable since it is analytical study to use a word verb determine than evaluate, but the depression is more common among diabetics as compared to healthy individuals.

(Refer Slide Time: 19:31)

## Asking yourself the right question

- Two ways to deal with a poor or irrelevant research question:
  - Try to answer it
    - The answer may be of no use of anyone
    - There may be no answer...
  - Try to reframe it
- If your research question is wrong:
  - No good hard work will save your work
- If your research question is right:
  - You have an opportunity to do a good job

NICER I01  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai  
  
nie.gov.in

Let me sum up, asking yourself the right question may end up in answering right way. Two ways to deal with a poor or irrelevant research question, first you may answer it, but then the answer may be of no use or there may be no answer. Try to reframe it, if you have a poor question. If your research question is wrong, no good hard work will save your work. If your question is correct you have an opportunity to do a good job, all the best wishes to do a good job.

Thank you.

**Health Research Fundamentals**  
**Dr. P. Ganeshkumar**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**Lecture - 03**  
**Literature Review**

Hi, welcome to Health Research Fundamentals, nicer 101. I am Dr. Ganeshkumar from ICMR School of Public Health, National Institute of Epidemiology. This ascent is about literature review. Literature review is an important step in any health research because this is an important link between what is known and what is not known. Basically, research is a systematic investigative process to increase our existing knowledge about your concern subject of interest or to revise our existing knowledge. So, it may be basic or applied, where basic is increasing an existing knowledge, where applied is that applying this basic research to develop new processes, new products, new knowledge or throw your light over an unknown area.

In this context, literature review is an important step in any health research, where it is going to tell us, where it helps and aids us to tell us to guide us, what is already existing knowledge about your concern subject. So, here in this session the most important learning objectives are why we require a literature review? What is a literature review? How a literature review is performed? And what are the steps of the literature review and certain ethical concerns in a literature review?

(Refer Slide Time: 01:33)



Why we require a literature review? As we already explained in the definition of a research that, it gives important link between what is known and what is not known. So, for to know that we require this important step called literature review, where it saves lot of time in your research and secondly, it know the subject matter better. So, this throws what is our existing knowledge about the concerned subjects and it suggests new research topics and questions.

For example, how it saves you from work? When you see when you want to develop a questionnaire for a physical activity. So, you review a literature, you can come across an already existing standardized, regionalized questionnaire for physical activity. So, you can use that which you no need to spend lot of time to develop your new questionnaire to measure the physical activity of an individual. Of course, it paves away and throws the light of, to know the subject matter better and third by reading the existing articles you may find certain lacunae in the existing language, which makes you to carry out your new research. So, it will aid you to carry out a new question, it aids you to develop new research questions, it aids you to develop new methods of an existing known subject.

(Refer Slide Time: 03:07)

## Lit review : Not just a summary



### *Information seeking*

Scan the literature efficiently using manual or computerized methods to identify a set of potentially useful articles and books



### *Critical appraisal*

the ability to apply principles of analysis to identify those studies which are unbiased and valid.



Is it a just a summary? So, we are going to read, whatever it is known and we are going to summarize it and tell, when this is what it is there in the existing literature or existing evidence. Now, it is not just a summary, it is first step, it is basically an information seeking. It scans the literature efficiently using manual or a computerized methods to identify a set of potentially useful articles, say like how we usually do in literature review is that, the whatever kind of a resources, valid resources over an existing subject, we need to retrieve the information from that. So, it may be a text book, it may be a manuscript, it may be a published article or it may be a conference proceeding. So, from there we need to go through that and we need to retrieve the kind of information from that and second thing is that after we have collected all those things, what we basically do is, not just summarize it.

We need to critically appraise, whatever the article we have collected so far. So, the critical appraisal is a most important step in literature review, which is an ability to apply principles of analysis to identify which is useful for you. So, these will be useful for me, this will not be useful for me, this is valid for this concerned subject, this is not valid for my point of view. So, likewise you need to critically appraise and existing collected literature.

(Refer Slide Time: 04:38)

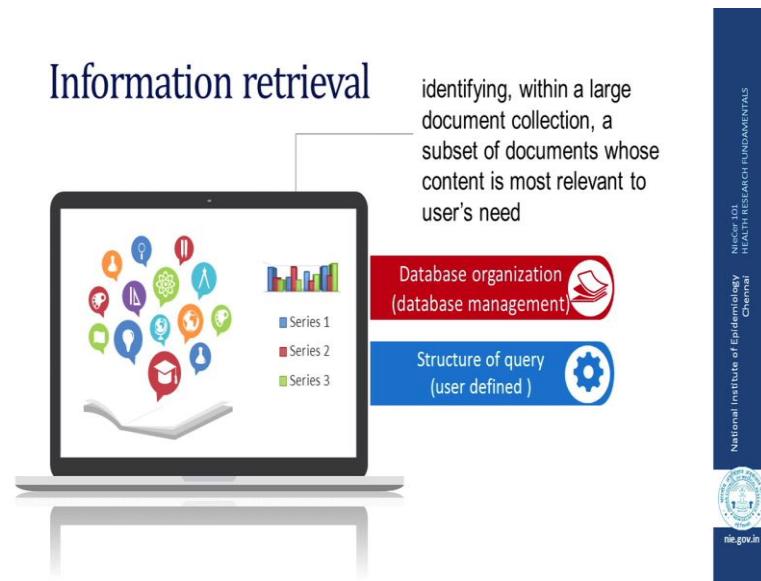


This is not just summary. So, what it is then? It is the organized research, how organized it is? So, organized research is a systematic process, say like the first important thing is that, you need to organize the information whatever you have collected so far and you have to relate it to the concerned research question which you are trying to develop or which you have already identify it and second thing is that it helps synthesize the results of it, from this existing whatever information which we have collected. So, in that collected information you have to summarize it, what is and what it is not known. This is all summarizing that is called synthesizing the results.

The third step is that you need to identify the lacunae, when you synthesize the results, when you organize or when you synthesize the results, the third thing there you can see is that, which is the lacunae here as appears in the literature; that means, for example, after when you are trying to review the existing diagnostic tools of tuberculosis and after you have reviewed that there are many tools are available and this is the different cast of the different diagnostic tools and these are the different work easy friendliness, quicker the results and you can identify that there is no quicker tool, which is not available for specific type of tuberculosis. Now, this gives you lacunae of existing literature. This may be helping you and this will lead you to identify a new research question or identify a new research path in it. So, that is how it is an organized research, it is not just

summarizing this existing articles and writing a kind of a paragraph.

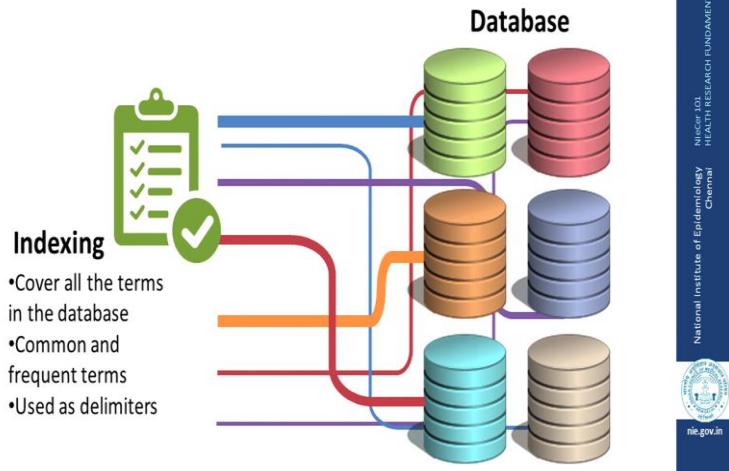
(Refer Slide Time: 06:23)



When you have seen that, you can see there is an information retrieval mechanism is inculcated in the literature review. Information literature, information retrieval in literature review what it is? It is basically identifying from a large database and from that by a set of queries, you are going to identify what are the documents which is a relevant for your needs. So, that is called information retrieval, that it happens from a big data base or it may be a huge database where all these articles are archived, collected or whichever is stored into that and from that you user by means of your set of query is seeking an information from this database. Now, this query system is user defined, it basically according to the need which a user is developing into that over the particular research question.

(Refer Slide Time: 07:26)

## Database structure and management



What actually is this database? And what actually is this database management? What is this a set of queries? So, you see here, all this information about a scientific knowledge in this information era is currently stored mostly in an electronized form. Even certain non electronized form is also stored in term of books or in terms of printed journals or conference proceedings. All these things are collectively called as a database. For example, it may be e-library, for example, it may be a collection of citations.

This data base will be huge, from this, like a last page of your text book, you can see index which helps you to identify what you are trying to search. Likewise, there are certain indexing mechanisms for this database from where you are going to collect required information by means of set of queries. Now, that is called indexing. This set of query is what, user is defined to retrieve the required information from this huge data bases.

(Refer Slide Time: 08:37)



Where we have to search? Which is appropriate place to search the required information?

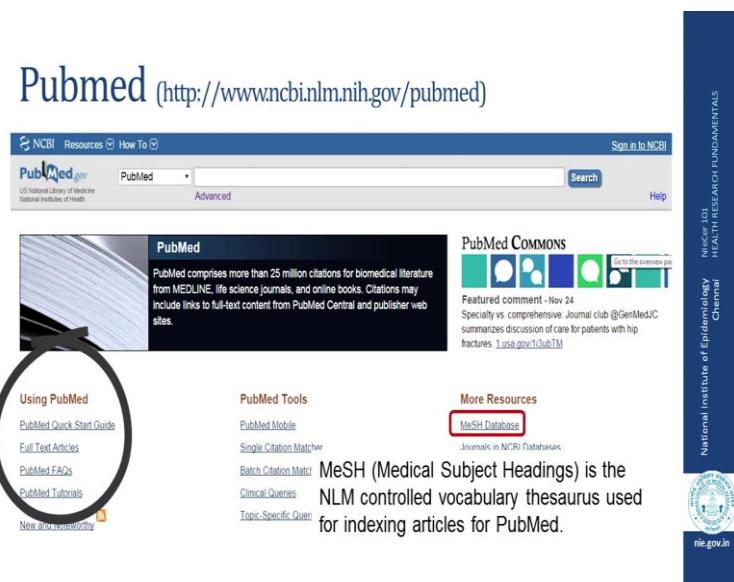
Where we need to search? It may be a general information about your specific health related event or a disease. For this, already we know that Google is a huge search engine, which has a huge database of the require information in it, but from all the Google results, we cannot rely entirely on that. So, there is a kind of an accreditation mechanism by a non-profit organization called Health on Net, this is called HON certified websites. This HON certified websites are those accredited websites, where you can get certain means of an authorized, certain means of reliable information about your specific health related events that is about general information, but by a professional, by a researcher.

What is a specifically when you want to retrieve an information, for example, I need to retrieve an information about all clinical trials about diagnostic test on tuberculosis. So, when I am going to do this kind of a specific query, this specific query has to happen over the specific scientific databases, which are PubMed, which is a huge database of scientific citations, biomedical citations. It may be a Scopus, which also have huge data base or even Google Scholar is also another URL, which is another portal where a huge collection of citations and abstracts across literature is available. So, from this specific query, you can retrieve the required article or the required manual script which you are trying to search.

Third, it may be from an archived full text articles, the archived full text articles are usually available in free open access, directories of journals or it may be from your fee based libraries. And this libraries may be e-library or may be a physical library, which is present where full text articles are there. From there, we can able to get the required information, which we are intended to research over a concerned topic and finally, you want to take the articles or you want to review those articles which is evidence based, which was the highest level of evidence currently available is a systematic review and meta analysis. So, from this collection of systematic reviews and meta analysis also you can retrieve that information, which is available in Cochrane library and even certain database also have these collection like Map of Medicine is also a collection of evidence based medicine articles in that portal.

These are the places, where you can do a query search specifically by an indexing mechanism and the required information can be retrieved from these databases. The most popular among all these things is PubMed. So, PubMed is a huge database, which has a 25 million citations is available and another commonly used portal is Cochrane library, which is a collection of systematic reviews and meta analysis.

(Refer Slide Time: 12:05)



This popular PubMed, let see about it, what is this PubMed? So, PubMed, which

comprises more than 25 million citations of biomedical literature. This PubMed is maintained by National Center for Biotechnology Information situated in US National Library for Medicine under National Institute of Health and this PubMed is free and open to access. Each and every PubMed, it is a collection of abstracts and full text articles. Each abstract has a link out resource, where the full text article is available and there are certain articles where you can freely access in PubMed, through a portal called PMC, which is called PubMed Central. So, through this PubMed Central portal you can see the free full text articles about your concerned topic of interest or your research question.

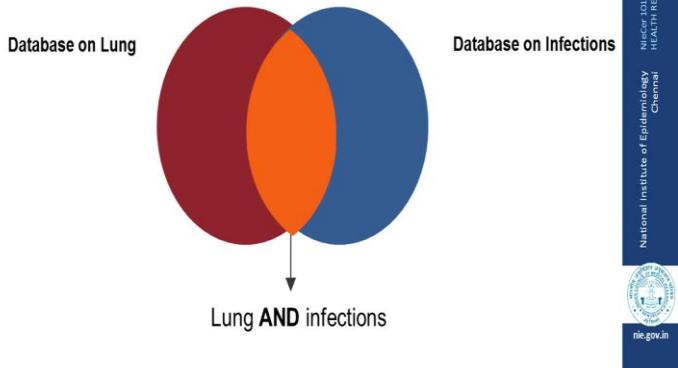
There is a good tutorial which is available in the landing page of the PubMed, where you can see in the portal itself. Apart from that, this US National Library of Medicine is maintaining their own set of defined-predefined vocabulary called MeSH, which is called Medical Subject Headings. Medical Subject Heading is a National Library of Medicine controlled vocabulary thesaurus and which is, you can see as called keywords under any abstract. If your keyword is in concordance with NLM defined vocabulary called MeSH, then your article has higher chances of getting identified by a set of systematic query mechanism. So, that is why MeSH terminologies are very important, which you can access the entire MeSH database in the left side of your portal, a PubMed landing portal and when you see in the PubMed portal, there you can have a tutorial which is available.

This entire tutorial explains you, that using PubMed, how to make a search in the PubMed? How to make a search through MeSH terminology? How to search through a single citation manager? And how to search through an author, year wise search, article wise search, study design wise search? So, it aids you this entire tutorial, which has a quick tour video tutorial, is also available in the landing portal and when you access it, it gives you a good detailed form of how to perform a research in PubMed portal.

(Refer Slide Time: 14:46)

## Searching a database

- Boolean query: AND

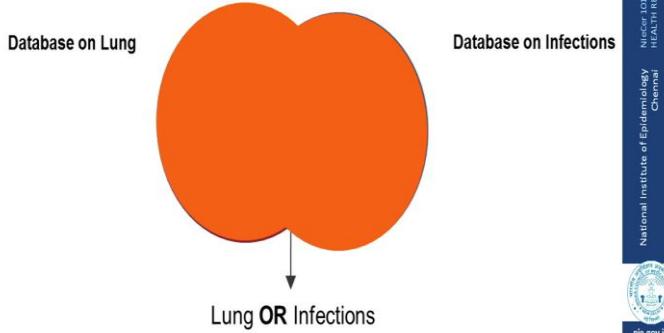


Now, I will explain you, basic search in your data base. How to do a basic searching in a database and this kind of search is called Boolean query, which is very common in any search engine, it may be web search engine, it may be Google or it may be Embase, it may be a PubMed. This Boolean query is standardized one, which uses this most important connections called and, or, not queries. In these Boolean queries, I will just give demonstration about, how actually when you are putting a query and how the information is retrieved. For example, here I am showing you two different database, one database is about lung another database is about infections, which encompass all infections and if I am putting a query, a Boolean query of lung and infections and how the query retrieve the result. It retrieves the result of lung and infection, which is those articles, those items which are specific to lung and infections.

(Refer Slide Time: 15:52)

## Searching a database

- Boolean query: /OR/

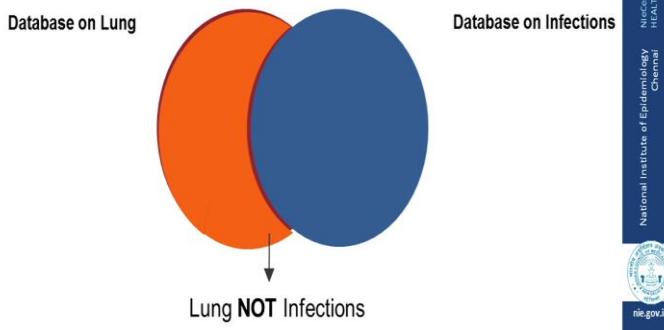


Whereas, when I am putting a query about lung or infections, what it gathers is, it gathers information about the entire items on lung, entire items on infections and even the items which is common that is lung and infections, which is lung infections. So, we get a huge search of results, when you are putting a query called lung or infections.

(Refer Slide Time: 16:16)

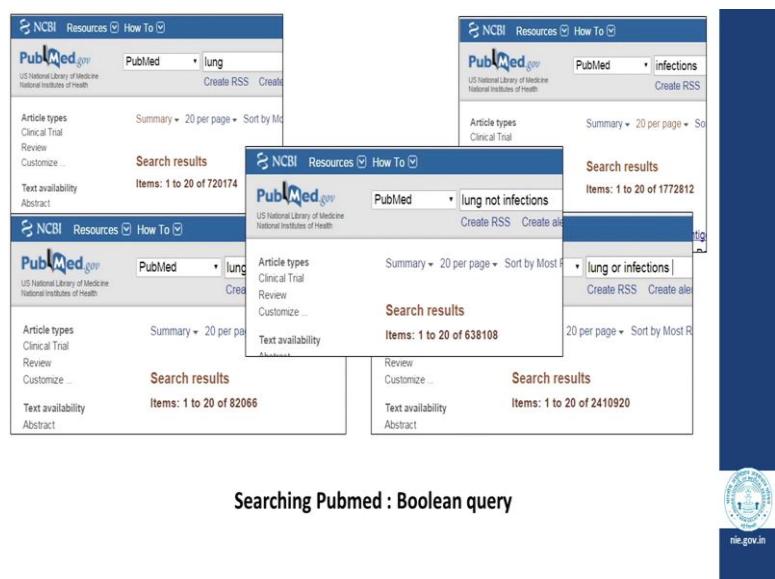
## Searching a database

- Boolean query: NOT



When you put query called lung not infections, so in this lung not infections, it totally remove this entire information related to infections from the whole search. Again, the infections related articles and items from the lung. So, what you get is results are the items, which is related to all items, which is related to lung except infections. So, this is how, this Boolean query works and, or, not.

(Refer Slide Time: 16:51)

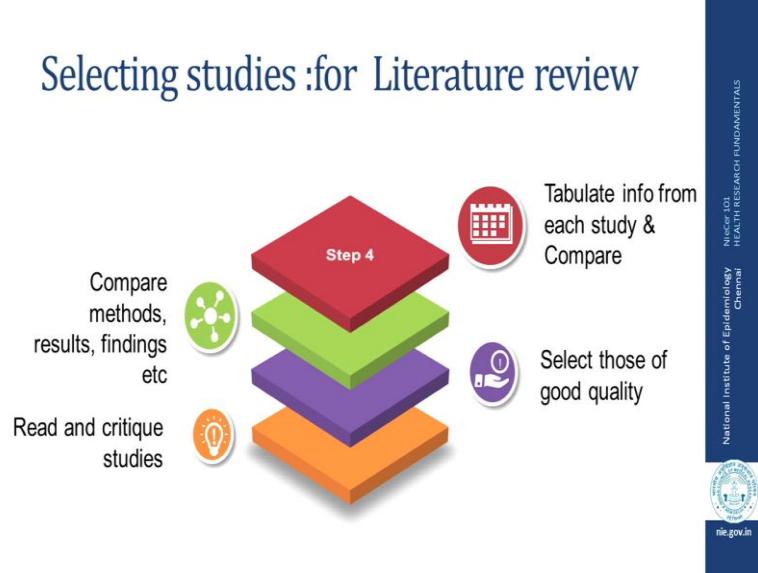


I will give a basic demonstration, similar to that how I did a Boolean query in PubMed. So, here you see, these are all the screen shots which I have cropped. When you see, when I made a Boolean query of only lung, what are the total numbers of items, which I have retrieved? The total number of items I have retrieved is nearly 7.2 lakhs. Only infections, the total number of items, which I have retrieved is 17.7 lakhs. When I put a Boolean query of lung and infections, it resulted me, only 82000. So, 82000 articles are related to only lung and infections. So, it removes all other lung related, all other infection related and it resulted specifically to lung and infections.

Let say, this gives another screen shot, when I make a search of Boolean query of lung or infections, here you can see, where we get information of entire information on lung, all the items about infections and all the items related to lung and infections. So, total is that nearly 24 lakhs. So, that is how depends upon your Boolean query, say like lung not

infections. In lung not infections, it removed all the infections related items and resulted only say like nearly 6.3 lakhs, which is only article related to lung except infections. So, this small demonstration which gives you an idea about how a Boolean query works; however, this is not a structured or a systematic way how to do a PubMed search. This session is not going to cover entirely, how to do MeSH search and all. So, the tutorials are already available in the PubMed dot com landing page. This is how a Boolean query works in any search engine including PubMed.

(Refer Slide Time: 18:53)



Now, what are the steps, how we have to do a literature review? Already, we know that after we have organized information is that, which are the studies is related to you? How we are selecting the studies? The step one is that by means of your Boolean query or by a structured query system, you have collected your specific articles related to do. For example, you are searching a Boolean query of same lung and infections; I want only RCT's about lung and infections, which is published in the past 5 years. Now, I have organized everything. So, what you need to do is that after you have organized it you need to read all those articles and critically appraisal it and how to do this, will be explaining you at the second step.

The second step is that, you have to select which is of those good qualities. Remember,

here this critical appraisal should be scholarly, it should not be too critical. It should be scholarly, for example, an article is explaining about the randomized controlled trial about a therapeutic regimen over lung infections and which you have seen is that these RCT's which you have collected is addressing over only the western population and you have this lacunae. So, what you are going to critical appraisal is that the existing evidence is available over the western population and Asian population is not available and Indian population is not available. Those articles whichever is reported about Indian populations are only quasi experimental studies, here it is scholarly critically appraisal.

You are not too critical; you are not finding a fault on it. Now, when you do a scholarly critical appraisal on it, you will identify your new area of research, you will identify what is existing lacunae, existing gap, which you can try to fill up by doing your new research. Third step is that, you need to compare the methods, results, findings of all those articles which you have organizing it. When you compare it, you can understand, what is an existing knowledge, about the different methodologies they have conducted in this particular topic?

So, that throws you a light, more about what is the methodology they have followed? So, far, in those kinds of your search query, for example, here when you want to do therapeutic regiments on lung infections, what are all the methodologies they have applied, while they want to test your therapeutic regimen for a lung infection? Likewise final step is that, you need to tabulate in your form, each study which has to be tabulated when you tabulate it, it helps you to organize. Number 2, it helps you to compare. Number 3, it helps you to compare the individual studies by itself and itself with other studies and it will give you good idea about and it will helps you to critically appraisal it. This is how, your model table looks like this.

(Refer Slide Time: 21:53)

### Example table of literature

Citation	Design	Objectives	Study population	Sample size	Measurable outcome & results	Authors Conclusion

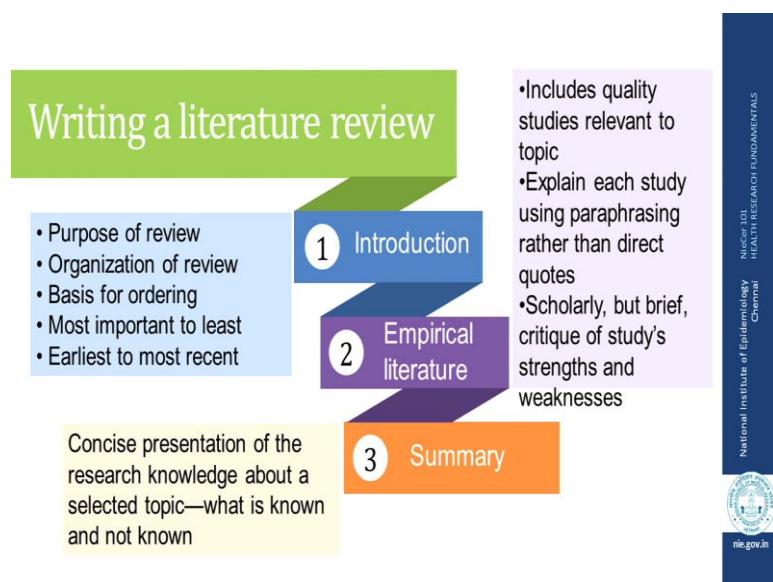
Chronologically placed  
RECENT

PAST



An example table of literature, where you can see, you can put up, add and delete your required columns. I made an example table; you can see citation, design, objectives, study population, sample size. What are the measurable outcomes? What are the main findings? And what are the authors' conclusions? You can organize it in kind of your table, which gives you, helps you to critically appraise it and it is better when you organize this table chronologically, which is from recent to past. So, your recent studies can be tabled, can be made in their first initial rows. It helps you to select those recent evidences or recent existing knowledge about it.

(Refer Slide Time: 22:37)



So, after you have organized all these, since you have tabulated it. Now, you need to write this literature review. In the literature review, in the write up part, these are the other three important parts. The first part is introduction, when you are writing a literature review, in the introduction; it gives you what is the purpose of the review? Why you are doing this? And how you have organized this review organization of the review? How you made the queries? How you have collected this information? And how you have organized it? That explains and basis of the ordering, most important to least and as well as what is the earliest to recent or recent to the earliest? How you have organized it? It gives you an introduction, how you have written? How you did this literature review?

Second is that, where you are actually writing all your existing information called empirical literature. So, it includes the quality studies, which is relevant to your research question or the topic and it explains each study about it is positive and in paraphrasing manner and your critical appraisal should be a scholarly way. So, that explaining it to that, these are all the identified lacunae, these are all the strengths, these are the novel methods. This you need to write in this part. Finally, you need to summarize your existing this entire literature review and when you summarize it, it should be a concised way, where you are going to tell that, this is about your existing knowledge about your research question or your concerned topic and this is known and this is not known and

this is lacunae of an existing literature. So, this is how you have to end a literature review right up and there are certain ethical issues, which we need to concern when you are doing a literature review.

(Refer Slide Time: 24:23)



Mostly, this problem is that when you are retrieving information from your specific manuscript or an article, the content from the study should be presented honestly. It should not be distorted, that you should not be read in between lines and taking only a part of your result as such. That is a most important thing. Second is that, any weakness of the study, again I am emphasizing it, that it should be very scholarly. You should not be too critical and it has to be addressed in a research point of view, that is scholarly point of view and finally, the source should be accurately documented.

There is a way how a source should be documented. There are different styles that are available, a very popularly followed style is Vancouver style, even Harvard style is available and based on that it has to be cited accordingly those sources. Finally, this is how a literature review can be done and this gives you a basic view about, what is a literature review? Why this literature review has to be done? What are the steps in doing this literature review? And how you can write a literature review?

Thank you very much.

**Health Research Fundamentals**  
**Dr. R. Ramakrishnan**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**Lecture - 04**  
**Measurement of Disease Frequency**

Welcome to this session of Health Research Fundamentals. In this session, we are going to see some measurements we commonly use to measure the Disease Frequency.

(Refer Slide Time: 00:20)

**Population at risk**

- Portion of a population that is susceptible to a disease
- Can be defined on the basis of demographic or environmental factors

NICER IITM  
HEALTH RESEARCH FUNDAMENTALS

National Institute of Epidemiology  
Chennai



nicer.gov.in

Before we go in for appropriate measurement of disease frequency, we need to understand certain concepts like a Population at Risk. The population at risk is the portion of population that is susceptible to a disease. That can be defined on the basis of demographic or environmental factors.

(Refer Slide Time: 00:49)

## Population at risk: Examples

- Population at risk of developing carcinoma of the cervix:
  - Female population
  - Age > 30 and < 70 years
- Population at risk of hepatitis B
  - Those individuals anti-HBc negative

NICER ICD  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

Now, let us look at some example. The population at risk of developing carcinoma of cervix is female population and in the age group of 30 to 70 years. Population at risk of hepatitis B is those individuals were at risk of developing hepatitis B, but were negative.

(Refer Slide Time: 01:14)

## Prevalence – (P)

- Number of existing cases (old and new) in a defined population at a specified point of time  
$$P = \frac{\text{# people with disease at a specified time}}{\text{Population at risk at the specified time}} \times 10^n$$
- In some studies the total population is used as an approximation if data on population at risk is not available

NICER ICD  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

Now, let us look at a measure which is very often used in epidemiology in any health

research. At (Refer Time: 01:24) prevalence, what you mean by prevalence? The prevalence is the number of existing cases both old and new, in a defined population at a specified point of time. P is equal to is the number of people with disease at specified time and that is divided by the population at risk at this specified time and that quantity is multiplied by a factor by 10 or 10 power n.

In order to make that as, suppose whenever your population at risk is very large and you have number of cases were small, you will get a value P as 0.001. So, in order to make into a round number you multiplied by 1000, 10000 or 100000 depending on what the value of P you get. In some studies, the total population used as in approximation if data on population at risk is not available because with the philosophy that everyone in the population are at risk of developing a particular disease.

(Refer Slide Time: 02:39)

## Point prevalence

- Measures the frequency of disease at a given point in time
- Applies when the data has been collected at one point in time
- $P = C / N$ 
  - C = # of observed cases at time 't'
  - N = Population size at time 't'

NATIONAL INSTITUTE OF  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai  
  
nier.gov.in

Again, this prevalence are broadly classified into say 2 different categories. We called as one as a Point Prevalence and another one as a Period Prevalence. By point prevalence what we mean is, this measures the frequency of disease at a given point of time, it is like a snapshot. This applies when the data has been collected at one point in time. It is denoted as P is equal to C by N, where C is the number of observed cases at that particular point of time t and N is the population size at time point t.

(Refer Slide Time: 03:21)

## Example of point prevalence

- 150 children in a school
- Screening for refractory errors at time “t”
- 15 children require glasses
- Prevalence of refractory errors
  - $15 / 150 = 10\%$

NICER IRI  
HEALTH RESEARCH FUNDAMENTALS  
Chennai



nie.gov.in

The example of point prevalence is supposed, if there are 150 children in a school and you are screening then for refractory errors, at a particular point of time t. And you find 15 children they require glasses, they have problem. Then the prevalence of refractory errors is 15 divided by 150 which is equal to 10 and called a 10 percent of the school children they have refractory errors or the point prevalence of refractory error in this particular school is 10 percent.

(Refer Slide Time: 03:56)

## Period prevalence - (PP)

- Measures the frequency of disease over some time
- Applies when the data has been collected over a period of time
- $PP = C + I / N$ 
  - C = # of prevalent cases at the beginning of the time period
  - I = # of incident cases that develop during the period
  - N = size of the population for this same time period

NICER 101  
HEALTH RESEARCH FUNDAMENTALS

National Institute of Epidemiology  
Chennai



nie.gov.in

The Period Prevalence is that measures the frequency of disease over a period of time. This applies when the data has been collected over a period of time and it is denoted as PP, which is equal to C plus I divided by N. What is C? C is the number of prevalent cases at the beginning of the time period and I is the incident cases, that is the new cases that develop during the period of your survey and sum of these two are divided by which is the size of the population for this same time period point.

(Refer Slide Time: 04:39)

## Exercise

- Scenario
  - Population of 150 persons
  - Follow up for one year
  - 25 had a disease of interest at the beginning
  - Another 15 new cases developed during the year
- Calculate:
  - Point prevalence at the start of the period
  - Period prevalence for the year

$P = C/N = 25 / 150 = 0.17 \text{ (17 \%)}$

$PP = (C+I)/N = (25+15)/150 = 0.27 \text{ (27 \%)}$

NICER IRI  
HEALTH RESEARCH FUNDAMENTALS

National Institute of Epidemiology  
Chennai



nie.gov.in

Example of this period prevalence is you know, you take a scenario of a population of 150 percent and you follow up for 1 year. At the beginning of your survey the 25 had disease of interest and another 15 new cases developed during the year. When we calculate it point prevalence at the start of the period and period prevalence over the period of time. Point prevalence is given by C by N there is 25 by 150 or 0.17 or 17 percent. Period prevalence is 25 plus 15 there is 40 over 150 it comes 0.27, that is 27 percent.

(Refer Slide Time: 05:21)

## Factors influencing prevalence

- Number of new cases
- Duration of the illness
  - If the disease is short, the prevalence is reduced
    - The prevalence of sudden infant death = 0
  - If the disease is long, the prevalence is increased
    - Rare lifelong disease can accumulate to build up a large prevalence

NICER IRI  
HEALTH RESEARCH FUNDAMENTALS



nie.gov.in

So, now there are several factors that influence the value of prevalence, the number of new cases, the duration of illness. Suppose, if the duration of illness is short, the prevalence is reduced. Say suppose, somebody gets disease and immediately you know either a cures or it dies then it goes out of our calculation, with thus when we go there are no disease persons. So, the prevalent of sudden infant deaths cut by times would be even 0 because when you go if there are no deaths that are there, all the children who had the disease had died. So, there are no cases at that particular point of time, so your prevalence may be 0.

And if disease is very long duration, the prevalence you know it goes suppose it in chronic deceases, a rare lifelong diseases it can accumulate to build up a very large prevalence.

(Refer Slide Time: 06:15)

## Causes of increase and decrease of prevalence

<u>Increase</u>	<u>Decrease</u>
<ul style="list-style-type: none"><li>• Long duration<ul style="list-style-type: none"><li>• Low cure rate</li><li>• Low case fatality</li></ul></li><li>• Increase in new cases</li><li>• Immigration of patients</li><li>• Improved detection</li><li>• Emigration of healthy people</li></ul>	<ul style="list-style-type: none"><li>• Shorter duration<ul style="list-style-type: none"><li>• High cure rate</li><li>• High case fatality</li></ul></li><li>• Decrease in new cases</li><li>• Emigration of patients</li><li>• Improved cure rate</li><li>• Immigration of healthy people</li></ul>

**Conclusion:** Changes in prevalence may have many causes and are difficult to interpret

NINCH IRI  
National Institute of Epidemiology  
Chennai  
  
nie.gov.in

We have to be careful, what causes an increase or decrease in prevalence? An increase in prevalence could be caused by long duration, just low cure rate or low case fatality. And there are more number of new cases that have come back or there are some immigration population patients who have with particular disease, if they immigrate into a particular area they could rather enhance the prevalence. Prevalence could also be increase, if there new improved detection mechanism you try to you know detect more cases because you have more sensitive test in your hand.

And prevalence could also increase, if there are healthy people going out of a particular region. So that your denominator is low and your numerator is all unhealthy or disease people are there and so your prevalence may increase. The decrease in the prevalence could happen exactly you know the opposite causes, the shorter duration, high cure rate, high case fatality, the decrease in new cases, immigration of patient, improved cure rates and immigration of healthy people. All these could bring down the prevalence. So, to conclude the changes in prevalence may have many causes and are difficult to interpret. So better we need to have a checklist of all these items and then look at them all before we try to rather say that the prevalence has increased or decreased over a period of time in a particular region.

(Refer Slide Time: 07:56)

## Uses of prevalence data

- Assessing health care needs
- Planning health services
- Measure occurrence of conditions with gradual onset
- Study chronic diseases

NICER IODI  
HEALTH RESEARCH FUNDAMENTALS  
Chennai



nie.gov.in

What are the uses of prevalence data? Prevalence of data is used to assess health care needs. It is very useful in planning health services because it measures the burden of disease, and it measures occurrence of conditions with gradual onset and prevalence is very useful in the study of chronic diseases.

(Refer Slide Time: 08:18)

## Incidence – (I)

- Number of new cases in a given period in a specified population
  - Time, (i.e., day, month, year) must be specified
- Measures the rapidity with which new cases are occurring in a population
- Can be expressed:
  - In absolute numbers
  - In terms of cumulated incidence
  - In terms of incidence density

NICER IODI  
HEALTH RESEARCH FUNDAMENTALS  
Chennai



nie.gov.in

Another important measure in epidemiology is the Incidence. The incidence is defined as the number of new cases in a given period in a specified population that is time is an important component, whether measure it by day or month or year, that must be specified when you are mentioning the incidence. This measures the rapidity with which or the speed of occurrence with which new cases are occurring in a population. This can be expressed in absolute numbers, in terms of cumulated incidence or in terms of incidence density.

(Refer Slide Time: 09:00)

### Cumulated incidence - (CI)

$\frac{\# \text{ of new cases}}{\text{Population at risk at the beginning}} \times 10^n$

- Also known as:
  - Attack rate
- Assumes that the entire population at risk at the beginning was followed-up for the time period of observation

NATIONAL INSTITUTE OF  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai  
  
nier.gov.in

Let us rather look at the different aspects of the incidence. Let us rather take the cumulated incidence first. The cumulative incidence is CI, is a number of new cases divided by population at risk at the beginning that is multiplied by a factor of 10. This is known as attack rate and it assumes that the entire population at risk at the beginning was followed-up for the time period of observation.

(Refer Slide Time: 09:39)

## Risk

- Probability that an individual will experience a health status change over a specified follow-up period
- This assumes that the individual does not:
  - Have disease at the beginning
  - Die from other causes during follow up
- Corresponds to cumulated incidence

NICER IED  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

Now, let us look at what do you mean by Risk? The risk is the probability that an individual will experience a health status or change over a specified follow-up period. This assumes that the individual does not have disease at the beginning and die from other causes during follow-up. This corresponds to a cumulated incidence.

(Refer Slide Time: 10:01)

## Incidence density - (ID)

# of new cases

$$ID = \frac{\# \text{ of new cases}}{\text{Total person-time of observation}} \times 10^n$$

Total person-time of observation

- Also known as:
  - Incidence rate
- Reflects more exactly the person-time observed

NICER IED  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

Now, let us look at the other measure which is called incidence density, denoted as ID. The incidence density is the number of new cases divided by total person-time of observation and that is multiplied by a factor of 10. This is also known as incidence rate. This reflects more exactly the person-time observed.

(Refer Slide Time: 10:25)

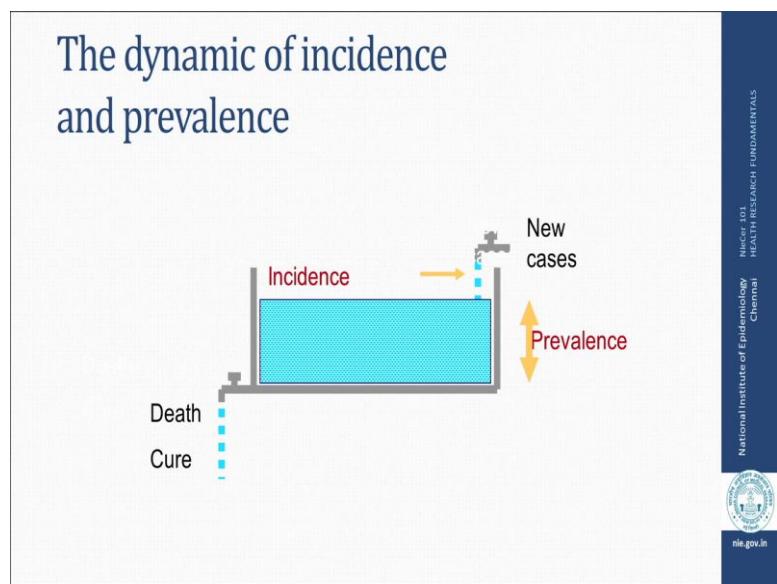
## Uses of incidence data

- Describe trends in diseases
- Evaluate impact of primary prevention programmes

National Institute of Epidemiology  
NIECHI, IIT  
HEALTH RESEARCH FUNDAMENTALS  
Chennai  
  
nie.gov.in

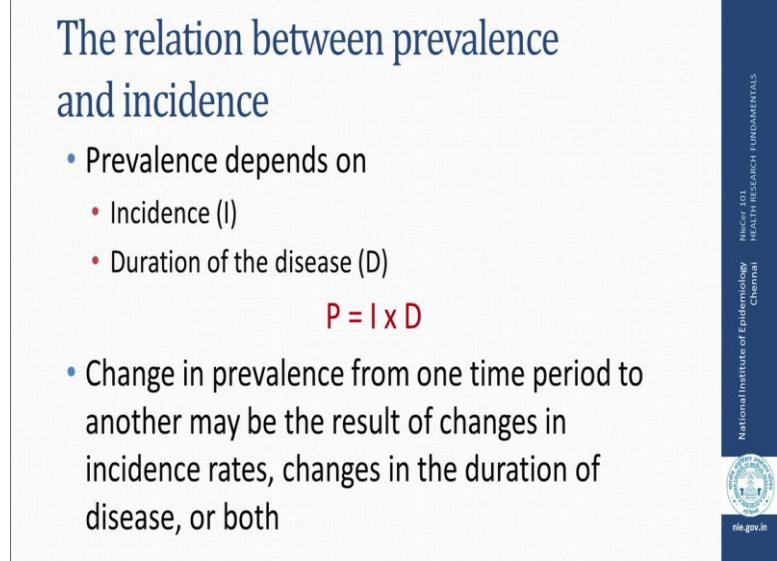
Uses of this incidence data is, this incidence data describes trends in diseases, whether they trend over a period of time, how the particular disease changes? And it evaluates impact of primary prevention programs.

(Refer Slide Time: 10:44)



The dynamic of incidence and prevalence can be depicted with a diagram like this see. There are new cases pouring in, there are cases, there are go in and going out because they are dying or they are getting cure. Incidence cases are the new cases and the cases which are remaining in the tap are the prevalence cases.

(Refer Slide Time: 11:13)



The relationship between the prevalence and incidence could be the prevalence depends on the incidence and the duration of disease. And it is denoted as prevalence is equal to incidence into duration, P is equal to I into D. Change in prevalence from one time period to another may be the result of changes in the incidence rates, changes in the duration of disease or could be both.

(Refer Slide Time: 11:43)

## Patterns of incidence and prevalence

- High prevalence and low incidence
  - e.g., Diabetes Mellitus
- Low prevalence and high incidence
  - e.g., Common cold

National Institute of Epidemiology  
HEALTH RESEARCH FOUNDATION  
Chennai



nie.gov.in

Now, let us look at the pattern of incidence and prevalence. High prevalence and low incidence, there are disease like Diabetes Mellitus. Low prevalence and high incidence are the examples are common cold.

(Refer Slide Time: 11:57)

## Case fatality

- Place in relation the number of deaths from a disease to the number of cases
- Reflects severity
- Can be expressed as:
  - Proportion
  - Ratio
- Not as rate (Although often referred to as case fatality rate)

NIEC/IIT  
HEALTH RESEARCH FUNDAMENTALS



nie.gov.in

Now, let us define a term called say the Case Fatality. In case fatality the place in relation to the number of deaths from a disease to the number of cases, how many cases you have? And how many of them they died? It reflects the severity of the case. This can be expressed as a proportion or a ratio not as a rate, though it often referred to as case fatality rate.

(Refer Slide Time: 12:30)

## Summary

- Prevalence is a static measure taken at a point in time
- Incidence is a dynamic measure taken over a certain time
- Mortality is calculated using population denominators to reflect burden while case fatality is calculated using cases as denominators to reflect severity

NIEC/IIT  
HEALTH RESEARCH FUNDAMENTALS



nie.gov.in

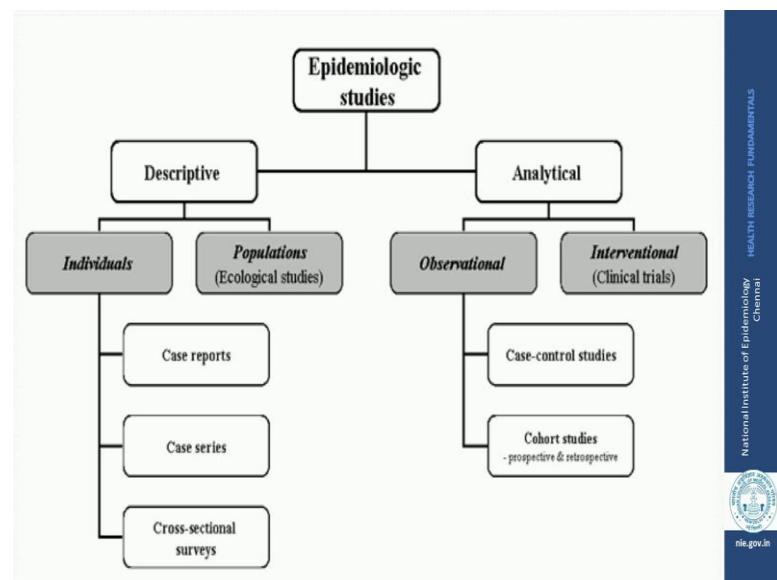
So, let us summarize these basic disease frequencies that we are finishing, one is the prevalence and prevalence is a static measure taken at a point in time or over a period of time. If you take it at a point in time it is called Point Prevalence, over a period of time it is called the Period Prevalence. Incidence is a dynamic measure taken over a certain time and the mortality is calculated using population denominators to reflect the burden, while the case fatality is calculated using cases as denominators to reflect severity. These are all the measures of a disease frequency.

**Health Research Fundamentals**  
**Dr. Prabhdeep Kaur**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**Lecture - 05**  
**Descriptive Study designs**

Welcome to this session of Health Research Fundamentals. Today, we will be discussing about Descriptive Study Designs.

(Refer Slide Time: 00:13)



Before we go on to the descriptive study designs, let me walk you through an outline of what are the various designs in the health research. So, as you can see in this slide broadly, the health research studies or the epidemiological studies can be divided into 2 categories; Descriptive studies and Analytical studies. Further, the descriptive studies can be of 2 kinds; one is the studies which are done for individuals and other is the studies which are done for populations. The studies which are done for populations are called Ecological studies and there are various types of studies that can be done at the individual level that include, Case reports, Case series and Cross-sectional surveys.

Now, in this lecture we will not be talking about the analytical study designs which are the other kind of studies, which will be covered in the different lectures. Now, within analytical, the other lectures will walk you through thus observational study designs,

such as Case-control studies and Cohort studies.

(Refer Slide Time: 01:20)

## Types of descriptive studies

- Case reports
- Case series
- Ecological studies
- Cross-sectional study

Now, coming to the descriptive studies, in today's lecture, we will be discussing about the 4 main types of descriptive studies that are, Case reports, Case series, Ecological studies and Cross-sectional studies.

(Refer Slide Time: 01:36)

## Case reports

- Detailed presentation of a single case
  - New or unfamiliar diseases
  - Rare manifestations
  - Generate hypothesis regarding pathophysiological mechanism

As many of you might be working in the clinical settings or may be clinicians, you might have come across when you read various journals, what is called as Case reports? Some of you may also be presenting the case reports, during your clinical training as well as in

various clinical meetings and conferences. So, what is a Case report? Case report is nothing but a detailed presentation of a single case; what does it mean? It means that, you may find a particular clinical manifestation or you may find some interesting findings in your imaging, interesting findings in your lab or you may find a new kind of manifestation of a disease, which you have never seen before and you may want to document that and share with your peers and this is called a case report.

So, case reports are extremely useful when you want to teach others about or share your experience with others about new diseases, unfamiliar diseases, rare manifestations, though you have seen the disease before, but this particular type of clinical scenario for this disease you have not come across. These kinds of case reports may be extremely useful either to generate hypothesis about patho-physiological mechanism or just your fellow clinicians to be aware that they may come across this type of clinical manifestation during their clinical practice.

(Refer Slide Time: 03:06)

CASE REPORT

Adenocarcinoma arising from a gastric duplication cyst with invasion to the stomach: a case report with literature review

K Kurooka, H Nakayama, T Kagawa, T Ichikawa, W Yasui

J Clin Pathol 2004;63:408–411 doi: 10.1136/jcp.2003.013946

This report describes a rare case of adenocarcinoma arising from a gastric duplication cyst, with invasion to the stomach wall, in a 40 year old Japanese man. A cystic lesion was found between the stomach and the spleen. The cyst had a well circumscribed smooth muscle layer, corresponding to the muscularis propria of the stomach and the mucosa of the alimentary tract. A well differentiated adenocarcinoma was found within the duplication cyst, invading its serosa. Well differentiated adenocarcinoma was independently found in the fundus of the stomach; the tumour of the cyst was connected by fibrous tissue. Microscopically, there was neither adenocarcinoma *in situ* nor preneoplastic lesions, such as epithelial dysplasia, suggesting that the carcinoma derived from a gastric duplication cyst that invaded the stomach. Duplication cysts should be included in the differential diagnosis of cystic masses of the gastrointestinal tract, and the possibility of malignancy within these cysts should be considered.

fibrofibrinous peritonitis. The cyst was strongly adhered to the stomach and retroperitoneum and did not communicate with the gastric lumen. The patient's postoperative course was uneventful. One month after surgery, a protruding and ulcerative tumour was found at the gastric fundus, where the cyst had adhered, by endoscopic examination. Biopsy specimens from the tumour revealed a typical, well differentiated, tubular adenocarcinoma. A proximal gastrectomy was performed. Seven months after surgery, the patient had multiple liver metastases and received chemotherapy.

METHODS

The surgical specimen from the patient were fixed in 10% buffered formalin and processed for paraffin wax embedding. Multiple sections of different fragments were stained with haematoxylin and eosin. Serial sections were immunostained by the avidin-biotin-peroxidase complex technique with the primary antibodies listed in table 1. Antigen retrieval was carried out using high temperature incubation in citrate buffer (0.01 mol/litre, pH 6.0).

National Institute of Epidemiology  
Chennai  
nie.gov.in

As you can see here, this is an example from a journal how the case reports looks like. The case report is usually published in most of the peer reviewed journals, in a separate section, wherein they would mention what the clinical condition is and what the particular case report is covering. In most of the case reports, will also talk about what is the existing evidence or what is a literature about this and whether this type of manifestation has also been observed by other clinicians.

(Refer Slide Time: 03:37)

## Case series

- Study of larger group of patients (e.g > 10) with a particular disease
  - Larger number may allow the investigator to assess the play of chance
  - Common way of delineating the clinical pictures of a disease
- Suffers from the absence of a comparison group

Now as a case reports cover one or two patients, case series is a step ahead of that, wherein you study a relatively larger group of patients who have a particular disease for example, you may study 10 or more patients with some rare type of cancer or with some rare type of tuberculosis or any other disease, how does it help you? It helps you to understand whether these kinds of findings that they are observing, are they really due to the disease or is it due to the chance. It may help you understand or develop what would be the clinical picture of patients presenting with a particular type of disease. The only problem with case series is that you do not have a comparison group.

(Refer Slide Time: 04:24)

### *Pneumocystis* Pneumonia —

Los Angeles

Reprinted with permission from *Morbidity  
and Mortality Weekly Report*  
June 5, 1981

*(Editors note: The following is "Document Zero" — the first mention in the medical literature to suggest ... "the possibility of a cellular-immune dysfunction related to a common exposure". These patients comprise cases one through five of the AIDS epidemic in the United States.)*

In the period October 1980 — May 1981, 5 young men, all active homosexuals, were treated for biopsy-confirmed *Pneumocystis carinii* pneumonia at 3 different hospitals in Los Angeles, California. Two of the patients died. All 5 patients had laboratory-confirmed previous or current cytomegalovirus (CMV) infection and candidal mucosal infection. Case reports of these patients follow.

Patient 1: A previously healthy 33-year-old man developed *P. carinii* pneumonia and oral mucosal candidiasis in March 1981 after a 2-month history of fever associated with elevated liver enzymes, leukopenia, and CMV viremia. The serum complement-fixation CMV titer in October 1980 was 256; in May 1981 it was 32. The patient's condition deteriorated despite courses of treatment with trimethoprim-sulfamethoxazole (TMP-SMX), pentamidine, and acyclovir. He died May 3, and postmortem examination showed residual *P. carinii* and CMV pneumonia, but no evidence of neoplasia.

Patient 2: A previously healthy 30-year-old man developed *P. carinii* pneumonia in April 1981, after a 5-month history of fever each day and of elevated liver-function tests, CMV viremia, and documented

So, just to give you an example. This is a very interesting case series that emerged on pneumocystis pneumonia, which was observed in 5 gay men and this led to an understanding that, how is it that the pneumocystis pneumonia, which is not a very common condition is actually occurring in same kind of individuals, which subsequently led to the discovery of AIDS. Coming to the third type of study design.

(Refer Slide Time: 04:49)

## Ecological studies

- Group as the unit of analysis
- No individual-level information on the distribution of exposure and disease
- Relate whether populations with high rates of disease also have high frequency of the suspected exposure

Now, this study design is slightly different. In what way is it different? The difference here is, here individual is not the unit of study or unit of analysis. This is a study that you do for groups, that you can do at a country level, you can do for populations, you can do a particular region of the country and you try to understand a particular problem for a group and you try to relate problem as well as, what could be the possible reason, why that problem is occurring also at the group level. So in this type of study, group is the unit of analysis.

The limitation is that you do not have individual-level data. You really do not know what the individual level exposure is. However, you are able to look at a particular population let us say; we can talk about what is an average intake of fat among people living in a state of Tamilnadu? Or, what is the average intake of carbohydrate among people living in a particular state? And then, you may look at what is the incidence of cardiovascular disease. So, both the measures you only know for the population. But you are not able to have any data at the individual level. How does it help you? It helps you to generate

hypothesis because what you can see here is you can try to correlate. Let us say, is it so that the states which have high per capita consumption of fat also have higher rates of coronary artery disease or not, if you find that then you could think of doing a further detailed study to actually understand whether there is a correlation between the fat intake and the cardiovascular disease.

(Refer Slide Time: 06:38)

## Cross sectional surveys

- Observation of a cross-section of a population at a single point in time
  - Unit of observation and analysis: The individual
- Collect information about disease burden
  - Also known as “prevalence studies”
- Recruitment of study participants
  - Population
  - Population sample
- Observation for the presence of:
  - One or more outcomes
  - One or more exposures

National Institute of Epidemiology  
HEALTH RESEARCH FUNDAMENTALS  
Chennai



nis.gov.in

Now, I am going to come to the most important study design and I think this is a study design that many of you might have used in your research studies and obviously, if you are going to be in research, you will end up using this study design for sure. This is the commonest study design and I think you all are aware, you all are aware of census, right? We all have every 10 year census is done for the whole country. Census is nothing but a cross sectional survey, what do you mean by cross sectional survey? It is in observation of cross section of a population at a single point of time.

Now, what happens during census? In the year 2011, somebody would have visited your house they would have come and asked you, how many members are there in your house? How much each person has studied? Where do they go for work? Where do you live? And, this information is for year 2011; however, it may not be the same information in 2012. So, what cross sectional survey captures is, a particular set of information that you want to collect from a population or an individual at a particular point of time.

Now, these kinds of studies are extremely useful to know the magnitude, means how big

the problem is, like if suppose if you want to know how many people in the community have high blood pressure? Or how many people have tuberculosis? Or you may even not about diseases even general information, how many people in this particular area are actually professionals? Any kind of information, any kind of health or non health information, when it is collected at a particular point of time, it is called cross sectional survey. Now, for a cross sectional survey, you will be learning in other lectures about, what kind of sampling method you can use? How you can select the participants?

But, broadly speaking, census is one example of a survey which covers the entire population; whole country is covered in this kind of survey. However, all surveys are not done like that, most of the surveys they will select a sample. It is not possible to survey one billion people for everything, so you might have come across something called National Family Health Survey, wherein a sample is selected and then the data is collected. Similarly, if you are a clinician, you may want to plan a study in your patients in the clinic. However, it may not be possible for you to include all the patients in your study, and you can think of using a sample. A sample allows you to understand, what is the magnitude of problem in your study population by doing survey at one point of time.

Now, here we can collect data on various type of exposure or outcome, what does it really mean? The exposure here means different types of risk factors for example, you may do a study, wherein you want to know about high blood pressure; high blood pressure could be your outcome, but in addition to that you also want to know whether people are overweight? Whether they are having any behavioral risk factor? Whether they smoke? Whether they drink? So, these are called exposures. So, your survey may include asking questions about their history of blood pressure or doing measurements or asking various questions about what you think are the potential exposures.

(Refer Slide Time: 10:06)

## Uses of cross sectional surveys

- Estimate prevalence of disease or their risk factors
- Distribution of health problem by time, place and person
  - Plan health care services delivery
- Set priorities for disease control
- Generate hypotheses
- Examine evolving trends
  - Before / after surveys
  - Iterative cross sectional surveys

Now, what are the possible uses of cross sectional surveys? First of all, most importantly cross sectional surveys are used to measure burden of disease or prevalence of disease. Let us say, out of 100 people if I am going to survey 100 people, how many actually have high blood pressure? It can also be used to measure the burden of risk behaviors, you may go and survey 100 men and ask how often do you take alcohol? What is the quantity of alcohol you consume? So, that is a prevalence of a risk factor. It helps you understand distribution of a health problem by time, place, person, what does it mean? It means, when you go and let us take, you take a particular community you went to that area and you did a survey and you ask people history of diarrhea, and you ask the history of diarrhea in this village as well as in the adjacent village, you ask who gets diarrhea and it allows you to understand how common the diseases in that particular village; are there any particular kind of people who are effected more?

Your study might have included people from all age groups, you may have under 5 children, you may have adolescents, you may have older people and what you can find out from this survey is let us say, of all the people I surveyed I may find the diarrhea is more common in children. So, how it helps the service provider? Cross sectional survey may provide very useful information for planning your health services, at the bigger level. What I mean is at the state level, at the country level or the global level. Cross sectional surveys help us in setting priorities, for disease control.

Now, where should I invest my resources? Which is the most important disease? Should I invest it in a study on hypertension? Should I invest it on a study on depression? To understand that you need to know, what are the big health problems in your country or in your state or whatever be the unit of study? The cross sectional survey allows you to measure the magnitude of the health problem. Now, cross sectional surveys are not ideal for testing hypothesis, you will be introduced to the analytical study designs, where you can do that; however, they help you generate hypothesis.

You can think of what are the potential risk factors, if you find you did a study and you found that the prevalence of hypertension is very high. Let us say, 40 percent people have hypertension and simultaneously, you might have asked various questions regarding their eating habits, regarding smoking, regarding alcohol used, regarding over weight and you could do a preliminary analysis and understand whether which of these risk factors might be playing an important role. And this part of cross sectional survey is called analytical cross sectional survey.

However, to test your hypothesis you need to use the other study designs. The other use of cross sectional survey is, suppose you want to know after having introduced a particular type of intervention, it could be a clinical intervention for example, you introduced a new drug. A new drug into your regimen of a particular disease for example, for malaria, you are using a particular drug or for TB. You want to know before and after, after introducing your intervention, whether the patient outcomes are better, you can do that using cross sectional surveys. So, this cross sectional survey design is extremely useful design in trying to measure the burden, understand the magnitude, understand the distribution, and sometime even to understand the effectiveness of your interventions.

(Refer Slide Time: 13:52)

## Examples of research questions to be addressed through surveys

- What is the prevalence of hypertension in a city?
- How satisfied are patients attending government hospitals in Chennai?
- What is the prevalence of physical inactivity among school children?

National Institute of Epidemiology  
HEALTH RESEARCH FUNDAMENTALS  
Chennai



nie.gov.in

Now, these are the some of the examples of research questions that can be answered through surveys. What is the prevalence of hypertension or prevalence of any disease in an XY city? The other type of study, in many time many of you might be in the health services and you may want to do surveys to understand your patient satisfaction or you may want to even just do a survey to know, what kind of utilization pattern is there of your health facilities? So, you can do cross sectional survey of patients. Many times they are also called exit surveys.

Now, thus cross sectional surveys can be done in any setting. It could be done in schools, it could be done in health facility, it could be done at the community level, it could be done at the clinical level. So, another example would be that you want to know how common is the physical inactivity? As you know, you may be interested in understanding childhood obesity and you may want to know, what is the prevalence of physical inactivity among school children? So, all these kind of questions can be answered using cross sectional survey study design.

(Refer Slide Time: 14:59)

## Advantages and limitations of cross sectional survey

- Advantages
  - Fairly quick and easy to perform
  - Less expensive
- Limitations
  - Not useful to study disease etiology
  - Not suitable for the study of rare diseases

Now, let us come to what are the advantages of this and what are the limitations? Now advantage, I think you all must have realized by now, they are quick you can plan and complete a cross sectional survey in a very short period of time; may be two weeks, one month depending on how big your sample size is. And, as you know I think the biggest cross sectional survey census even that is completed in less than a year where the whole billion population is covered. If you are doing a smaller survey they tend to be less expensive, unlike some other study design where you may have to keep on doing data collection for much longer period of time.

There are few limitations of cross sectional survey that should be kept in mind while designing your study. First of all, they are not useful to study the disease etiology, as I told earlier these are descriptive studies and here you cannot test hypothesis. They are also not suitable in study of rare diseases, let us say if your disease, let us say cancer; now cancer incidences 20 per 100,000 and if you want to measure the burden of cancer, just to get those 20 cases you will have to survey 100,000 people, it is not practical. This kind of situation, it may not be an ideal study design.

(Refer Slide Time: 16:20)

## Cross sectional survey: major limitation

- Prevalent cases  
(Old and new cases)
- Exposure and outcome examined at the same time. e.g.
  - Obesity and diabetes



National Institute of Epidemiology  
Chennai  
HEALTH RESEARCH FUNDAMENTALS  
[nie.gov.in](http://nie.gov.in)

Now, coming to the most important limitations and this you should all be very familiar and keep in mind when you design a cross sectional survey. Now, if you go and collect data on hypertension, there are patients who will tell you that; yes, I have hypertension I am taking treatment for last one year. Now, you do the blood pressure measurement in addition to these old cases, you are also going to pick up some of the new cases, who were not aware at the time of survey that they have hypertension. In a cross sectional survey, you measure all the prevalent cases, I think in the earlier sessions you might have been introduced to these terms, terms such as prevalence, incidence.

So, cross sectional surveys measure prevalence. Now which is fine, but what is the problem here? Now, if you go and survey an individual and you ask, do you have a hypertension? They say, yes and then you ask them, are you a smoker? They say, yes. What it does not tell you? It is the chicken and egg story, you do not know which happened first, whether hypertension happened first or smoking happened first. Similarly, if you take obesity and diabetes, you go in a cross sectional survey you may find people who are diabetic as well as overweight. How do you know which happened first? Whether diabetes happened first or overweight happened first? So, when you do a cross sectional survey, you need to be very cautious in interpreting your results, you need to keep in mind that these exposure and outcome cannot be linked very well in this kind of study design and if you want to do that, you will have to use different kind of study designs that you will be learning in the upcoming lectures.

(Refer Slide Time: 18:06)

## Take home messages

- Case reports and case series are useful for uncommon clinical manifestations
- Ecological studies can be used to relate group level data and generate hypothesis
- Cross sectional surveys help to measure the burden or magnitude of health condition

National Institute of Epidemiology  
HEALTH RESEARCH FUNDAMENTALS  
Chennai



nie.gov.in

Now, just to summarize, case reports and case series are extremely useful to document uncommon clinical manifestations in a certain set of patients. And, these are extremely useful for the clinicians. Ecological studies are useful when you want to relate a group level data and generate hypothesis. Cross sectional surveys, most common study design and the very useful study design, it is useful to measure burden or magnitude of health conditions.

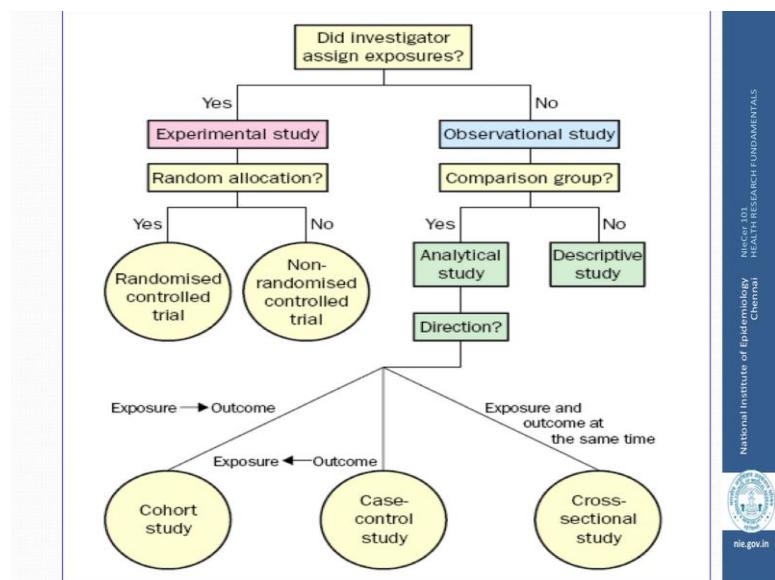
Thank you very much.

**Health Research Fundamentals**  
**Dr. Manoj Murhekar**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**Lecture – 06**  
**Analytical study designs**

Hello and welcome. My name is Manoj Murhekar and in today's lecture I will give you an overview of cohort and case control studies. In earlier lectures, my colleagues have given you an overview of different study designs.

(Refer Slide Time: 00:27)



Let us have quick recap. Epidemiological studies are broadly divided into 2 categories; first is Experimental studies and second is Observational studies. And this categorization is based on these questions; did the investigator assign the exposure? So, in experimental studies investigator assigns the exposure and this exposure could be in terms of new intervention, new drug or vaccine. These studies are further classified into Randomized and Non-randomized studies, based on Random allocation of exposure.

On the other hand, in observational studies investigator does not assign the exposure. If there is no comparison group in observational studies, such studies are called as

Descriptive studies. And in these studies, we described health event in terms of time, place and person. If there is a comparison group in observational studies, the studies are called as analytical studies, which are further divided based on the direction of the studies. Cohort studies they progress from exposures to outcome, whereas case control study progress from outcomes to exposure and in cross sectional studies we measure exposures and outcomes at same time.

(Refer Slide Time: 01:53)

## Analytical studies

- Investigator does not assign the exposure
  - Makes careful measurement of patterns of exposure and disease in populations
- Comparison group
  - Make inferences about exposure and disease

National Institute of Epidemiology  
Chennai  
  
nie.gov.in

So, in short analytical studies are the one in which investigator does not assign the exposure, there are no randomization. So, what investigator does essentially is he carefully measures the pattern of exposure and disease in populations. There is a comparison group in analytical studies and using this comparison group the investigator next inferences about exposure and the disease.

(Refer Slide Time: 02:20)

## Cohort study

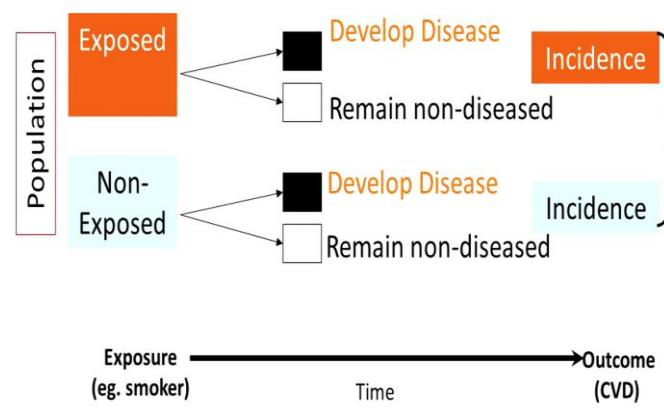
- Cohort
  - 300 to 600 man unit in Roman Army
- Cohort
  - Group of people sharing some common characteristics (ex. Birth cohort)



Let me first talk about Cohort study. The word cohort has a military origin, military routes rather than medical routes. In Roman army, a 300 to 600 man unit was called as cohort, whereas in epidemiology, the word cohort is a group of individuals sharing some common characteristic; one such example could be a birth cohort, all the children who are born today will form today's birth cohort.

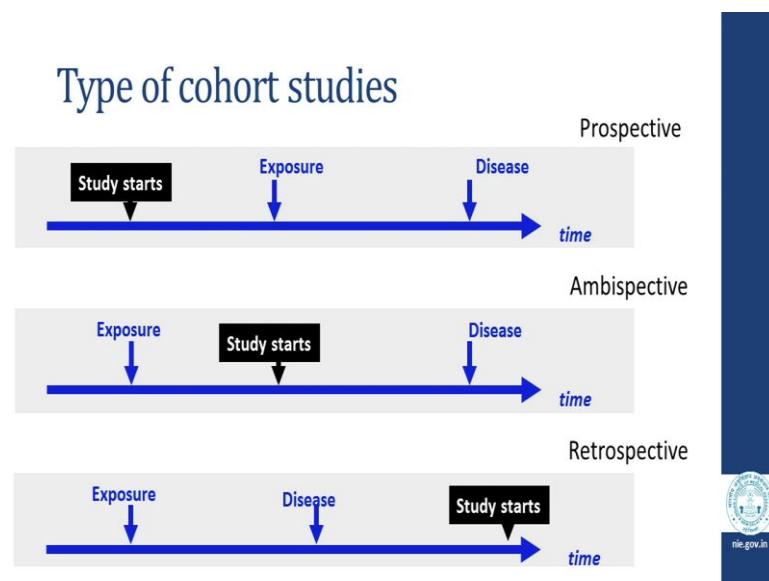
(Refer Slide Time: 02:55)

## Design of cohort study



Let us see, how the design of cohort study is. As we know, cohort study will progress from exposure of outcome and in this particular example exposure is say, cigarette smoking and outcome development of cardio vascular disease. Cohort studies begin with selection of Exposed and Non-exposed cohort, and in this example, it would be people who are the cigarette smoking and those are not smoking. Once we identify this cohort, these cohorts are followed in time some of this exposed and non-exposed individuals will develop the disease that is cardio vascular disease, whereas the remaining people would remain non diseased. We will then calculate the incident of cardio-vascular disease in exposed population and in unexposed population. And we will compare this incidence using a measure of association called as relative risk. I will talk about this relative risk later.

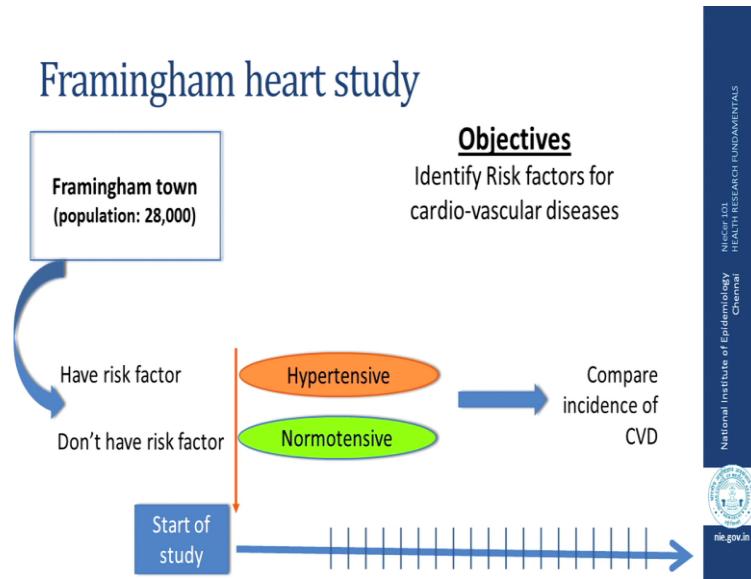
(Refer Slide Time: 03:59)



There are 3 types of cohort studies; the first is Prospective cohorts study. In the prospective cohorts study, by the time your study starts exposure and disease has not yet occurred. Whereas in case of your Retrospective cohort study, both the exposures and disease has already occurred when you start the study. And there is a combination of these two approaches, which is called as Bidirectional study or Ambispective study, wherein when your study starts, the exposure has already occurred and then you follow this exposed and then exposed individuals, till they develop the outcome. Let me explain

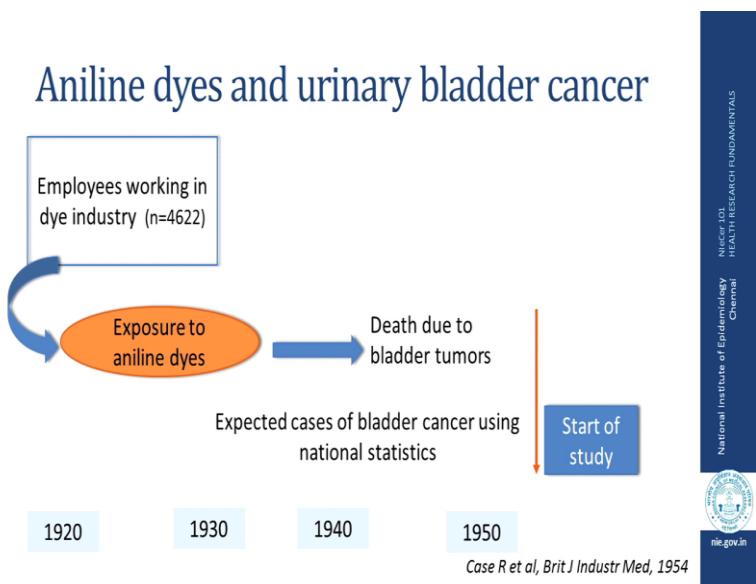
these different types of studies by giving some example.

(Refer Slide Time: 04:51)



First example is of Framingham heart study, which is one of the oldest cohort study initiated in 1940s. The objective of this study was to identify risk factors called cardio-vascular diseases. This study was conducted in a town of Framingham, which had a population of about 28,000. So, this population; in fact, the sample of this population was then divided into 2, based on those having risk factor and those who was not having the risk factor. And the investigator considered several risk factors, one of which was hypertension. So, for the purpose they classified this population into those who had hypertension and those who did not have hypertension. This cohort was then followed up in time and the incidence of cardio-vascular diseases was compared into cohorts.

(Refer Slide Time: 05:50)



This is an example of a Retrospective cohorts study. The objective of this study was to evaluate the role of Aniline dyes or exposure of aniline dyes and development of urinary bladder cancer. So, the investigator for this study recruited about 4622 workers who were working in dyes industry between 1920 and 1951. So, this recruitment was based on available records in those factories. Investigator also revived the death records of this 4622 individuals and essentially, they looked about any mention of urinary bladder tumors on their death records, and then they compared death rates in these population with that of expected number of deaths of bladder cancer using national statistics. So, by the time the studies started, both exposure and outcome had occurred.

(Refer Slide Time: 06:58)

## Elements of cohort study

1. Selection of study populations
2. Gathering baseline information
3. Follow-up
4. Analysis



So, these are the 4 important components of cohort study. First is selection of study population, second is gathering baseline information, third is following up this cohort and fourth is doing analysis.

(Refer Slide Time: 07:15)

## Selection of study population

- General population cohorts or a sub-set
  - Framingham heart study
  - Nurses health study
- Special exposure cohorts
  - Occupational groups



There could be 2 approaches of selecting study population. You could select your cohort

from general population as was done in case of Framingham study or you could select a subset of general population as was done in nurses' health study. The second approach could be selecting a special exposure groups, such as occupational groups.

(Refer Slide Time: 07:39)

## Gathering baseline information

- Objective
  - Valid assessment of exposure status of members of cohort
    - Identification data
    - Exclude individuals having disease at baseline
    - Define individuals at risk
    - Obtain data on co-variables (other exposure variables)



Once you select this study population, the next important step is collecting baseline information from this population and the objective of this step is to have a valid assessment of exposures status of members of cohort. And by doing this baseline information we can also collect identification details of the study population, we can exclude those individual, who are having the disease of interest at baseline so that the population which remains is at risk of developing the disease and we can also obtain the data about other risk factors or other exposure variables.

(Refer Slide Time: 08:18)

## Choice of comparison group

- Internal comparison group
  - Unexposed persons in the population
- External comparison group
  - When internal comparison group not available
  - Ex: Observed number of bladder cancer deaths in aniline dye industry compared with expected cases



As we have seen in analytical study there is a comparison group and we have 2 options having comparison group in case of cohorts. One is internal comparison group and the unexposed person in the population is taken as an internal comparison group and example could be Framingham cohort study. Wherein, those who had hypertension were considered as exposed population and those who were normotensive was unexposed population.

Sometimes it is not possible to have internal comparison group. As we have seen in the case of aniline dye example, everybody in those factories were exposed and it was not possible to have a internal comparison group. And therefore, the investigators compared the death rates with that of general population, so you could take an external comparison group in such situations.

(Refer Slide Time: 09:18)

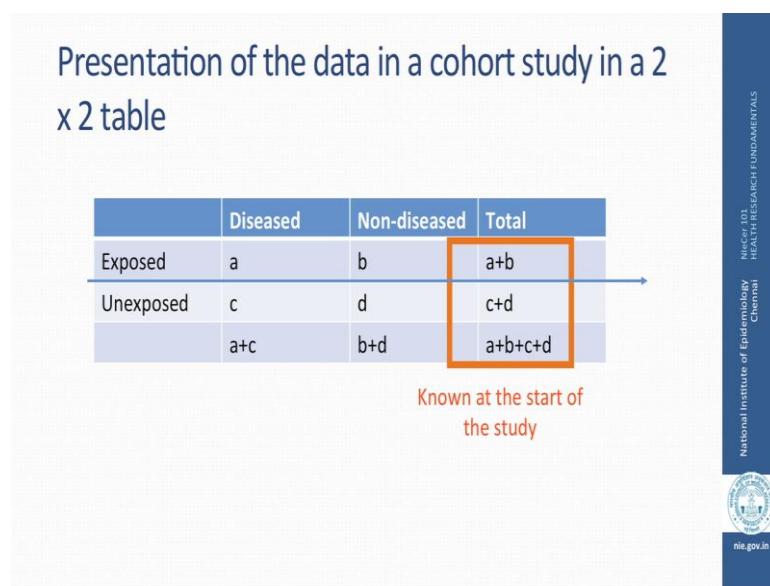
## Follow-up

- Objectives
  - Uniform and complete follow-up of all cohort members
    - Uniform surveillance in exposed and unexposed groups
  - Complete ascertainment of exposures and outcome/s
  - Standardized diagnosis of outcome events



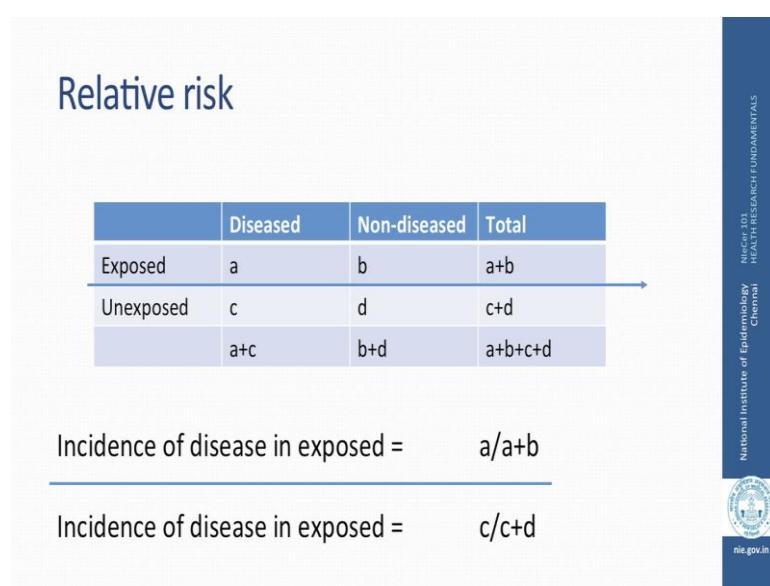
Once you recruit your exposed and unexposed population, the next very important step is doing a good follow-up of these populations. There could be 3 principles of having good follow-up, first is having a uniform surveillance in exposed and unexposed group; having complete ascertainment of exposures and outcomes and third is using a standardized diagnosis of outcomes, especially since the cohort studies can last for a long period of time.

(Refer Slide Time: 09:50)



This is how the data in cohort study would look like this is how the 2 by 2 table in cohort study would look like. We started the study with selecting people who are exposed, which is a plus b and people who are unexposed which is c plus d. And we followed this people so a plus c developed the disease and b plus d remain non-diseased. So, at the beginning of the study we know who were expose and who were not exposed.

(Refer Slide Time: 10:22)



So, we can calculate incidence of disease in exposed population which can be given by formula a upon a plus b and incidence in unexposed population would be in c upon c plus d and the ratio of these 2 incidence is a relative risk.

(Refer Slide Time: 10:42)

## Interpreting Relative risk

- RR=1
  - Incidence in exposed and unexposed is same
  - Exposure is not associated with disease
- RR > 1
  - Incidence in exposed is higher than unexposed
  - Exposure is positively associated with disease
- RR < 1
  - Incidence in exposed is lower than unexposed
  - Exposure is negatively associated with disease



How do you interpreting this relative risk? There could be 3 possible scenarios of relative risk, one is relative risk is equal to 1. If your relative risk is 1, it means that incidence of disease in exposed and unexposed population is same and we can interpret that the exposure is not associated with the disease. Relative risk could be more than 1, which means that incidence of disease is higher in exposed population as compared to an unexposed population and we can interpret that the exposure is positively associated with the disease. Relative risk can also be less than 1, which means that incidence of disease in exposed population is lower than unexposed population and here we can interpret that exposure is negatively associated with the disease.

(Refer Slide Time: 11:39)

## Cohort study – Strengths and weaknesses

- Strengths
  - Allows calculation of incidence
  - Examine multiple outcomes for a given exposure
  - Clarity of temporal sequence
  - Good for investigating rare exposures
- Weakness
  - May have to follow large numbers of subjects for a long time.
  - Expensive and time consuming.
  - Not good for rare diseases.
  - Not good for diseases with a long latency.
  - Differential loss to follow up can introduce bias.



Cohort studies have certain strengths as well as certain weaknesses. So, what are their strengths? They allow calculation of incidence because when we start the study, we start the study with selecting exposed population and unexposed population and we follow them in time therefore, it is possible for us to calculate the incidence of disease. We can examine multiple outcomes for a given exposures. We are very confident about temporal clarity in case of cohort study, and last is that this studies are especially useful for rare exposures.

So, what are the weaknesses? The sample size for cohort study could be very large; we need to follow these people for a very long time and therefore, cohorts studies could be expensive and time consuming. There are not recommended for diseases which are rare or diseases which have very long latency, and if you do not have good follow up or differential follow loss in exposed and unexposed population it could introduce certain amount bias in your study.

(Refer Slide Time: 12:53)

## Case control study

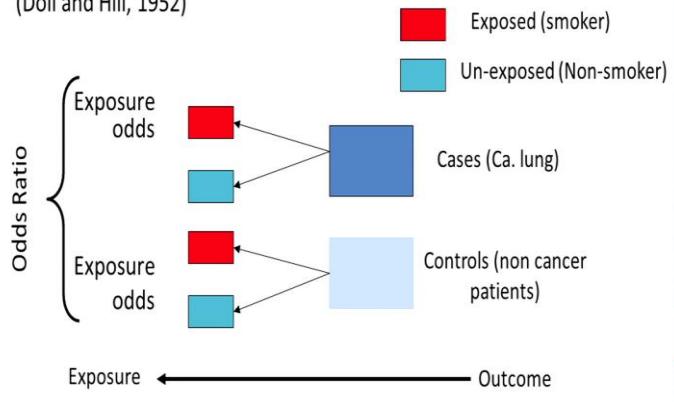


Let us now see Case control study. Case control studies are exactly opposite to that of cohort study when it comes to the direction or logic of the study.

(Refer Slide Time: 13:06)

## Design of case-control study

Objective: Test association between cigarette smoking and lung cancer  
(Doll and Hill, 1952)



Let us first see the design of cohort study. And to explain the design, I will use the example of one of the very old case control study conducted by Doll and Hill. The

objective of this study was to test the association between cigarette smoking and lung cancer. As the name suggests case control, we start with selecting cases and control is the one who does not have the disease in question. So, the first step is selecting cases, so for this study Doll and Hill selected lung cancer cases, who are admitted in hospital in about 20 hospitals in London. These cases were all histopathologically proven cases of lung cancer. So, for each case they selected a control which was a non-lung cancer patient admitted in the same hospital and these cases are controls and then intrigued to find out their prior exposures.

So, Doll and Hill found out how many of the cases where cigarette smokers, they had a detailed questionnaire to ask about history of cigarette smoking. They asked how many of them were smoking. What is the age of starting smoking? What type of cigarette they were smoking? And so on and so forth. So, we find out how many of the cases are exposed? How many of the controls are exposed? And same way how many of the cases are unexposed? And, based on this data we calculate what is called as Exposure odds among cases and Exposure odds among controls and then we calculate what is known as odds ratio as a major of association between exposure and outcome.

(Refer Slide Time: 14:50)

## Elements of case control study

1. Selection of cases
2. Selection of controls
3. Information on exposure
4. Analysis



Like cohorts study, there are four important elements of case control study. First is

selecting cases, second is selecting controls and third is collecting valid information about exposures and then doing the analysis.

(Refer Slide Time: 15:03)



## Selection of cases

- All people in source population who develop the disease of interest
  - Sample of cases
    - Independent of the exposure under study
- Clear definition of outcome studied
- Prevalent vs. incident cases
  - Prevalent cases may be related more to survival with disease than to development of disease

Coming to the selection of cases, theoretically all people in the source population who develop the disease of interest could be included in your study or you could sample these cases. However, one thing you should keep in mind that, the selection of cases should be independent of exposure under study. We need to have a clear definition of outcome to be studied. One also need to decide whether to include prevalent cases or should include only incident cases.

Prevalent cases mean those cases which already occurred in the past, whereas incident cases are the cases newly occurring cases. So, if you take a prevalent case they are readily available and by including them we can save our time and money. But, in spite of this obvious advantage it is generally recommended to include incident cases, mainly because prevalent cases maybe related more to the survival with the disease than the development of disease.

(Refer Slide Time: 16:13)

## Sources of cases

- Hospital/clinic based cases
  - Easier to find
  - May represent severe cases
- Population based (cancer registry)
  - not biased by factors drawing a patient to a particular hospital



Where from I can select these cases? Again, there could be two important sources; one is from hospitals or clinics, it is easier to find cases in this hospitals and clinics; however, it is quite possible at cases which are admitted are more severe cases and may not represent the cases in community. The other approach could be a population based selection of cases and one such example could be cancer registry and these cases are more likely to represent the source population, primarily because they are not biased by factors drawing patient to a particular hospital.

(Refer Slide Time: 16:52)

## Selection of controls

- Represent the distribution of exposure in the source population of cases
  - Selected from the same source population that gives rise to the cases
- Selected independently of their exposure status



Here from I can select the controls. As I mentioned earlier, control is the one who does not have the disease under investigation. Why do we need control? Controls essentially represent the distribution of exposure in source population. So, they generally tell you the background rate of exposure in the population from which cases have come. Like cases, they also need to be selected independently of their exposure status.

(Refer Slide Time: 17:30)

## Selection of controls

- Population based
  - Sampling of the general population
- Health care facility based
  - Patients with other diseases
- Case-based
  - Friends, Neighbourhood



There again could be 3 sources of controls, first is the population based controls and you could sample from general population. Second is you could select controls from health facility and in case of Doll and Hill study they selected control from the health facility, but we could select patient with other diseases. And the third source of controls could be case-based controls that are from friends or neighborhood.

(Refer Slide Time: 17:59)

## Collecting good data on exposure

- Objectively
  - Reproducibility of exposure measurement
- Accurately
  - Information reflecting as closely as possible the effect of exposure
- Precisely
  - Quality management in exposure measurement



Once you select cases and control, the next important step is collecting data about past exposures. And again, there are 3 important principles collecting the data on exposures objectively so that your measurements are reproducible, accurately and precisely.

(Refer Slide Time: 18:19)

Presentation of the data of a case-control study in a 2 x 2 table

	Cases	Controls	Total
Exposed	a	b	a+b
Unexposed	c	d	c+d
	a+c	b+d	a+b+c+d

Known at the start of the study

NICER FOR  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai  
  
nier.gov.in

Once you collect the data on exposures, this is how your 2 by 2 table will look like. This is exactly the same table which we saw for the cohorts study. However, in case control studies, when the study started we knew who were case and we knew who were controls. So, a plus c, were cases to start with and b plus d were the controls. And we found out that of a plus c cases, a were exposed and c were unexposed and same way b plus d controls b were exposed and d were unexposed. In case control study we cannot calculate incidence of disease, like what we could calculate in case of cohort studies.

(Refer Slide Time: 19:06)

## Odds ratio

	Cases	Controls	Total
Exposed	a	b	a+b
Unexposed	c	d	c+d
	a+c	b+d	a+b+c+d

Odds that case was exposed =

$$\frac{\text{Probability that case was exposed}}{\text{Probability that case was not exposed}} = \frac{[(a/a+c)]}{[(c/a+c)]} = a/c$$

Odds that control was exposed =

$$\frac{\text{Probability that control was exposed}}{\text{Probability that control was unexposed}} = \frac{[(b/b+d)]}{[(d/b+d)]} = b/d$$

$$\text{Odds ratio} = [a/c]/[b/d] = ab/bc$$



So, what do we do? Then what we do is, essentially we calculate a measure of association called Odds ratio. This odds ratio is the odds at cases was exposed is given by this formula; probability that case was exposed, upon probability that case was not exposed. And we know that probability a case was exposed is a upon a plus c and probability that the case was not exposed is c upon a plus c and the ration of these 2 probabilities is a by c. Same way we also calculate the odds that control was exposed, which comes out to be b by d and the ratio of these 2 odds becomes odds ratio which is ab by bc, which is nothing but a gross product ratio.

(Refer Slide Time: 20:03)

## Interpreting Odds Ratio

- OR=1
  - Odds of exposure among cases and controls are same
  - Exposure is not associated with disease
- OR > 1
  - Odds of exposure among cases are higher than controls
  - Exposure is positively associated with disease
- OR < 1
  - Odds of exposure among cases are lower than controls
  - Exposure is negatively associated with disease



How to interpret this odds ratio? Again like relative risk there could be 3 scenarios, one is odds ratio equal to 1. If odds ratio is equal to 1, it means that odds of exposure among cases and controls are same and we can conclude that exposure is not associated with disease in such situation. Odds ratio could be more than 1, it means that your odds of exposure among cases are higher than that of controls and we can conclude that exposure is positively associated with the disease. If odds ratio is less than 1, it means that odds of exposure are among cases are lower than that of controls and we can conclude that exposure is negatively associated with the disease.

(Refer Slide Time: 20:51)

## Case control study: Strengths and weaknesses

- Strengths

- Good for examining rare outcomes or outcomes with long latency
- Relatively quick to conduct, inexpensive
- Requires comparatively few subjects
- Multiple exposures or risk factors can be examined

- Weaknesses

- Susceptible to recall bias
- Selection of an appropriate comparison group may be difficult
- Rates of disease in exposed and unexposed individuals cannot be determined



Case control study also has certain strengths and weaknesses. These studies are especially good, if the outcome is rare or the diseases have the long latency period. They are fairly easy or quick to conduct and hence inexpensive. Requires relatively less subjects than that of cohort studies and multiple exposures or risk factors can be examined at the same time. The weakness of cohort study include that they are susceptible to several biases, the recall bias one of the most important bias. Sometime selection of control could be a problem, selection of an appropriate comparison group may be difficult and we cannot calculate incidence or disease in these studies.

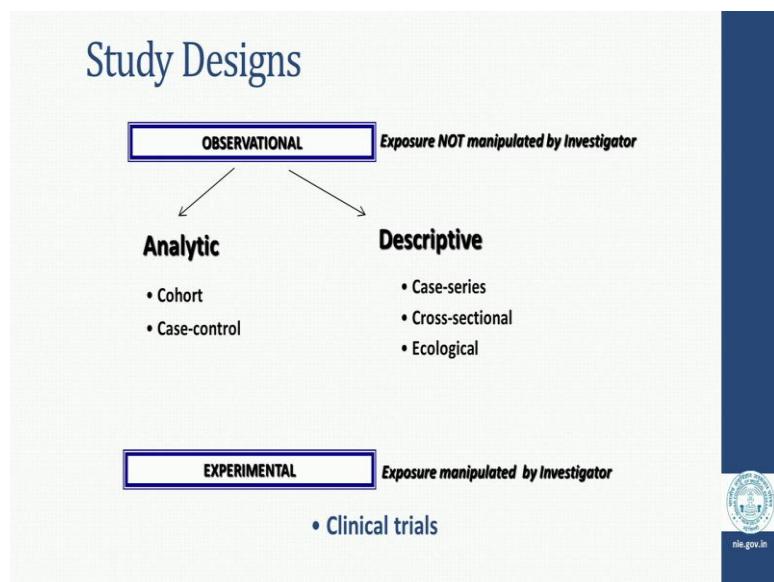
Thank you.

**Health Research Fundamentals**  
**Dr. Sanjay Mehendale**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**Lecture – 07**  
**Experimental study designs – Clinical trials**

Hello. In the course of Health Research Fundamentals, today I am going to discuss the Experimental study designs or Clinical trials.

(Refer Slide Time: 00:16)



As has been discussed earlier, the various types of epidemiology study designs are classically described as observational designs and experimental study designs. Observational study design is where the exposure is not manipulated by the investigators and there are two different types of observational studies. There could be descriptive studies like a series cross sectional studies or ecological studies and analytical studies include case controls studies, which are retrospective in nature and cohort studies, which are prospective in nature. But, considered as more advanced are the experimental study designs, where the exposure is manipulated by the investigator and a classical example of this is clinical trials.

(Refer Slide Time: 01:05)

## Randomized Controlled Trials

One of the main scientific advances in methods of clinical research in the 20th century. They are considered as the methodologic standard of excellence and gold standard for scientific experiments.



nie.gov.in

These clinical trials, these kinds of studies where one of the main scientific advances in the last century, they are considered as the methodological standard of excellence are also described as gold standard for scientific experiments.

(Refer Slide Time: 01:25)

## Significance of clinical trials

- Clinical trials translate results of basic scientific research into better ways to prevent, diagnose, or treat disease

National Institute of Epidemiology  
NIHUE-TBI  
HEALTH RESEARCH FUNDAMENTALS



nie.gov.in

Their main significance is that they are essential for translating the results of basic

scientific research into better ways to either prevent the disease, diagnose a disease or a condition or to treat a particular disease and they have a huge translational value. But, primarily randomized controlled clinical trials and we are going to discuss this at length today.

(Refer Slide Time: 01:41)

## Randomized controlled clinical trials

- A clinical trial is a planned experiment designed to assess the efficacy of prophylactic / diagnostic / therapeutic agents, devices, regimens, procedures etc. applied to human subjects
- It essentially involves comparing the outcomes in a group of patients treated with a test treatment with those observed in a comparable group of patients receiving a control treatment where patients in both groups are enrolled in a prospective study, treated or exposed to intervention and followed over the same period



These are planned experiments, primarily designed to assess the efficacy of prophylactic, diagnostic or therapeutic agents. It also helps to test out the new devices or different types of drug regimens or new procedures that are being introduced including the investigative procedures, etcetera, particularly in human subjects. Basically, it involves comparing the outcomes in two groups of individuals and when we talk about doing a therapeutic trial, we talk about people suffering from a particular disease, who are grouped into two different groups.

Wherein, one group requires a new kind of a treatment and the other group requires the standard treatment that is being available at that point of time. Then, what is done is over a period of time, they are evaluated, which we call it as a prospective study and we find out how many of them get effectively cured for example. So, this is the prospective nature of the study, all the participants are essentially followed for a certain period of time and that is why it is a planned experiment. Why? Because here is where, as we have

discussed earlier in the definition. The environment has been modified, manipulated or modified by the investigator because the investigator has decided that some people will be placed in one particular arm of the clinical trial, wherein the other group will be placed in the other arm of the clinical trial and that is what is the process called randomization, which we will be discussing sometime later.

(Refer Slide Time: 03:18)

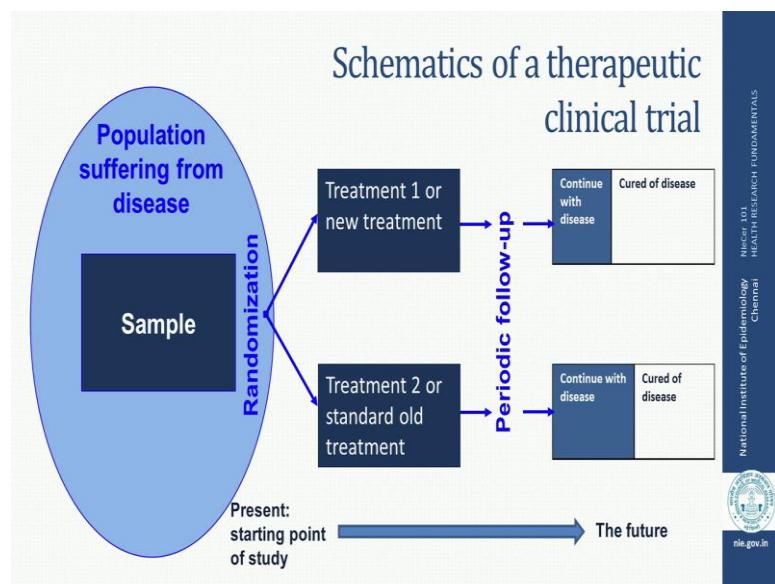
## The objectives of clinical trials

Clinical trials are normally conducted to evaluate new forms of therapy or prevention methods such as

- New drugs/ treatment
- New medical / health care technology
- New organization/ delivery system of health care
- New methods of primary prevention
- New programs of screening or early detection

Clinical trials can be broadly used for a variety of reasons as we briefly discussed. New drugs, new treatments, new medical or health care technologies, new organizational or delivery systems of health care, new methods of primary prevention or new programs for screening of early detection of diseases. They could be employed in a variety of scenario.

(Refer Slide Time: 03:44)



Normally, clinical trials are equated with drugs and pharmaceuticals. So, this essentially have intent of therapeutic benefit. So, if we just think about how this could be done is? We have a large number of individuals, who are actually suffering from a particular disease. What we do here is? We take out a sample of people, who are, say eligible to apply eligible to, say participate in a clinical trial. Here is where the problem is, that these are the people who should be willing to participant in the study. So, we cannot say that, this particular sample is essentially generalized or it is generally representative of the total population that we are dealing with.

Now, in a process of randomization, which we will talk about they will be sent out to two different arms of the clinical trials. Whereas, one arm will receive the new treatment or one type of treatment, whereas the other arm will receive the standard old treatment or the second type of treatment. Then all these individuals will be prospectively followed up for a defined duration and with a defined frequency. Maybe it is 1 year, once in every 3 months or 2 years, once in every 6 months or whatever and then we assess them again to find out because it is a therapeutic trial. What we want to figure out is how many of them have actually got cured of the disease? And how many continue to have the disease?

The reason why we are testing out a new treatment, we are having an expectation that the

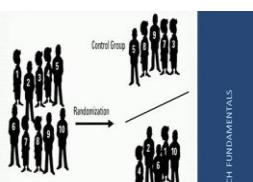
newer treatment will have more benefit. So, the number of people being cured of this particular disease would be more compared to that in the other arm and this can be statistically analyzed.

In case of a prevention scenario, say if we talk of new vaccine for example, this can also be tested in a clinical trial design. Here, the type of population will obviously change. Here, we will talk of people or the population, which is susceptible or at risk of developing that particular disease. Here, again we will take out a sample of people who are willing to participate in the study. Like in the therapeutic trial, these will be randomized into two arms, one arm or one group of people in them will develop, will receive the vaccine and the other will receive a placebo. This is a critical decision which needs to be really discussed at length, whether it is a placebo or some other type of a vaccine or what can be given in this particular scenario as a comparator arm.

Again, these two groups of individuals are followed for a certain period of time with a certain periodicity and what we try to judge is how many of them in either of the two arms actually acquire the disease? The expectation is that if the vaccine is effective, less number of people in that particular arm who are receiving the vaccine would acquire the disease compared to those who have not received the vaccine in the comparator arm or the placebo arm. This can also be statistically analyzed and so we decide, whether this particular vaccine has been effective in preventing occurrence of that particular disease in susceptible populations.

(Refer Slide Time: 07:07)

## Randomization



Randomization ensures that participants have an equal chance to be assigned to one of two or more groups:

- A. One group gets the most widely accepted treatment (standard treatment/ gold standard)
- B. The other gets the new treatment being tested, which researchers hope and have reason to believe will be better than the standard treatment

Randomization provides the best way to prove the effectiveness of a new agent or intervention by ensuring that

- A. All groups are as similar as possible
- B. Confounding, co-interventions and bias in outcome ascertainment is minimized

NATIONAL INSTITUTE OF EPIDEMIOLOGY  
CHENNAI  
NIE GOVT OF INDIA  
nie.gov.in

The most critical step in a clinical trial is what we call as randomization. Randomization is a process, where we say that the participants have an equal chance of being assigned to any study group. There may be one study group, there may be two arms in the trial and there may be multiple arms in the trial. Accordingly, the original numbers of participants that are found eligible that get placed into various arms in a pre decided manner. In the number, in each arm is decided and candidates get assigned or the participants get assigned to different kinds of arms here. What is important here is, the participants do not decide that he or she or they want to participant in a particular arm. It is done through a neutral process. Also, the investigators do not decide, whether a particular participant goes in arm a or arm b or arm 1 or arm 2. It is a process of randomization, which is a third party procedure which decides that.

So, one group gets the most widely accepted treatment, which we call it as the standard treatment or the gold treatment. Whereas, the other group get us the new treatment and here, the hope and expectation is that the newer treatment is better than the standard treatment. This is the process of randomization. So, it ensures that we can test the effectiveness of new agent optimally here because all groups are most likely to be similar, as similar as possible and confounding, co-interventions and bias in outcome ascertainment can be minimized.

(Refer Slide Time: 08:50)

## Blinding is another quality improvement technique in clinical trials

- Blinding can be at the level of
  - Participants [single blinding]
  - Participants and investigators [double blinding] and
  - Participants, investigators and analysts [triple blinding]
- Blinding helps to eliminate
  - Co-intervention: participants use other therapy or change behavior or study staff, medical providers, family or friends treat participants differently
  - Biased outcome ascertainment: participants may report symptoms or outcomes differently or physicians or investigators may elicit symptoms or outcomes differently

NICER FOR  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

Similarly, blinding is another quality improvement technique, which is often used in clinical trials and blinding is a procedure when participants, if they do not know, what they are getting, whether they are getting drug a or drug b? This is called single blinding that is here participants are not aware as to what kind of drug is being given to them, investigators may be aware.

Sometimes participants as well as investigators do not know what kind of drug allocation is being done because the type of treatments that are been given, the nature, the color and the appearance of the interventions is exactly identical. This is a double blinding procedure and even when the analysts also are kept blinded, while they analyze the results, this is called as triple blinding. Basically, this stepwise blinding from single to double to triple blinding that eliminates the kind of subjectivity in judgment of outcomes in a large number of instances.

There are two problems that generally arise, one is a core intervention. What happens is participant, if they realize that one particular group is getting a particular type of drug, the other group is getting the other type of drug. They try to do some kind of an exchange and decide something between themselves. Sometimes, this is also facilitated by some kind of study staff or medical providers or family or friends, etcetera and hence

this kind of a co-intervention that take place, can vitiate the results of the study and hence blinding helps to eliminate that.

Also, if for example, if a patient or a study participants knows what he is getting, if he is getting a new drug, may be some of the minor elements he might want to exaggerate as compared to the, if he knows that he is getting the standard treatment, he may not want to report that. Similar thing may happen in case of those, who are actually evaluating that study participant and hence these kinds of biases in outcome ascertainment that can come in, can be minimized in blinding.

(Refer Slide Time: 10:59)

### Phases in clinical trials and objectives

Trial phase	End-points/ objectives	Sample size and participants
Phase I	Safety Acceptability	Up to 50 Healthy volunteers
Phase II	Long-term safety Dose and schedule Early indications of efficacy	100 to 500 Low risk
Phase III	Effectiveness	1000 and more High risk
Phase IV	Post-marketing surveillance	1000 and more Community based

NIEHTS  
National Institute of Epidemiology  
Chennai

nie.gov.in

Typically, all the clinical trials are done in 4 phases. The phase I, is the step 1 in clinical evaluation of any new intervention that comes in and here is where the trial is done in a very small number of individuals, generally below 50 and they are mostly the healthy volunteers, who are the part of phase I study and goal is to do safety and acceptability evaluation here. When this particular product is found to be safe and acceptable, it passes on to the phase II trial, which is generally done in larger number of individuals, generally 100 to 500 and who may be having some kind of low risk of a particular disease and here is where we studied the long term safety. We also try to study the dose and schedule, in case of vaccine for example or in case of a vaccine again some early indications of

efficacy. This is a phase II design.

A phase III design, essentially is a large trial, which looks at the efficacy of a particular intervention and under controlled clinical conditions, generally they involve thousands of individuals and who are more at risk of acquiring that particular disease or who are suffering from that particular disease and we try to look at in a therapeutic scenario, how that particular drug is effective in curing or in prevention scenario how that particular prevention measure is helpful in preventing the occurrence of that particular disease. Once, all these phases I, II and III are completed, the product goes for licensures in the country, it gets a license and once it gets the license, it gets marketed in the country and then phase IV trials are undertaken, which are considered as post-marketing surveillance and they are done in again thousands and thousands of individuals, but it is generally collection of information in a community based manner not in a thorough clinical kind of a clinical trials set up. This is how clinical trials are conducted.

(Refer Slide Time: 13:11)

## Example of a therapeutic trial

To study if a new drug regimen [NDR] effectively lowers viral load and improves CD4 counts in HIV infected persons compared to standard therapy [HAART]

1. Identify HIV infected persons, define inclusion and exclusion criteria
2. Randomize patients into 2 groups, one receives NDR, the other HAART
3. Follow-up periodically, estimate viral load and CD4 counts periodically
4. Use statistical methods to see if there are differences between viral load and CD4 counts in the two groups

Just, I will give you an example of a therapeutic trial. If you want to study a new drug regimen, whether it can effectively lower the viral load and it also results in an improvement in the CD4 counts in HIV infected persons. So, I am talking about HIV AIDS scenario, where people are eligible for treatment. Now, currently there is a

standard HAART, Highly Active Antiretroviral Therapy; which is available, which comprises of 3 drugs, which is being given at ART centers, Antiretroviral Therapy centers, which are run by the government in our country, all of us know that.

Somebody comes up with the new drug regimen, which probably is a cheaper regimen. Probably where the number of tablets required are much smaller than the standard HAART regimen and so, one wants to evaluate whether this is an optimum regimen to be used in the country or not. So, how would we go about it? We will identify HIV infected individuals and define the inclusion and exclusion criteria for participation, whether they are eligible to be put on antiretroviral therapy. Then they will be randomized, all those who are found to be eligible in screening will be randomized into two groups or arms. One will receive the new drugs regimen, the other will receive the gold standard that is available now; which is the highly active antiretroviral therapy will follow them up periodically.

The expectation will be both these sets of drugs will effectively lower the viral load and help to improve the CD4 counts. This will be evaluated periodically, once in every 6 months for example and if the total period of study is 1 year or 2 years, we just see, how these two groups have reacted or responded to these two different types of the regimens and statically we will try to find an answer, whether this new therapy that we are talking about has provided any kind of an additional benefit. This can be done using standards statistical methods.

(Refer Slide Time: 15:15)

## Example of a Prevention trial

A new vaccine candidate has been developed that has generated laboratory and animal data supporting its safety and ability to generate immune response. It is being considered a promising candidate for humans.

1. Decide at the national, regional and local level whether this vaccine is appropriate for the country and the population
2. Develop a Phase I trial design
3. Find healthy volunteers [adults/ children, men/ women]
4. Carry out screening followed by enrollment: Randomize patients into 2 groups, one receives vaccine, the other placebo
5. Follow-up the participants periodically, record safety and estimate immunogenicity periodically
6. Use statistical methods to establish safety and immunogenicity periodically

NATIONAL INSTITUTE OF  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai  
  
nie.gov.in

In case of a prevention trial, new vaccine candidate has been developed and there is a generally data available in the laboratory, also in the animal studies which you have, which support that, it is a safe vaccine candidate and it is able to generate some kind of a immune response against a particular disease and it is felt that it can be used as a candidate in humans for prevention against that particular disease. So, how we will go about this particular thing is, first is essentially to decide at the national, regional or local level, whether this vaccine is appropriate for the country and whether we want to evaluate this or not. Then develop a phase I trial design, here is where because we are talking about a phase I study now. A new vaccine is being developed; we have to start with a phase I trial design. We will have to find healthy adult volunteers, then depends on what type of vaccines. If it is a primarily disease in children, sometimes children could also be recruited in the study.

With then, we carry out screening to find out whether, who are the people who are willing to participate in the study are actually eligible for enrollment and then we randomize them into two groups. One group will receive the vaccine, the other will receive the placebo and then we will follow them up to find out the safety, how will the safety be evaluated? We will see what kind of reactions are occurring immediately and in the long run, say for a period 1, 2, one and half years. What kind of adverse events are

reported in both these kinds of individuals and then try to compare between the people who are receiving vaccine verses those who are receiving the placebos. Also, we will do some blood test, to find out what kind of immunogenic response has been stimulated in either of these two groups and compare them using standard statistical methods. So, that is an example of a prevention trial.

(Refer Slide Time: 17:18)

## Advantages & disadvantages of RCTs

### Advantages

- The only effective method known to control selection bias
- Controls confounding bias without adjustment
- Facilitates effective blinding in some trials
- Maintains advantages of cohort studies

### Disadvantages

- May be complex and expensive
- Lack representativeness - volunteers differ from population of interest
- Ethical challenges are immense



icmr.gov.in

So, even if there are impediments here, the clinical trials are the only way for making a progress in medical science because, if there are no clinical trials, no new drugs will come. If there are no clinical trials, no new technologies will be tested, no new vaccines will be tested and so they must be supported and the adequate information about clinical trials must be disseminated. The advantages of a clinical trial include that this is the only effective method known to control the selection bias of participants. Also, it controls confounding bias without any adjustment, facilitates effective blinding in some trials and maintains advantages of cohort study or a prospective study, but there are certain disadvantages of this. It is a very complex procedure, it requires thorough training all the investigators and also any clinical trial is very expensive.

As I mentioned earlier, we essentially use a sample from the total population that is eligible to participate in the trial and hence, there is some level of problem with respect

to generalizability of this particular finding. But, that is something which one has to accept and there are ethical challenges which are immense. So, we know that this is a difficult study to do, but we also know that progress will essentially be made if we have new drugs, if we have new therapies, if we have new technologies and if we have new interventions.

Thank you for your attention.

**Health Research Fundamentals**  
**Dr. Tarun Bhatnagar**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

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

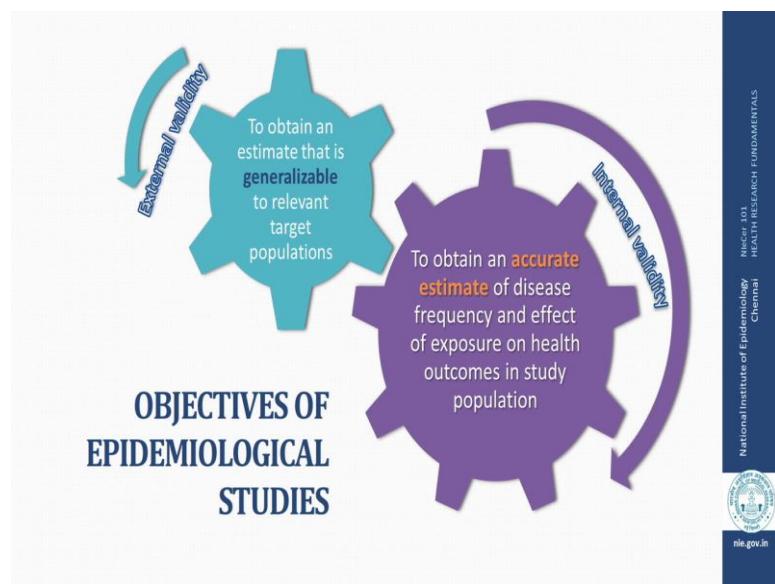
Hello, welcome to this section of Health Research Fundamentals. Today, we are going to talk about Validity in Epidemiological Studies.

(Refer Slide Time: 00:17)



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

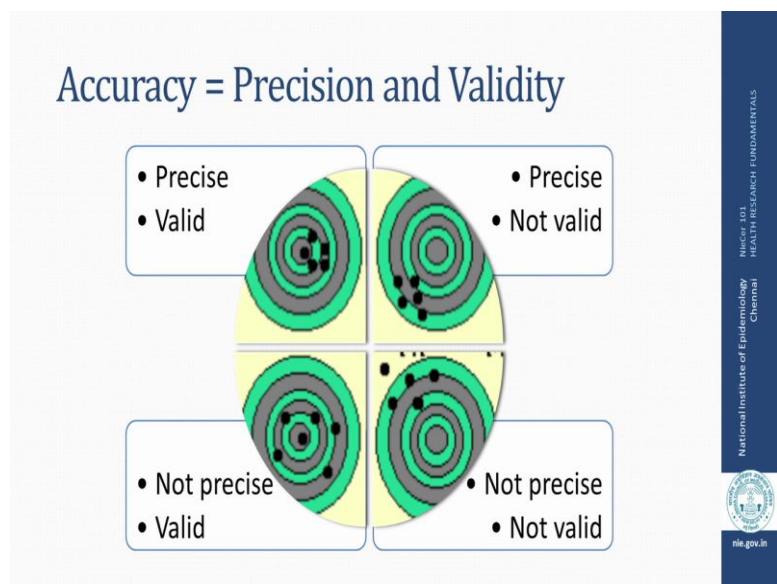
(Refer Slide Time: 00:42)



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 methodology that is being used to either estimate the frequency or determine the effect of an exposure?

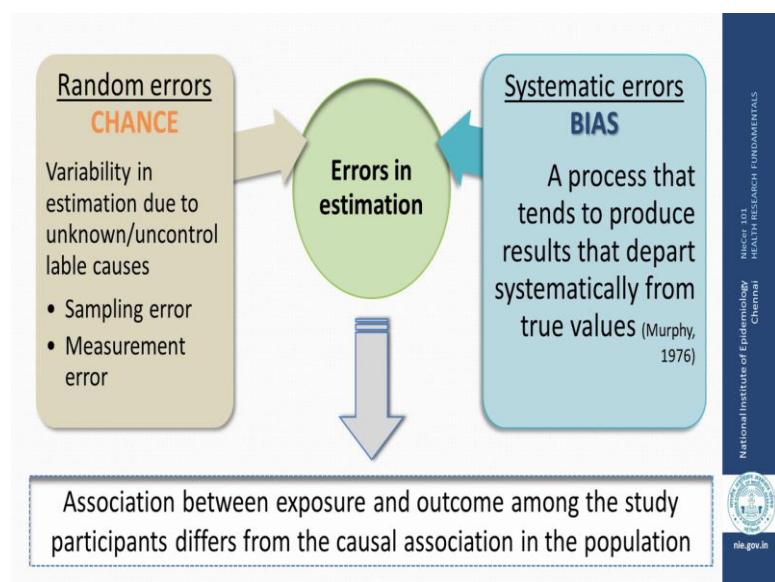
Now, in the long run what we would want is we would want that this estimate is generalizable to the relevant target populations among which the study is being done. Now, this aspect of any epidemiological study is known as external validity, so that the results of the study can be extrapolated to the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

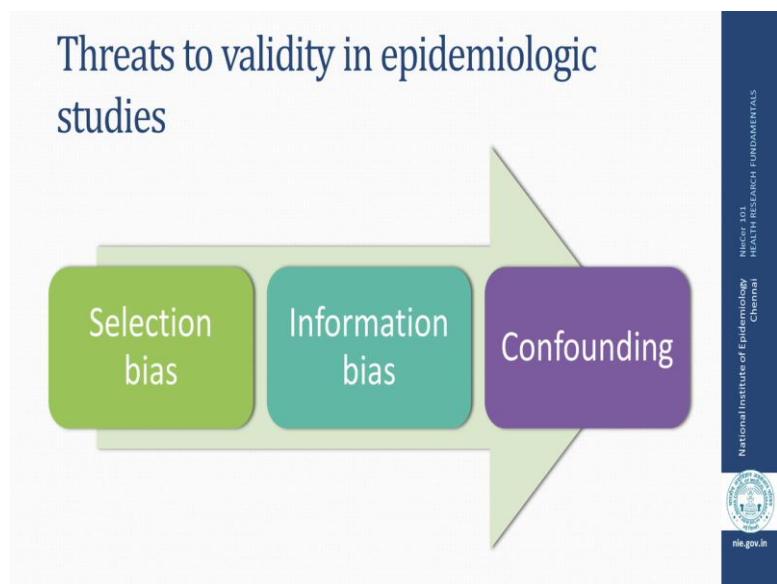


As I told you in epidemiological studies all that we are doing is, basically estimating. We are estimating, either the frequency of a disease or a health outcome or we are estimating what is the effect of an exposure on an outcome?

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. The more problematic error that we may face in any study, were are called systematic errors or biases. These are the errors that are basically a threat to validity of epidemiological study.

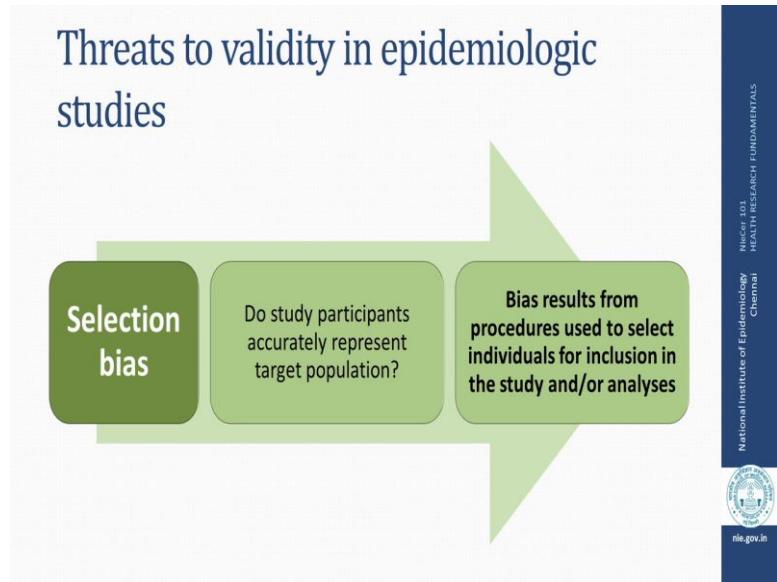
And, how do these errors happen in any study? Basically, the way we do the study, the methodology that we use to do epidemiological study and if it is done in a certain way that tends to produce results, that are not the true results, then that leads to errors which are called biases. Ultimately, what we would see is that, either the estimates or the associations that we are trying to access 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



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 making 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 have, 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 follow. Remember that, we have to follow a people over a period of time in cohort studies and if it is likely that people, who are less exposed or more exposed, they are more likely to be lost or people who are at more risk or at lesser risk. If they are more likely to be lost follow up, that eventually can lead to results that are biased and that would be attributable to selection bias.

(Refer Slide Time: 07:01)

## Dealing with selection bias

### Designing stage of a study

- Use incident cases, not prevalent cases
- Case control studies
  - Use population-based design
  - Apply same eligibility criteria for selecting cases and controls
  - Both cases and controls undergo the same diagnostic procedures and intensity of disease surveillance

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 controls studies are more prone to this, two selection bias and various ways in which, to deal with selection bias in case controls studies is to use population base design rather than

hospital base 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 in 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)



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, people 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

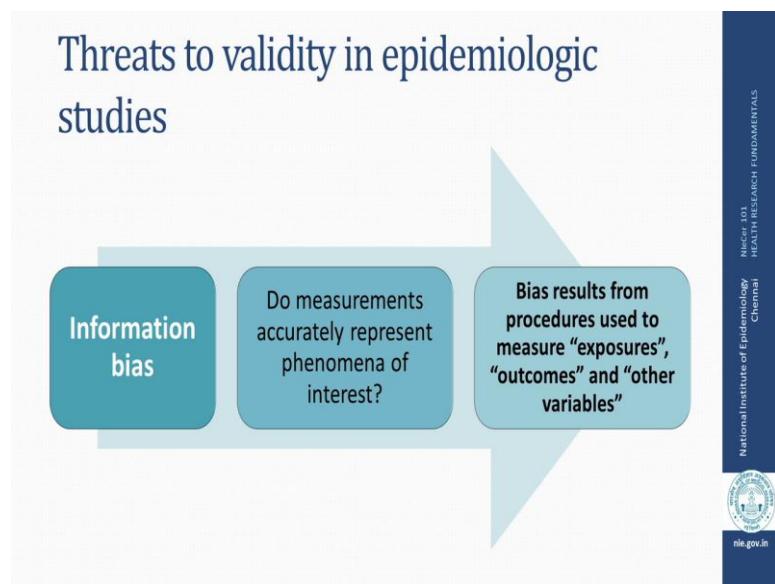
### Analysis stage of study

- 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  
ENVIRONMENTAL  
HEALTH SCIENCES  
NIEHS  
National Institute of Epidemiology  
Chennai  
nie.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. Large difference, 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 they 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 measures 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 outcomes present or absent is accurately measured and other variables, which is socio demographic age, gender, other 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

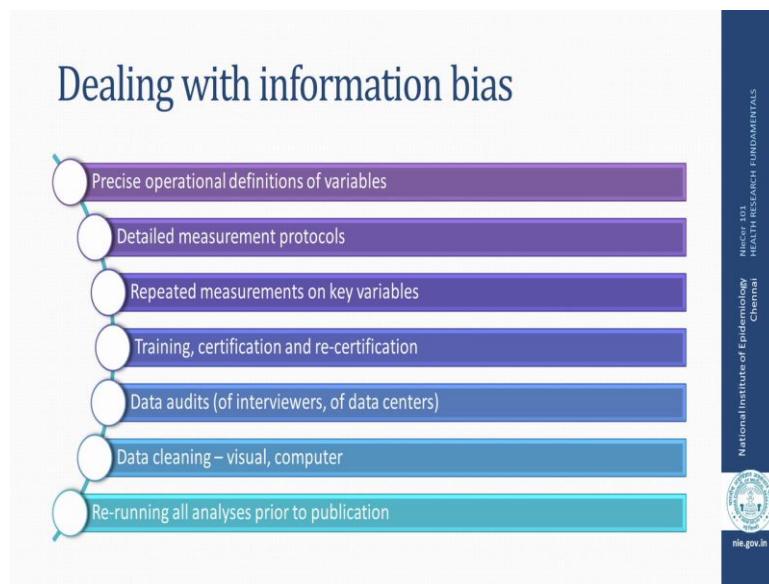
- **Investigator** - Systematic collection of information supporting expected conclusions (Unconsciously or Consciously)
  - May be examined in the analysis
- **Prevarication** - Systematic distortion of the truth by subjects

Now, how does this happen? In case controls 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 controls studies is through the process of recall. Remember that, we have cases and controls and we are trying to re 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 outcomes 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 study, 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 and 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, an observation epidemiological studies, we are dependent on what are study participant 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)



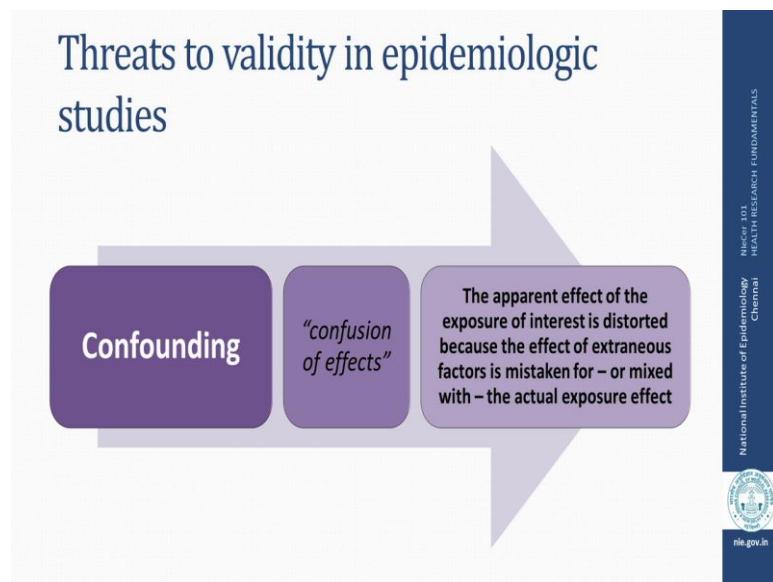
Now, how do we deal with information bias? First of all because we are measuring the exposure variable, the outcome variable and the other variables, we need to setup precise operational definitions of what we are going to measure and how much is it going to be. We need to have detailed measurement protocols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

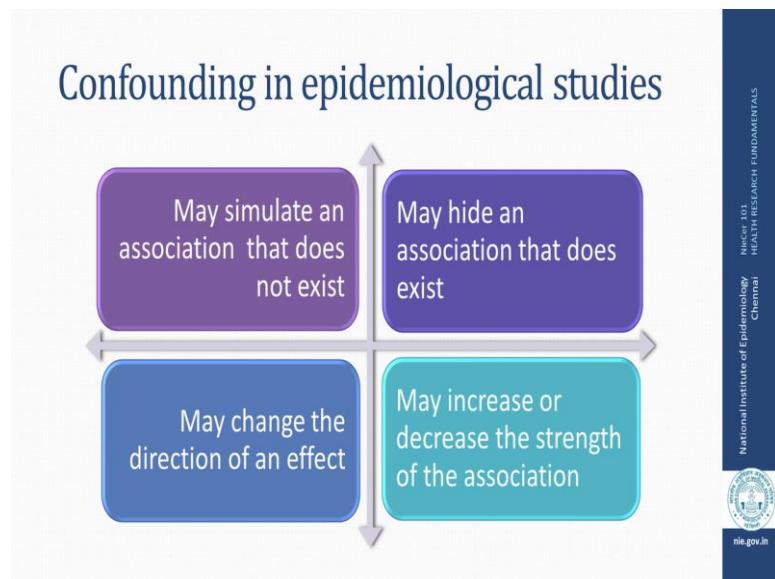
(Refer Slide Time: 15:22)



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

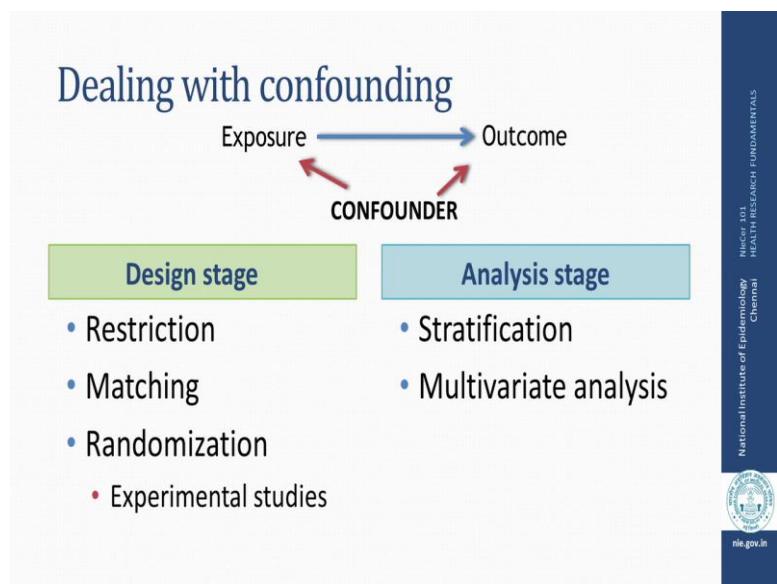
you are more exposed are you more likely to get the disease or vice versa. Now, what we want to know is the effect of this exposure on a particular outcome. This effect can be confused with the effect of a third factor, which can have an influence both on the outcomes as well as the exposure and this is what leads to what 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 associations 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 associations. So, you may say that an exposure is more associated with the outcomes or less associated with the outcomes 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 associations between the exposure and outcomes, this associations is influenced by this third factor. Now, we can deal with confounding, both at the designs stage and at the analysis stage. It is always better to deal it with the design stage than to take care of the analysis stage. So, at the design stage, we can do several things.

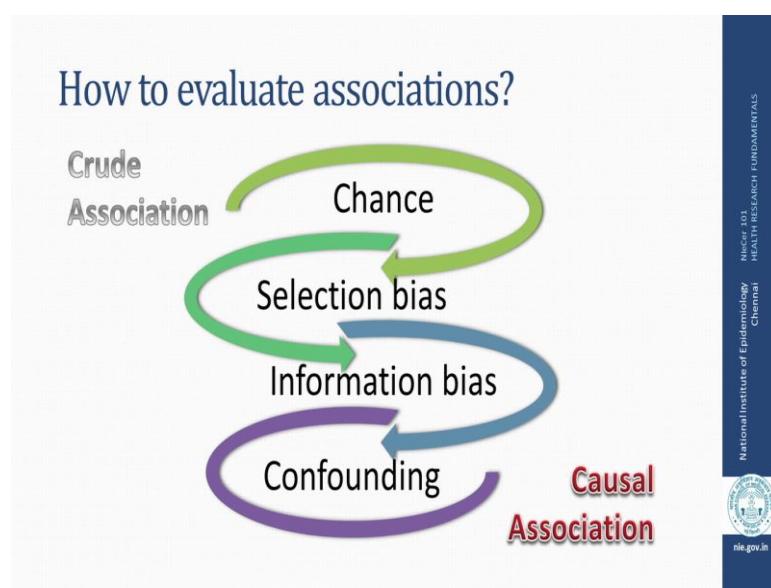
One, we can do what is called restriction; we can restrict a study participant to only those people who are in one straighten of the confounders, so that the confounders cannot play a role in the association between exposure and outcomes.

Secondly, we can do, what is called matching, is 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 make sure that the two arms in a randomized trials are similar 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 strata 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. 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.

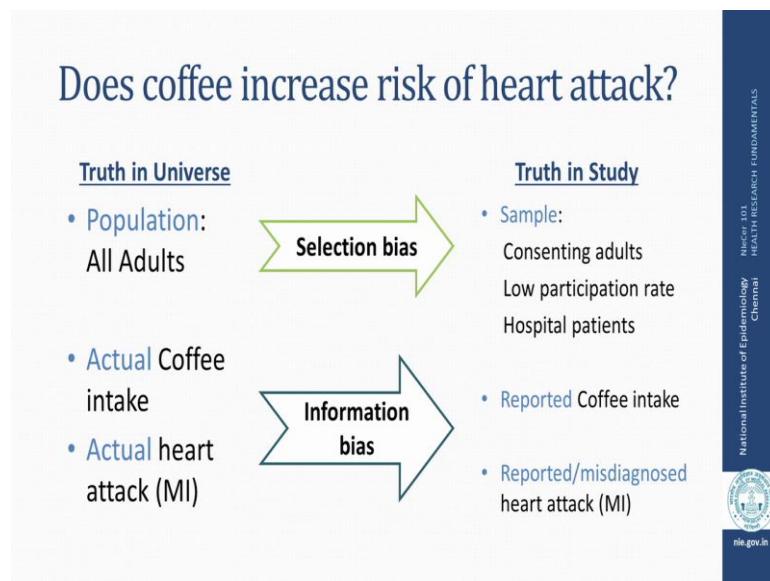
(Refer Slide Time: 19:49)



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 associations 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 and 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 taking care of. Only after going through this process, we would be able to say that whether the crude association is actually the causal 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 you agreed to take part in the study. Now, these people could be more 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 been misdiagnosed of a heart attack. There could be other, the chest pain that the study participants may report as heart attack, may actually have been due to other causes and that is reported as heart attack. So, actually what we may see, 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?

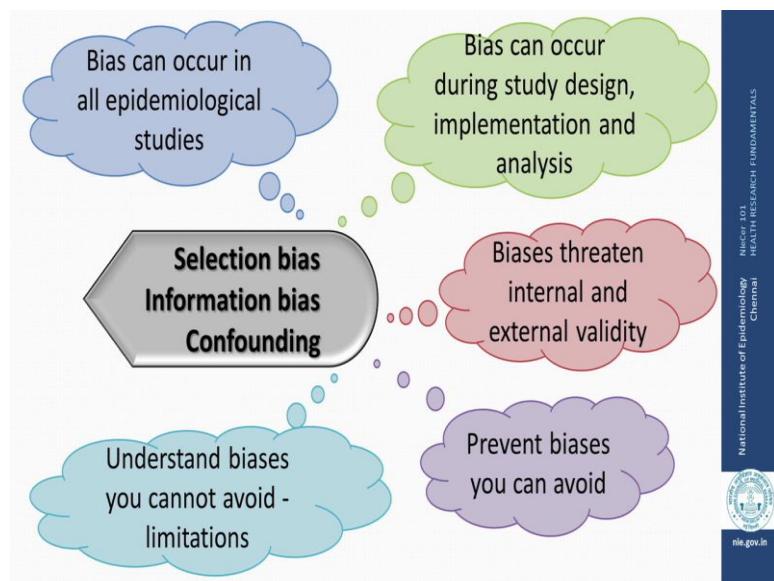
- Was the association between coffee and MI due to **CONFOUNDING** by smoking?
  - "A confounder is associated with both the exposure (coffee) and the outcome (MI)."
  - Smoking in
    - coffee (+) 86%
    - coffee (-) 27%
    - MI (+) 80%
    - MI (-) 40%

National Institute of Epidemiology  
Chennai  
nie.gov.in

Then of course, there is confounding. Could it be that this association that we saw between coffee and heart attack what we called myocardial infarction inside in medical terminology? Could it be confounded by smoking? Is it possible that those, 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 a heart attack. The association that we have seen 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, there is 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 epidemiological study and these biases can occur in all epidemiological studies more. In observational study, 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 threaten our study and design it appropriately and try to prevent as many biases as possible for 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 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.

**Health Research Fundamentals**  
**Dr. Tarun Bhatnagar**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**Lecture – 09**  
**Qualitative Research Methods: An overview**

Welcome to this session of Health Research Fundamentals. Today, I am going to give you a glimpse of another method of research and what we generally call as Qualitative Research Methods.

(Refer Slide Time: 00:21)

Quantitative versus Qualitative research methods		
	Quantitative	Qualitative
Data	Numbers	Text
View of the world	Social reality - objective, measurable, external to individual <b>ETIC</b>	Social reality – subjectively interpreted and experienced <b>EMIC</b>
Logic of enquiry	Deductive – testing formal hypotheses	Inductive – understanding of social processes derived from data
Research design	Ensures repeatability	Interpretation of responses
Validity	Objective (reliability)	Subjective (credibility)
Cross-cultural generalizations	Application of the same observation method to different cultures	Require conversion in abstract inter-cultural categories

National Institute of Epidemiology  
Chennai  
  
[nie.gov.in](http://nie.gov.in)

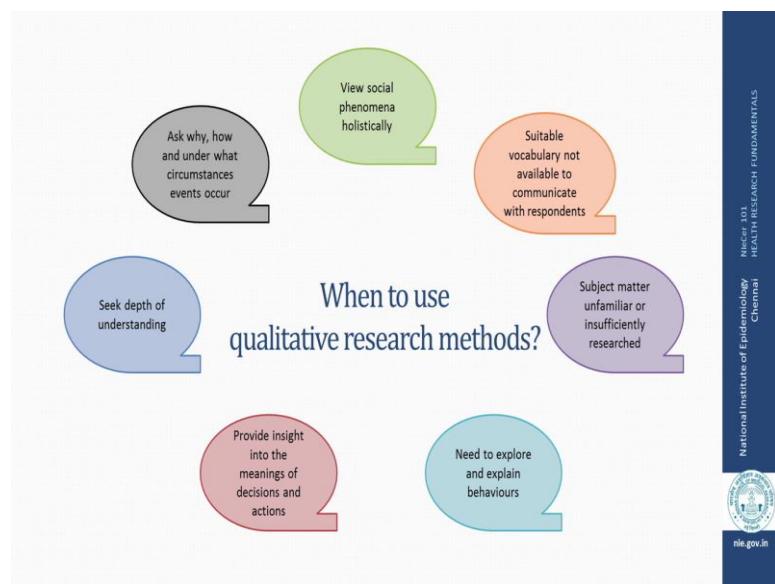
I am going to give you a snap short of what these methods are. In the paradigm of health research or any research for that matter, we generally divide methods into quantitative and qualitative. Qualitative methods generally find their origin in the Science of Anthropology, Sociology, Physiology, wherein which is more about dealing with human beings understanding their behaviors. When we look, sort of try to compare the quantitative with the qualitative methods some of the basic difference is that you would see is that the qualitative is always concerned with words, with text. So the data that we get in qualitative methods is actually text and not necessarily numbers that you usually see in quantitative.

In qualitative research, what we try to do is to interpret the social reality more from the participant point of view and their experience rather than measuring it objectively from the investigators experience. So, this is what is known as the EMIC perspective. In terms of the logic of inquiry, again qualitative methods are more inductive and they are used in the way to understand the processes that are derived from the data compared to the quantitative methods, wherein which is more deductive and where we try to test our formal hypothesis using the data.

In terms of the research design, qualitative methods are mostly about interpreting the responses of the participants, rather than trying to see whether what they are saying is what the investigator feels. Whereas quantitative methods ensure repeatability of the data and the methods are such that if there is a same studies done by somebody else we would try to expect similar kind of results.

Qualitative methods, the validity is more in terms of credibility of the responses that we get from the respondents and also the credibility of the investigator, who is doing this. Again, in terms of the generalization of the qualitative data that we get from these methods, it is more abstract and it is something that transits boundaries of cultures and we can understand different cultures in terms of what the participant wants to say. Whereas, in terms of quantitative methods, it is more application of the same methodology to different cultures trying to understand different cultures using the same methods.

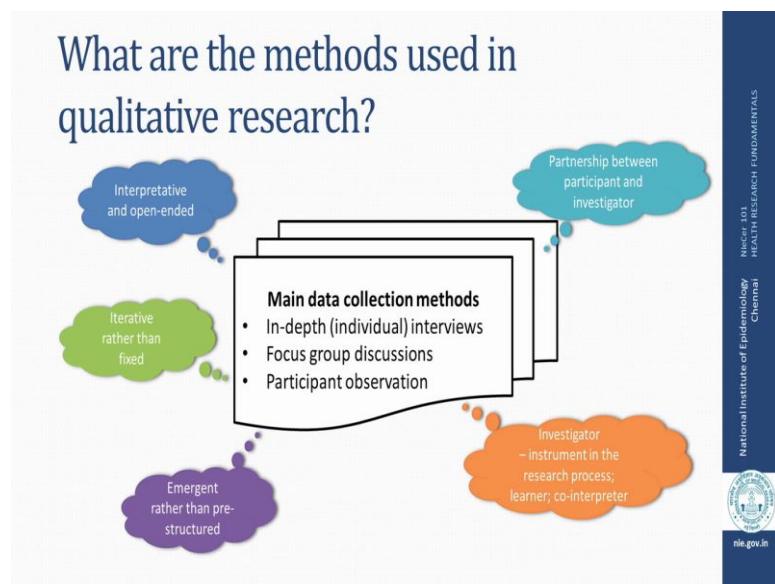
(Refer Slide Time: 02:56)



Now, when do you use qualitative research methods? Well, we would want to do that when we really want to understand, why the events happened? How did they happen? And what circumstances made them to happen? We would like to get more understanding of why these health phenomena are occurring? We want to provide insight into the meanings of the way people behave. We would like to know, why people are behaving in a certain way.

Sometimes the subject matter that we are studying is insufficiently researched and qualitative research methods help us to understand these methods. Many a times, if we get into new areas, maybe the subject matter vocabulary is something that we need to understand as an investigator and that is again where qualitative research would help us to do that. In terms of social sciences, qualitative research methods actually help us to view the social phenomenon more holistically in terms of how the participants see it through their eyes.

(Refer Slide Time: 04:02)



What are the various methods that we use in qualitative research? The various methods that we use are little different from the methods that we tried to use in quantitative research. In qualitative research, the methods are more interpretative and open ended, they are more iterative rather than fixed. The methodology emerges as the research goes on rather than using a pre structured format. In fact, qualitative research methods are done, so that it becomes a partnership between the participants and the investigators and the investigator actually becomes himself or herself, an instrument in the whole research process and it is more like a co-interpreter of what the participant wants to say. The main methods that we use in qualitative research include in-depth interviews, focus group discussions and participant observations. We will go through one by one through each one of them.

(Refer Slide Time: 05:01)

## In-depth (Individual) Interviews

NIHUEI IED  
Health Research Fundamentals

National Institute of Epidemiology  
Chennai  
nie.gov.in

- Open-ended interviews
  - Discover the interviewee's own framework of meanings
  - Obtains rich, contextualized, in-depth information
  - Avoid imposing the researcher's structures and assumptions
- Technique
  - Follows interview guide
  - Probes
  - Reflecting on remarks made by the informant
  - Collects respondent's perspective and words
  - Level of structure varies
- When to use?
  - Complex subject matter and knowledgeable respondents
  - Highly sensitive subject matter
  - Geographically dispersed respondents
  - Peer pressure an issue - social desirability a threat

So, in-depth interviews are basically open ended interviews, which are done one to one between the investigator and the participant and these are done to discover the interviews own framework of what they mean in their language. It is useful to obtain rich contextualized information and it also avoids imposing the investigators assumptions over what the participants wants to say. In terms of the techniques that are used in-depth interviews, although it is open ended and unstructured, we still follow what we call as an interview guide, which is a set of items that we would like to discuss and get information from the participants.

We use probes to get more in-depth understanding of the responses from the participants. In general, we collect the respondents prospective and the data as I told you before is more of the words and that which are recorded or noted down by the investigator. The level of structure how we go on with the interview varies depending on how the response comes from these participants and as I said these is more an emergent kind of methodology rather than a fixed pre-structured method.

When would you like to use in-depth interviews? Well, when the subject matter is complex and we want to know more about it from the respondents, when the respondents probably are more knowledgeable about the study that we are doing. Many a times for

highly sensitive subject matters like sexual behaviors or family planning issues, drug use, alcohol use things like that, in-depth interviews are a good way to interact more freely and get more in-depth information from research participants. Many a times, the research participants may be geographically dispersed and it is good to talk to them one by one in different time and place and to get more information. Of course, each method has its own advantages and disadvantages.

(Refer Slide Time: 07:03)

In-depth (Individual) Interviews	
Advantages	Disadvantages
<ul style="list-style-type: none"><li>• Most in-depth - Why behaviors are practiced?</li><li>• Data on how people think and talk; conceptualizations of behavior</li><li>• Exact words &amp; language people use amongst themselves</li><li>• "Emic" perspective = insider's perspective</li></ul>	<ul style="list-style-type: none"><li>• Based on a few people, usually not systematic sample, but purposeful or convenience sample</li><li>• Not generalizable</li><li>• Interviews very long, lots of data! Time consuming to analyze</li><li>• Researchers opinions of what the data means</li></ul>

Advantages, the in-depth interviews as per name are most in-depth. We really get to understand, why certain behaviors are practiced by individuals and why not? It gives us data on how people think? How do they conceptualize their own behavior? What is their context in terms of the way they do or think? We are able to get the exact words and language that people may use about the subject matter that we are trying to understand and it really gives us an insider prospective of the subject matter that we are trying to research on.

In terms of the disadvantages, since we do few in-depth interviews and we do not do it on many people like 100 or as we may do in quantitative methods, so there is no specific sampling. The sampling is more convenient and purposeful and we try to get people who are more knowledgeable and who will be able to give us information. This makes the

findings from these interviews not generalizable in the strict quantitative sense. Many a time, the interviews usually go very long may be 40 minutes, 40 even go, sometimes may go to more than an hour or so and you get a lot of data which means a lot of words and it could be time consuming to actually analyze these words. Since, the analysis is more interpretative, sometimes there is a possibility that what the data means the interpretation could depend on how the researcher feels and interprets this data rather than how the participant would have wanted it to be known. So these are some of the pros and cons of doing in-depth interviews.

(Refer Slide Time: 08:55)

## Focus Group Discussions

- Open-ended group interviews that promote discussions between participants on specific topics
- Usually 6-8 'similar' participants
  - Similar age, gender, socio-economic status, education, others...
    - Similar cognitive structures
    - Similar perceptions of their social environment
    - Similar normative beliefs
- Moderator and note-taker
- Flexible interview guide
- When to use?
  - Group interaction important
  - Cost and timing issues
  - Idea generation
  - Problem identification and definition goal
  - Identify local/group specific vocabulary/terminology
  - Evaluating messages for an intervention

The other method is Focus Group Discussions. Again these are open-ended interviews, but instead of individuals, now we have a small group of people who discuss a certain topic amongst themselves. Usually, we have 6 to 8 participants, who are a homogeneous group, which means they are supposed to be similar in terms of various characteristics like age or gender or socio economics status or education, occupation and this helps us so that the discussion goes on in a more cohesive manner.

Usually, in a focus group discussion apart from, you would have a moderator, who moderates the whole discussion and another individual, who works as a note taker in order to take down the notes that are being discussed. We could also have recordings,

video or audio recordings of this focus group discussions, which could be used to analyze it later. Again, as in in-depth interviews we use an interview guide, which is basically a set of items that we would like in the participants to discuss and we make sure that all the items are discussed in the way that we want them to be. However, the interview guides are flexible in the sense of, we do not have to follow question one, question two, question three, it basically flows as the discussion flows and the moderator is able to guide the discussion in order to obtain all the information that is there to obtain.

When do we use Focus Group Discussions? These are used in areas where group interactions give us a lot of rich information. Again, cost and timing may be issues if you were to in-depth individual interviews, we could get information from a lot many more participants in smaller time compared to in-depth interviews. It helps us to generate ideas because we have a group of people, who more than one person whose able to respond to the various ideas that we are trying to generate. It helps to identify problems and define goals. Many a times depending on what you want to understand local terminologies and the vocabularies, then these focus groups are good way to understand these issues. And if you want to again look at, if there has been an intervention that has been put into place and we would like to evaluate, whether it has worked or not worked then focus group discussions are a neat way of understanding these issues.

(Refer Slide Time: 11:31)

## Focus Group Discussions

### Advantages

- Some people more comfortable and talk more openly in group settings
- Natural way for some people to talk about problems and personal issues in some cultures
- Collects information on social norms
- Can provide lots of data in a limited amount of time

### Disadvantages

- Difficult to access practice of very personal or sensitive behaviors in groups
- Not GENERALIZABLE
  - Subject to dominant personalities
  - Sensitive to biased analysis
- Transcribing time consuming - often 30-40 pages each
  - Difficult to identify speakers
- Analytic challenge!

Of course, the pros and cons, we know that there are some people, who are more comfortable talking openly in group sittings, and it may be a natural way for some people to talk and so it is a nice way to understand their problems and personal issues. It is good way to collect information on social norms, where people can discuss these issues amongst themselves. And as I said earlier it provides a lot of data in a limited amount of time compared to doing individual in-depth interviews.

However, the disadvantages are that sense we are talking to a specific group of people, the data that is generated will depend on the actual make-up of this group and it may be difficult to access the practice of some very personal or sensitive behaviors because people may not like to talk about them in groups. Again, since these are groups of few people, the data may generally not be generalizable to the bigger target population because what we get in the information is more subject to dominant personalities may be people, who are more talkative their views may be heard more than people who do not like to talk as much. Of course, as with in-depth interviews there would be a lot of data that is generated, many a times it runs into pages and transcribing this can be time consuming, and of course it gives a challenge in terms of how we are able to analyze this data.

(Refer Slide Time: 13:07)

## Participant Observation

- The researcher becomes participant in social event or group under study and records observations
- Advantages
  - Data is very deep and detailed
- Disadvantages
  - Difficult to systematically collect; especially in middle of important moment - hard to take notes so details may be forgotten
  - Analytic methods for observation notes not well described

The third more common method that is used in qualitative research is what is called Participant Observation. Now, this comes more from ethnographic and anthropological domain, wherein the researcher himself becomes a participant in a social event or group that we are trying to study. The good thing is that we get very deep and detailed data because you as a researcher, you are part of whatever is happening in that group. However, it sometimes becomes difficult systematically collect the data because you are right in their and it may be hard to take notes or do recordings and again the analytical methods that are used for participant observation data are still evolving and that can again be a challenge in terms of the analysis.

(Refer Slide Time: 13:57)

## Qualitative data (text) analysis

<b>Grounded theory</b> <ul style="list-style-type: none"> <li>• Transcripts of interviews</li> <li>• Potential analytic categories— themes</li> <li>• Coding text into categories</li> <li>• Relations among categories</li> <li>• Build theoretical models</li> <li>• Exemplars - quotes from interviews</li> </ul>	<b>Content analysis</b> <ul style="list-style-type: none"> <li>• Theoretical framework</li> <li>• Set of codes for variables in the theory</li> <li>• Applying codes systematically to a set of texts</li> <li>• Unit-of-analysis-by-variable matrix from the texts and codes</li> <li>• Statistical analysis of matrix</li> </ul>
--	--

Bernard HR. Research methods in anthropology : qualitative and quantitative approaches. 2006

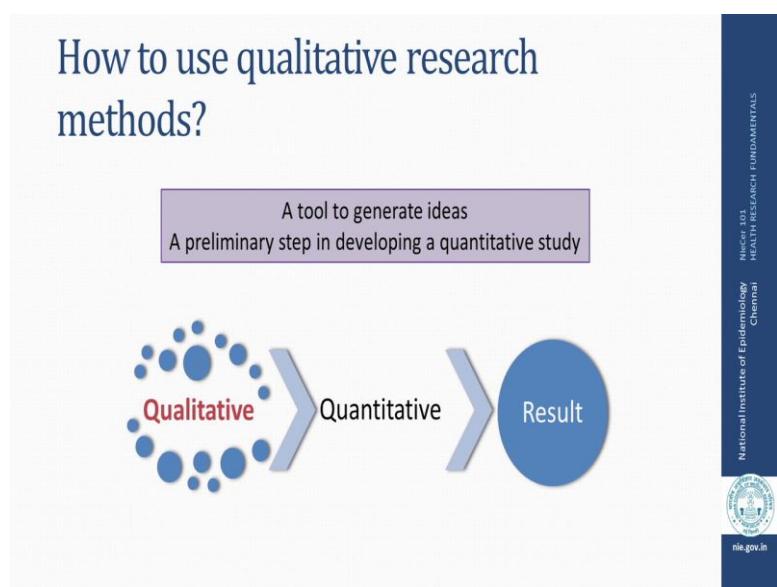

  
 National Institute of Epidemiology  
 Chennai  
[nie.gov.in](http://nie.gov.in)

Now coming to analysis, there are various ways in which qualitative data can be analyzed and as I told you earlier, qualitative data means text. The basically there are two ways in two methodologies that have been known to be used in terms of analysis, one is called a Grounded theory approach and then there is a Content analysis approach. The grounded theory approach goes from understanding the data and then developing theories based on that. Whereas content analysis is when you start of with the theoretical framework and then you try to analyze the data to understand the theory.

What is common in both of these methods is that, first of all you need to transcribe the

interviews into text, maybe from local language into English or whatever is the language of their investigator. There is a coding of themes and categories and then relationships are built based on these themes and categories and these are used to understand the perspective of the participants. And then codes are used, just as we have tables with data in quantitative research. In qualitative research, if you may have seen certain, may be papers from that, you would usually see codes from the interviews which are used as examples to illustrate the main theme that is generated from this data.

(Refer Slide Time: 15:31)



How do we use qualitative research methods? We could use qualitative research methods in different ways, it could just be used as a tool to generate ideas and basically we try as a preliminary step in developing a quantitative study. For example, if say you are trying to understand why people go for say open defecation, you could, you may be developing a survey to understand this process. But, in order to understand the various categories of questions, we may do a small qualitative study to understand what could be the probable reasons and then use it to modify the questionnaire for the quantitative survey on this subject.

(Refer Slide Time: 16:14)

## How to use qualitative research methods?

To help understand the results of a quantitative study

The diagram illustrates the relationship between quantitative and qualitative research methods. A large blue arrow labeled 'Quantitative' points towards a large blue circle labeled 'Result'. Below this arrow, a smaller blue arrow points upwards from a cluster of blue dots labeled 'Qualitative'.

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

It could be used to help understand the results of quantitative studies, so you done a qualitative study first and then you get results which you need to get an in-depth understanding of why you are getting these kinds of results. And again qualitative methods could help you to do that or you could use qualitative research methods as the primary data collection method and this would be where, what you the objective of the study is to actually understand behaviors, get an in-depth understanding of why people do or not do certain things, why certain behaviors are practiced or not practiced? And you get a lot of rich data in terms of using qualitative data itself.

(Refer Slide Time: 16:31)

## How to use qualitative research methods?

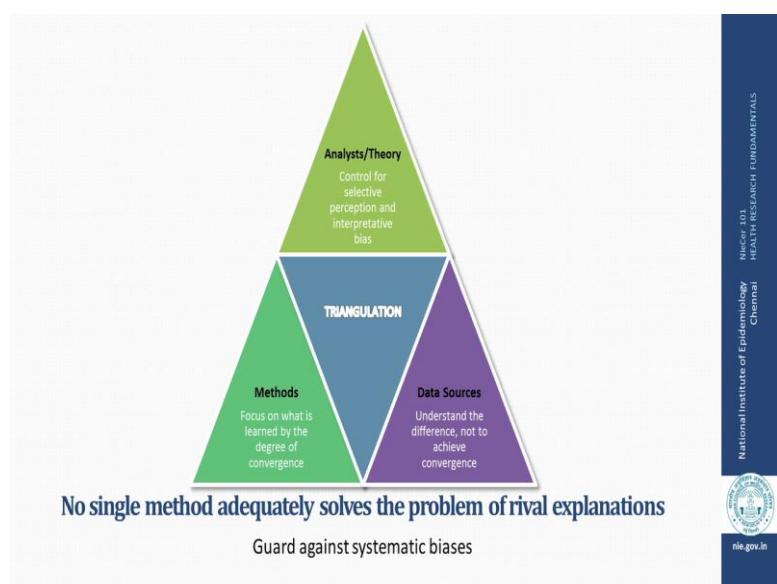
The primary data collection method  
– sometimes but not necessarily along with a quantitative study like a survey

A diagram illustrating the relationship between qualitative and quantitative research methods. On the left, the word "Qualitative" is written above a cluster of blue dots. On the right, the word "Quantitative" is written below a cluster of blue dots. In the center, there is a large blue circle labeled "Result". Two arrows point from the clusters towards the central "Result": one arrow points from the "Qualitative" side, and another arrow points from the "Quantitative" side.

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

Sometimes, but necessarily it can be done along with the quantitative study also.

(Refer Slide Time: 17:02)



Now, what is more important is that, no single method, whether it is quantitative or qualitative, would be able to give you all the explanations. They could, even you may get different explanations or different results when you use different methods plus as an

investigator as it is important in any research study, we need to guard ourselves against systematic biases in terms of data collection and in terms of actually interpreting that data. So, it is a good thing to what we call as what we do as triangulation, which is basically trying to understand the same topic from different angles through different methods using different theories and using different data sources.

There could be a triangulation of analysis or theory, which helps us to control for any selective interpretative biases, we could use multiple methods and then focus on what information that, what is the same kind of information that we get from these methods or we could use different data sources for the same study and then try to understand the difference is that emerge to these data sources and then put of these together to actually get comprehensive understanding of the research topic that we are trying to do.

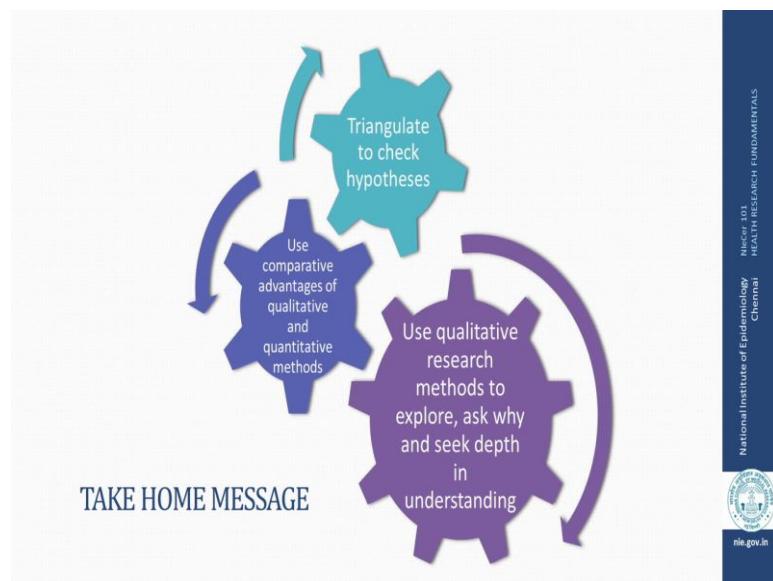
(Refer Slide Time: 18:24)



How are these methods useful? In general, qualitative research methods are very good to actually identify the determinants of health, why do people behave in a certain way? What are the attitudes? What do they perceive in terms of what they are doing or not doing? If there have been certain interventions already done quantitative research methods are a good way to understand, why these interventions have been successful or not successful. Qualitative research methods are again a good tool to explain may be

various problems that may have a reason in terms of, why people make certain choices? Or why people use certain services or not use certain services? And in general overall again there are good way to understand the context in which certain decisions are made whether at the policy level or a social or a legal level.

(Refer Slide Time: 19:22)



So, just to sum up, we have the 2 paradigms of research, qualitative research and quantitative research. Qualitative research is a very good way to explore, to understand why and to seek a more depth in understanding of the topic that we are trying to study. While we are doing that it is always good to use the advantages of both qualitative and quantitative methods to enrich our research process. And of course, triangulate the various methods to check whether the analysis that we are getting is something that can be enriching and fruitful.

Thank you.

**Health Research Fundamentals**  
**Prof. Dr. R Ramakrishnan**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**Lecture – 10**  
**Measurement of study variables**

Welcome to the session of Health Research Fundamentals. We often hear statements like I have data on 100,000 Leprosy patients, I have data on climatic change in Chennai city, I have data on road traffic accidents and so on. In all these statements, there is a word data. In this session, we are going to rather see what this data means, what are the different types of data? And how to convert these data into pieces of information?

(Refer Slide Time: 00:49)

## Types of Data

- Qualitative
  - Nominal
    - Eg. Color of Eyes
  - Ordinal
    - Eg. Stages of disease condition
- Quantitative
  - Discrete
    - Eg. Family size
  - Continuous
    - Eg. Height / Weight

NICMAR IIT  
HEALTH RESEARCH FUNDAMENTALS

National Institute of Epidemiology  
Chennai



nicmar.gov.in

Data can broadly be classified into Qualitative and Quantitative data. Qualitative data as the names just, we cannot quantify them, it is on some sort of a quality. Again, this qualitative data could be a Nominal data or an Ordinal data. The nominal data, the examples are the color of eyes, the different regions of a city and so on and the ordinal data are the data which can be arranged in a sort of an order like examples you know stages of disease condition. Quantitative data again are of two categories, one is a discrete data which essentially is a full number, a number of siblings, family size,

etcetera and the other one is a continuous data, where is a continuous measurement like height and weight. So, these are all different types of data which requires different type of an identical skill.

(Refer Slide Time: 02:03)

## Describe - Central Value

- Data is not information.
- Summarize
  - Average
    - Mean
    - Median
    - Mode

Now, our aim is to get some information out of data. The large set of data, it is very essential but still looking at just the data, you cannot aggregate any information. So, we need to summarize them. One of the ways of summarizing the data is to get a value of as sort of an average. Now, average you mean, the first average that comes to our mind is the Mean.

(Refer Slide Time: 02:37)

## Arithmetic Mean (AM)

- Most commonly used; Simply called MEAN
- Add all the observed values ( $\text{Sum} = \sum X_i$ )
- Mean = Sum / n
- Sample Mean is denoted by  $\bar{x}$
- Population Mean is denoted by  $\mu$

NICER IOT  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

A mean which is also called an Arithmetic Mean, this is a most commonly used and simply it is called Mean. You add all the observed values, we call that as sum which is sigma  $X_i$  in a mathematical notation and mean is nothing but divide the sum by the number of observations you have used in your calculations, which is  $n$ . The sample mean is denoted by an  $x$  bar, line on top of  $x$  and population mean is denoted by  $\mu$ .

(Refer Slide Time: 03:13)

## Example

- Age of 10 Pregnant women
- 26, 31, 25, 21, 26, 26, 27, 25, 27, and 26
- $$\text{Sum} = (26+31+25+21+26+26+27+25+27+26) = 260$$
- $$n = 10$$
- $$\text{Mean} = \text{sum} / n = 260/10 = 26 \text{ years}$$

NICER IOT  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

Let us rather see an example. Suppose, there are 10 pregnant patients who had visited an ANC clinic and their ages are 26, 31, 25 and so on and what is a mean age of these pregnant woman? The mean is got by summing up all the ages which comes to 260. There are 10 observations, so divided it by 10. It is 260 by 10 which is equal to 26. We say the mean age of pregnant woman who visited the ANC clinic is 26 years.

(Refer Slide Time: 03:50)

## Describe - Central Value

- Data is not information.
- Summarize
  - Average
    - Mean
    - Median
    - Mode



NIE 1981  
National Institute of Epidemiology  
Chennai  
[nie.gov.in](http://nie.gov.in)

Now, one of the problem with this average mean is, some extreme values either big or small even one or two, if they are present in your data set, that could influence on their average because you are adding all and one big value, if we add the whole mean becomes an over estimation. In order to control this or in order to avoid this, we have another measure which is called Median.

(Refer Slide Time: 04:22)

## Median

**The Median describes literally the middle value of the distribution**

**Divides the distribution exactly into two halves  
(i.e. 50% of the data will fall on either side)**

**Useful when there are extreme values**



The median is, literally the middle value of the distribution. It divides the distribution exactly into two halves that is 50 percent of the data will fall on either side. This is a very useful measure, especially when you have extreme values.

(Refer Slide Time: 04:42)

## Example

**Duration (days) of hospital stay of 11 patients**

**1, 2, 3, 4, 5, 6, 7, 8, 8, 9, 77 (Arranged in ascending order)**

**Median is the middle value (6<sup>th</sup> value) = 6**

**(Mean = 11.8)**

**If n is even; then take average of middle two values.**



Let us just see this example. Suppose, you have a data on the duration of stay in hospital

of 11 patients, the duration is 1 day, 2 days, 3 days, and 9 days for 10 patients and then the 11th patient, it is 77 days. Of course, I have arranged this data in an ascending order. The median is the middle value, which is a 6th value, the value you get  $n + 1$  divided by 2, 11 plus 1, 12 divided by 2 is 6. So, the 6th value is the value 6, which means the median is 6 here. Whereas, when we really compute the mean for this it comes out to be 11.8. As you could rather see 6 is more appropriate measure of average in this case rather than the mean 11.8. If  $n$  is given, then you take the average of middle 2 values.

(Refer Slide Time: 05:41)

## Describe - Central Value

- Data is not information.
- Summarize
  - Average
    - Mean
    - Median
    - Mode

NIECR 101  
National Institute of Epidemiology  
HEALTH RESEARCH FUNDAMENTALS  
Chennai  
  
niecr.gov.in

Now, there is another measure which is called Mode.

(Refer Slide Time: 05:48)

## Mode

**The Mode** is the value that occurs most frequently

Mode is the only location statistics to be used –  
for nominal data - not measurable  
characteristic

Epidemiology – Describing an epidemic with  
respect to TIME



Mode is the value that occurs most frequently. In fact, mode is the only location statistics which we can use for nominal data, which are not measurable. In epidemiology, we do use more quite often. In an epidemic curve with respective time, we look for the modal class and then that gives an idea of the incubation period of the pathogen.

(Refer Slide Time: 06:17)

## Example

- Colour preference of people for their car

<u>Colour preference</u>	<u>No. of persons</u>
Green	354
Yellow	852
White	310
Red	474

Mode = Yellow



The example for a mode is the color preference and the number of persons, 354 people they prefer green, 852 prefer yellow, 310 prefer white and 474 prefer red. So maximum number of people they prefer yellow, answer the mode of class is yellow. As you could rather see as in the respect of rather mode, there can be multiple modes, there cannot be a mode at all in a (Refer Time: 06:42). Suppose, if all the values are 354 here, then there is no mode. So, mode can exist, there can be multiple modes in a data set.

Now, we have seen Mean, Median, Mode are 3 good measures of summarizing your data to get an average value.

(Refer Slide Time: 07:08)

## Describe – Dispersion

- Is it enough to know the average?
  - Example of swimming pool.
- Measures of variability
  - Range
  - Inter-quartile range
  - Mean deviation from mean
  - Variance / Standard deviation

So it is not enough you just rather know the average value. Say for example, you go to a swimming pool and you do not know swimming and you are 5 feet 7 of inches and then, if the pool managers says the average depth of the swimming pool is 4 and a half feet you feel very comfortable and you jump, and suppose a place where you jump is 9 feet then may know the thing that you missed to ask is, yes the average is 4 and a half feet, but what is the variability? There maybe you know place where it is shallow as 3 feet and as depth as 9 or 10 feet. So, you need to rather ask, what is the variability? One of the measures that comes to our mind is the Range.

(Refer Slide Time: 07:57)

## RANGE

### Definition:

The difference between the Minimum and the Maximum value of the observations

### Advantage:

A quick and easy indicator of dispersion.

### Disadvantage:

Influenced by extreme values; Uses only two data points

NIECR IDI  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

The range is the difference between the minimum and then the maximum value of the observations. An advantage of this measure is it is very quick and easy indicator of dispersion. But, has I had said about the mean, the range also is influenced by extreme values and also we consider only two values, the first and then the last and in between we are not using the data at all and that is the great disadvantage of range. There is another value which is called Inter-quartile range.

(Refer Slide Time: 08:30)

## Describe - Dispersion

- Is it enough to know the average?
  - Example of swimming pool.
- Measures of variability
  - Range
  - Inter-quartile range
  - Mean deviation from mean
  - Variance / Standard deviation

NIECR 101  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

(Refer Slide Time: 08:34)

## INTER-QUARTILE RANGE

### Quartile Deviation

#### Definition:

Defined as the interval between the value of the upper quartile (Q3) and the lower quartile (Q1)  
 $\text{Inter Quartile Range} = Q_3 - Q_1$

NIECR 101  
HEALTH RESEARCH FUNDAMENTALS  
National Institutes of Epidemiology  
Chennai



nie.gov.in

#### Advantage:

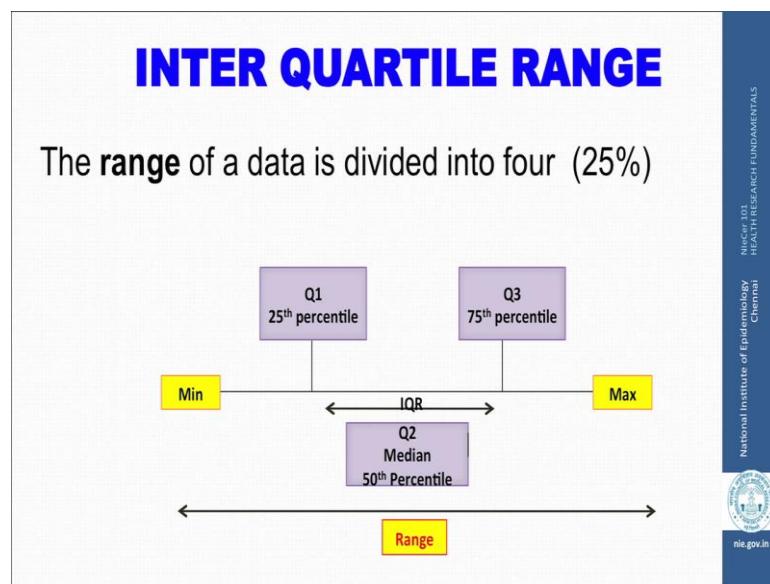
Unaffected by the extreme values

#### Disadvantage:

Covers only the middle 50% observations

This to a large extend take care of this extreme values in the sense, we divide the data sets into 4 quarters.

(Refer Slide Time: 08:38)



And we try to remove the first quarter and then the last quarter and consider only the middle 50 percent of values and this inter-quartile range is the Q3 minus Q1 and a great advantage of this is, this value does not rather get affected by extreme values. But again the disadvantages is, it covers only the middle 50 percent of the value and then the same disadvantage that we had for range that uses only two values and in between values are not made use of and that is a great disadvantage of this value.

(Refer Slide Time: 09:15)

## Describe - Dispersion

- Is it enough to know the average?
  - Example of swimming pool.
- Measures of variability
  - Range
  - Inter-quartile range
  - Mean deviation from mean
  - Variance / Standard deviation

NICER ICD  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

Another measure of variability is Mean deviation from mean. What you mean by that? Say for example, from your data set, every data point we try to subtract a mean and then we try to take average of these mean deviation, which is called Mean deviation from mean.

(Refer Slide Time: 09:40)

## MEAN DEVIATION

**Definition:** The mean deviation is the average of the absolute (ignoring the sign) deviations of the observations from the arithmetic mean.

**Advantage:** It is based on all the observations in the group. It is easy to grasp the meaning of the procedure.

**Disadvantage:** It ignores the sign of the difference of the value of the observation and arithmetic mean.

It is not widely used because of the availability of a more advantageous measure.

NICER ICD  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

One of the problems with these is, if you rather do with that what happen is there are some values which are less than the mean, some values which are more than the mean and if you do the summation of all these, you get a value 0. So, mean deviation from mean is always 0. In order to get over that, what we do it is, we ignore the sign and then we just take the difference and then we take the average. This called Absolute mean deviation, an advantage it is based on all observations in the group it is easy to grasp the meaning of the whole procedure. But the disadvantage is, it ignores the signs of the difference of the value and it is mathematically it is not very vigorous to use this value.

(Refer Slide Time: 10:37)

## Describe – Dispersion

- Is it enough to know the average?
  - Example of swimming pool.
- Measures of variability
  - Range
  - Inter-quartile range
  - Mean deviation from mean
  - Variance / Standard deviation

So, in order to get over that we have another measure, what we do is? We do take the difference of each observations from mean and instead of ignoring the sign, we square them because square takes care of even the minus and then the plus everything becomes plus and then we take an average of that, that value is called Variance.

(Refer Slide Time: 10:59)

## STANDARD DEVIATION - SD ( $\sigma$ )

**Definition:** The SD is the square root of the average of the squared deviations of the observations from the arithmetic mean

The square of the SD is called variance

**Advantage:** The SD is the most important measure of distribution. While the variance is in unit squared, the SD is expressed in the same units of measurement as the observation. It is suitable for further analysis

The SD together with arithmetic mean is useful for description of the data

NICER IRI  
HEALTH RESEARCH FUNDAMENTALS



nie.gov.in

And since variance is, we are squaring and then the measurement also squares we take a square route at the end and that is called Standard Deviation. The standard deviation, which is denoted as SD is a square route of the average of the squared deviations of the observations from the arithmetic mean. The square of the standard deviation is the variance. Advantage of standard deviation is most important measure of distribution, while the variance is in unit square, the standard deviation is expressed in the same units of the measurement and it is suitable for further analysis. So, standard deviation together with arithmetic mean is useful for describing the data and these two measures are extensively used for further treatment of your data set.

(Refer Slide Time: 11:51)

## Coefficient of Variation (CV)

Purpose: To compare the relative variability in different groups

Definition: The coefficient of variation is the SD expressed as a percentage of the arithmetic mean (AM).

$$CV = \left( \frac{SD}{AM} \right) \times 100$$

NICER IRI  
HEALTH RESEARCH FUNDAMENTALS



nie.gov.in

I am going to introduce to you one more measure which is called Coefficient of Variation. The purpose of this measure, suppose if you have different groups, different data sets to compare and then you want to rather compare the relative variability in different groups. So, the coefficient of variation is the standard deviation expressed as a percentage of arithmetic mean because the standard deviation by arithmetic mean, what happens this is, they both are the same units of the measurements, so units of measurement get canceled, so what you get is a pure number and that number expressed in terms of percentage that is multiplied by 100, you get coefficient of variation.

(Refer Slide Time: 12:38)

## Summary

- Choose appropriate central / dispersion value
  - Mean / SD – if no extreme values
  - Median / IQR – if there are extreme values
  - Mode /Range – for qualitative variables/ time distribution in epidemic curve
- Mean and SD are used the most.

NICER IRI  
HEALTH RESEARCH FUNDAMENTALS



nie.gov.in

In summary, we have to choose an appropriate central or dispersion values. The Mean and Standard Deviations are the most appropriate central and dispersion values especially, if there are no extreme values. If there are extreme values, there are methods of still using mean and standard deviation using some transformations of your data that requires (Refer Time: 13:06) expert handling of your data. Otherwise, you go in for median and inter-quartile range and these two measures median, inter-quartile range do take care of extreme values. The mode and range is normally used for qualitative variables, time distributions in epidemic curve. The mean and standard deviations, as I said are the most used measures of variability and the summary statistics.

Thank you.

**Health Research Fundamentals**  
**Dr. R Ramakrishnan**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**Lecture - 11**  
**Sampling Methods**

Welcome to Research Fundamental Modules, the nicer course one. Today, we are going to talk about some aspects of Sampling Methods.

(Refer Slide Time: 00:22)

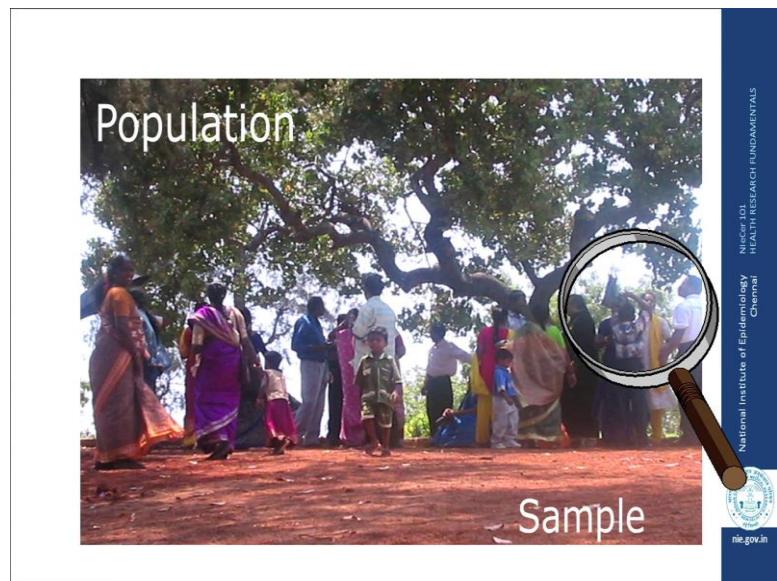
**Definition of sampling**

*Procedure by which some members of the population  
are selected as representatives of the entire  
population*

NIEC-IQI  
National Institute of Epidemiology  
Chennai  
nie.gov.in

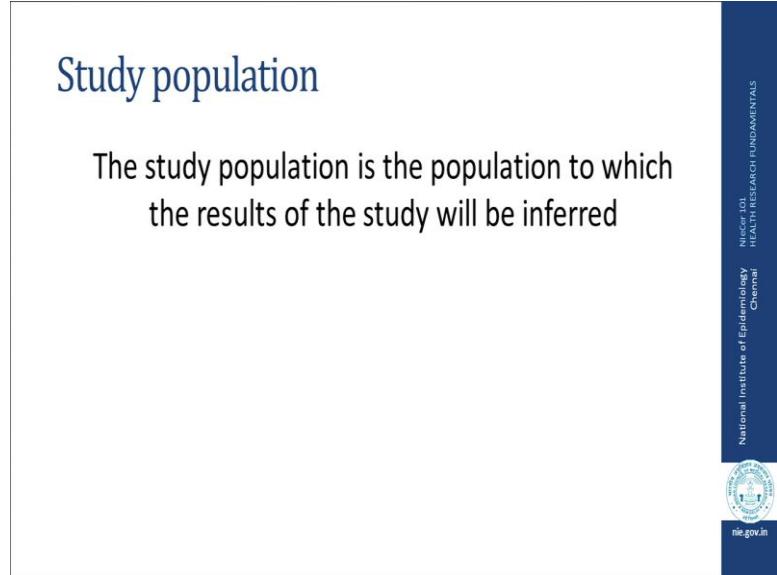
Sampling is really required, whenever you are dealing with a very large population and you want some quick information. Now, look at the definition of sampling. The sampling is a procedure by which some members of the population are selected and they are supposed to be the representative of the entire population.

(Refer Slide Time: 00:47)



See, if you have a population of people like this and you are looking at a portion of them and that is called a Sample.

(Refer Slide Time: 00:58)



I would like to introduce you to some concepts and one of them is the Study population. What you mean by Study population? The study population is the population to which the results of the study are to be inferred.

(Refer Slide Time: 01:18)

## The study population depends upon the research question

- How many injections do people receive each year in India?
  - Study population: Population of India
- How many needle-sticks health care workers experience each year in India?
  - Study population: Health care workers of India
- How many hospitals have a needle-sticks prevention policy in India?
  - Study population: Hospitals of India



Say for example, how many injections do people receive each year in India? The study population in this case is the entire population of India. Suppose, your research question is, how many needle-sticks health care workers experience each year in India? Then the study population becomes health care workers of India. Suppose, if your study question is how many hospitals have a needle sticks prevention policy in India? Then your study population in this case becomes hospitals of India.

(Refer Slide Time: 01:55)

The sample needs to be representative of the population in terms of time

- Seasonality
- Day of the week
- Time of the day

NIEER IOL  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

The sample which been select should be representative of the population for which we require an answer and this representation should be in accordance in, seasonality, the day of the week, the time of the week.

(Refer Slide Time: 02:11)

The sample needs to be representative of the population in terms of place

- Urban
- Rural

NIEER IOL  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

Whether it is urban or it is rural, or it is should rather match the composition of age, sex and other demographic characteristics of the population.

(Refer Slide Time: 02:14)

## The sample needs to be representative of the population in terms of persons

- Age
- Sex
- Other demographic characteristics

(Refer Slide Time: 02:24)

## Definition of sampling terms

- Sampling unit (Basic sampling unit, BSU)
  - Elementary unit that will be sampled
    - People
    - Health care workers
    - Hospitals
- Sampling frame
  - List of all sampling units in the population
- Sampling scheme
  - Method used to select sampling units from the sampling frame

Now, let us introduce you to some concepts or terminologies that are often used in the sampling parlance. What do you mean by Sampling Unit? Sometimes it is called basic

sampling unit, BSU. These are the elementary unit that will be sample that could be people or health care workers or hospitals as we had seen in our early example. What do you mean by Sampling Frame? The sampling frame is list of all sampling units in the population and what do you mean by Sampling Scheme? The sampling scheme is a method used to select sampling units from the sampling frame.

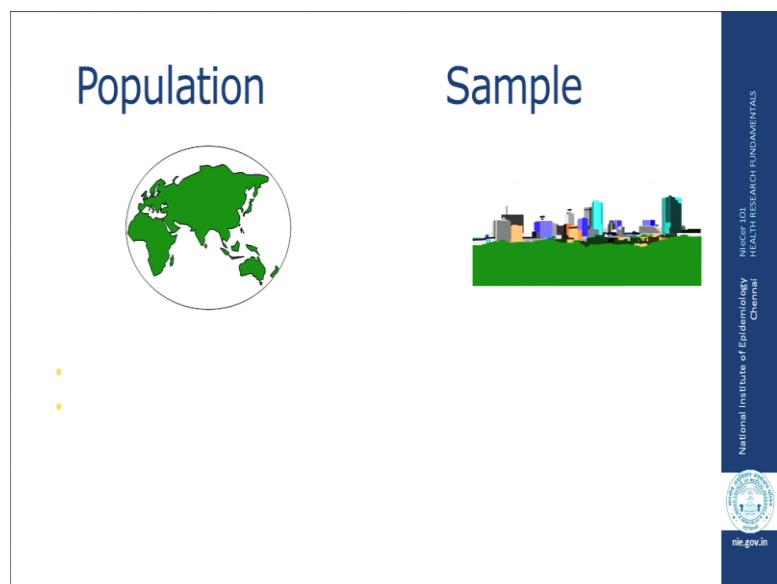
(Refer Slide Time: 03:12)

## Why do we sample populations?

- Obtain information from large populations
- Ensure the efficiency of a study
- Obtain more accurate information

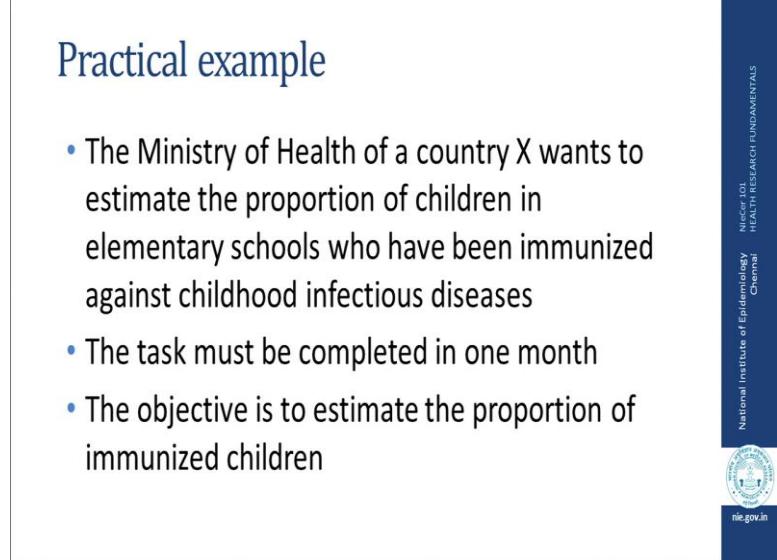
There are different ways how or why we should do the sample populations? If you have in enough resource, you can probably study the entire population. But still, even if have the resources, it is not wise to study the entire population because often the population is very large and a large population when you are going collect information one of the major constraint could be the time. You may require lot of time to collect information and you may say I will employ lot of people to do that, but what would probably happen is if you have lots of people collecting information there could be lot of inter observer variations which could rather add on to a tremendous amount of error and unfortunately you cannot measure the amount of such errors. So, it often happens that by doing a samples survey you often get accurate information. The information that you get from sample surveys are more accurate than the information you do on a large scale population studies.

(Refer Slide Time: 04:36)



So, a population could be an entire universe, whereas a sample could be as selected a small regions.

(Refer Slide Time: 04:44)



Let us look at a practical example. Suppose, the ministry of health of a country X wants to estimate the proportion of children in elementary schools, who have been immunized

against childhood infectious diseases. You could just imagine, you know the proportion of children of all elementary schools, who have been immunized against childhood infections of a country. So that is a task but one of the conditions that he has put is the task must be completed in one month. So, the objective is to estimate the proportion of immunized children and you want the results in a month's time.

(Refer Slide Time: 05:24)

## Type of samples

- Non-probability samples
  - Probability of being selected is unknown
  - Convenience samples
    - Biased
    - Best or worst scenario
  - Subjective samples
    - Based on knowledge
    - Time/resources constraints
- Probability samples

NICER RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai  
  
nier.gov.in

Now, let us rather look at the different ways; how you can rather get this information? Or in other words what are the different types of samples that could? See, broadly speaking the sample could be a Non-probability Sample or a Probability Sample. What you mean by a Non-probability Sample? Non-probability sample is the probability of being selected that is a sample, the probability of being selected for your study is not known. It could be a convenient sample or purposing sample, you just rather convenient whatever the region that is convenient to you, close by to your place, you can rather go and rather see first 100 people that you come across that could be a convenient sample.

What could rather happen? That sample could be biased or it can rather give either a best or a worst scenario, people you know it is a convenient location. You may get rather the results very different from a location, which is not very convenient or which is very remote and difficult to approach. And also, some of these are all very subjective samples

and to derive some objective criteria from a subjective sample is always difficult. But, nevertheless these non-probability sampling methods still are useful and that is being extensively used mainly to generate hypothesis or to prepare for more systematic probability samples. Now let us look at, what do you mean by Probability Samples?

(Refer Slide Time: 07:07)

## Type of samples

- Non-probability samples
- Probability samples
  - Every unit in the population has a known probability of being selected
  - Only sampling method that allows to draw valid conclusions about population

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

In a probability sample, every unit in the population has a known probability of being selected. What is the advantage? This is only sampling method that allows to draw valid conclusions about the population.

(Refer Slide Time: 07:30)

## Random sampling in probability samples

- Removes the possibility of bias in selection of subjects
- Ensures that each subject has a known probability of being chosen
- Allows application of statistical theory

NICER IOL  
HEALTH RESEARCH FUNDAMENTALS



nie.gov.in

It removes the possibility of bias in selection of subjects and also ensures that each subject has a known probability of being chosen. It allows application of statistical theory because many of the statistical text that you do it insist on a random sampling and these tests are valid only if the samples are a random sample.

(Refer Slide Time: 07:54)

## Sampling error

- No sample is a perfect mirror image of the population
- Magnitude of error can be measured in probability samples
- Expressed by standard error of mean, proportion, differences...
- Function of:
  - Sample size
  - Variability in measurement

NICER IOL  
HEALTH RESEARCH FUNDAMENTALS



nie.gov.in

I would like to rather introduce you to the concept call Sampling Error. No sample is a perfect mirror image of the population. Always you know when you pick a sample from a population and when you look at the results, it may not be exactly the same as the results in the population. But, fortunately the magnitude of error could be measured in terms of probability in the case of probability samples. This is expressed by standard error of mean or proportion or differences and that is a function of the sample size and then the variability in the measurement. So, sampling error is a very important component in sampling theory, which helps us in identifying the sample size and things so on.

(Refer Slide Time: 08:56)

## Methods used in probability samples

1. Simple random sampling
2. Systematic sampling
3. Stratified sampling
4. Cluster sampling
5. Multistage sampling

Now, let us look at some of the popular sampling methodologies that are employed in sample service. Let us rather look at the first Simple random sampling.

(Refer Slide Time: 09:11)

## 1. Simple random sampling

- Principle
  - Equal chance for each sampling unit
- Procedure
  - Number all units
  - Randomly draw units
- Advantages
  - Simple
  - Sampling error easily measured
- Disadvantages
  - Need complete list of units
  - Does not always achieve best representation

NIEIR IDI  
National Institute of Epidemiology  
Chennai



nie.gov.in

As a name suggests, it is a very simple sampling procedure, very easy to understand in which every individual sampling units have got an equal chance of being included into a sample. How do you do that? We number all the units and we randomly draw units. The advantages as I mentioned, it is very simple and sampling error is also very easily measured. Major limitation of this is, you need to have a compete list of all units many times it may not be available and also some times you may get a sample, which is very different from the whole population may not be very representative of the population.

(Refer Slide Time: 10:07)

### Example of simple random sampling

Numbers are selected at random

1	Albert D.	25	Monique Q.
2	Richard D.	26	Régine D.
3	Belle H.	27	Lucille L.
4	Raymond L.	28	Jérémie W.
5	Stéphane B.	29	Gilles D.
6	Albert T.	30	Renaud S.
7	Jean William V.	31	Pierre K.
8	André D.	32	Mike R.
9	Denis C.	33	Marie M.
10	Anthony Q.	34	Gaëtan Z.
11	James B.	35	Fidèle D.
12	Denis G.	36	Maria P.
13	Amanda L.	37	Anne-Marie G.
14	Jennifer L.	38	Michel K.
15	Philippe K.	39	Gaston C.
16	Eve F.	40	Alain M.
17	Priscilla O.	41	Olivier P.
18	Frank V.L.	42	Geneviève M.
19	Brian F.	43	Berthe D.
20	Hélène H.	44	Jean Pierre P.
21	Isabelle R.	45	Jacques B.
22	Jean T.	46	François P.
23	Samanta D.	47	Dominique M.
24	Berthe L.	48	Antoine C.

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

See, an example of a simple random sampling could be if you have the list of all say about these 48 names, you pick a random numbers of 9, 18, 32, and 40. So these are all the names that are selected as your sample.

(Refer Slide Time: 10:21)

## 2. Systematic sampling

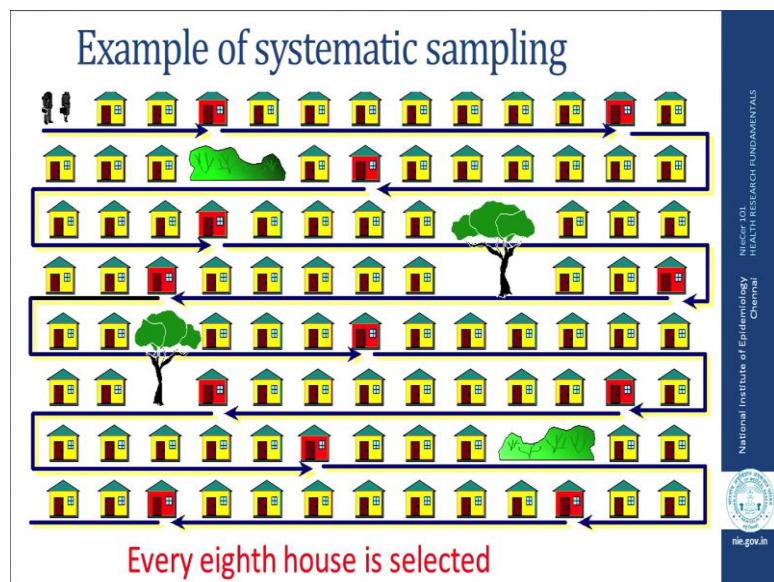
- Principle
  - A unit drawn every k units
  - Equal chance of being drawn for each unit
- Procedure
  - Calculate sampling interval ( $k = N/n$ )
  - Draw a random number ( $\leq k$ ) for starting
  - Draw every k units from first unit
- Advantages
  - Ensures representativity across list
  - Easy to implement
- Disadvantage
  - Dangerous if list has cycles

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

Now, the next sampling type is a Systematic Sampling. A systematic sampling rather done, it is initial sampling unit is picked by random and then every kth unit from that from your population are examined. A unit is drawn and every k units and every equal chance of being select for each of the unit. So, you calculate the sampling interval called k, which is divided by N divided by the number of sample size that you require. And you draw random a number which is less than or equal k for starting and draw every k units from the first unit.

What are the advantages? It ensures representatively across the list. It is easy to implement. You gave a worker that you say, you start from this house and every 10th house you go on rather see and cover all the houses, it is very easily it is been done. If there is some sort of a cycle of some specific characteristic that you are studying then you might probably get a sample, which is very typical in a systematic sampling and also some of the statistical measures that you are going compute. It is difficult when you are going to have systematic sampling, where you do not have an exact formulas, you may have to use some approximate formulas.

(Refer Slide Time: 11:53)



The example of a systematic sampling is you see, in the first, the red house is selected and then every eighth house from that is selected and all the red houses in these houses are your selected samples.

(Refer Slide Time: 12:13)

### 3. Stratified sampling

- Principle
  - Classify population into homogeneous subgroups (strata)
  - Draw sample in each strata
  - Combine results of all strata
- Advantage
  - More precise if variable associated with strata
  - All subgroups represented, allowing separate conclusions about each of them
- Disadvantages
  - Sampling error difficult to measure
  - Loss of precision if small numbers sampled in individual strata

There is a sampling method called Stratified Sampling. The principle of it is, who classify population into homogeneous subgroups, which are called 'strata' and you draw sample from in each strata combine the results of all the strata to get an idea of the whole population. The advantage of it is more precise, if variable associated with strata and all subgroups represented, allowing for separate conclusions about each one of them. Suppose, a natural strata could be male and female, so you have an estimate for male and you have an estimate for female and you can have an estimate for a combine male and female for the whole population. But the disadvantage is, sampling error is difficult to measure, and that could be loss of precision, if you are going rather have a lot of strata and for each strata you have small numbers in it.

(Refer Slide Time: 13:13)

## Example of stratified sampling

- Estimate vaccination coverage in a country
- One sample drawn in each region
- Estimates calculated for each stratum
- Each strata weighted to obtain estimate for country

Example of a stratified sampling it is, suppose if you want to estimate the vaccination coverage in your country. One sample drawn from each region north, east, south and west and the estimate calculated for each of the stratum and at the end you can weight the stratum according to the size of the regions.

(Refer Slide Time: 13:36)

## 4. Cluster sampling

- Principle
  - Random sample of groups (“clusters”) of units
  - All or proportion of units included selected clusters
- Advantages
  - Simple: No list of units required
  - Less travel/resources required
- Disadvantages
  - Imprecise if clusters homogeneous (Large design effect)
  - Sampling error difficult to measure

Another important type of sampling, which is very popularly used in the health surveys in research, is called Clusters Sampling. The principle of cluster sampling is that, a random sample of groups or a cluster of units, and all proportion of units are included in these selected clusters. Its advantages is, it is simple, we do not require a list of unit and less of travel or resources are required because you are going to collect a cluster and you are going see only within the clusters. And the disadvantage is, if the clusters of homogeneous then it may result in a large design effect. All the people in the sample may have very homogeneous results which could result in a design effect and sampling error is difficult to measure in a cluster sampling.

(Refer Slide Time: 14:43)

## Cluster sampling

- The sampling unit is not a subject, but a group (cluster) of subjects.
- It is assumed that:
  - The variability among clusters is minimal
  - The variability within each cluster is what is observed in the general population

*Sampling techniques*

The sampling unit is not a subject, but a group or a cluster of subject. The assumptions here it is, that variability among the cluster is minimal. The variability within each cluster is what is observed in the general population.

(Refer Slide Time: 15:00)

## The two stages of a cluster sample

1. First stage: *Probability proportional to size*
  - Select the number of clusters to be included
  - Compute a cumulative list of the populations in each unit with a grand total
  - Divide the grand total by the number of clusters and obtain the sampling interval
  - Choose a random number and identify the first cluster
  - Add the sampling interval and identify the second cluster
  - By repeating the same procedure, identify all the clusters
2. Second stage
  - In each cluster select a random sample using a sampling frame of subjects (e.g. residents) or households

Now, how these clusters sampling is usually done. It is done as two stage approach. In the first stage, a probability proportional to size, that is select the number of clusters to be included, compute a cumulative list of all the population in each unit with a grand total, divide the grand total by the number of clusters and obtain the sampling interval, choose a random number and identify the first cluster, add the sampling interval and identify the second cluster and so on and by repeating the same procedure, identify all the clusters.

Once your clusters are identify, then in the second stage in each cluster, you select this random sample using the sampling frame because as I had mentioned you earlier on simple random sampling when you want do a simple random sampling you need to have all the list of the your sampling frame. So, in a small cluster it is possible for you to formulate the sampling frame and you can select people from that sampling frame on a random basis.

(Refer Slide Time: 16:14)

## 5. Multistage sampling

- Principle
  - Several chained samples
  - Several statistical units
- Advantages
  - No complete listing of population required
  - Most feasible approach for large populations
- Disadvantages
  - Several sampling lists
  - Sampling error difficult to measure

NICER IODI  
National Institute of Epidemiology  
Chennai



nie.gov.in

Another important sampling methodology that is half an hour, employed is called a Multistage Sampling. In this multistage, especially in a very large, you want some estimates for at the national level, you need to do a sampling in several chains samples and several statistical units are there. The advantage is, there is no complete listing of the population is required and it is most feasible approach for large populations. The disadvantage is, there are several sampling units and sampling error at times, it is very difficult to unless you follow certain very specific methodologies for selecting at each stage.

(Refer Slide Time: 17:02)

## Key issues

- We cannot study the whole population so we sample it
- Taking a sample leads to sampling error, which is measurable
- Good design and quality assurance ensure validity and while appropriate sample size will ensure precision
- Probability samples are the only one that allow use of statistics as we know them

Some of the key issues that I would like to bring to you is we cannot study the whole population so we sample it. Whole population studying it is could impact a result in inaccurate results so the taking sample leads to sampling error, but which is easily measurable and we do not have a measure for non sampling error, whereas we have a measure for sampling error. Good design and quality assurance ensure validity and while appropriate sample size will ensure precision. The probability samples are the only one that allows the use of statistics as we know them and so it is always advantage to use a probability sample so that you can have a valid conclusion, a precise conclusion and also you can employe statistical test on them.

Thank you so much.

**Health Research Fundamentals**  
**Dr. R Ramakrishnan**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**Lecture - 12**  
**Calculating sample size and Power**

Hi. In this nicer course on Health Research Fundamentals, in this module, we are going to see some aspects of sample size, how much subjects require to do study and some of the concepts that goes behind the Calculation of Sample Size.

(Refer Slide Time: 00:32)

## Objectives

---

- Understand the relationship between sample size and power
- Determine sample size necessary to achieve a given level of power for estimating a simple proportion, and other measures of effect

National Institute of Epidemiology  
Chennai



nie.gov.in

The usual question most of the investigators they have in their mind when they want to rather start doing a research study is, how much subjects that I should recruit in to my study? How many patients I should see? How many households that I should cover? And a simple answer for this question is there is no simple answer. This requires a little bit of logical thinking like, and usually it depends on some of the information that we already should know before we start our study.

This module, we will try to understand, what the relationship between sample size is and I am going to introduce you, to a concept called power. And, we also try to rather

determine the sample size, which is absolutely essential or necessary to achieve a given level of power for estimating may be a simple proportion or any other measures of effect.

(Refer Slide Time: 01:58)

## Steps in Estimating Sample Size

- Identify major study variable
- Determine type of estimate (% , mean, ratio,...)
- Indicate expected frequency of factor of interest
- Decide on desired precision of the estimate
- Decide on acceptable risk that estimate will fall outside its real population value
- Adjust for population size
- Adjust for estimated design effect
- Adjust for expected response rate

NATIONAL INSTITUTE OF  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai  
  
nie.gov.in

As I was mentioning to you, there is no simple answer for estimating a sample size. We need to go in a systematic manner and let me rather take you step by step in the process of estimating sample size. First of all, you will have to identify, what is a major study variable you are planning to study? In any investigation, you will be seeing a number of things, you will have to identify which among them is the most important variable, which you want rather study above. Say for example, when you are studying on may be scrub typhus in a community; your aim may be to estimate a prevalence scrub typhus in which case, the variable whether a person has got scrub typhus or not is at a major variable.

Suppose, if you are interest, which is not on the prevalence but on some of the associated factors, whether a person has been expose to (Refer Time: 03:19) or something like that, in that case that becomes the major study variable. Then the second step is to determine the type of estimate. Are you going to rather study at mean or a ratio or a percentage or proportion because accordingly we need to rather re-frame or have a formula for computing the sample size?

Then one of the important things that comes out it is, you need to indicate the expected frequency of factor of interest, common sense says, suppose if you are going to rather study something very rare you need to rather have a large sample, unless you see a very large number of people you may not probably get sufficient number of people with the factor of your interest. On the other hand, if you are going to rather study something which is very common, in that case you do not need a large sample even and a small sample you may be able to rather give fairly a good sufficient and a precise estimate of your factor of interest.

Then the next factor is, the decide precision of the estimate. How precise you want your estimate to be? You want your estimate to be within 5 percent this side that side or within 10 percent this side that side. As you want your estimate to be more precise then naturally you need rather have a larger sample. If you are willing to rather give say plus or minus 20 percent, then probably your sample size will be small as compared to plus or minus 10 percent.

Then the next point is, I want plus or minus 10 percent but how sure I want that my estimate is plus or minus 10 percent. What is the amount of rather risk that I am willing to accept? Whether a 5 percent risk or a 10 percent risk. These are all some of the elements that are essential to compute the sample size and invariably these are all the elements that has to be rather given by the investigator to whoever is computing the sample size. The other three items that I have rather given are, you have to adjust what population size. Are you going to take your sample from a very large population? Or you are going to rather take from a small population? Because usually the sample size formula assumes that you are taking a sample from a very large population. If you are going to rather take your sample from a small population, you need to do some adjustment factor into it.

The next is, adjust for estimated design effect. See in my earlier lecture on sampling I talk to you about a clustered design effect, wherein you are going to rather select not individual subjects as your sample but you are going to rather select cluster of subjects as your sample. There could be a correlation between the subjects in the same cluster. So, in order to get over it you need to multiply your sample size by a factor called Design effect

so that you have a larger sample which takes care of this correlation within the subjects in a cluster.

Then the last bullet point in this, it is adjust for expected response. You have to rather you know, you decide that you want to do 300 and if you just go and study 300 maybe you know 10 percent of them they did not turn up and you have only 230. In order to adjust for that you have some may be 10 percent extra as your sample size so that assuming a non-response, you still have sufficient sample to answer your question.

(Refer Slide Time: 07:31)

## $\alpha$ and Confidence Level

- $\alpha$ : The significance level of a test: the probability of rejecting the null hypothesis when it is true (or the probability of making a Type I error).
- Confidence level: The probability that an estimate of a population parameter is within certain specified limits of the true value; commonly denoted by “ $1 - \alpha$ ”.

NICER for  
HEALTH RESEARCH FUNDAMENTALS



nie.gov.in

I am going to introduce you now to some concepts, which are essential to understand the computation of sample size. The first one is the alpha or the Type I error, the significance level of a test. What do you mean by that? It is the probability of rejecting the null hypothesis when actually it is true. In the statistical parlance, it is called Type I error. And the confidence level is the complement of that, that is 1 minus alpha and that is naturally the probability that an estimate of a population parameter is within certain specified limits of the true value.

(Refer Slide Time: 08:23)

## $\beta$ and Power

- $\beta$ : The probability of failing to reject the null hypothesis when it is false (or the probability of making a Type II error).
- Power: The probability of correctly rejecting the null hypothesis when it is false; commonly denoted by “ $1 - \beta$ ”

NIEIR ID:  
National Institute of Epidemiology  
Chennai



nie.gov.in

The next are Beta and Power. Beta is nothing but the probability of failing to reject the null hypothesis when actually it is false. So, if something is false you have to reject it, but you accept it and the probability of making this is called a Type II error. And the complement of that is 1 minus beta is commonly denoted as power and which is a correct decision and that is nothing but probability of correctly rejecting the null hypothesis when it is false.

(Refer Slide Time: 09:01)

## Precision

A measure of how close an estimate is to the true value of a population parameter. It may be expressed in absolute terms or relative to the estimate.

NICER IODI  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

Another important concept that goes into the computation of sample size is the Precision. By precision what you mean is? It is a measure of how close an estimate is to the true value of a population parameter. It may be expressed in absolute terms or relative to the estimate. They say plus or minus 10 percent or plus or minus 10 percent of the estimate.

(Refer Slide Time: 09:27)

## Sample Size Required for Estimating Population Mean

- Suppose we want an interval that extends  $d$  units on either side of the estimator

$$d = (\text{reliability coefficient}) \times (\text{Standard error})$$

- If sampling is from a population sufficiently large size, the equation is:

$$d = z \frac{\sigma}{\sqrt{n}}$$

- When solved for  $n$  gives:

$$n = \frac{z^2 \sigma^2}{d^2}$$

NICER IODI  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

Now, let us look into a scenario where you need to compute a sample size and your interest is to estimate the Mean of a Population. You have a sample, you have to compute a sample mean and you want to estimate the population mean from your sample mean. And what should be the sample size that you need? So, the general idea of the computation of sample size is, it is always a reliability coefficient into standard error is called d and through d, we can estimate the sample size n. Say, the d is a formula given here is  $z \sigma / \sqrt{n}$ . How do we get that? We get that mainly using the concept of a sampling distribution.

What do you mean by sampling distribution? Suppose, I take several samples of the same size and each of the sample given estimate of the population mean and if I have a distribution of all those sample means, theoretically, it is proved that that distribution is at normal distribution and also the standard error which is the standard deviation of that state distribution is given by  $\sigma / \sqrt{n}$ . So, what we normally we do it is? We try to rather have see the using the principles of normal distribution 2 sigma limit of the sampling error 95 percent of the values they lie. The z in the formula is nothing but the standard normal deviate for a particular level of significance. And suppose, the idea of sample size is you fix that and then when you fix that you have only one unknown namely, the n in the denominator and if we can solve for the n, which is nothing but n is equal to  $z^2 \sigma^2 / d^2$ , then you get an idea of what the n is.

(Refer Slide Time: 11:47)

## Example 1 (1/2) What Sample Size Do I Need If...?

A health department nutritionist , wishing to conduct a survey among a population of teenage girls to determine the average daily protein intake

What information is needed to estimate the sample size?

- The nutritionist must provide three items of information: the desired width of the confidence interval, the level of confidence desired, and the magnitude of the population variance

NICER IRI  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

I think you know these concepts will get clearer, if you see an example. A health department nutritionist, he wishes to do a survey among population of teenage girls to determine the average daily protein intake. So, that is the research problem. And what information is needed to estimate the sample size? So the nutritionist must provide three items of information. The first one, the desired width of the confidence interval, the next one, is the level of confidence desire and you should rather give a rough magnitude of the population variance.

(Refer Slide Time: 12:36)

## Example 1 (2/2) What Sample Size Do I Need If...?

- Solution: The nutritionist would like an interval about 10 units wide; that is, the estimate should be within about 5 units of the true value in either direction. A confidence coefficient of .95 is decided and on that, from past experience, the nutritionist feels that the population standard deviation is probably about 20 grams.
- Summarizing the information:  $z = 1.96$ ,  $\sigma = 20$ , and  $d = 5$
- Calculation:

$$n = \frac{(1.96)^2 (20)^2}{(5)^2} = 61.47$$



Assume, that he gives them all, say the nutrition feels that the 10 units this side and that side is what he is expecting, which means now 10 units on the whole, so 5 units this side and 5 units that side and the confidence coefficient of 95 percent is decided upon, and from his past experience from the literature review, the nutritionist feel that the population standard deviation is probably about 20 grams. Now, we have the information  $z$  is 1.96 because 95 percent confidence interval has got a corresponding  $z$  value of a normal distribution 1.96. Sigma is already given as 20 and then the desired length  $d$  is 5 units this area or that side. If you plug in all these value into the formula, it become  $n$  is equal to  $1.96^2 \cdot 20^2 / 5^2$  which comes to 61.47. Which means you need to have at least 62 teenage girls in order to get an estimate of the mean protein intake and the estimate that you gave 95 percent of the time will be within 5 units this side or that side of the true mean population mean of protein intake.

(Refer Slide Time: 14:08)

### A note on Population Standard Deviation $\sigma$

- The formulas for sample size require knowledge of  $\sigma^2$ . However, in general, the population variance is unknown and has to be estimated:
  - A pilot or preliminary sample. Observations used in the pilot can be counted as part of the final sample
  - Estimates may be available from previous studies
  - If thought that the population is approximately normally distributed, we may use the fact that the range ( $R$ ) is approximately equal to 6 standard deviations.

$$\sigma \approx R/6$$

NINCHI DOI  
NATIONAL INSTITUTE OF EPIDEMIOLOGY  
HEALTH RESEARCH FUNDAMENTALS  
Chennai



nie.gov.in

In this formula, we have sigma that is the population variance. And in most scenarios you may not probably know a value of sigma because this is study you are going to rather do sigma may not be available and how to get a sigma? One of the ways that you can get this variance is do a pilot survey, preliminary survey. Of course, you can even use this observation used in the pilot for your final sample tool. An estimate available from the pilot survey could be used or you can use an estimate which is available from previous studies, and suppose you know you have a large data available and you have the range of the data, assuming that it is normally distributed, you can get an approximate value of sigma as the range divided by 6. So, these are all ways of getting the value of sigma in your formula.

(Refer Slide Time: 15:17)

## Sample Size Required for Estimating Proportions

- The formula requires the knowledge of  $p$ , the proportion in the population possessing the characteristic of interest. However, this is what we are trying to estimate and is unknown
- A pilot or preliminary sample. Observations used in the pilot study can be counted as part of the final sample
- Estimates may be available from previous studies and the upper bound of  $p$  can be used in the formula
- If impossible to come with a better estimate, set  $p = 0.5$  in the formula to yield the maximum value of  $n$

NICER IRI  
HEALTH RESEARCH FUNDAMENTALS

National Institute of Epidemiology  
Chennai



nie.gov.in

Suppose, we are going to estimate a Proportion not a mean, the formula is more or less similar but what you need to give is you must rather have knowledge of  $p$  that is the proportion of the characteristic or the factor of interest in the population. This also may not be, see you are going to do a study to estimate a proportion and invariably when I ask this question to the investigator he says, Sir, I am going to do a study to find it, how can I have an idea of  $p$ , can probably as I had rather mentioned earlier he can probably do a pilot study to get an idea of  $p$  or you can get from the literature, what could be the value of  $p$  and if it is impossible then the best thing is to estimate to get a value of  $p$  as 0.5, so that it is the maximum value of  $n$ .

(Refer Slide Time: 16:19)

Sample Size Required for Estimating Proportions

The method is essentially the same as for population mean. Assuming random sampling and approximate normality in the distribution of  $p$ , brings us to the formula for  $n$  if sampling is with replacement, from a population sufficiently large to warrant ignoring the finite population correction :

$$n = \frac{z^2 pq}{d^2}$$

Where  $q = 1 - p$

NICER IRI  
National Institute of Epidemiology  
Chennai  
  
nie.gov.in

So the formula for that is very simple the  $n$  is the  $z$  squared and  $d$  square in the denominator are common and here instead of your sigma what you have is the  $pq$ , where  $q$  is nothing but 1 minus  $p$ .

(Refer Slide Time: 16:38)

## Example 2 (1/3)

### What Sample Size Do I Need If...?

- I want to estimate the true immunization coverage in a community of school children
- Previous studies tell us that immunization coverage should be somewhere around 80%
- Precision (absolute): we'd like the result to be within 4% of the true value
- Confidence level: conventional = 95% =  $1 - \alpha$ ; therefore,  $\alpha = 0.05$  and  $z_{(1-\alpha/2)} = 1.96$  = value of the standard normal distribution corresponding to a significance level of 0.05 (1.96 for a 2-sided test at the 0.05 level)

This also will be clearer if you see an example. Suppose, want to estimate the true

immunization coverage in a community of school children. Previous studies tell us that the immunization coverage should be somewhere around 80 percent. Suppose the absolute precision, we would like the result to be within 4 percent of the true values. Then the confidence interval which is conventionally taken as 95 percent and 1 minus alpha therefore, there are 5 percent alpha level,  $z_{\alpha}$  is 1.96. Then we have all the values that are needed for our calculation, d the absolute precision is 0.04, p the expected proportion of population is 0.8.

(Refer Slide Time: 17:20)

## Example 2 (2/3)

- $d = \text{absolute precision} = 0.04$
- $p = \text{expected proportion in the population} = 0.80$
- $z_{(1-\alpha/2)} = 1.96 = \text{value of the standard normal distribution corresponding to a significance level of } \alpha \text{ (1.96 for a 2-sided test at the 0.05 level)}$

NICED 101  
National Institute of Epidemiology  
Chennai  
  
nie.gov.in

So naturally the q must be 0.2 then  $z_{\alpha}$  is 1.96. And when we plug them all into a formula what we get is 384.

(Refer Slide Time: 17:36)

## Example 2 (3/3) Sample Size

$$\begin{aligned} n &= \frac{z^2 \cdot p \cdot (1-p)}{d^2} \\ &= \frac{(1.96)^2 (.80) (.20)}{(0.04)^2} \\ &= 384 \end{aligned}$$



So, you need 384 subjects to get an estimation of the immunization coverage, within 4 percentages this side or that side and you have 95 percent confidence that the true value lies in this particular interval.

(Refer Slide Time: 18:00)

## Design Effect

- A bias in the variance introduced in the sampling design, by selecting subjects whose results are not independent from each other; relative change (increase) in the variance due to the use of clusters.
- The design effect can be calculated after study completion, but should be accounted for at the design stage.
  - The design effect is 1 (i.e., no design effect) when taking a simple random sample.
  - The design effect varies using cluster sampling; it is usually estimated that the design effect is 2 in immunization cluster surveys.



So, we had seen something I talked about the design effect earlier. The define design

effect is caused because of a bias in the variance introduced in the sampling design by selecting subjects, whose results are not independent from each other because in a cluster there may be you know, suppose a child is immunized in the first house hold there is a firmly a large chance that the child in the next house also is immunized, you cant say that it would be absolutely independent. In order to account for that you need to multiply your sample size by a factor called Design Effect.

(Refer Slide Time: 18:45)

## What You Need to Calculate Sample Size for Analytical Studies

- Desired values for the probabilities of  $\alpha$  and  $\beta$
- The proportion of the baseline (controls or non-exposed) population
  - EXPOSED (for case-control studies), or
  - DISEASED (for cohort/intervention studies)
  - Often based on previous studies or reports
- Magnitude of the expected effect (RR, OR)
  - Often based on previous studies or reports
  - Minimum effect that investigator considers worth detecting
- Formula: different formulae depending on study design, research question, and type of data

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

In similar sort of scenario, suppose we are going to rather do a case-control study or a cohort study and how you should go about in estimating the sample size. What you need is the desired value of the probabilities of alpha and beta and the proportion of base line or there is a control or non expose population. In the case of case control studies, the proportion of exposed or in the case of cohort studies, the proportion of disease and you need to rather have some idea of them. These are all often based on previous studies or reports and also you should have some sort of an idea of the magnitude of the expected effect that is the magnitude of related risk or odds ratio.

This again is based on a previous studies or reports. And what is the minimum effect the investigators consider? What is detecting? These are all some of the information once it provided then there are very easy formula available, the different formula depending on

the study design, research question and the type of data.

(Refer Slide Time: 20:03)

### Example 3 (1/3)

#### What Sample Size Do I Need If...?

- Cohort study of oral contraceptive (OC) use in relation to risk of MI among women of childbearing age
- Previous studies
  - Proportion of non-OC users who are at risk of disease = 0.15
  - Proportion of OC-users who are at risk of disease = 0.25
- Conventional  $\alpha$  = 0.05 (two-sided)
- Conventional  $\beta$  = 0.20 (80% power to detect a difference if one truly exists)
- Assume equal sample sizes ( $n_1 = n_2$ )

National Institute of Epidemiology  
Health Research Fundamentals  
Chennai



nie.gov.in

Now, let us take few examples and this example should rather give you some idea of, how the sample size is computed in different scenarios? Take for example, a cohort study of oral contraceptive use in relation to the risk of myocardial infarction among woman of childbearing age. Previous studies I have indicated that the proportion of non-OC users who are at risk of disease is 0.15 that is 15 percent of non-OC users' women of childbearing age have a risk of myocardial infarction. So, proportion of OC users who are at risk of disease is 0.25 and the conventional alpha is 0.05. Suppose, the beta is taken as 20, 0.20 that is you want 80 percent power to detect the difference of it truly exists and assume that you are going to rather have equal sample sizes for your users and non users.

(Refer Slide Time: 21:18)

### Example 3 (2/3)

- $p_0$  = proportion of non-OC users who are diseased = 0.15
- $p_1$  = proportion of OC-users who are diseased = 0.25
- $q_0 = (1-p_0) = 1.0 - 0.15 = 0.85$
- $q_1 = (1-p_1) = 1.0 - 0.25 = 0.75$
- $z_{(1-\alpha/2)} = 1.96$  = value of the standard normal distribution corresponding to a significance level of  $\alpha$  (1.96 for a 2-sided test at the 0.05 level)
- $z_{(1-\beta)} = 0.84$  = value of the standard normal distribution corresponding to the desired level of power (0.84 for a power of 80%)

The formula is obtained using these following parameters you know  $p$  naught which is nothing but proportion of non-OC users who are diseased, which is given as 0.15.  $P_1$  is proportion of OC users who are diseased, which is given as 0.25 and your  $q$  naught which is a complement of  $p$  naught is 0.85. Your  $q_1$  is a complement of  $p_1$  which is 0.75.  $Z$  alpha is 1.96 we saw in the last example.  $Z$  beta is 0.84.

(Refer Slide Time: 21:53)

### Example 3 (3/3)

$$n \text{ (each group)} = \frac{(p_0 q_0 + p_1 q_1)(z_{1-\alpha/2} + z_{1-\beta})^2}{(p_1 - p_0)^2}$$
$$\frac{[(.15)(.85) + (.25)(.75)][1.96 + 0.84]^2}{(0.25 + 0.15)^2}$$
$$\frac{(0.315)(7.84)}{0.01} = 246.96$$

Therefore: 247 OC users (and 247 non-OC users)

We have all these values which can be plugged into the formula, which gives n is equal to 216.96 or 247. So we need to have 247 OC users and 247 non-OC users, follow them over a period to get a desired result.

(Refer Slide Time: 22:20)

## Example 4 (1/3) What Size Sample Do I Need If...?

- Case-control study of oral contraceptive (OC) use in relation to risk of MI among women of childbearing age
- Previous studies: 10% of women use OCs
- OR of MI associated with current OC use = 1.8
- Conventional  $\alpha = 0.05$  (two-sided)
- Conventional  $\beta = 0.20$  (80% power to detect difference if one truly exists)
- Assume equal sample sizes ( $n_1=n_2$ )



Now, let us take an example of a case-control design. How do you go about? In the case control study of oral contraceptives use in relation to the risk of myocardial infarction among woman of childbearing age. Previous studies says, 10 percent of woman use OCs and OR of MI associated with current OC use is 1.8. Then the other thing as conventional alpha as 0.05, conventional beta is 0.20 and assuming equal size for case in control.

(Refer Slide Time: 22:47)

## Example 4 (2/3)

- $p_0$  = proportion of controls who are current OC users = 0.10
- $p_1$  = proportion of cases who are current OC users = 0.18
- $q_0 = (1-p_0) = 1.0 - 0.10 = 0.90$
- $q_1 = (1-p_1) = 1.0 - 0.18 = 0.82$
- $z_{(1-\alpha/2)} = 1.96$  = value of the standard normal distribution corresponding to a significance level of a (1.96 for a 2-sided test at the 0.05 level)
- $z_{(1-\beta)} = 0.84$  = value of the standard normal distribution corresponding to the desired level of power (80%)



And see you have all these parameters  $p_0$  is equal to the proportion of controls who are current OC users which is 0.1.  $P_1$  is equal to proportion of cases, who are current OC users and that is 0.18.  $Q_0$  is 0.9 and  $Q_1$  is 0.82.  $Z_{\alpha/2}$  is 1.96,  $Z_{\beta}$  is 0.84.

(Refer Slide Time: 23:27)

## Example 4 (3/3)

$$n \text{ (each group)} = \frac{(p_0 q_0 + p_1 q_1)(z_{1-\alpha/2} + z_{1-\beta})^2}{(p_1 - p_0)^2}$$

$$\frac{[(.10)(.90) + (.18)(.82)][1.96 + 0.84]^2}{(0.18 + 0.10)^2}$$

$$\frac{(0.2376)(7.84)}{0.0064} = 291.06$$

Therefore: 291 cases and 291 controls



If you plug them on in a formula then you get 291.06 which indicate that you need to rather have 291 or 292 cases and 292 controls in order to get an estimate of your OR.

(Refer Slide Time: 23:43)

Sample Sizes: Case-Control Study of OC Use and MI	
OR	Required sample sizes
1.2	3834
1.3	1769
1.5	682
1.8	291
2.0	196
2.5	97
3.0	59

NATIONAL INSTITUTE OF  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai  
  
nie.gov.in

Now, this slide gives you the required sample size for various OR. Say for example, you want to detect an OR of 1.2 then you need a sample size of 3834. Whereas you have to estimate an OR of 3, it is enough you have 59 in each group. So, what it means it is, if you want to detect a very small difference then you need to have a large sample to identify that small difference. If you want to detect the large difference then it is enough you have a small sample, then you know you will be able to get an estimate of your OR which is 3 or more.

(Refer Slide Time: 24:32)

## The 10% Rule

- Note that sample-size estimates should be interpreted as providing merely a MINIMUM estimate of the sample sizes necessary for the study
- The formula takes into account only the overall crude association between exposure & disease; i.e., no confounders are considered
- 10% rule: increase the sample size 10% for each confounder/variable added

NIEHS IRI  
National Institute of Epidemiology  
Chennai



nie.gov.in

Then see in any of this analytical studies, when you are looking for an association of one variable, there could be a third factor which could be effecting the values of this association, which is in the epidemiological parlance, we call that as confounders. There could be one variable or two variables, which are confounders in a particular association. The general rule is, if you have some confounders in your studies you hike your sample size by 10 percent for every confounder variable that you have.

(Refer Slide Time: 25:09)

## SAMPLE SIZE : Free Soft wares for Sample Size

**OpenEpi**

**Supported by Centers for Disease Control and  
Prevention, Atlanta**

[www.openepi.com](http://www.openepi.com)

NICNET IRI  
HEALTH RESEARCH FUNDAMENTALS



nie.gov.in

**PS: Power and Sample Size Calculation**

**by Department of Bio statistics**

**Vanderbilt University**

<http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize>

Now, having seen different scenarios, where the sample size is computed and then the concepts behind it, I am going to introduce you to 2 softwares, which are free softwares in the open source, which can be very easily used to compute sample size of different study designs. One is called OpenEpi and this software is supported by the CDC Atlanta and the website for that is www dot openepi dot com and it is very, very simple software to use. There is a tutorial in for this particular software which gives you some examples and depending on what sort of a study design that you have, you can plug in the values that the software ask and you will get the desire sample size.

Another one, which is called PS that is Power and Sample size calculation, this is by the Department of Bio statistics Vanderbilt University, this is also an open source software. This is also fairly user friendly software where you can rather compute to your sample sizes. So, wherever you do an investigation and when you compute the software, you give the software in your reference saying that you use this particular software and these are all the assumptions or these are all the values that you had rather plugged-in in this software, so that this is my sample size. This should be reflected in your method section.

So, to recapitulate the module, sample size there is no magic number as sample size available. Sample sizes have to be computed using various parameters that are supplied

by the investigator. Investigator may have some idea of it ready made, if he does not have those ideas you may have to probably do a pilot study to get this kind of an idea. And then you know it depends on, how much risk that you are willing to take? How much precision that you want on your estimates? And this is usually there is no fixed number we can always negotiate depending on the resources that are available in terms of money and time. Suppose, you know I say you need to rather have 300 and you do not have that much resource and you do not have that much time to do, you can always rather they know reduce the sample size, but you should know what price that you are going to pay for reducing you may have to compromised on the precision or the risk that you will be taking on this sort of estimates.

Thank you so much.

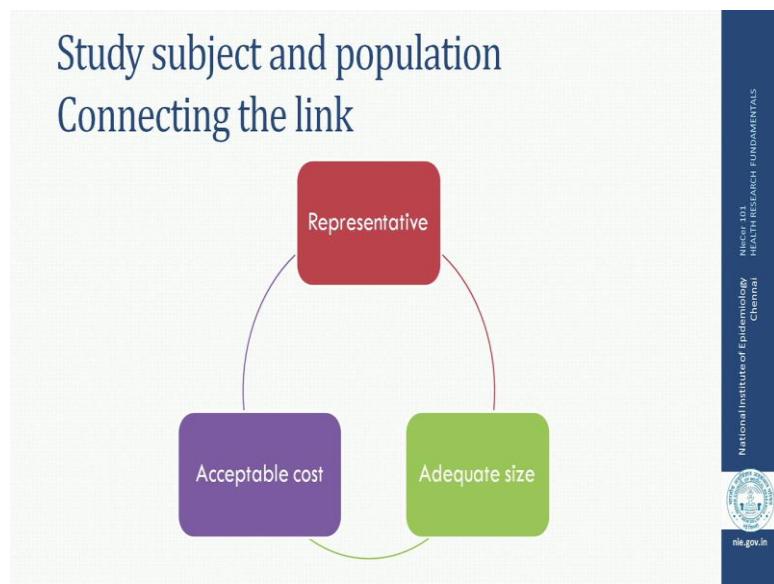
**Health Research Fundamentals**  
**Dr. P. Ganeshkumar**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**Lecture - 13**  
**Selection of Study Population**

Hi! I am Dr. Ganeshkumar from ICMR School of Public Health, National Institute of Epidemiology. Today, we are going to have a lecture on Selection of Study Population.

A good choice of study subjects serves the vital purpose of ensuring that the finding in the study accurately represents the population of interest. So, that is an important part, how the selection of a study population gathers the right information about the health research over the population of interest.

(Refer Slide Time: 00:43)

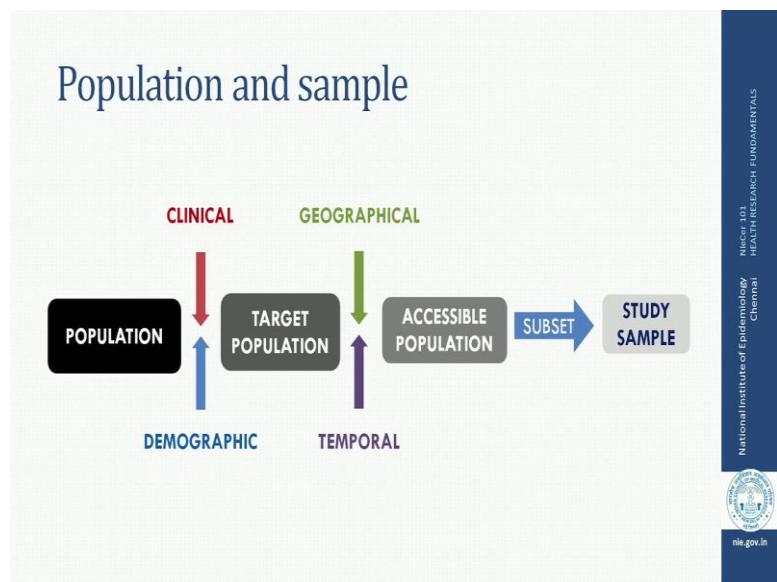


See, when we are going to select a study population or a selection of subjects from the population of interest. There are 3 important things which we have to remember particularly; first number one is that, we have to select the study subjects at an acceptable cost in terms of time and then money. Second, we need to have an adequate size of the study population so that it controls the random error, the adequate size is most important.

Number three, your study population; the study subject should be representative enough to the population of interest so that your findings can be generalizable to the population of interest.

That is how in this video lecture we will be seeing, how to select a study population with a good representativeness? And we will be also discussing about what are all the issues and what are all the solutions of selecting a study population? And what are the recruitment strategies in a clinical research? Whereas, how to achieve your adequate sample size and the process of sampling in a research will be dealt in the separate lecture. So, this lecture will not be covering about what is adequate size? As well as, what is that techniques or sampling techniques in sectioning the study population?

(Refer Slide Time: 02:14)



In a health research, in a clinical research or a public health research, when you are going to select a study population, these are all the terminologies which we need to understand, which ideally happens over the process of doing a research. For example, in a layman term, a population is a larger area for example, population of India or the population of Tamilnadu, a large geographical area. From that population by setting up or for defining certain clinical under demographic characteristics, what you are trying to derive is called

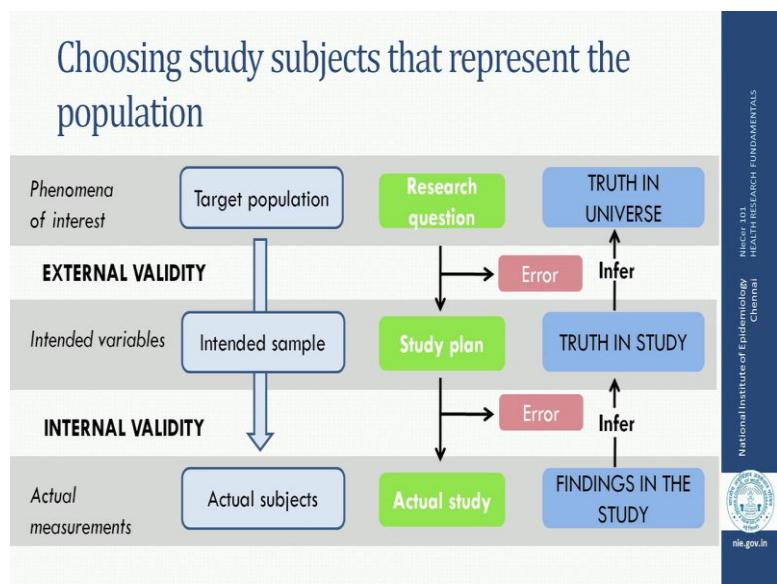
a target population. So, your target population is determined by certain clinical and demographic strategies, characteristics.

By applying certain geographical and as well as temporal characteristics over there temporal target population, what you are trying to derive is a called accessible population. So, if from these accessible population by means of certain subset of this accessible population, what we are trying to derive is called Study sample. So, this is how a process of deriving a study sample from the population. So, from this step you can able to understand that two important terminologies; one is called Target population which is defined by clinical and demographic characteristics, so again is a Accessible population which is defined by geographical under temporal characteristics of the subset of the target population and from the subset of the accessible population what we derive is a study sample.

I will give one example for these; for example, our target population which based on the research question, if the research question is like I want to study about what is a low dose of metformin to reduce the dysmenorrhea among PCOS females in reproductive age group. So in that here, when you specifically see that the clinical and demographic characteristics are those PCOS females in the reproductive age group and they are sufferings from polycystic ovarian disease and they should be having a clinical feature of dysmenorrhea. And, the geographical and temporal characteristics are defined by the accessibility of the population; this is the subset of the target population.

For example, when I am going to conduct in a city and those patients who are attending my OPD are my accessible population and I am recruiting them for this study. So, this defines the geographical and temporal is that from January 1 to December 31st of a year, a specified year. So, all the patients who are attending my OPD's, with this clinical and demographic characteristics, in this period will be recruited for my study. And from the, I estimated sample size I will be recruiting those subset of accessible population called study samples.

(Refer Slide Time: 05:10)



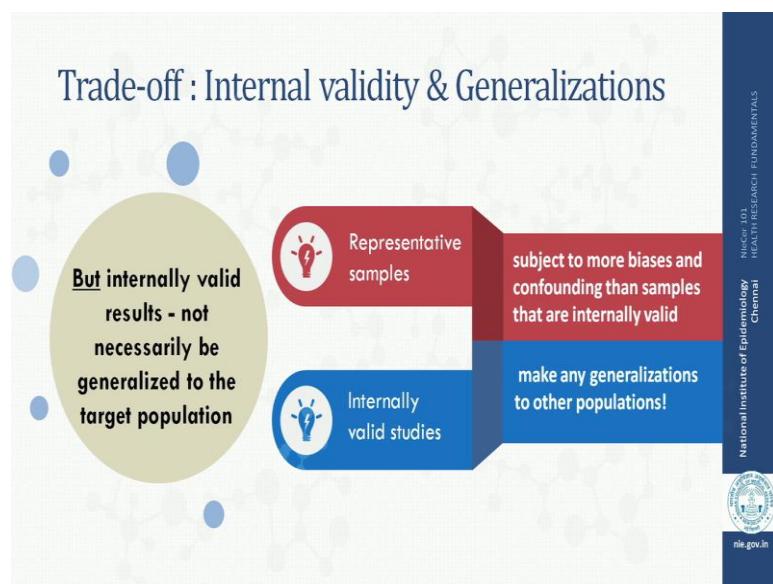
So, let us see how the journey of choosing the study subjects that represent the population happens. So, I am just throwing a kind of an algorithm here which will explain you certain two important terminologies called External validity and Internal validity. For example, by means of a phenomena of interest; that means, by your research question your setting up certain clinical and demographic characteristics what to derive is a target population, and from that you have an intended sample and that intended sample you are measuring certain intended variables and what you derive is a actual subjects, error study subjects, and where you do your actual measurements. So, same how you do is that, you do your research, you choose your research question of the phenomena of the interest and you have a study plan of conducting the study and measuring it with an actual measurements and that is your actual study here.

So, how it happens in reverse? Say, from the findings of your study, from those actual subjects what you infer is called a truth in the study and this truth in the study is what you are trying to infer over the truth in the universe, so this goes in a reverse. So, you are trying to generalize your study findings over the population of interest and that is where you have depicted as truth in the study and truth in the universe. So, here you can clearly note that there are certain errors, which may happen when you are deriving the subjects

from the target population to the actual subjects. And previously, I explained that these kinds of random errors can be handled by an adequate size of the sample population.

Now, you can clearly see here that, where this external validity and internal validity lies. So, the internal validity is a kind of a degree that how far the dependent variable influenced by the independent variables and this is very much consistent within the study subjects. If it is very much consistent and that has a good internal validity and same, when this derive, mean this findings or we are going to generalize to the population the generalizability is called that external validity, where the findings of the study are applied over the population of interest. So, internal validity is very, very essential. So, that is a most important part when we are trying to generalize the results.

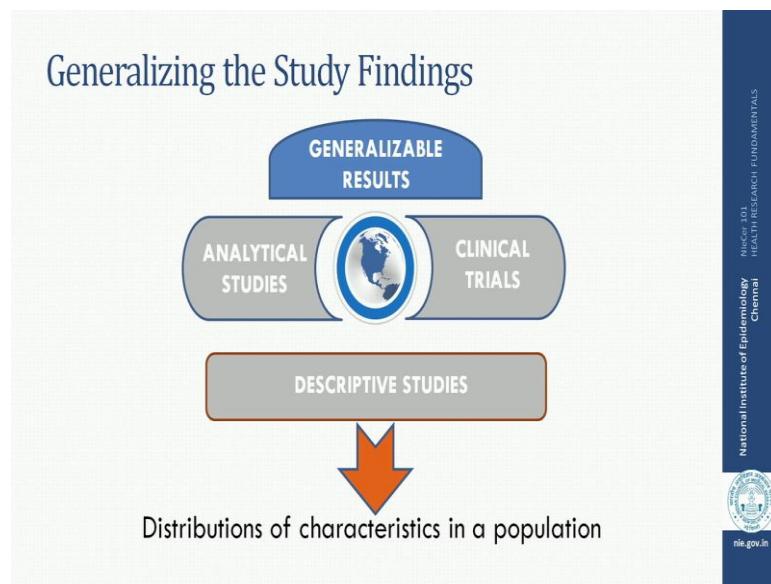
(Refer Slide Time: 07:50)



So, now you could able to understand that there is some kind of trade-off between internal validity and generalizations. So, what happens when we are going to select representative samples from the study population? When we are choosing representative samples from population then they are subject to more biases and confounding than the samples that are internally valid. Then what happens when those studies with good internal validity? So, it makes any generalizations to other populations, but there is a hinge in that, what it is?

The internal valid results, are not necessarily be generalized to target population. So, that is how and I am trying to explain here that there is always a trade-off between internal validity and generalizations. So, generalizability hardly categorical answer of yes or no, it is very rare. It has many trade-off, it is kind of a mix where scientific unpractical decisions have been taken when we are choosing the study subjects and when we are going to infer the study findings and applicable to the population of interest. So, it depends upon the design, it depends upon the method and also it depends upon the specific research question and where we are applying to derive the findings.

(Refer Slide Time: 09:18)



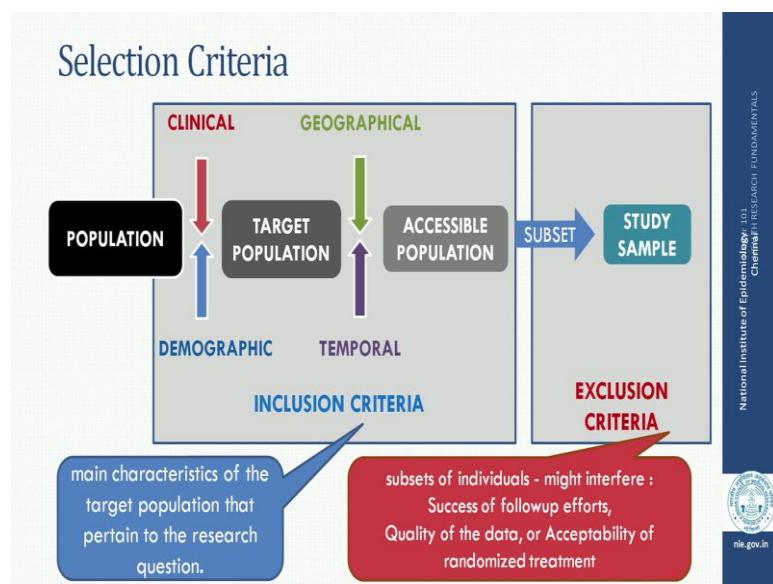
In general, generalizing the study findings can be very valid can be easily happen in clinical studies and clinical trials and analytical studies. So, generalizable of these study findings is very common. So, very popular example is say Framingham Cohort Study. So, in that Framingham cohort study what they have selected is from your individuals from your county called Framingham. From that, when they are selecting it whether the findings, whichever the findings of the study can be generalizable to the whole population of United States of America? No, it depends upon the study findings.

What they have identified in Framingham is that, the association between those cardiovascular risk factors or the association, the strength of association of hypertension has a

cardiovascular risk factor is consistent not only to the population but also to other kind of genetically different ethical group called Americans, Afro-Americans and all those things. Whereas, when we are trying to generalize certain other findings the prevalence of hypertension, which is identified in this Framingham study may not be generalizable to other ethnicity.

So, that is how descriptive studies always have an issue that, because there the study results cannot be generalizable as such. It is specific to the population which it has been studied and because descriptive studies mainly study the distribution of the characteristics of the population which may not be generalizable as like that when we are doing in analytical studies or a clinical trials. So, now, you can understand that generalizing the study findings is based on even the study design and the type of method and how the findings depends upon the findings, whether it is an analytical study findings or it is a descriptive study findings, it is very important.

(Refer Slide Time: 11:31)



Let us see that, the other important part of selecting the study population with health research is called selection criteria. So, what is that selection criteria? So, this flow chart which you have understood already, which we have discussed already is about that how we derive a target population by a set of clinical and demographic characteristics and

how we derive the accessible population, which is the subset of the target population by means of geographical and temporal characteristics. So, the terminology called inclusion criteria and exclusion criteria is based on this, how? See, inclusion criteria is the main characteristics of a target population that pertain to the research question. So, that main characteristics means here the clinical characteristics, demographic characteristics and often the selection criteria in terms of geographical and temporal characteristics or defined in the inclusion criteria, who has to be included in my study. So, it depends upon these characteristics.

Now, what is that external exclusion criteria. So, in exclusion criteria we are deciding in our study that who are those subjects, which we are not going to include because they might interfere in the success of the follow-up efforts or they might interfere with the quality of the data and they may interfere with the acceptability of the study or even there will be interfering with the ethical concerns. So, those study subjects or those groups of people with those characteristics, we do not want to include in our study is defined as that exclusion criteria.

So the step as such is that; first, we need to define a specific inclusion criteria based on the specific inclusion criteria, what you derive is a accessible population and by means of an exclusion criteria you are neglecting or you are avoiding to include certain individuals, so that what we are trying to derive is a subset of your accessible population called study samples, with an estimated sample size in it.

(Refer Slide Time: 13:52)

Designing Selection Criteria for a Clinical Trial of Low Dose Metformin to reduce dysmenorrhea in females with poly cystic ovary		
<b>Inclusion criteria</b> (Specifying populations relevant to the research question and efficient for study)	Demographic characteristics	Females in reproductive age group ( 15 -44 years)
	Clinical characteristics	Females with Poly cystic ovary with dysmenorrhea
	Geographic characteristics	Patients attending OPD of the hospital in that region
	Temporal characteristics	Between Jan 1 – Dec 31 of specified year
<b>Exclusion criteria</b> (Subset of population will not be studied because of)	Interfere with loss to follow-up	Chances of moving out of location or marriage
	Interfere with quality of data	Patients already on metformin therapy for other cause
	Being at high risk of possible adverse effects	Hypersensitivity to metformin / Renal dysfunction

National Institute of Epidemiology  
Chennai  
  
nie.gov.in

So, more in that inclusion criteria and exclusion criteria, here I am giving an example what we have discussed already. See, the study, designing a selection criteria say for a clinical trial of low dose metformin to reduce dysemenorrhea in females with the polystic ovary. Here, let us see what how we can fix up our inclusion and exclusion criteria? So, the inclusion criteria, as we discussed already that the main characteristics of the population those who have to be included in my study, which is relevant to the research question and which will be very efficient to the study. So here demographic characteristics, I will include females in the reproductive age group of 15 to 44 years; yes, it defines that particularly and the clinical characteristics are those reproductive age group females with polystic ovary and suffering from dysmemorrhea.

And the geographical characteristics, patients attending the OPD of the hospital where the study is going on in the particular region and temporal characteristics is that, between the period, January 1 to December 31 of the specific year. So, temporal is that time. With this set of inclusion criteria, now in this study what are the exclusion criteria I am fixing up, a subset of the population which should not be included, which cannot be included in my study. See, when I am concerning with those subjects who will interfere with the loss to follow up or chances of moving out form the study area may have a higher chance of

getting drop out, loss to follow up or some amount by means of marriage, those who are potential, in this particular study period the loss to follow up.

Number two, interfering with the quality of the data, how? For example, here in this example patient who are already in metformin therapy for some other cause may be due to diabetes. So, they may not be included in my study and what is in terms of being on high risk of possible adverse effects with this kind of a design feature, I will in exclude those individuals who are hypersensitive to metformin therapy and contraindication for metformin is called renal dysfunction. So, those individuals with the renal dysfunction will also be excluded from this metformin therapy. So, that is how this gives you an idea about, in terms of selection criteria, how to fix our inclusion criteria and how to fix our exclusion criteria in the selection of study subjects as selection criteria.

(Refer Slide Time: 16:32)

## Clinical versus Community Populations

- If the research question involves patients with a disease – hospitalized or clinic-based patients
  - a specialty clinic at a tertiary care medical center → patients with serious form disease - distorted impression
- For research questions that pertain to diagnosis, treatment, and prognosis of patients in medical settings, sampling from primary care clinics can be a better choice.
- True population based samples are difficult and expensive to recruit, but useful for guiding public health and clinical practice in the community

With this understanding, what will be your source of the population, Clinical versus Community Population?

(Refer Slide Time: 16:39).

## Clinical versus Community Populations

The diagram illustrates the differences between Clinical and Community populations through two circular icons and associated text boxes.

- Clinical Population:** Represented by a blue circle containing an illustration of a patient lying in a hospital bed on a trolley. The text box next to it states: "Research question involves patients" and "Source of study subject: Hospitalized or clinic-based patients".
- Community Population:** Represented by a green circle containing an illustration of several people standing in front of small red houses. The text box next to it states: "True population based samples" and "Difficult and expensive to recruit But useful - public health and clinical practice in the community".

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

Say like, clinical population where we will include those studies where we have a research question involving the patients, but the previous example those are all the PCOS reproductive age group, females with PCOS with dysemenorrhea. If it is so, then obviously, thus source of study population obviously, will be from hospitals or from the clinic based patients. So, it is preferred that best when we need to have a good internal validity and as well as, so that to be generalize these findings to the population of interest it is preferred with good quality of data the source of clinical population is from the primary health care clinics.

For example, when we are studying a true population; true population based sample which are of a huge geographical area, a community based populations and which will be a very good source for healthy subjects, when we are trying to study over a healthy subjects, over a healthy individuals. For example, Vaccine - efficacy studies obviously, it has to be from the community based studies.

So, in this term, what is a problem here in terms of true population based study is that, it is very difficult. House to house of enumeration have a more difficulty of including them and getting the study subjects within the specific time period is another difficulty and it is very expensive to recruit also. But, in terms of deriving a good public health and

clinical decisions among the community then obviously, population based sample has to be done, which I have given the example called vaccine efficacy studies. So, what will be your source of population? It depends purely upon your research question.

(Refer Slide Time: 18:38)

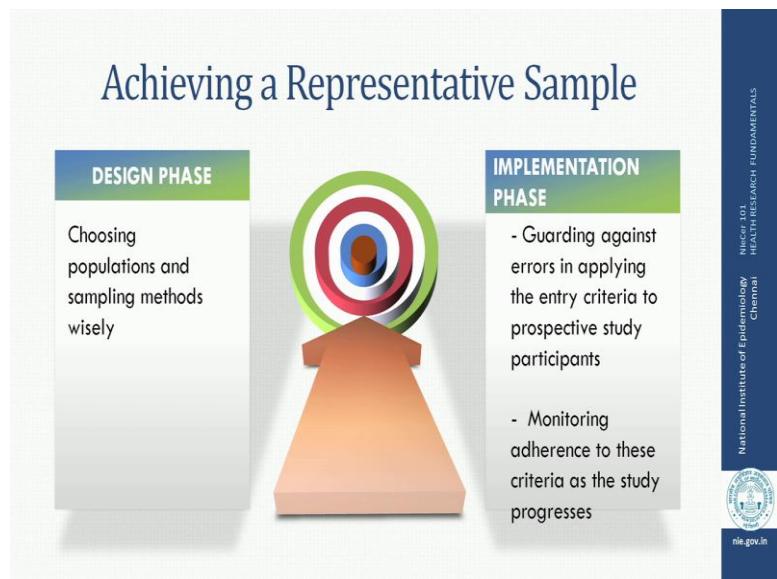


So, let us see, what are the recruitment strategies? And what has to be considered when we are doing the recruitment of study subjects in your study? So, what is that recruitment basically? Most important factor whenever we are going to recruit the study subject is that feasibility, because feasibility is an important factor when we consider to choose those accessible population and when we are trying to do a sampling procedure. Feasibility decides both, what is our sampling procedure? And how we are going to choose this accessible population from the target population? Or how we are going to choose this subset of samples from this accessible population when the feasibility is an important factor?

And, there are two important goals in term of recruitment, what are the goals of the recruitment? Number one goal is that the subject should be adequately representing the target population because you are trying to select a subset of a target population by means of geographical and temporal characteristics with predefined clinical and demographic characteristics. So, these subjects should adequately represent it, whatever

the kind of findings you are trying to infer and apply is over this target population. And second important goal is that there should be enough subject to meet the sample size requirement because previously we discussed that it should have an adequate size to counter the errors the random error generated in the study so that adequate size should be there, that are the 2 important goals.

(Refer Slide Time: 20:18)



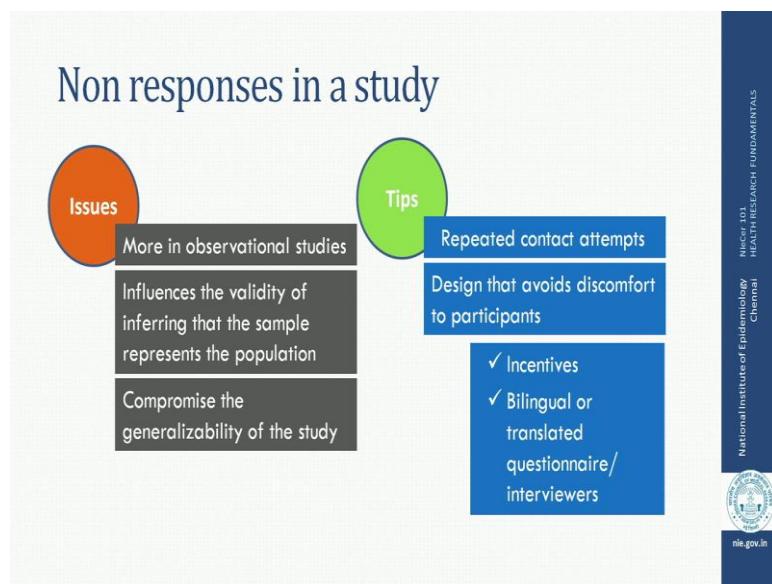
Let us see that one by one, how to achieve a representative sample? Achieving a representative sample, it happens at the beginning of the design phase and ends at the implementation phase. How it happens in the design phase? And the design phase itself, we will decide, what is the choosing the population accessible population from the target population and choosing those sampling methods wisely from the target population can happen at the designing phase itself, Whereas, in implementation phase how it ends? Number one is that, when you are trying to guard the errors by applying certain criteria over a period of time; that means you have these selection criteria.

By means of selection criteria what you can do in the implementation phase is that, you can achieve this representative sample and you can exclude those individuals, who cannot give a consistent or representative finding to the study. And second is that, unit monitor this entire study; monitor out this study subjects are adhere to this entire period,

there should not be any loss to follow up and if there is a loss to follow up how it has to be guarded? And, that is what we are trying to discuss in the next slide.

So, as an overall what you can see is that, when you are achieving a representative sample which begins at the design phase and it ends in the implementation phase.

(Refer Slide Time: 21:48)



So, as I told that something called Non responses in the study that is called loss to follow up and non responsive rate. This non responses in the study is basically happen more in observational studies and very least in clinical trials and in this, it influences the validity of your findings, which represents the study population and most important thing, third is that it compromise your generalizability. So, two things are affected very important; number one is that, internal validity will be affected when there is a non-responsive rate there is a loss to follow which reduces your sample size requirement and second is that, the representativeness and generalizability will also be effected. So, these are the two important things because of this non-response and trade.

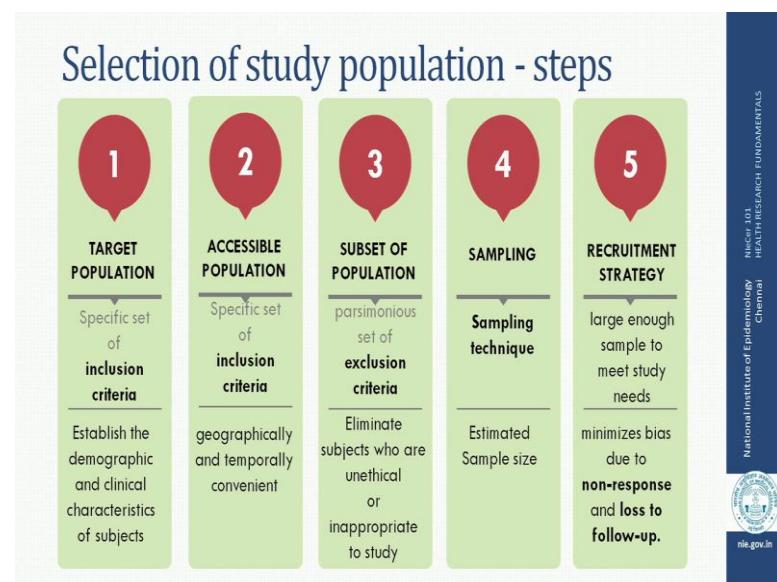
But, how we can able to tackle these non-responses in the study? Then say, at the initial instance where you could able to get the adequate size and the person is not responding or not available, then repeated contact attempts may give you or may give you an access

to the recruit them again into your study. And second is that, design that avoids discomfort to the participants by after they got recruited in between, if they have a loss to follow-up how to tackle this, that.

You can play some kind incentive mechanism, you can take away certain discomforts like, there should not be too sensitive an instrument should be, which should not be invasive. So, you can try to not disturbing in your other technique in the study or methodology in the study, you can still think of designing certain study instruments which has to be in the local language of the them, which it has to be understandable and those interviewers should talk in their local language to the participant. So, the bilingual staff with the translated questionnaire may reduce the discomfort of the study subjects.

So, by these mind of a mechanisms, it is where we could able to tackle of recruiting this adequate samples at the beginning itself and as well as where we could able to tackle to stop them loss to follow up while the study is going on. And finally, I am trying to give a summary of how this selection of study population in a step by manner can happen.

(Refer Slide Time: 24:11)



So, in the first step what happens is that, we define a target population by a specific set of inclusion criteria by means of a clinical demography, geographical and temporal

characteristics. Second is that, this accessible population is again by specific set of inclusion criteria as said by geographical and temporal characteristics. Third step is that, what we are deriving is a subset of this accessible population, which we are excluding by setting up some kind of exclusion criteria because we are eliminating the subjects which are unethical and also inappropriate to study.

Fourth step is that, we are doing a sampling procedure here, so by defining the sampling technique and by then estimating sample size which is large enough to control the random error will be done in the fourth step. And finally in the fifth step, is a recruitment strategy where we are trying to recruit those adequate subjects and as well as those subjects by those recruitment strategies to reduce the non response and rate and as well as last to follow-up. So, that is how this is where the important considerations has to be thinking of about selecting the study population for your health research.

Thank you.

**Health Research Fundamentals**  
**Dr. Sanjay Mehendale**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**Lecture - 14**  
**Study plan and project management**

Hello. In our course of Health Research Fundamentals, today we are going to discuss Study plan and project management. It is really critical that lot of energy and time is spent by all who are involved in the project or in the research study plan development in planning it really well.

(Refer Slide Time: 00:16)

### **Principles of project management**

To ensure that the defined objectives are met

To also ensure that products/ deliverables are delivered within the defined timeframe and budget at the expected quality standards

The end result should be to provide directions for future implications .. Basically for better tomorrow



Because, only a well planned study succeeds. Most of the times if we have not thought of all the eventualities, will land up in certain situations where there are difficulties which are faced while implementing the study and also while interpreting the study. And today, what we are trying to discuss is what kind of systematic process or approach can be taken to ensure that the project is implemented appropriately or the study is implemented appropriately.

What is important in project management is to ensure that the defined objectives are

adequately met and there are certain deliverables, which are defined right at the beginning of the study that this is what we want to achieve, which is our definition of objectives, and so they should be reached with in the defined time frame and with the available budget not compromising the quality at all. So, all this is achieved through effective project management. Primarily, because what we do is the project if it ends successfully the result should be able to provide direction for future implications, basically what it means is whatever research we do it should help us to do something better tomorrow.

(Refer Slide Time: 01:46)



It is important that we understand that any kind of implementation process involves some kind of underlying principle, with a primary objective of achieving a specific goal. Let me take an example of the process of resource allocation and resource management. Well, the principle here is appropriate time management. The resources have to be allocated in a timely manner, if it is going to be a long term study it will be on multiple occasions that the resources will have to be mobilized. So, that timing really becomes very critical because only then we are able to achieve the goal of efficient, we can progress efficiently towards achievement of our goals.

Another critical aspect in the process is planning and scheduling the activities. This is the

important principle which undergoes behind that and to ensure that, this happens is the monitoring and supervision. Every single detail has to be planned out appropriately so that when we reach the goal, we reach it with the best possible quality standard and hence it is important to keep this whole process in mind and the principles in mind.

(Refer Slide Time: 03:03)

## Ad hoc approach to conducting a research study is often non-productive

### The confusion at the beginning of the study

- I want to do a study, but I am not clear about the objectives
- I have prepared a questionnaire, but I am not clear about exact information I need
- I will collect data, but I am not clear how I will use that

### The disastrous end result .....

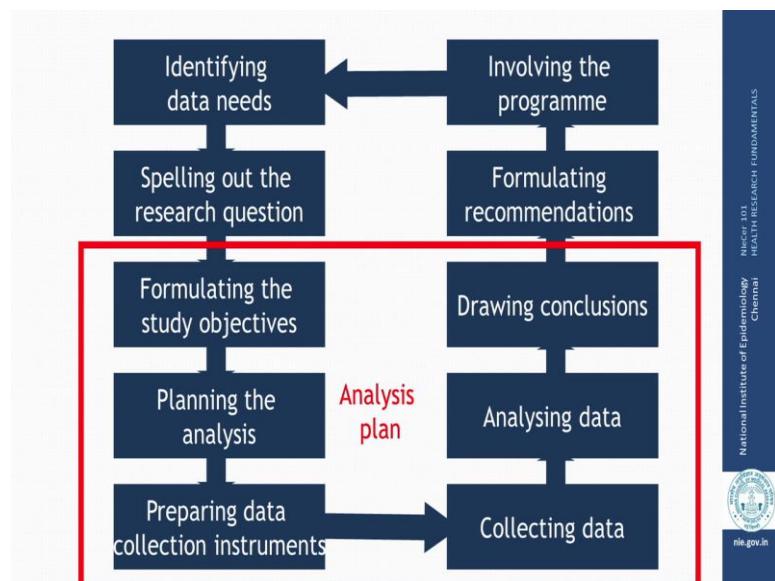
- I have data that are difficult to analyse
- I have analyzed the data but finding it difficult to interpret
- The interpretations are difficult to use in programs or for policy making

Ad hoc decisions which are taken or ad hoc approach which is taken while conducting research study often becomes non-productive. So, there can be lot of confusion at the beginning of the study in the minds of people. So, like I want do a study, but I am not clear about what the objectives are. I have prepared a questionnaire, but I am not clear about exactly what information I need. I also feel that I will be able to collect the data, but I am not clear how will I use it. All these kinds of confusions arise because the investigator himself has not understood what the research is all about.

What is going to happen as a result of this? This is going to result in a disaster, because this will be a situation which will lead to production of data or compilation of data which is difficult to analyze or then the analyzed data becomes difficult to interpret and even if the interpretations are made probably they are of no use to the program or also in the policy making. So, eventually what it means is any kind of an ad hoc approach, ad hoc approach which is taken without proper thought being given to that is not likely to

succeed.

(Refer Slide Time: 04:21)



Any research process typically starts with identifying the need for that particular research, then correctly verbalizing the research question or spelling out the research question, formulating the study objectives, planning the analysis, then preparing data collection instruments, then collecting data, analyzing data, drawing appropriate conclusions, making the specific recommendations to the concerned people and eventually again accessing, whether our needs that we had initially identified have been fulfilled or not or whether there is any need to do anything else.

If we look at this whole process, in terms of identifying data needs and spelling out the question this all is the planning stage or the initial stage even while the study is being conceptualized. This is a pre-planning phase. The steps of formulating the objectives and the analysis plan and deciding about the study instrument methodology, where the way in which data will be analyzed and then the way it will be interpreted, this all is a part of analysis plan. And, what we do after that is the dissemination of these findings to the concerned stake holders so that they can use it for appropriate programmatic absorption or policy making.

(Refer Slide Time: 05:48)

## A road map to study planning and management

- Formulate appropriate objectives for the study
- Choose the right design to determine key indicators
- Identify parameters needed for the key indicators
- Prepare the analysis outline
- Estimate sample size



So, basically the road map to study planning and management involves multiple steps. It all starts with formulating appropriate objective for the study, then choosing the right design to determine the key indicators. Please understand, I am going to walk you through this particular thing but proper decision about what kind of study is required to answer the objectives that we have framed is very critical and important one.

(Refer Slide Time: 06:37)

## Framing the study objectives is critical

- Fewer the better ..
- May be mentioned as primary and secondary
- Should be clearly phrased:
  - Aimed at testing a hypothesis: **Determine** whether a contaminated well caused an outbreak
  - Aimed at measuring a quantity: **Estimate** the prevalence of diabetes

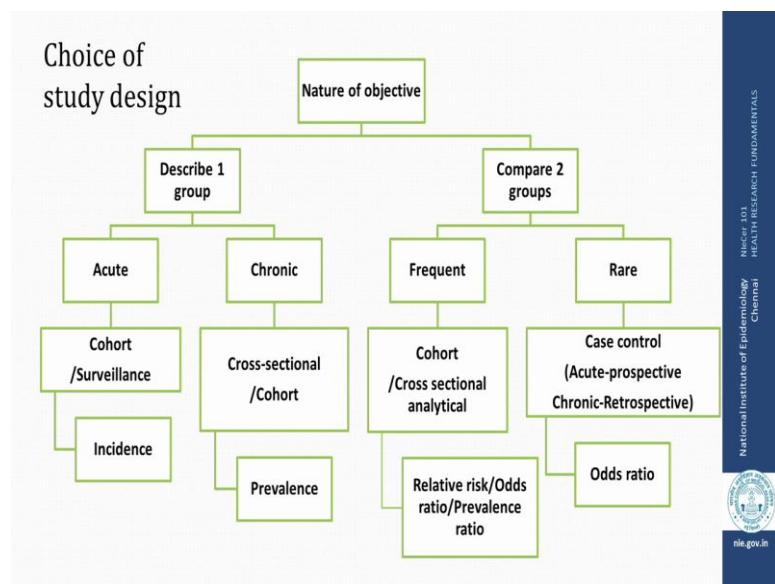


We have to also identify the parameters that are needed to estimate the indicators that we find, that we have decided, that are important for this particular study and then prepare the analysis outline. Also, it is important to estimate the sample size before the study is initiated because the study conducted on a small sample also will not be generalizable. When we talk about the study objectives, the basic principle is fewer the better. Most of the studies with a long list of objectives often become very confused studies because many of these objectives remain unfulfilled, because they become complex, the data collection tools increase, there are lots of variations that come in while collecting the data and in general there is a kiosk. So, fewer the better is the principle.

They can be described. Objectives can be described as primary objectives and secondary objectives. It is important, the primary objective is important because that generally decides the sample size for that particular study. Often in any study, sample size is calculated based on the presumption that we should be able to achieve the primary end point at least. Secondary end points are the analyzable issues, which are the additional information, pieces of information that we obtain in any research study. But it is important that the objectives are clearly phrased.

Normally, they could be of 2 types, they could be more of an exploratory type or what we call it as aimed at testing a hypothesis and here is, where we use the word determine. So, determine whether a contaminated well caused an outbreak that is an example. Or they could be say sort of confirmatory in nature or estimating in nature, say to actually decide the prevalence of a particular condition, for example, diabetes in a population. So, we have to keep in mind and use the appropriate verbs while defining the objectives.

(Refer Slide Time: 08:33)



There are different types of study designs, which are adopted to answer various research questions. Whenever we talk about or whenever we think about descriptive objectives, whenever we are exploring the acute conditions like for example, pneumonia case is occurring in children, the right kind of designs or diarrheas occurring in children and etiology.

The right kind of designs would be to do cohort studies or surveillance studies which are may be hospital based surveillance or community based surveillance. And normally, the major that you derive out of it is, incidence; this is true in case of acute conditions. But when we talk about chronic conditions, once that conditions occurs it persist for a long time either on treatment or not on treatment. So, the major that we normally get out of such kinds of studies is prevalence. And here, the right kind of designs to be used are either the cross-sectional studies or the cohort studies.

Well, many of the epidemiological studies deal with comparing 2 groups say for example, people who are exposed to a particular condition not exposed to a particular condition, people suffering from a particular disease not suffering from a particular disease, and so whenever we are moving from the variable or the exposure to the outcome, we call it as cohort study. As we know it is a prospective assessment and

whenever we try to look at the exposure after the outcome has already developed it is called as a retrospective approach and often the commonest study design which is employed is a case control study here.

But, what it is important to understand and remember here is a cohort study can be undertaken only when the outcome is more likely to occur frequently, because if it is going to take a long time to happen then probably the study will be of enormously long period and the adequate number of outcomes may not be achieved. So, for a frequently occurring outcome which, wherein from exposure to the outcome the length is likely to be minimal, cohort study is a good approach to take or there alternatively a cross sectional analytical study can be undertaken; but in case of rare exposures and where the duration between the exposure and the outcome is likely to be very long, then it is better to go for a case control approach and in these situations the relative risk and odds ratio. In case of cohort studies, it is the relative risk which we obtained which is a more definitive say indicator of relationship and odds ratio is also is a strong indicator of association.

(Refer Slide Time: 11:31)

## Identification of information needed to calculate the indicator

- Decide the indicators that the study will generate
  - Rates, ratio, proportions or quantitative variables
- Identify the information elements that will be needed to calculate the indicators
  - Numerators
  - Denominators
- Also list information elements that will be used to calculate indicators
  - Outcome variable(s)
  - Covariate
    - Potential risk factors
    - Potential confounders

It is important that the discussion in the planning stage focuses a lot on the information needs that with respect to the indicators. There are rates that we calculate, sometimes we

calculate ratios or proportions etcetera, but for all of these indicators we do need a numerator and we do need a denominator. We have to understand exactly how we are going to collect the information that is going to be required to determine the numerator also the denominator. But sometimes, this relationship between the exposure and the outcome also is affected by lot of other co-variates, they are called as risk factors or confounders and I will discuss some of them.

(Refer Slide Time: 12:21)

## Principles to be followed while collecting the information elements

- Use the variables that will best reflect the information element – it is important to review the available evidence
- Use validated or standardized methods and criteria
- Adopt standardized case definitions and laboratory criteria/ normal ranges
- Decide the most accurate way of collecting information on various elements – Observation, interview or laboratory methods

NATIONAL INSTITUTE OF EPIDEMIOLOGY  
NIEHS RESEARCH FUNDAMENTALS  
Chennai  
  
nie.gov.in

But the basic principles that we have to follow while we collect the information elements is that, we must use the variables which will be actually analyzable, this information we can obtain by reviewing the literature fairly scrupulously because it provides us a lot of evidence of which variables co-variates are important. It is important to also use validated or standardized methods because then the chances that this particular study will be accepted globally are maximum. We must adopt standardized case definition, for example, when we are going to talk about pneumonia; what is pneumonia? We should define it properly.

If we are going to talk about smoking, this as an exposure variable, what is that we are going to consider a smoking? Is it the frequency of smoking? Is it, yes or no? Or the number of cigarettes smoked per day? We have to have clarity on these matters.

Sometimes, we also use laboratory criteria and so we have to have also well defined definitions there for example, if you have to define anemia; how do we define anemia here? Does it depend on age? Does it depend on gender? All these have to be specified right in the beginning; we have to then decide which is the most reliable and accurate way of collecting that information. Sometime it could be just the observation or it is a questionnaire through which we collect this information or it could be actually laboratory essay through which we get this information.

(Refer Slide Time: 13:56)

**Outcome measurement for iodine deficiency**

Outcomes	Information element	Data collection method to obtain the variable
Chronic iodine deficiency	•Goitre	•Physical examination
Current exposure to iodine	•Urine iodine excretion	•Laboratory
Access to iodized salt	•Testing household salt for iodine	•Field spot test

NIEHS IRI  
 NATIONAL INSTITUTE OF  
 ENVIRONMENTAL  
 HEALTH RESEARCH FUNDAMENTALS  
 National Institute of Epidemiology  
 Chennai  
  
 niehs.gov.in

For example, if we are going to talk about as an outcome, whether there is an evidence of chronic iodine deficiency. The way to look at it is, we would look at what is the say prevalence of goitre in a specified community, and how we will do it is by actually doing physical examination. But if our objective is to find out what is the current exposure to the iodine here? What we would try to do is, at an individual level try to estimate the urine iodine excretion and for this we would require some laboratory methods to actually estimate this.

But sometimes, it also becomes important to go one step behind and find out what are the dietary patterns? Is there an adequate iodine being provided through the diet? And so, what one would want to do is to test the household salt for iodine and this would

involves some kind of field level spot test which are done to figure out, whether the salt which is consumed in the various households actually has enough iodine in them.

(Refer Slide Time: 15:03)

## Covariates in iodine deficiency

- Potential risk factors
  - Income
  - Community (e.g., minorities)
  - Caste
  - Education
- Potential confounding factors
  - Age
  - Sex
  - Residence

NATIONAL INSTITUTE OF  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai  
  
nie.gov.in

I did mention about the risk factors, about the confounding factors. There are certain risk factors like income and the community which are related to access for example, or the level of literacy, the practices, cultural and social which are observed by the community, the dietary patterns all these also influence the outcome. Then they also have to be appropriately analyzed when we do the interpretation of our results. Similarly, with age the risks sometimes vary with gender, the risks vary the residence, they also vary. They are considered as confounding factors, if they affect both the exposure as well as the outcome variable. There is no harm, even if the confounders are there provided we have collected information on all of the confounders as a part of our study and in the questionnaire. We can always analyze the effect of the confounders.

(Refer Slide Time: 16:06)

## Advantages of making an analysis plan

- Helps to focus on the objectives of the study
- Start by preparing dummy tables
- Helps to avoid comparisons for which the study has not been designed
- Makes sure that only data that can be analysed is collected
- Saving time: quick publication, dissemination and policy feedback



It is important to make the analysis plan because it helps to focus on the objectives of the study. This all thing can start by once, if you have clarity in our mind, what study we are doing. We can also prepare dummy tables, right in the beginning of the study because then we know what we must do and what we must not do. We also know what data we should collect and what data we no need to collect because it saves time. It can result in to quick publication and quick dissemination of findings and early policy feedback. So, this is important to make a good analytical plan.

(Refer Slide Time: 16:44)

## Sample size ....

- The analysis plan helps to determine the sample size
  - Measurement or testing?
  - Study design: Cohort, case control or survey
  - Level: Descriptive or analytical



Sample size is really critical because it is decided by what exactly is the type of outcome assessment that we are doing; whether it is by measurement or by testing. What kind of study design we are doing? Whether it is a cohort study, case control study or a survey and whether it is a descriptive study or an analytical study. This is itself is a sort of a big lecture point and so what is important to understand is when a study is being planned, it is important to involve a statistician, who would help you to analyze, help you to determine the sample size for the study.

(Refer Slide Time: 17:30)

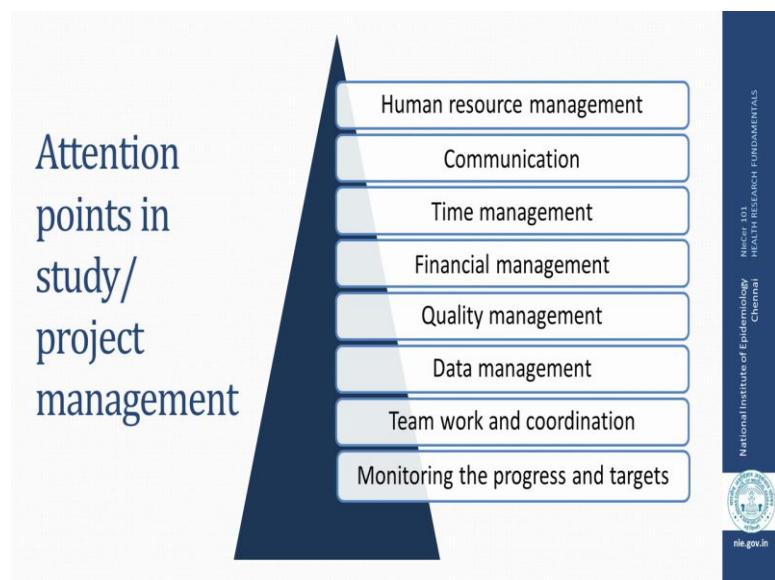
## Common reasons for study failures

1. Badly defined research question and objectives
2. Unrealistic timescales - too short or too long
3. Inappropriate and incompetent staff: Lack of direction, motivation and training
4. Inadequate monitoring, failure to respond to contingent situations and carry out mid-course corrections



Often it happens that the studies fail, why do the studies fail? It is because, either they are badly defined research questions there or objectives are not correctly defined. The time skills that have been decided either are too short or too long. Sometimes, the staff is inappropriate, incompetent. This might be because of lack of correct direction, lack of motivation or lack of training. So, all these have to be taken care of. Or sometimes, it can also result from everything else is right, but there is no adequate monitoring and there is a failure to respond to contingency situations and carry out mid-course corrections that is where monitoring and supervision becomes important.

(Refer Slide Time: 18:12)



So, for success of any kind of a study there are certain attention points one has to look at. And they start with human resource management, a very critical aspect. The study staff has to be carefully chosen, appropriately trained and with appropriate communication there should be a good dialogue. Only, generally the observation is a team succeeds, but the individuals fail. And so, between the various members of the team, there should be good bonding, extremely good communication and the leader has to ensure that this often team meetings take place and this rapport between individuals builds strongly.

Time management essentially is the responsibility of the leader, and one has to ensure that this is taken due care off. Time management in terms of appropriately scheduling various activities ensuring that they are done in time, this is really important. Financial management is also critical. Sometimes it so happens that the study starts well with the funding being given, but suddenly some kind of glitch develops by which the finances are not being granted or given in a continuous manner, suddenly the activities of the project stop. And hence, this also is an important part of the planning. This has to be planned well in advance at what stage what kind of money will be released and it must be ensured by the researchers that targets which are defined well in advance are appropriately met. So, the financier does not find it difficult to release the money what is earlier decided on.

Quality management at all levels is critical. Quality management in data collection, various clinical procedures, data management, various laboratory procedures, in supervisory visits, every single aspect of a study that we can think of quality is really critical and if that is maintained then often the studies are very successful.

Data management is important. Often the studies that take care of data management in a timely manner, where concurrent data management is planned they are able to give away the results in a timely manner. If the researchers have not planned it well and then they decide to do the data management at the end of this study, often it is disaster because if there are some issues that are happening in the way the data is being collected, if the data is being managed in a timely manner, somebody is looking at it, finding the faults in it, there is a possibility to make a change, do necessary corrections this opportunity gets lost if we are handling this whole issue at the end. I did talk about team work and coordination, which is really critical and important.

And monitoring the progress and target, here it is to be decided depending on the budget whether it is an internal mechanism that is set up for the supervision and monitoring of the study or an external monitor is brought in to take care of what is happening in this particular study. But, this is again a very important step. Whether it is a research project or it is any other project, there are various aspects that we have to think about; it is a teamwork, it is about communication, it is about human resource management, it is about time management, all these factors are important and we succeed as a team.

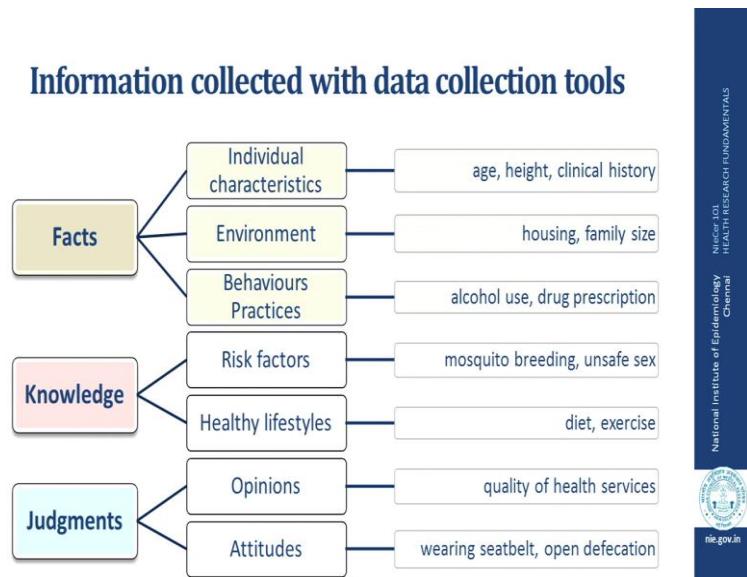
Thank you for your attention.

**Health Research Fundamentals**  
**Dr. Tarun Bhatnagar**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**Lecture - 15**  
**Designing data collection tools**

Hello and welcome to this session of Health Research Fundamentals. Today, we are going to talk about designing data collection tools or the instruments that we use to collect data in health research.

(Refer Slide Time: 00:22)

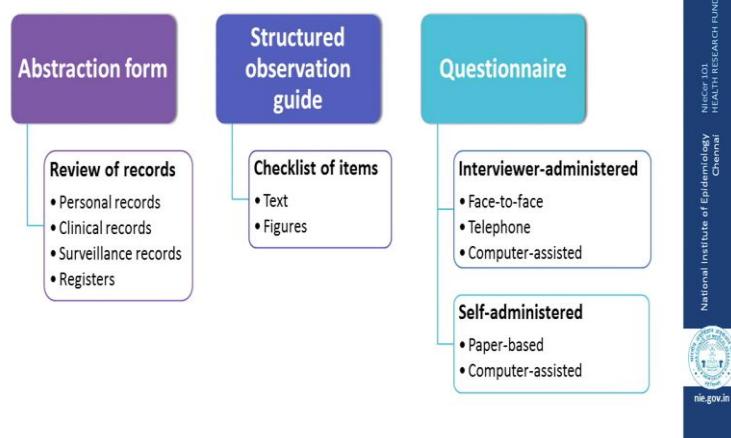


Generally, in health research, the kind of information that we would like to collect can be basically divided into three areas. We want to know facts such as, the characteristics of our study participants, the environment that they live in and their behaviors or practices. Secondly, we might want to know their level of knowledge for things such as, risk factors for getting disease or knowledge about healthy lifestyle, so as to prevent diseases and thirdly and very importantly we might want to collect information on what we call the domain of judgments, basically what are the opinions of the research participants on a certain issue may be such as quality of health services.

We would also may, we may want to know about the respondent's attitudes towards certain things for example, could be something as wearing seat belts, use of open defecation and so forth. Now, how do we collect data on all these different kinds of information that we want to collect.

(Refer Slide Time: 01:31)

## Different tools to collect data



For this purpose, we can make use of a variety of tools depending on what kind of data we would like to collect. One we have what are called abstraction forms, which is basically doing a review of records of the study participants, which could be their personal records, the forms, if you want to get information on their disease conditions, signs and symptoms, a treatment given, then we could look at the clinical records. We could look at data in general the data that is collected through disease surveillance and we could also look at registers, wherein some information, which may be there and all of this information can be called out into what we want in the form of a data abstraction form.

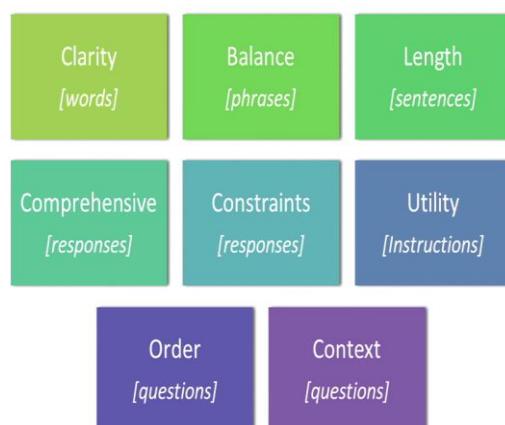
Secondly, another tool that we have for collecting data is a structured observation guide. This is very useful when we would like to document certain processes, whether they are happening, the way they are happening, that they are happening in time or not, are the objectives met and for this purpose we may use a check list of items that we would like

to collect data on which could be the textual or figurative and third and most importantly, the tool that is used most commonly is a questionnaire, wherein we would like, we talk to the person and get information.

Now, again questionnaire can be divided into two kinds it could be interview or administer, where the data collector actually administers the questions to the respondents, which could be done either face to face, it could be done on a telephone and now even have computers assisted technologies to do face to face interviews. The questionnaires could also be self administered, if the study participants can read and write and they are knowledgeable enough to understand what the investigators want. These could again be either paper based or now we also have computer assisted self administered questionnaire, which can help the respondents to directly put their information into a computer database.

(Refer Slide Time: 03:55)

## Key elements of data collection tools



Aday LA, Cornelius LJ. Designing and conducting health surveys : a comprehensive guide. 3rd ed. CA: John Wiley & Sons, Inc. 2006



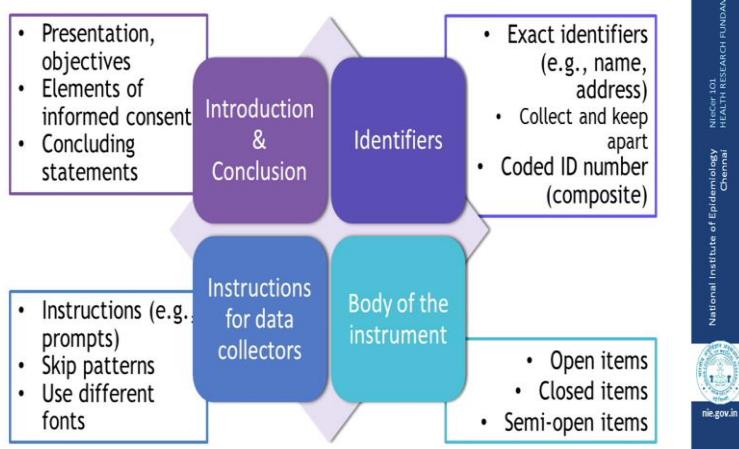
Now, whatever the tool of, in data collection we use, what we want is that we need a valid response from the study participant. The response should make sense and we should be able to use that information effectively for our health research needs. In order to do that, every data collection tool needs to make sure that there are some key elements that make up this data collection tool. Some of these elements are the clarity of words

that you use in a data collection tool, the balance of phrases and sentences, the length of sentences; how long are these questions, the comprehensiveness of the responses in terms, if you are giving categories of responses that you expect from the study participants.

These categories can actually even post constraints in terms of what information could or could not be collected through a data collection tool, of course the utility of the data collection tool and specially the utility of the instructions; which I will elaborate in the next slides. The order in which the questions are asked is very important element which can decide how your respondents answer your questions and of course, the context in which you frame these questions and the tools are used.

(Refer Slide Time: 05:18)

## 4 components of the data collection tool



So, if we look at any data collection tool generally it has 4 components, there is the introduction part at the beginning and a conclusion at the end. There are what are called identifiers, then each question may have linked instructions for the data collectors and of course, the whole body of the instruments which is basically the question items. In terms of the introduction, the introduction is used to present the study, to the study participants, state out the objectives and probably get informed consent or so forth. It is also good to always have a concluding statement at the end of your questionnaire and

thank the study participant for their time and effort that they have put in to answer your questions.

Now, every data collection tool will also have what we called identifiers, which could be either the actual identifiers, the information such as the name and address of the study participant, which can identify who the person is, it is always the good idea from an ethical and a human subject protection prospective. If you are collecting this data, to collect this in a separate sheet of paper and which can be referred to later on, if need be.

On the other hand, in order to maintain confidentiality investigators also use coded id numbers to give identifiers to the study participants and these id numbers could be composite in such that they could have numbers which denote say, the state which the person belongs to the district the village and then the household and then the individual id. So, it could be a mix of all these numbers, all these codes and then you get a composite code looking at which you can actually identify who the respondent is, but would not be able to actually get an exact identification and which is good from the confidentiality prospective.

In terms of the instructions, there could be general instructions for the data collectors such as prompts in terms of for example, do not read out all the responses, tick only the one that the study participant mentions. There could be instructions for skip patterns, now the question is may have skip patterns in the sense that there may be some questions which based on the response to those, the subsequent question may not be relevant and then there is an instruction, which says that you skip this from question 2, you may go to say question 19 and skip rest of the questions because they are not relevant to this study participant.

It is also always a good idea to may be used different fonts, so as to make it clear that what is the actually question and which part of the item is actually an instructions for the data collector and then of course, we have the whole body of the instrument which is basically the question items. Now, these question items could be of various types, we could have what are called open questions, we could have close ended questions and we could have somewhere, something in the middle.

(Refer Slide Time: 08:39)

## Open questions

- Answers are not suggested
- Respondents must generate an answer
- Advantages
  - Give freedom of response
  - Stimulate memory
  - Can be useful to generate closed responses later
  - Useful at a hypothesis raising stage
- Inconvenient
  - Difficult to code and analyze
  - May be incomplete and / or unfocused

### Examples:

- What disease can you acquire from tobacco?
- What places did you eat at in the week preceding the disease?



Let us see what are these different types of questions? As the name suggests open questions are the ones where the answers are not suggested to the study participants and the respondent has to generate an answer. The good thing about these questions is, it gives total freedom to the respondent to give the answer of what they want. They are not constrained by the categories of answers that already exist. It helps to stimulate the memory of the research participant and gives you a more better answers who as to speak and it is also useful at hypothesis raising stage, wherein we are really not sure what the appropriate answer is and you can generate a lot of responses from the study participants.

Of course, when you generate a lot of responses, open questions the inconvenience is that it may be difficult to code and analysis. You may have a long list of responses and then to categorize them later may be an issue and sometimes, if it is open, the responses may be unfocused or incomplete and that can pose a challenge in terms of the analysis. Now, to overcome this problem, we can have open questions, but then we can have them with closed end answers. Although, there is a category of answers given for there for those that question, but their data collector does not suggest an answer from these categories to the study participant. So, when the answers are freely mentioned by the respondent the interview will spontaneously take those that are specified from the list of

categories of responses given in the question here. So, it is expressed as an open question, but you finally analysis this as a closed ended question.

(Refer Slide Time: 10:29)

## Closed questions:

### 1. Dichotomous options

- Suggested answers include 2 options only
  - Yes and No
  - Male and female
- Advantages
  - Forces a clear position
  - May be useful for key, important, well framed issues
- Inconvenient
  - May oversimplify issues

#### Examples:

- Did you eat at restaurant X between 1 and 28 February?
- Have you ever consumed tobacco products?
  - ✖ A dichotomous question here is likely to over-simplify, unless it is used as an introduction

What are close questions? Close questions are the once, where you have a question and you have a set category of answers only which are the once that are acceptable to this investigator. These could be two types, we could have close questions with only two options, dichotomous options such as yes-no, male-female and so forth. So, these kind of question (Refer Time: 10:54) is a clear position for being respondent to take and it is very useful to get key information specially for important issues and it is; which is very focused. Although sometimes depending on the question it may actually over simplify some of the issues, where a yes-no answer is not something that is going to give you a very good information.

(Refer Slide Time: 11:19)

## Closed questions: 2. Multiple options

- Multiple options of answers are suggested
- One or multiple answers acceptable
- Advantage
  - Larger choice of answer options
- Inconvenient
  - May be difficult to choose only one option

### Example:

- Where do you go to seek treatment for fever?
  - Government Hospital
  - Private clinic
  - Pharmacist
  - Traditional healer
- Do you wear a helmet when riding a bike?
  - Always
  - Sometimes
  - Never

NICER IOL  
HEALTH RESEARCH FUNDAMENTALS



nie.gov.in

Then we have close questions with multiple options, so more than two options. Now here again, they can be two kinds of close questions with multiple options. We could have questions where although there are multiple options, only one of the option; is acceptable. So, depending on what the respondents says, one of this option is ticked by the data collector. On the other hand, we could have closed ended questions with multiple options, wherein even multiple responses by the study participants may be acceptable.

The important thing to note is that, while you are designing the questionnaire you need to put a clear instruction for the data collector, whether only one answer is acceptable for this question or more than one answer may be acceptable. So, we have a large choice of answer options; again depending on how the question is framed and what the actual question is, sometimes it may become inconvenient and difficult to choose only one option, if there is a possibility of more than one option, but that possibility is not provided in the questionnaire.

(Refer Slide Time: 12:39)

## Closed questions:

### 3. Quantitative answers

- The respondent must provide a quantified answer
- Advantage
  - Allows creation of continuous variables
- Inconvenient
  - May require validation:
    - Some “quantified” answers might be limited in the way they can be handled as continuous variables

#### Examples:

- How many times did you visit the clinic in the last 12 months?
  - True continuous variable
  - Four visits is the double of two visits
- How would you describe your pain on a 1-10 scale where 1 would be the minimum and 10 would be the maximum?
  - In fact a qualitative variable with 10 options
  - Requires validation
    - Six may not be the double of three on the scale

So, we have to be mindful of this when you are designing your questionnaires and thirdly, we could have close questions which have quantitative answers, where the respondent has to provide a number such as age, such as may be one example here, if you see; how many times did you visit the clinic in the last 12 months? These kinds of questions allow the creation of continuous variables and then measuring and doing the analysis for continuous variables, if you need we can always categorize these variables later on in the time of analysis if needed.

However, sometimes it can become inconvenient to give a quantitative answer because some quantified answers may be limited in the way they can be handled as continuous variables and where the number itself is difficult to interpret. So, we need to be careful in how we are framing these kinds of questions.

(Refer Slide Time: 13:33)

## Semi-open questions

- Suggested answers
- Possibility to create another answer
  - Other, specify: \_\_\_\_\_
- Advantage
  - Leaves the door open to unplanned answers
- Inconvenient
  - Difficult to analyze

### Example:

- Did your child have complication following measles?
  - None
  - Pneumonia
  - Diarrhoea
  - Eye problems
  - Other, specify: \_\_\_\_\_



Then there we could have something called, which are called semi-open questions, where you basically have a question with several responses, the answers here are suggested, but there may be one option where, which is kept open and the most common one that we see in data collection tools and questionnaires is others. For example, did your child have complication for measles, it been that could have not have been any complication, that could be pneumonia, diarrhea, eye problems or they could be some other complications, which may not be so common so has to be put in a category, but then you gave an option to the respondent tools even say things other than what is in the list of categories of responses. So, it leaves the door open for unplanned answers; however, if there are too many of these responses, it may sometimes become difficult to analyze.

(Refer Slide Time: 14:34)

## Formulating questions

- Write short and precise questions
  - Full and complete phrases
  - Avoid ambiguities
- Use simple words of every day language
- Avoid negatives and double negatives
  - ✗ Do you sometimes care for patients without washing hands?
  - ✓ Do you systematically wash hands before caring for each patient?



Now, let us look at, how some of the principles and do's and don'ts of formulating the questions. It is always the good idea to actually short and precise questions say for example, if you want to know the age of your study participant, just writing age is not a good idea. You should always use full and complete phrases, so what is your age? So avoiding ambiguities, it is a good idea to use simple words and not use very complicated academic language, use everyday language in terms of questions because again remember that your respondents are lay people.

When you are formulating questions, again it is good idea to avoid negatives; especially double negatives. So, one example that we have here is do say, do you sometimes care for patients without washing hands. If you see carefully, there is a negative connotation here and there are sort of two parts of the questions, one is caring for patients and the other one is washing hands or without washing hands. So, a better way to phrase this question could be to ask it directly and more positively, so do you systematically wash hands before caring for each patient, which makes it clear and unambiguous.

## Formulating questions

- Ask only one question at the time
  - ✗ Did you refuse treatment because you feared side effects?
  - ✓ Did you refuse treatment?
  - ✓ If yes, was this because you feared side effects?
- Be specific
  - ✗ Are you aware of the modes of transmission of HIV?
  - ✓ Among these practices, can you tell me those that could expose you to HIV?
- Use neutral tone to avoid influence
  - ✗ Have you been promiscuous in the last six months?
  - ✓ How many partners have you had in the last six months?

Again when you ask a question, it should be only one question at a time; say one example here they we have here, so did you refuse treatment because you feared side effects. Now, here actually we have two questions; one is asking; did you refuse treatment? And the other is trying to find out the reason of why, if the person refuse treatment, why did they do so. It may be that the respondent may not have refuse treatment then how does that respond and answer this question. So, it is a good idea to actually split such questions into two questions, where in one could be first; the first part say could be, did you refuse treatment? Depending on yes or no, if the answer is yes, the following question could be; was this because you feared side effects? It makes things very clear.

Again, the questions need to be specific and not weighed, so an example that we have here is say, you want to know from the people about how HIV is transmitted and the question is; Are you aware of the modes of transmission of HIV? Which is sort of an open ended, it lives space for people who answer whatever they may want to answer. But if you really want to know whether HIV is transmitted through sexual route, through heterosexual route, homosexual route, blood transfusion, drug use, etcetera, then it is better to actually put these as categories of responses and then phrase the question as, among these practices can you tell me those that could expose you to HIV.

So, you know that you are going to get the proper answer to the question in which you wanted to have. It is also good idea to always use a neutral tone and avoid judgmental tones which can influence the response of the study participants. Remember that you are there as a data collector, we just collect data and not be judgmental of what the respondent is telling you. One example that we have here, so if you want to know about say the sexual practices of people, instead of asking them have you been promiscuous in the last 6 months. Again, the words promiscuous has a negative connotation , so instead of that it could be more neutral and more academic kind of a question, wherein you just be direct and specific and ask about; How many partners have you had in the last 6 months without being too judgmental.

(Refer Slide Time: 18:36)

## Sorting the order of questions

- From simple to complicated
- From general to specific
- From casual to intimate
- Group together questions related to the same topic
- Identification questions at the beginning or at the end
- In chronological order, if questions related to sequence of events
- Introduce simple questions as a break if the questionnaire is complex
- Triangulate through multiple questions on the same topic if the subject is important



Now, the next thing that you would want to do, when you are designing your questionnaire is, to actually sort out what would be the order of the questions that you have. Remember, that the way in which you ask questions should be such as if your having a conversation with a study participant and it should be a smooth flow of questions one after the other link to one another.

Some general principles to keep in mind; is that always ask simple questions first and keep the complicated questions for the later part. You can ask more general questions,

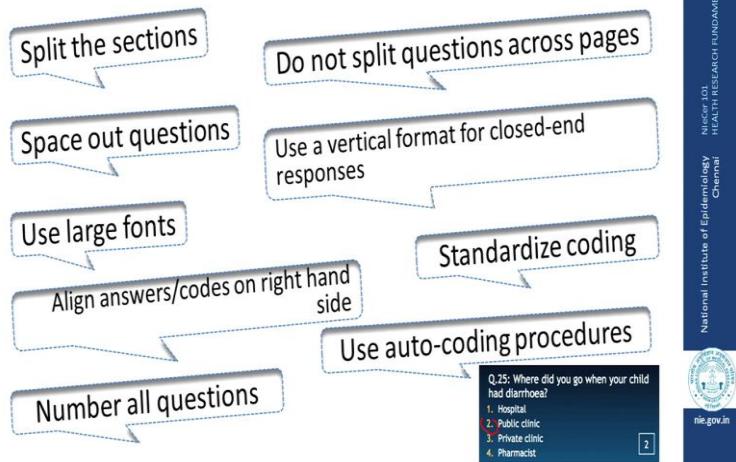
socio demographic characteristics like that and then go on to the specific questions of what you are study is all about. Ask more casual questions in the beginning, which the respondent will be easy to answer more of factual questions and then more intimate questions, sensitive issues, questions about attitudes and opinions could be something that could come later on.

It is always again a good idea remember to actually group together all the questions which are related to the same topic of inquiry. Instead of having them spread across different parts of the questionnaire, which can actually confuse the respondent. In terms of asking the identification questions either they can all be asked at the end in terms of the name, age, gender, say the address and so forth or they could even be asked at the end so as to get to the specifics right at the beginning.

If you are trying to collect information about a sequence of events, then your questions should be in that proper chronological order of how things would have happened in real time, which will actually help the respondent to recall the responses in a better way and also in a more logical way. Again, if your questionnaire is complex, there are lots of questions, it is always good to give a break in the middle and maybe have some simple questions and then come back to your complex questions. Many a times, we may ask the same kind of question in different ways in the same questionnaire and then that is used, if that is the subject matter that is really important for the study and you really want to know that the respond what the respondent is telling you make sense is valid is true and then. So, the multiple questions on the same topic could be asked at different places in the questionnaire and then when you are trying to analyze it, you can triangulate these responses to get to what the information you would like to extract from this questions.

(Refer Slide Time: 21:27)

## Laying out the data collection tool



nie.gov.in

Once you have sorted out the order, now what is needed is to actually lay out all these questions in the questionnaire. Again remember, it is a laying out the format the structure of the questionnaire is again critical because the way the respondents look at the questionnaire, the design and look of the questionnaire influences or can influence the response of the participants. So, if you have different sections in the questionnaire, it is a good idea to split the sections, have one sections, may be have line and then have the next section.

Do not try to cramp questions all together, have spaces between the questions, so there it is readable clearly. Try to use large fonts, not too small fonts so maybe font size of 11 or 12 would be ideal. Again do not split questions across pages, if you have a question and then if half of it goes to the next page, it becomes difficult for the data collector to actually read the question, you will have to turn the page and so forth. So, if that is happening, make sure that you sort of bring the whole question on one page. In terms of formatting and aligning, alignments are again gives you a nice, neat look for the questionnaire. So, it is a good idea to actually align your questions on the left hand side and your answers and course on the right hand side, which gives a neat two column kind of a look to your questionnaire, which makes it more appealing.

Do not forget to number your questions starting from 1, 2 ever that you have. Again, one very important thing to keep in mind is coding. Remember that ultimately what you are going to do is use this data and to enter it into software, give it codes to analyze it. So, it is always a good idea, to actually have a coding system inbuilt in the questionnaire itself. So, you need to standardize your coding so wherever say, the simplest example I could give you is, a yes, no. So, you could have a code of 1 for yes and 0 for no for example. Make sure that every question, every response where you have yes, no; you have coded it as 1, 0 say you have male, female. So, you always code the male may be as one and the female as two or something like that.

Another way to actually simplify this coding is to use what we call as auto-coding. So, the numbers that you give to the categories, say you have four response categories. So, you would number them 1, 2, 3, 4 and if say the response is number 2, then you use the same number 2, as a code for this question items. So, these are some of the ways in which you make sure that the layout of your questionnaire is neat and it is presentable and it something that helps both the data collector as well as the data entry operator and the person who is going to enter and analyze the data at the later stage.

(Refer Slide Time: 24:44)

## Finalizing the data collection tool

Checking the instrument against the objectives/analysis plan	Reviewing the instrument	Language of the instrument
<ul style="list-style-type: none"><li>• Suppress unnecessary questions</li><li>• Add missing questions</li></ul>	<ul style="list-style-type: none"><li>• Colleagues</li><li>• Experts</li><li>• Statisticians (Coding)</li><li>• Field workers</li><li>• Data entry operators</li></ul>	<ul style="list-style-type: none"><li>• Write in the language in which they will be administered</li><li>• Translation is required<ul style="list-style-type: none"><li>• Initial formulation (e.g., in English)</li><li>• Translation (e.g., in Hindi)</li><li>• Back-translation (e.g., back to English)</li></ul></li></ul>



When you are finalizing your data collection tool, make sure that the questions that you have are something that is relevant to the study that you are doing. So, as may have been mentioned earlier sessions, the investigator needs to be a slave of the study objectives and what the analysis, the analysis that is already pre-planned for the study. So, make sure that the questions that you have are relevant for this to answer those study objectives. Do not put unnecessary questions just because you are going out in the field and doing a study does not mean that you can ask anything and everything and if you feel that there are certain missing questions make sure to add them.

Once you have done all this, it is also a good practice to actually review your instrument before you take it to the field. The reviews could be done by your colleagues and experts in the field, you could also give it to the statisticians, to actually review look at the coding; whether that is going to be something that is going to be useful for them and then you could even the field workers or the data collectors, who are going to actually collect the data, they can be your key informants to actually go through the questionnaire and tell you whether the flow is appropriate, whether the questions make sense, is there any ambiguity or is there something that is not understood and so forth.

Keep in mind that the language of the instrument; the questionnaire or the data abstraction form or so forth has to be in the language in which you are going to you interview the study participants. So, if your study participants speak Tamil, then the questionnaire should in Tamil, if it is Hindi, it should be in Hindi. Generally, as investigators, English is the common language, so we may be your initial formulation of the questionnaire would be in English. Then what you need is a translation, you need to translate it into the local language and then very importantly have somebody else to back translation into English, so as to make sure that the translated version in the local language makes the same sense as you wanted it to be and when you frame those questions in English.

(Refer Slide Time: 27:05)

## Pilot testing the data collection tool

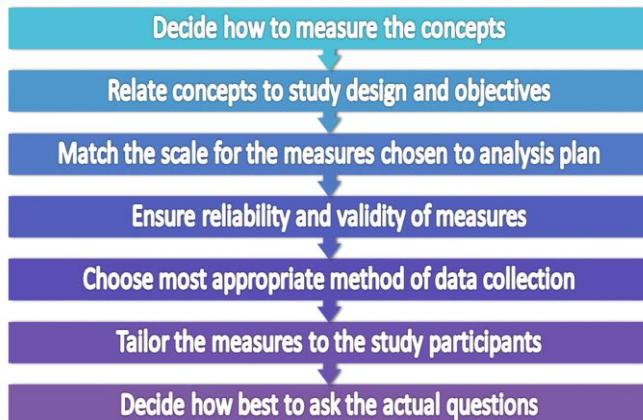
- Check that the instrument is:
  - Clear
  - Understandable
  - Acceptable
- Check flow and skip pattern
- Check pertinence of coding
- Estimate the time needed to ask all the questions
- Pilot test with a few volunteers
  - Persons similar to the study population
  - Persons who are not to be included in the study



Before going to the field, it is always important to pilot test your tools. You need to make sure that your study instrument is clear, the questions that you have asked are understandable to the people and they are acceptable, people are not vary of answering those questions. You need to check the flow and the skip patterns make sure that the coding works and it also gives you a sense of how much time it is going to take for you to actually finish the questionnaire. All this can be done by doing the pilot testing by actually administering this questionnaire to a few volunteers who are similar to the study population that you are going to do, but remember that these people on which you pilot test your questionnaire, should not be included later on in the main study.

(Refer Slide Time: 28:57)

## Designing health research tools



So, when we are designing the health research tools, we need to keep certain principles in mind. You need to first make sure, what is it that you want to measure, remember epidemiology is all about measure. Then you need to relate these concepts to your study designs and the study objectives, you need to match the scales, how you are going to measure these to and then how you are going to do the analysis. Make sure that the scales, the questions, the questionnaires, the data collection tools that you are using are reliable and valid for the population that you are going to apply them to. Taking all these things in mind, choose the most appropriate method of data collection; whether it is a data abstraction form or a structure observation guide or a questionnaire and the type of questions that you are going to put in these data collection instruments.

Keep in mind your study participants, in terms of the language of the questionnaire and also the way in which you are trying to measure the concepts that you are doing and then decide finally, how best you are going to ask the actual question in the study questionnaire. Remember, a study questionnaire can make or break the study; this is something, once you have collect a data you may not have the opportunity to go back. So, it is essential that the data that you collect is valid and reliable and in order to do that, it is key that the data collection instruments that you develop are something are totally valid and appropriate to the study that you are trying to conduct. That is it for today.

Thank you.

**Health Research Fundamentals**  
**Dr. Prabhdeep Kaur**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**Lecture - 16**  
**Principles of data collection**

Welcome to this session of Health Research Fundamentals. Today, we will be discussing about Principles of data collection. As you all might have gone through various parts of this course, wherein we had been teaching you about, how you develop your research proposal? How you write your protocol? And now you have gone through all that and you are all set to collect your data. Let me tell you that, this is one of the most important components of your research study because this is what is going to determine, what you get out of your data.

(Refer Slide Time: 00:40)

## Data quality

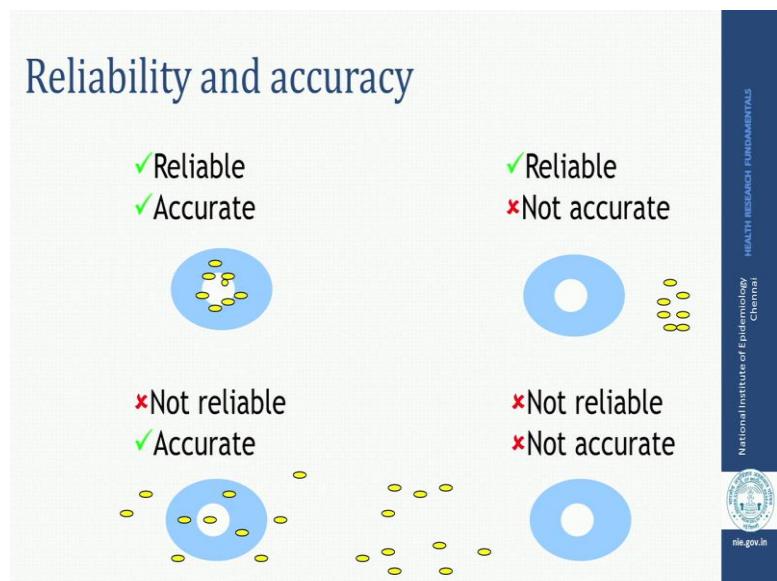
- Reliability
  - Reproducibility/repeatability/precision
  - Ability of a measurement to give the same result or similar result with repeated measurements of the same thing
  - Refers to stability or consistency of information
- Accuracy
  - Ability of a measurement to be correct on the average

HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai  
  
niit.gov.in

So, today as we go through this lecture let me discuss about two main principles of data quality, which is reliability and accuracy. So, reliability, what does it really mean? It means that your study should have repeatability and precision. What it means? If this same study is repeated by different investigators or the same measurement is done by

different at different time points, you should still be able to get more or less similar results. It also refers to stability and consistency of the information. Now, reliability does not ensure that your data is accurate. Accuracy refers to ability of a measurement to be correct. It could happen that both, the attributes may not happen at the same time or they may happen. So, let me give you an example of what would be the various scenarios in terms of data quality.

(Refer Slide Time: 01:42)



As you can see in the slide, your study might give you reliable and accurate result, which means your data is giving the measurement that you really want, the measurement is accurate and repeatable. This is an ideal case scenario. The worst case scenario is that, your measurement may not be reliable at all and may not be accurate also. Now, there can be other scenarios, where your measurement is reliable, that is, if you repeat this measurement again and again, you get the same measurement, same kind of result repeatedly. However, this could still not be accurate and there can be another scenario that you may get the accurate result, but when somebody tries to repeat this experiment you may not be able to repeat it at all. So, an ideal study should ensure that whatever outcome you are measuring is repeatable as well as it is accurate, to ensure that, we can follow certain principles of data collection.

(Refer Slide Time: 02:44)

## Six steps in data collection

1. Draft question-by-question guide
2. Train staff members who will collect data
3. Initiate data collection and ensure quality
4. Review collected data for quality and completeness
5. Debrief to trouble shoot difficulties
6. Validate

HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

In today's lecture, we will be going through six broad principles which are listed here. I will be walking you through each of these principles. So, the first is, as you have your questionnaire ready and you are all set to start your data collection. One more step you need to do is, you need to develop a guide, wherein for each of the questions you can have a small explanation as to how this data actually need to be collected. The second component that we will talk about is how you could do a good training of your staff members to ensure that the data quality is good. The third component is how do you ensure data quality, when the work has already been initiated. It could be in a clinic setting or it could be in the field and the next component is how you need to do periodic reviews to ensure good quality of your data.

During this process, it may happen that your staff will come across various difficulties in data collection and you need have periodic debriefing for that and finally and very, very important component is how you validate that your study results are actually correct. So, let us go over each of these components one by one.

(Refer Slide Time: 03:58)

## 1. Draft question-by-question guide

- Short document to be understood as a guide for field workers
- Consider each question, number by number
- Provide guidance as to how the data should be collected
- Used as a road map for good data collection
  - Drafted initially
  - Revised as issues arise and are addressed

HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

So, what do you really mean by a question by question guide? It is nothing but a short, simple document, which is a guide for your field worker or your clinic research staff or whoever is involved in the data collection. Now, why this is an important component? Everybody may not understand the question in the same way. Different individuals or your staff members may have their own interpretations as to how these questions need to be asked and how it needs to be explained to the respondent. The q-by-q guide ensures that all the investigators uniformly ask the question and they all understand the question in the same way. What you do here? You just take your questionnaire and write a short explanation under each of your question, as to how this question should be asked and what could be the possible responses and in certain situations you may have some probes. You would like the investigator probe little more on a particular question. If you want them to do so, you would like to make a mention of that.

Similarly, it allows you to explain, where they should skip? Where they should give more emphasis? All this guidance can be given through this small document. This not only provides a guide as to how data should be collected, but it is your road map and at any point of time, whenever there is a doubt, whenever there is a lack of consistency with an investigator you can always ask them to go back to this document to ensure that the data collection is being done in a uniform way. Now, it could happen that as you start

collecting data, you may come across certain difficulties in the field and you may have to revise this guide, which is all right and what is important is whatever changes are done, they should be documented in this q-by-q guide at that point of time.

(Refer Slide Time: 05:52)

## Example of Q by Q guide

- **Question 6 (Housing):**
  - Observe the house and note if made of mud or bricks
- **Question 12 (Household income) :**
  - Identify all the person with financial income in the household
  - Estimate each source of income
  - Sum up to generate household income

National Institute of Epidemiology  
Chennai

nie.gov.in

This is just to show you an example. For example, you have a question on, what is a type of house like we could give several options, this is a kutcha house, it is semi pakka house or a pakka house. Everybody may not understand, what is a kutcha house? So, you may want to give an explanation. Observe the house and note, if it is made of mud or bricks and then you could write an explanation. If you find that the house is made of mud, then mark it as a kutcha house.

Similarly, another example when you are asking a question regarding household income? Now, it could happen that in a particular house hold there is more than one member, who are earning members of the family. And in this situation, you would like your investigators to first of all inquire, how many earning members are there in the family? Then they need to find out each one of them, how much they earn? And what is their monthly income? They need to add up all, that amount and then write the answer as what is a total household income. So, similarly for each question you can have the detailed explanation.

(Refer Slide Time: 06:58)

## 2. Train staff who will collect data

- Select good, experienced investigators
- Present the study and its objectives
  - Slide presentation
- Distribute the q-by-q
- Walk people through the q-by-q
- List tasks to be conducted
- Answer questions
- Simulate interviews within the team

HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

Now, once your q-by-q guide is ready, you are all set to start your data collection. Before you get on the ground, whether it is a clinic, whether it is a field base study, it is a household study, whatever will be the type of study, training, good training is essential to ensure that the quality of data is good and in that, the first and foremost step is choosing right kind of people. You need to understand, what your study requirements are and depending on the study requirements, you need to recruit or select right kind of investigators. Now, this could be depending on, suppose you are doing a study, which on clinical research. Then you need to make sure that your investigators know how to ask the clinical questions. Are they knowing, how to use the clinical terms, when they are interacting with the patient? If it is a field base study, you need to make sure that your investigators are familiar with the local language. They know how to culturally interact with the people at the community level. So, selecting the right kind of investigator, who will be suitable for your data collection, is the most important step.

Once you have done that, you need to have a classroom session with them. This is nothing, but walking them through your, first of all you introduce them to what your studies all about, why are you doing this study? What are your objectives? And what is that you would like them to know the basics of, what are the various definition you are using in the study? What are the various component of the study? What is the kind of

data that needs to be collected? For example, your study may have different component, there could be a questionnaire, there could be different measurements and maybe you want them to measure something like weight, height. So, you need to familiarize them, what are the various kinds of a data collection that they need to do?

Once you have done that and given them the overviews then you need to share this q-by-q guide, question by question guide. Walk them through each question, allow them to ask questions and let them go through it and see whether they are able to interpret those questions, whether they are able to understand those questions. Once, they have done that, you need to tell them what exactly they need to do in the field? How they need to ask those questions? And what should be the explanation that they should understand before they ask the questions?

Now, as a first step having gone through this exercise, you would like them to probably do this interviews, first of all in the classroom, wherein one of the investigator could act as the interviewer and the another investigator could be the respondent and they should simulate this in front of you, which allows you to guide them, to tell them, how the question need to be asked. Whether the question is being understood by the respondent, whether question is being asked in the appropriate language that is given in the questionnaire. Having done that you are all set to start your data collection, but there is one more step.

(Refer Slide Time: 09:50)

### 3. Initiate data collection and ensure quality

- Do pilot on site interviews under supervision
  - Note issues that may come up, resolve them as a group
  - Continue until the procedure is clear to everyone
- Plan data collection process with a supervisor and investigators
  - Ensure study forms are verified by the supervisor every day for any errors
  - Be available to answer questions
  - Do onsite visits
  - Do not press for quick completion

HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

Now, having done something in the class room does not mean you will be able to do it on the ground because the real situation may be very different. Some of your investigators may be working in that kind of a situation for the first time. For example, you may hire a person who is trained in clinical research, but they have never been in a clinical setting. Similarly, you may hire a field investigator, who is a graduate in social work or many other any other background. However, they may never have been in the community setting. So, once they get on to the ground, they may not feel the same level comfort. So, before you get on to the main data collection, you need have pilot. A pilot, where they do the interviews exactly in a similar setting, what your study setting is, but it may not be the same place, it may not be the same clinic or it may not be the same village.

In this situation you will see that, they may come up with different kind of questions. They may not have anticipated, how the respondent will respond. They may not have anticipated what difficulties might be there in making the respondent understand the questions. So, as you through this pilot, you explain them the process, clarify their doubts and this procedure you could continue unless everybody is very, very clear about what the questions are and how the questions need to be asked.

Now, having gone through this, you are all set now to do your final data collection. Here, what is very important is that, you need to plan your team structure. So, there has to be a supervisor, who could be present all the time or who could be present periodically along with the investigators. Ensure that the study forms are verified by supervisor every day. Now, why is this important because once you have left the clinic or once you have left the village or the place where you are doing data collection, some of the errors cannot be corrected, during, as the data collection starts they may come up with different queries. So, you need to be available to answer those questions. This could be over the phone; this could be over messages or whatever way you can use.

Now, once the data collection is ongoing, the next step that as a principle investigator you need to do is few on site visits, to ensure that the data collection is done as per the protocol and the questionnaire is being actually used in the field the way you had perceived it. I think one of the important think you need to keep in mind is very often what happens is there are time pressures. You want to finish your study quickly, you want to be done with it, you may pressurize them, you may say, I would like you to do ten questionnaire every day. However, you need to make sure that the quality is not diluted because you are pressing them to work under very tight time lines. Having done this, having your data collection, as a data collection goes on at various time points you may want to review.

(Refer Slide Time: 12:52)

#### 4. Review collected data for quality and completeness

- Each team checks the data collection instruments before the respondent leaves
- The supervisor checks the instruments before leaving the location
- All take responsibility for the instrument:
  - Names and signatures
- Principal Investigator checks instruments as they come

Suppose, let us say you have already collected now, 50 forms. At this stage, you may want to make sure that the data collected is of good quality and the data is complete. How can you do that? The first step can be done in the field itself. As your supervisor finishes the day, he can collect all the forms from various team members and check them on that day itself. The next step is the forms will reach you as a principle investigator and you need to go through them to make sure that you are satisfied with the quality of data.

(Refer Slide Time: 13:35)

## Checks to conduct

- Completeness
  - Did the field worker fill all items?
- Readability
  - Is the writing readable?
- Consistency
  - Do the answer make sense?
  - Is there internal consistency?

National Institute of Epidemiology  
Chennai  
  
nie.gov.in

Now, what kind of checks you can do, when you have got the forms in your hand? I think the first and foremost and the easiest one, but the most important one is ensuring there are no blanks. It could happen that when they are asking multiple question, since your questionnaire may have about 20 questions or 25 questions, they may just forgot to ask a few questions or they may have asked the question, but they did not mark it in the questionnaire, but in either case you will see them as a blank. So, you need to ensure completeness of the data.

The second is readability, sometimes the way it is marked it may be hard for you to even understand and some of the answers could have some actual explanation to be written. Some kind of clinical symptoms or some kind of narratives has to be written, ensure that those things are readable. The next is the consistency, does the answer make sense? Do you feel that this could be the way the people would have answered or you got this answer because may be your respondents were not understanding the questions.

(Refer Slide Time: 14:41)

## 5. Debrief to trouble shoot difficulties

- Regular meetings
  - Evening or morning
- Facilitate a discussion about
  - Issues identified
  - Clarification needed
- Make note of decisions on the q-by-q if needed

HEALTH RESEARCH FUNDAMENTALS

National Institute of Epidemiology  
Chennai



nie.gov.in

Periodic reviews are essential, in case of any study, whether it is a clinical study; it is a field study or any type of research study. Now, these regular meetings, one it could be done by the supervisor, which could be in the field itself or after coming back or you, can have a periodic review once in a week, once in a month, depending on what is duration of your study. What is that these review meetings will be useful for? First of all to clarify, if there are any queries they have about the questionnaires, if they understand the questions well, if they come across something which they had not anticipated. Now, it may happen that during these meetings you may end up doing some changes in the way your answers are drafted or in the way questions are asked. If you are doing that ensure, that these are well documented and they are added it in your q-by-q guide.

(Refer Slide Time: 15:36)

## 6. Validate

- Select a number of study participants at random
- Conduct a second interview
- Compare results
- Debrief discrepancies with:
  - Individual worker if the errors are made by a particular investigator
  - Whole team if the issue is relevant for all

HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

Now, coming to the last, but not the least very, very important component of data collection, which is validation of your data. How can this be done? So, for any study a small sub sample, it could as small as even 5 percent should be selected and an independent second interview should be done. This is to ensure that the data collected by your investigator is actually valid. By comparing the results, you will be able to find out, if there are any discrepancies, if somebody has made any major errors or if it could be that a particular investigator might have been repeating certain errors or it may happen certain errors are repetitive across the team. So, depending on whatever the problem is you may want to discuss this with the individual team member or with all the study team.

(Refer Slide Time: 16:27)

## Take home message

- Understand the concepts of data quality
- Good training off site and onsite is essential
- Supportive supervision and team work are key to good quality data collection

HEALTH RESEARCH FUNDAMENTALS

Chennai



nie.gov.in

So, just to sum up, what we have gone through today. Before you get on to your data collection understand the concepts of data quality, good training both, in the classroom as well as on site, in a similar setting where you are going to conduct the study is essential for a good research study and supportive supervision, team work are key to good quality data collection.

Thank you very much.

**Health Research Fundamentals**  
**Dr. Manickam Ponnaiah**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**Lecture - 17**  
**Data management**

Hello friends. Welcome to this session, the course Health Research Fundamentals. We are in the last stages of our course. We have already covered important topics such as Importance of data collection, Importance of validity, Importance of measurements and now I think we are in a position to understand the Importance of the Principles of Data Management and Analysis.

(Refer Slide Time: 00:41)



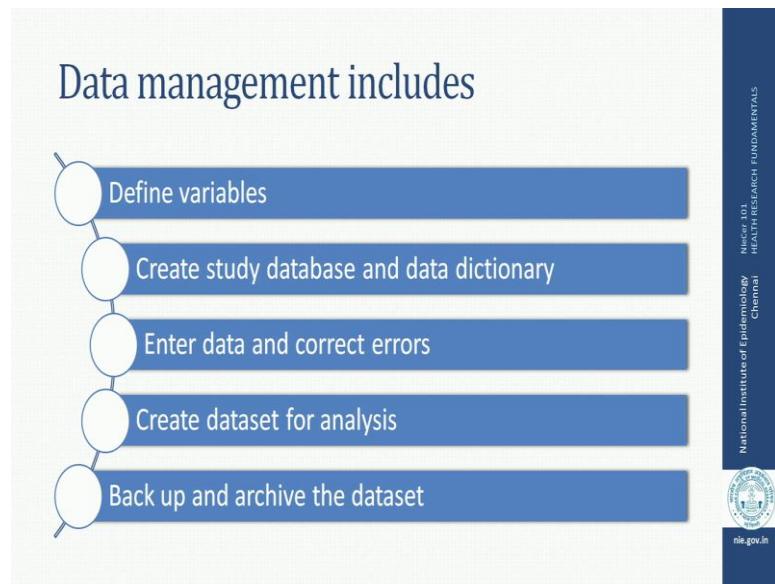
**Key areas**

- Database management
- Data analysis strategy

National Institute of Epidemiology  
Chennai  
NIE GOVT OF INDIA  
nie.gov.in

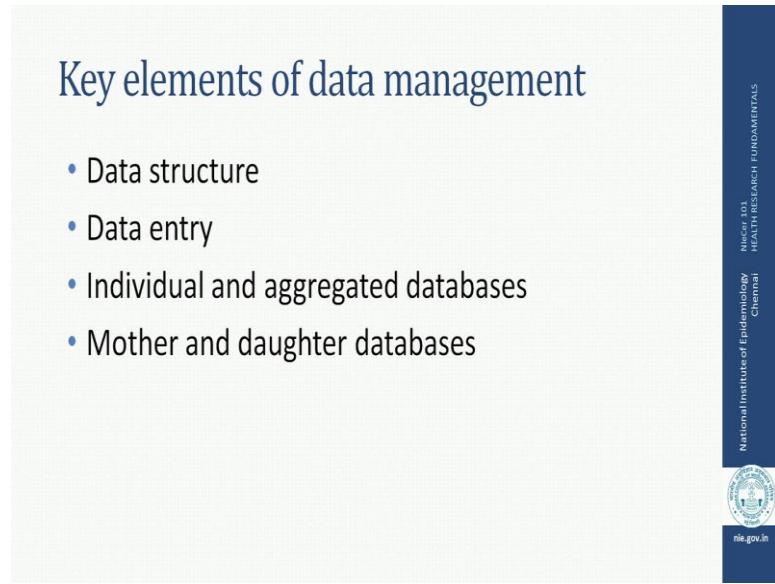
Today, we are going to cover 2 areas; one, database management; the other one, data analysis strategy.

(Refer Slide Time: 00:49)



Data management includes the following; Defining variables, creating a study database and data dictionary about which I am going to expand, entering the data and correct the errors, create a data set for data analysis, backing up the data and archiving the data set.

(Refer Slide Time: 01:10)



Today, we are going to cover the key elements of data management, including Data

structure, Data entry, Individual and aggregated databases, Mother and daughter databases.

(Refer Slide Time: 01:22)

## Basic structure of a database

- Lines represent records
- Columns represent variables

	Identifier	Variable 1	Variable 2	Variable 3	Variable 4	Etc...
Record 1						
Record 2						
Record 3						
Etc...						

National Institute of Epidemiology  
HEALTH RESEARCH FUNDAMENTALS  
Chennai



nie.gov.in

What is data structure? For some of you, who may be new to the word database this is an example of a database. Each of these lines, these horizontal lines represent records, but I need to one particular individual, each of these columns represent a variables, the information that is collected on certain variables based on the study (Refer Time: 01:51).

(Refer Slide Time: 01:53)

## Data documentation

- Structure
  - Name, number of records etc
- Variables
  - Name, values, coding
- History
  - Creation, modification
- Storage information
  - Media, location, back up
- Additional information

Structure

NATIONAL INSTITUTE OF EPIDEMIOLOGY  
HEALTH RESEARCH FUNDAMENTALS  
Chennai  
  
nie.gov.in

We need to initially formulate the entire plan of how the data is going to be managed. That can come in the form of what is called data documentation. It can talk about the structure, meaning the name, number of records and other relevant information about the structure. The variables in terms of the name, what values that are assigned to the coding, etcetera and the history of this database in terms of when it was created? When it was modified? The storage related information, in which media it is going to be stored? Where and how it is going to be backed up? And any other relevant additional information is recorded in terms of the structure.

(Refer Slide Time: 02:37)

## Identifier in the database

- Unique
- Maintained by a computerized index
- Secured by quality assurance procedures

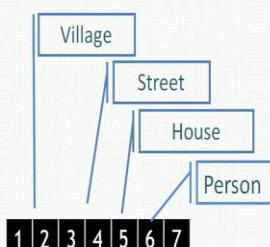
NIECER IDI: HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai  
  
nie.gov.in  
Structure

Let us look at the first of the important elements in database, this is called Identifier. This identifier has to be unique, that is why it is called unique identifier. It is maintained by a computerized index and this particular unique identifier has to be secured by a quality assurance procedure that guarantees that each of this data has its own internal validity.

(Refer Slide Time: 03:09)

## Using codes within the unique identifier

- Unique identifier may contain all information about that particular ID
- Each digit or set of digits refer to specific information
  - Example:
    - First and second digit: village
    - Third and fourth digit: Street
    - Fifth digit: House
    - Sixth and seventh digit: Person



NIECER IDI: HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai  
  
nie.gov.in  
Structure

The code can comprise information that will talk about that particular individual. For example, it can have 7 digits; each of this digit or set of digits can refer to a specific identifier information about that particular individual, about which, about whom the data was collected, for example, in this example the first and the second digit can denote the village or area, the next set of 2 digits the third and fourth may denote street, the digit number 5 may indicate the house or you know flat or residence, door number, the last 2 digits can denote the persons sequential number. Therefore, the 7 digit may represent about that particular individual and by parts it can give information.

(Refer Slide Time: 04:08)

## Structure of the variables in the database

- Integer
  - Specify the number of digits
- Numeric
  - Specify the number of decimals
- Alpha-numeric
  - Specify length
  - Turn all letters to capitals
- Dates (specific format)

NATIONAL INSTITUTE OF ENVIRONMENTAL SCIENCE AND TECHNOLOGY  
NIEHS RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai  
nie.gov.in

You need to specify certain things about variables. The variable as you might have seen in the lecture on measurement, there are different types of variables that do exist and therefore, it requires your attention in the beginning itself. You need to specify whether the variable will be entered as digits or if it is a numeric, whether the number of decimals are important, the variables can be entered as length, in which case you need to specify the length and it is preferable that when you are entering as both text and number, you turn all the letters into capitals to avoid errors, which can cause a lot of problem in data analysis. And finally, you can have dates in specific format; when I say specific format you need to specify whether it is entered as the Indian format date, month and year in 4 digits or month, month, date and then year; this has to be specified in the structure of the

variable.

(Refer Slide Time: 05:19)

**Creating variable names**

- Clear
  - Need to refer to the questionnaire item
  - Understandable (e.g., "EXERDAILY" for "Exercise daily")
- Short, no space
  - Most softwares require less than 10 characters
- Consistent
  - "EXERPAST" for "Exercise daily in the past"
  - "EXERCURRDLY" for "Exercise daily in the current"
  - "EXERPASTOCC" for "Exercise occasionally in the past"
  - "EXERCURROCC" for "Exercise occasionally in the current"
  - "VARIAB" for all crude variables (EXERCISE)
  - "VARIAB\_12" for all dichotomized variables (EXERCISE\_12)
- No duplicate
  - Trimming of names by software can create duplicate name

National Institute of Epidemiology  
Chennai

Structure

nie.gov.in

While creating the variable names, which pertain to your data collection instruments about which you will have a clear idea later part of the course; you need to be very clear. The name, variable name should refer to an item in your data collection instrument. It has to be understandable format for example, if the questionnaire item is about exercising whether the individual exercises daily or not. The variable name could be EXERDAILY; exercise daily that clearly denotes what questionnaire item it refers to. The second important criteria are in terms of keeping it short, leaving no space between the letters of the variable name. Most softwares may require less than 10 characters therefore, you have to be very choosy at the same time it is self explanatory.

The third important aspect is, Be consistent for example, for different types of response to a question on how frequently somebody exercises, exercising daily in the past can be denoted by ECERPAST, exercise past that clearly know self explanatory about the particular questionnaire item. If it is currently daily then it is coded accordingly, if it is past occasionally those words are given in the variable names so that by looking at the variable name the investigator can easily identify; this is a variable, this is the questionnaire item it refers to.

And finally, you may have variables collected as such these are called crude variables for example, it may refer to number of times one exercises; it could be 3 times, 4 times, 5 times a day. And finally, you may decide to regroup them into 2 categories; exercised or not, in which case you can denote that variable as exercise in the crude variable. As such, when the data was collected you can change into EXERCISE underscore 12, which denotes it is dichotomized. It is dichotomized into exercised or not. So, you have to be consistent in the pattern by which you create variable names. And finally, it is very important that you assign a variable name, otherwise if the software is left to assign a variable name by itself, it can create lot of confusion including duplicates of similar items within the questionnaire.

(Refer Slide Time: 07:58)

## Design data entry-friendly data collection instrument

- Outline
  - Identifiers
  - Demographics
  - Outcome (Health problem/disease)
  - Exposures (variables, including third factors)
- Auto-coding function



When you are designing a data collection instrument, about which my colleague is going to expand in the later part of this course; it is important that the design itself you are very clear about broad sections of the questionnaire, so that when it is converted into database you know that there are sections that you have to enter for example, there is a section called identifiers; there is a section called you know demographics, which means you know one talks about age, gender, community and family related issues; and then the third section called outcome. It is about the problem in the question and or disease if it related to typically clinical related information.

And finally, another section called exposure in which you can talk about all the variable that you are going to measure including, what Dr. Tarun might have already talked to you about third factors including confounders. Finally, the instrument should allow an auto-coding, if you collect information on exercise daily; Yes - No; if it is written already written as 1 2, we need to enter into the database 1 or 2, we do not have to code it again. So, that is what is meant by auto-coding. So, the data collection instruments should be designed in a way that it facilitates a data entry design easy.

(Refer Slide Time: 09:23)

## Coding

- Prefer numerical coding
- Decide on
  - Missing values (.) or (9, 99, 999)
  - Not applicable (8, 98, 998)
- Avoid cumbersome codes
  - WALKING (1) and CYCLING (2)
  - Doing WALKING and CYCLING (12)
- Use as "1" or "0" ("1" or "2") as baseline for gradients (Yes/No or Present/Absent) as appropriate depending on software for analysis

Entry

NIHRC IRI  
National Institute of Epidemiology  
Chennai  
nie.gov.in

An important aspect of data entry is all about coding. It is always preferable to have numerical coding. Of course, you would have seen, if it is textual information in the form of qualitative there is a different way of dealing with it but with reference to this section we are talking about quantitative data analysis. So, it is preferable to have numerical coding. In particular, you need to decide on how you will code missing values; it could be in the form of a dot or a depending on the field, you may choose to enter as 999, triple 9, be careful you do not enter you know a missing value for age is 99; that can mean differently. And, if it is not applicable you enter with a constantly with the particularly coding in the data collection instrument. For beginners, with inadequate experience in handling databases it is advisable do not create cumbersome codes. It is equally advisable for senior researchers.

For example, if you have a field do you walk everyday? Walking, as a variable. Do you cycle everyday? As another questionnaire; you have 2 variable names walking and cycling. Do you do both? And then there is a coding, and somebody very innovatively thought walking and cycling if somebody is engaged we will give a coding as 12, which is basically 1 and 2 combined, but that is not going to be helpful when you analyze information.

And last, but very important this is very critical because most of the times you may be dealing with dichotomized variables. So, you need to be very clear where you are going to give 1 for Yes, 0 for No or 1 for Yes, 2 for No or 1 for present and 2 for absent, as a base line for all the gradients. Some of the softwares have a different understanding of this 1 and 0. So, when it comes to analysis you need to be careful about your software related details also.

(Refer Slide Time: 11:35)

## Constructing a data dictionary

- Contains, for each variable:
  - Variable name
  - Description of questionnaire item
  - Various values of variable (e.g., 1, 2, 3)
  - Meaning of each value (e.g., 1=Yes, 2=No)
- The catalogue is particularly useful:
  - When a database is shared with others
  - If the researcher has to get back to the database later

Question	Variable name	Type	Format	Values	Logical checks
1	EXERDAILY	Integer	Yes No	=1 =2	Skip pattern
2	EXERTYPE	Integer	Walking Cycling	=1 =2	
ETC...					

*Some softwares create variable catalogue automatically; ideally investigator constructs the same*

NATIONAL INSTITUTE OF EPIDEMIOLOGY & CHENNAI  
National Institute of Epidemiology & Chennai

Entry  
nie.gov.in

Finally, when it comes to data entry you need to have what is called the catalogue. Before the data entry is made, you create what is called data dictionary or variable catalogue in which you talk about each of these variables which questionnaire item it refers to? What are the values there will be assigned to this variable? What is the meaning of the each of these values in a particular format?

Some of the softwares generate on their own, this data dictionary as a variable catalogue, but then it is preferable that you develop your own data dictionary for your study, in which you refer first to the question item, the variable name that you have given, the type of variable, the format in which data is collected, the values that are assigned to and some logical checks, if any. This is written so that if this database is shared with others, the person can make use of the data dictionary and can do the analysis on his or her own. It is equally important for you, if you after some time as lapsed you go back to the database, it gives an idea what you have done and what is it all about for each of this variables and it helps you later when you want analyze your data again.

(Refer Slide Time: 13:00)

Check specifications before data entry

- Minimum and maximum values
- Legal codes
  - Set of values that will be accepted  
e.g., 1, 0 and 9 for "Yes", "No" and "Missing"
- Skip patterns
- Automatic coding
- Copying data from preceding record
- Calculations

NATIONAL INSTITUTE OF  
ENVIRONMENTAL  
HEALTH SCIENCE

National Institute of Epidemiology  
Chennai

nie.gov.in

Entry

Before data entry, one makes sure that there are checks and balances. This is also very important from ensuring internal validity. You specify minimum and maximum values that can enter into a particular field for example, if you study is about children up to 5 years, the age column will not entertain anything more than 5 years at the time of data entry itself. So, that it minimizes the errors that can come in even at the time of data entry, so that these will be acceptable at the time of entry. You may specify skip patterns for example, if you ask a question do you exercise and that person says, no; then you skip a lot of questions about type of exercise, frequency, nature, intensity and things like that. So, skip patterns are very useful even at the time of data entry.

And then of course, we talked about automatic coding, when you enter that code it automatically denotes something that is refer to in the data collection instrument and that can be analyzed immediately. There may be certain times, need for copying data from the preceding record you know for example, lab results if they have to be carried forward to another section it can get copied by itself; this can be specified in the database. And finally, some calculations for example, you may collect height and weight data but you may not calculate body mass index by your own in the data collection instrument. You can ask the database do it when you enter height and weight it automatically calculates BMI. So, these specifications are necessary before data entry is made.

(Refer Slide Time: 14:41)

## Data entry

- Use as opportunity for partial data cleaning
  - Write comments
  - Seek clarification
- Use checks
- Mark each paper as data entry is completed
- Validate after data entry

NATIONAL INSTITUTE OF ENVIRONMENTAL HEALTH SCIENCES  
NIEHS FOR HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai

nie.gov.in

While you enter the data, it is important that we see these as an opportunity for cleaning the database. For example, you enter a data and you find that there are some notes, you need to write, there are some clarifications that you need because you do not think that there is appropriate. So, it serves a purpose of cleaning, the data entry person refers this back to the investigator for additional inputs to clean up the data. You can use checks, while entering the data which we also discussed as an automated check within the database. You have to mark each paper as and when the data entry is completed so that the duplicates are not entered and after the data is entered you may have to validate by different means. So therefore, data entry is one step in the data cleaning aspect.

(Refer Slide Time: 15:42)

## Individual and aggregated databases

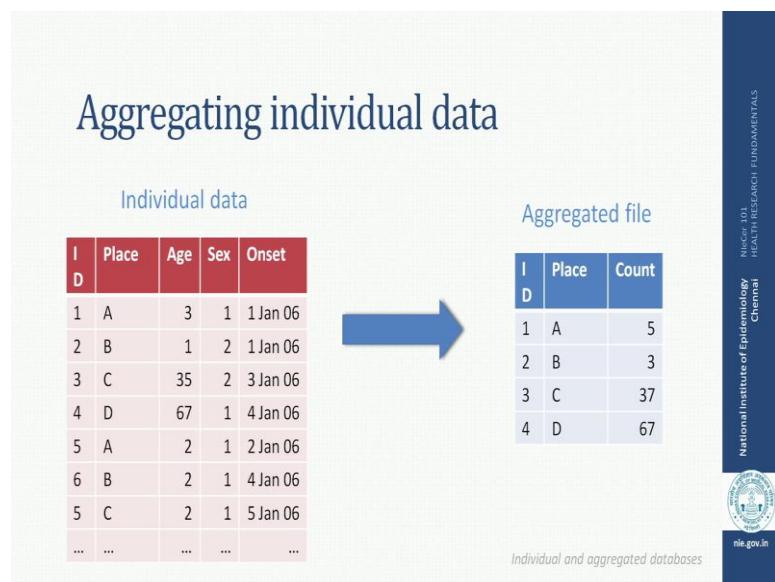
- Individual databases
  - Each record is an observation
- Aggregated database
  - Records contain counts
  - Normalized database
    - Only one count by record
    - Facilitates further aggregation

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

*Individual and aggregated databases*

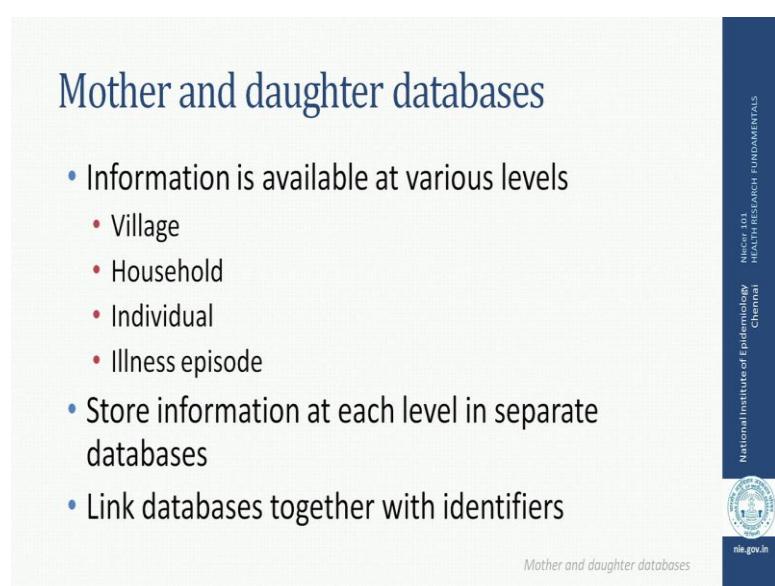
We talked about individual and aggregated databases. We showed you a database that is an individual database, each record in the horizontal line is an observation. There could be instances in which you may have aggregated databases, where you may enter counts in each of the records. If you enter only one count by record that it is ideal, that is called normalize database because normalize database in which each of the roads contain only one count for that particular record. It facilitates aggregation by a (Refer Time: 16:20), I will show you what it is.

(Refer Slide Time: 16:23)



For example, on the left hand side you see an example of individual data about people in whom place, age, gender and onset. So, each of these records indicate an individual and at the same time, we have aggregated data for example, by place one can also get number of people affected, number of people having problem and things like that. So, this is an aggregated data by place, whereas the red color database shows the individual data.

(Refer Slide Time: 16:56)



There can be an instance in which you may have what is called, Mother and daughter database. You may collect information at various levels; you may collect information at the village level, you may collect the information at the household level, you may collect information at the individual in the household, you may collect within an individual information about several episodes of illness or different problems in that individual using different questionnaire.

So, in essence you may have information about different levels, that does not mean you know you repeat the information in all the levels for that particular individual, for example, for that individual you will write information about the house; for that individual you write information about the village, not necessary you can keep them at their level. And then, at the time of data analysis you can link the database that comes from village with database that comes from household, you can link the database that comes from household to the individuals so that you can sensibly analyze, you do not have to worry about keeping everything together and get confused at the time of data analysis.

(Refer Slide Time: 18:08)

### Mother and daughter databases

Household level data			
HousID	Location	Community	HousInco
1	A		3
2	B		1
3	C	35	2
4	D	67	1
5	E		2
6	F		2
5	G		1
...	...	...	...

Individual level data			
HousID	PersonID	Diseased	Exposed
1	101	1	1
1	102	2	1
2	201	2	2
2	202	1	2

- Each database has its own unique identifier
- Link these relational databases using a common index identifier
- Merge files when needed

*Mother and daughter databases*


  
 National Institute of Epidemiology  
 Chennai  
[nie.gov.in](http://nie.gov.in)

For example, this is a household level data where you have information about the house ID, location; the house has such its community status and its income. The individual in

the house may have information pertaining to whether they have a disease or not or they are exposed to particular factor or not. So, you can see here the house ID is repeated here, the person ID for that particular first household is indicated here and disease or not or exposed or not are here. These are 2 different databases entered differently; one is a household database, another is an individual house database. We can link them as and when necessary using this connection called house ID, which is common to both the databases. There is a procedure by which we can do these in softwares, you can even merge this files if needed.

(Refer Slide Time: 19:04)

## Summing up on data management

- Code database numerically
- Enter data using quality assurance procedures
- Store information at the level where it needs to be stored
- Relate/Merge files when needed and as required

NATIONAL INSTITUTE OF ENVIRONMENTAL  
HEALTH SCIENCE  
NATIONAL INSTITUTE OF EPIDEMIOLOGY  
Chennai  
niehs.gov.in

Summing up on the data management, you need to code database numerically, you need to enter data using quality assurance procedure which I outlined, need to store information at the level, where it needs to be stored and we can relate or merge files when needed and as required.

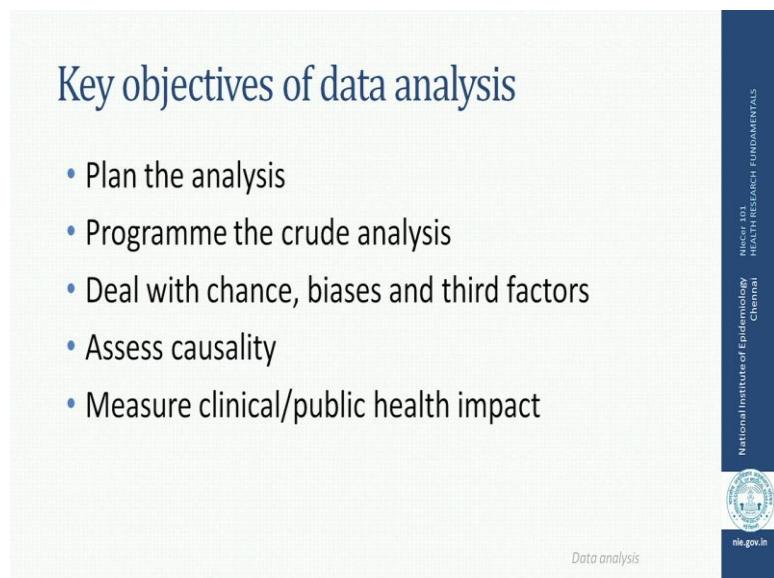
Thank you very much.

**Health Research Fundamentals**  
**Dr. Manickam Ponnaiah**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**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**

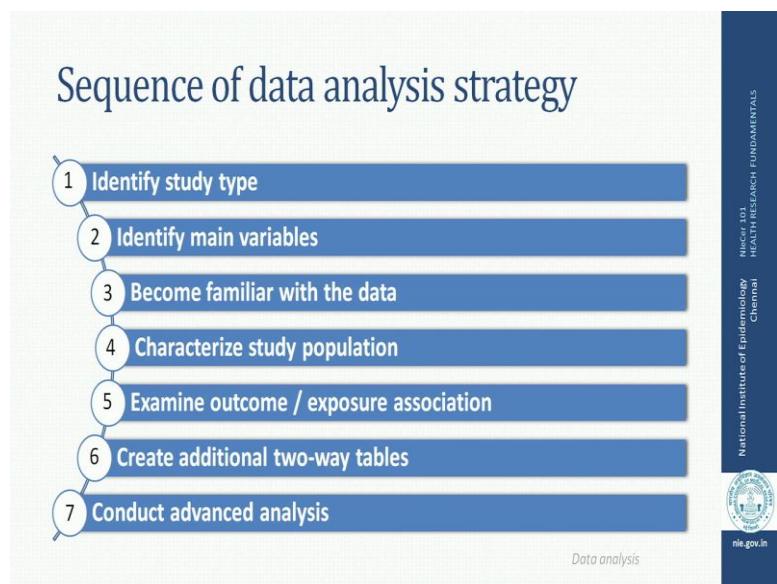
- Plan the analysis
- Programme the crude analysis
- Deal with chance, biases and third factors
- Assess causality
- Measure clinical/public health impact

*Data analysis*

NATIONAL INSTITUTE OF EPIDEMIOLOGY  
HEALTH RESEARCH FUNDAMENTALS  
Chennai  
  
nie.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 way 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)

## 1. Identify study type

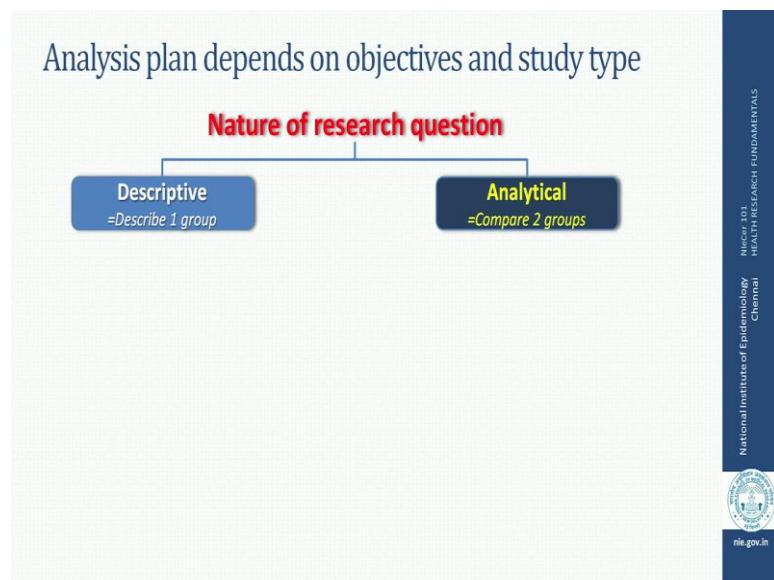
- Establish main analysis framework
  - Descriptive study {Estimation of a quantity}
  - Analytical study {Testing hypotheses}
- Get familiar with the study
  - Review protocol for study objectives and study type
  - Review questionnaire
  - Review analysis plan
  - Review data collection procedures
  - Obtain electronic database(s)
  - Decide on the software for analysis\*



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 text.

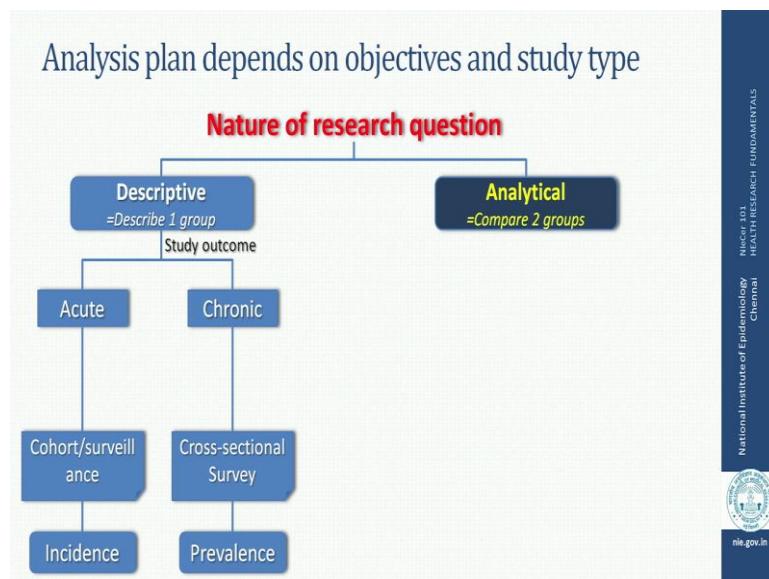
For that clarity, you need to be familiar with the study, you need to look at the protocol especially study objectives and study type, you need to review the entire questionnaire, you need look at the analysis plan, you need to look at the data collection procedures, you need to obtain the electronic 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)



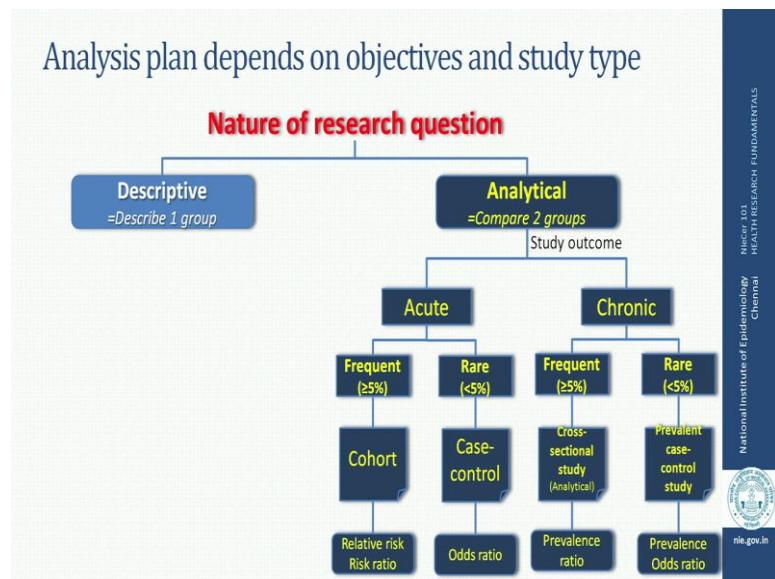
This analysis depends on the statement of objectives and study type. Therefore, the nature of research question has to be very clear, whether it is descriptive involving, describing one group measuring a quantity or is this analytical involving some intervention comparing two groups and hypothesis.

(Refer Slide Time: 02:52).



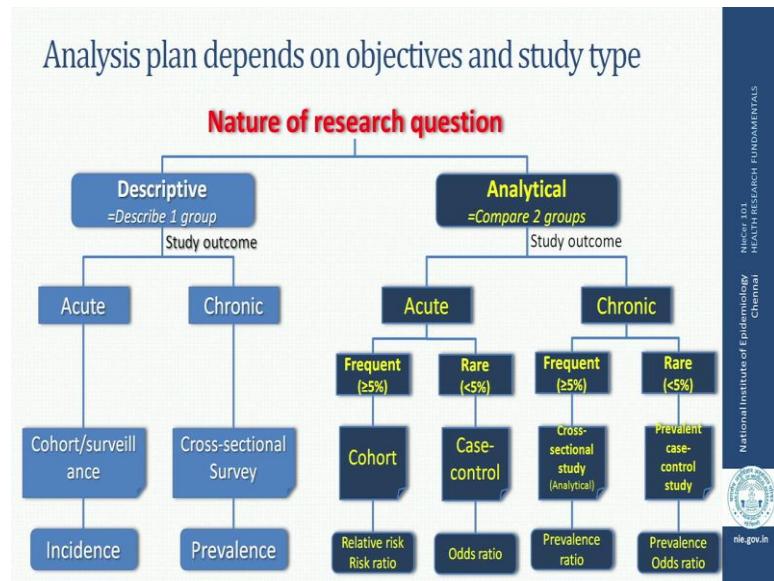
If it is descriptive, you need to ask this question whether it is acute or chronic outcome and then accordingly, you will zoom into a study design and then measure either incidence or prevalence.

(Refer Slide Time: 03:04).



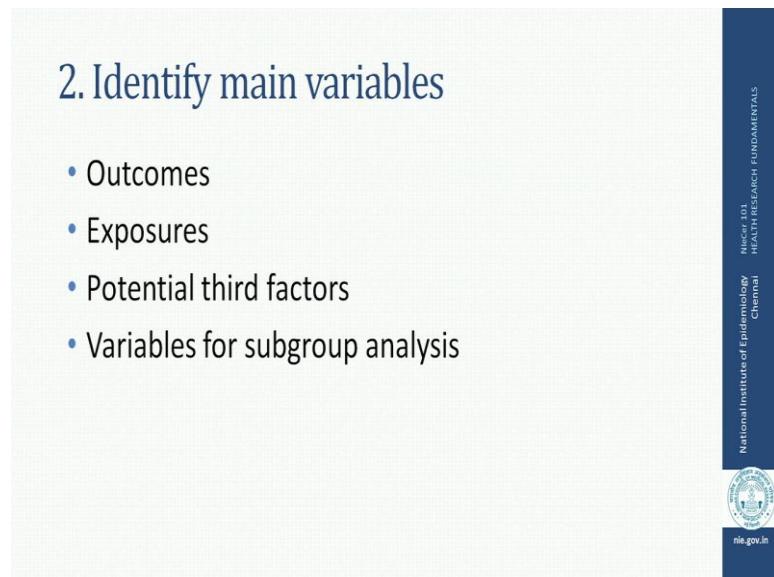
If it is analytical and again you will ask question what is the study outcome? You may decide on acute or chronic outcome and then choose an appropriate measure of indicator, it could be relative risk or risk ratio, if a cohort study or it is odds ratio based on case-control study or prevalence ratio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

National Institute of Epidemiology  
Health Research Fundamentals  
Chennai  
  
nie.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 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

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 known 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

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 have 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



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



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 drenching



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  
NIECH 101  
HEALTH RESEARCH FUNDAMENTALS  
  
nie.gov.in

As an example, we talked about exercise and diabetes 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 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

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 variable 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 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](http://www.cdc.gov/epiinfo)

NICER IRI  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai  
nie.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.

**Health Research Fundamentals**  
**Dr. Sanjay Mehendale**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**Lecture - 19**  
**Ethical framework for health research**

Hello, in this course of Health Research Fundamentals. Today we are going to talk about and discuss about some of the ethical issues in health and biomedical research. It is important that, we as researchers know what these issues are and how to protect the welfare and safety of the research participants that are going to be as a part of our research.

(Refer Slide Time: 00:31)

**Ethical foundation is crucial for research,  
including health research**

- Any research involving human participants should follow international standards of ethics
- Indian national standards are not less exacting and Indian ethical guidelines are on par with international guidelines
- Ethics review is also expected in situations involving no risk when available data are used or minimal risk such as when only questions are asked, no samples/ other specimens are collected



It is always important to remember that ethical foundation is considered as implicit for conducting any kind of research and it is applicable not only to health research, but to any research in general. There are certain international guidelines that have been set in, there are certain international standards and we have to be within those standards. But we also have to remember that, there are Indian national standards that are available, which have been developed by Indian Council of Medical Research and they are not at all less exacting and we have to follow these guidelines as well.

Ethics review is expected in situations sometimes there is a feeling that ethics review is important only in cases where there is a significant risk involved, some invasive procedures involved, but that is not true. Even when we are using available data, where we say that there is no risk involved to human participants, ethics review is required. Also, there are sometimes some situations when minimum risk is involved say like only questions are asked to people were no samples and specimens are collected. Even in these situations ethics review is considered important and mandatory.

(Refer Slide Time: 01:44)

### Evolution of various guidance documents has greatly improved the practice of ethics in biomedical research

INTERNATIONAL	
1947, Nuremberg Code	Initiated discussion on rationale and justification of research risk benefit analysis, competence of investigators and voluntary consent in research
1964, Helsinki Declaration, Revised 1983, 1989, 1996, 2000, 2008, 2013	Commitment to individual rights to make informed decisions, investigators' duties, research participants' welfare, vulnerability
1978-79, Belmont report	Described the basic ethics principles of autonomy, justice and beneficence, emphasized informed consent and review by ethics committee
1992-93, CIOMS guidelines [Council of International organizations on Medical Sciences and WHO], Revised 2002	Reporting of adverse drug reactions and safety of research participants, benefit-risk balance, need and principles of pharmacovigilance,
1996, ICH [International Council on Harmonization]	Good Clinical Practice



In the past so many decades, many international documents have been developed and that have helped for improving the practice of ethics in biomedical research. This all started with some kind of experimentation that was done in the World War II, among the captives and after that whatever happened, the people felt that lot of atrocities got committed and human subjects where used for research in a improper way and lot of discussions started on that. One effort in this direction which began very early was the development of the Nuremberg Code in 1947, which started some discussion related to the rational and justification of research, basically whether that particular research is necessary? What is the risk benefit analysis? How it is important in deciding whether that particular research is important or not? Also, looking at the competence of investigators and also initiated some discussion on the voluntary consent in any kind of research.

Thereafter, for the first time many countries came together and met and they signed on a document, what is called as Helsinki Declaration in 1964, which got revised several times after that and the latest revision came up in 2013. Helsinki Declaration, basically talks about commitment to individual rights or to make informed decisions, but at the same time also emphasizes duties of investigators. Also, talks about patients rights research participants welfare and in addition talks about certain groups that are considered as vulnerable and it is necessary that certain steps are taken to protect their interest.

In the United States, in 78-79 Belmont report was published, which described the basic ethics principles of autonomy, justice and beneficence and we are going to cover this during this particular session a little later. And it also reemphasized the importance of informed consent in research and here was for the first time the importance of review by ethics committees, which are called as Institutional Review Board in the west, it was emphasized. Again in 1992-93, the Council of International Organization on Medical Sciences and WHO called CIOMS, they developed the document called as CIOMS guidelines and which was revised in 2002. This is another important international document, which provides guidelines regarding reporting of adverse events and safety of research participants. This is particularly significant in case of clinical research and clinical trials, where experimentation is done, in case of new drugs and new vaccines and so on.

It also talked about benefit and risk balance and need and principles of pharmacovigilance. So, it was stress for the first time that by doing phase 1, 2 and 3 studies our responsibility does not end, but probably we need to do a continued surveillance in the population to figure out, what is happening on as far as the long term safety of these interventions are concerned.

In 1996, the International Council on Harmonization, ICH as it is popularly called as developed basic guidelines for good clinical practice. Subsequently, taking the basic clues from this particular document, good clinical laboratory practice document has been developed; good clinical epidemiological practice the document has been developed. So, it found lot of applications in different spheres of health related research.

(Refer Slide Time: 05:44)

**Indian Council of Medical Research introduced Ethical Guidelines for Research on Human Participants**

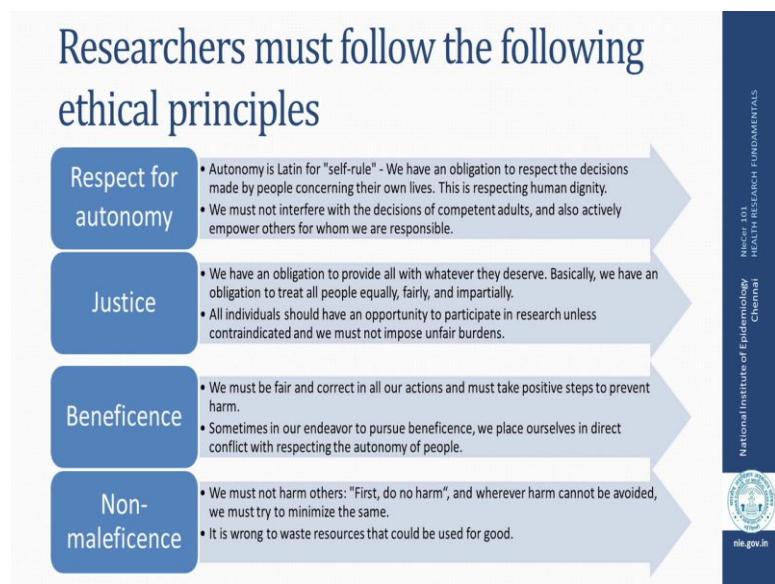
NATIONAL GUIDELINES	
2000, ICMR guidelines 2006, Revised ICMR guidelines <a href="http://www.icmr.nic.in">http://www.icmr.nic.in</a>	All institutions in the country which carry out any form of biomedical research involving human beings should follow these guidelines in letter and spirit to protect safety and well being of all individuals.
There are several other national guidelines available	Genome Policy and Genetic Research [2000], Indian GCP [2001], Amendment of Drugs and Cosmetics Act [2002], Assisted Reproductive Technology [2005], Stem Cell Research and Bio-banking [2006]

NICER ID:  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai  
  
nier.gov.in

India was not far behind in developing its own documents. The Indian Council of Medical Research in 2000 for the first time introduced Ethical Guidelines for Research on Human Participants. This was a major consensus document, which was produced and got revised in 2006 and is again in the process of a revision, which might happen in the next year that is 2016. This document is available on the website of Indian Council of Medical Research. Basically, it gives the guidance for all the institutions in the country, which carry out any kind of biomedical or health research, which involves human beings and provides the guidelines that people have to follow, the researchers have to follow to protect the safety and well being of all individuals.

But in addition to that there are several other national guidelines, which are also available which include the document on genome policy and genetic research. There has been an amendment to Drugs and Cosmetic Act in the early 2000, which is available. There are guidelines available for stem cell research, assisted reproductive technologies, bio banking and researchers working in these areas have to be aware of the guidelines, which have been provided and stick to those.

(Refer Slide Time: 07:10)



I earlier mentioned that there are certain important basic ethical principles. Among them respect for autonomy, justice, beneficence and non-maleficence, these are the ones, which have to be considered. Autonomy is considered as basically is a Latin word for self-rule, which means we have to respect individuals for what they are and this is like respecting human dignity and so we must not interfere in what people feel like doing or what peoples thought processes are. But at the same time also indicates that all those people who are not adequately aware have to be empowered to understand, what this is all about as well. Basically, what it means is people should be clearly informed that they have a right to decide to participate in research or not to participate in research.

The next principle, which is of justice, it emphasizes that we have an obligation to provide all with whatever they deserve. Basically, what it means is the participants or if there is an obligation to treat all people equally, fairly and impartially. So here, what is required is the benefit of research should be extended to everybody and except in certain situations there are certain groups like say, for example, condition like pregnancy, when women cannot participate in research. So, unless contraindicated, all groups should have an opportunity to participate in research, but this should never be imposed on anybody.

The next two principals, like they just go hand in hand. Beneficence means we must do

everything which is fair and which is correct and we should be correct in our actions and in our deeds also and we should take only positive steps to prevent any kind of harm, this is an important thing. Non-maleficence is the other side of it, we must do everything to do things to help people, we should also not cause any harm to others. So, do no harm is the principle or the explanation of what we called by non-maleficence. So, whenever harm is evident, see whenever we talk about any new drug trial for example, there is always expectation that some kind of a side effect would always be there. What we have to really ensure as researchers is that we take appropriate steps to ensure that this harm would be minimum and if at all it occurs appropriate care is taken care of. So, whenever we conduct any kind of a research, we have to ensure that these basic ethic principles or ethical principles are followed.

(Refer Slide Time: 10:00)

## What is informed consent?

- Informed consent is the process of informing the potential participants about the proposed research in a systematic manner and empower them to take an informed decision to participate in the research study
  - Understand study procedures and risks and benefits
  - Get all questions and concerns answered
  - Take a learned and informed decision to [or not to] participate
- This process can be repeated several times during the research study if necessary
- Although group consent is desirable [e.g. tribal studies], it cannot replace individual consent

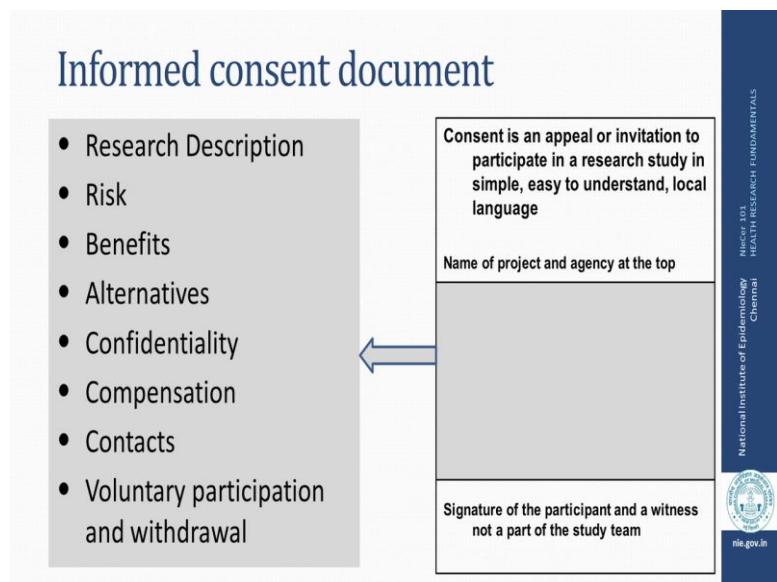


And one way to assure this is through a process is what is called as informed consent. This is a process as I very specifically mentioned because sometimes it is not a onetime event, it is not a tick mark. It is a very systematic way in which the proposed research is explained to the potential participants in a systematic step by step process and the potential participants empowered to take an informed decision to participate in the research study. What does it involve actually? So, that the participants have to understand, what the study procedures are and what are the risks and benefits of their

participation? They can ask they can have the liberty to ask all kinds of questions and raise their concerns which have to be appropriately answered and then finally, the participants take a very appropriate learned and informed decision to participate or not to participate in the study.

So, there could be several sessions that could be involved in completing this process. A certain individual may understand the whole process in one single sitting. For an individual, it may require multiple sittings and multiple sessions, but the researchers have to be persistent and perseverant to take the potential participants through this process meticulously. In certain situations, like when we work in a certain tribal populations, institutional setups, it is important to take a group consent or consent from the concerned authorities, but one has to remember that this kind of a consent does not replace the individual consent.

(Refer Slide Time: 11:41)



The informed consent, it is classically a document, which explains various things and so the document has at its header, the name of the project and the agency that is conducting this particular research and the main body of this particular document talks about various things like, it describes the research study in brief, it talks about the potential risks, it talks about the benefits. Sometimes, it is possible that when research is conducted,

benefits are not necessarily individual oriented. The research might cause benefit to the whole community and that has to be explained correctly. Their alternatives have to be explained in the sense that participants have to be explained that, even if they decide not to participate in research, it is just fine and they will continue to get us services as they would otherwise get even if they decide not to participate.

In addition, the researchers have to commit and give the assurance of confidentiality, keeping their records confidential because sometimes some sensitive information gets collected and participants may be worried about the information that their sharing with the research team and so it is important that they have to be given this assurance of confidentiality. Sometimes, some harm that results as a part of research participation has to be appropriately compensated and that clause also is essential as a part of the informed consent processes. We have to also give some basic important contact information as to whom the research participants can contact for any additional information, for any concerns that they may have during participation of the study and this information should be clearly included.

In addition, one of the most important clauses that get added is about voluntary participation. The document has to emphasize that every person has a right to decide whether to participate in the research study or not. Also, during the part of the research study and this becomes particularly important when it is a long term follow up study. A person may decide to drop out of the study at a certain point of time and it is perfectly within the rights of the individual to do so and this has to be explained. So, all these constitute the body of the informed consent and towards the end of this particular form, the signature of participant and if the participant is illiterate, then the signature of a witness who is not a part of the study team has to be obtained.

(Refer Slide Time: 14:30)

## Stakeholders in informed consent process

- Researchers and institutions:
  - Information – discussion and explanation – comprehension – voluntary decision
- Participants:
  - Informed, free and independent consent without coercion or force
- Sponsors, monitors, regulators:
  - Assess fairness of consenting procedure
  - Verify consent documentation of research participants

NICER IRI  
HEALTH RESEARCH FUNDAMENTALS

National Institute of Epidemiology  
Chennai



nie.gov.in

There are various stakeholders in this informed consent process. In research, please understand that this informed consent procedure is considered as a very important procedure and it is a very sacrosanct procedure. We as researchers have to understand that we and our institutions have an obligation to provide the required information, which is good in length and also in-depth to the participants. They should be given a chance to discuss their issues, their problems. They should be provided with adequate explanation. We have to also ensure that they have understood the information that is being given to them with respect to the research study appropriately and then also ensure that they take a voluntary decision; there is no coercion or coaxing on part of the researchers.

On part of the participants, who are the other kinds of stakeholders in this whole process, they have to themselves ensure that, they have understood whatever has been told to them. They just should not sign the informed consent form without understanding what goes behind that or what is included under that. They have to understand their rights, they have to understand various provisions, the researchers or scientists are going to make as a part of the informed consent document. Basically, they should sign the form freely, independently and without any coercion or force.

The third kind of stakeholders in this process includes the sponsors, the monitors or regulators here. These are the kind of agencies that assess the fairness of the consenting processes at various levels. This starts right from the beginning of the research study before it gets approved by the institutional ethics committee right up to monitoring throughout the procedure by the monitors and regulators as well. They also have an authority to verify the consent documentation of research participants. So, every person who is involved in research has a combined responsibility to ensure that informed consent is appropriately taken.

(Refer Slide Time: 16:36)

## Issues related to informed consent

- **Whom does informed consent benefit?**
  - The research participant
  - The investigator
- **Is the research procedure adequately explained in the IC form?**
  - The language, simplicity and clarity
  - Translations and back-translations, certification of translations
  - Test of understanding
- **Issue of witness to consent procedure**
  - Impartial witness
- **Can there be different types of informed consents**
  - Traditional written IC form
  - Audio consent and video consent: Mandatory for investigational new drug [IND] trials in India
  - Pictorial consent



There are certain issues around that, whom does the informed consent benefit? It benefits everybody really; I talked about some stakeholders here. So, from the research participant point of view, it gives him lot of information and it therefore, it is important for him, for the investigator it is documentation that this particular process has been completed in the best possible way, in the most appropriate way.

What is important to understand is whether the research procedure and various other aspects of research are adequately explained in the informed consent form. So, the language has to be simple. There has to be lot of clarity on various issues, it should be in the local language. So, it is also advised that the consent form which is developed in the

local language is back translated and then either it is certified or checked with the original English consent form for its accuracy. Some investigators have tried doing test of understanding also and this is considered as one of the good practices. Once, the informed consent form is done, some kind of a small objective type of test is quickly given to the research participants to assess, whether the understanding on the informed consent form has been adequate or not.

The issue of witness can become a critical issue. This becomes particular important in scenarios, where the people are not literate and they cannot sign the informed consent form. So, here is where, there is a need to have an impartial witness. It is important to ensure that the principal investigator, researcher himself does not sign as the impartial witness. It should be somebody, who is not connected with the research process.

There has been a lot, which has been talked about, whether the oral consent is valid or not. Typically, there should be two informed consent forms to be signed and then one has to be returned back to the participant, which the participant can keep with him for record because it also provides lot of answers to the questions that might arise in the mind of the participant subsequently. There has been some discussion going on with respect to audio consent and video consent, well the regulatory authorities in India have now made it mandatory to record the consent procedures in case of investigation in new drugs in India. So, this is an important regulation which we must keep in mind.

(Refer Slide Time: 19:11)

## Importance of scientific review

Explores the scientific novelty, rationality and relevance

1. Justification for conducting the trial in the context of national priorities
2. Scientific merits of the research project and feasibility: Review of toxicological studies, laboratory and animal data
3. Technology transfer and capacity building at sites

Soundness of the study design:

Inclusion-exclusion criteria,	Sample size,
Randomization/ blinding procedures	End-point assessment
Study procedures and follow-up schedule	Pharmacy plan

*Scientifically well-planned research studies are more likely to correctly address human subjects and ethical issues*

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

Different types of reviews have to take place before any kind of study becomes ready to be undertaken there actually. There is a scientific review and regulatory review, which takes place before the actual ethics review happens. The scientific review looks at the novelty, the rationality and relevance of the study. Basically, what is the justification of doing this particular study? What are the scientific merits? Whether appropriate study procedures have been done, have been taken into consideration? What are the inclusion exclusion criteria? Whether the sample size has been calculated appropriately? How will the endpoint or the outcomes going to be assessed? So, these aspects have are normally looked at in the scientific review. Why this is important is because scientifically well planned research studies are often likely to be correctly addressing the human subject issues and also the ethical issues. So, if the science is good often it takes care of ethics which is behind that.

(Refer Slide Time: 20:12)

## Objectives of regulatory review

- Evaluate pre-clinical trials data
- Assess in-country regulatory requirements for drug/ vaccine/ product import
- Ensure national requirements for special situations -genetically engineered products, stem cell research, research on reproductive technologies, organ transplantation etc.
- Sample shipment and transfers, transfer of raw data: IPR issues
- Exchange of scientists or visitors
- Budget: Foreign funding
- Research in border or high-security areas

*Careful regulatory review results in answering some of the ethical concerns*

Similarly, regulatory reviews look at the various aspects like the pre-clinical trials data that has been done. This is particularly important in case of newer drugs, newer vaccines, the various animal data that is there, toxicology data that is there, then there are certain in country regulatory assessments for various drug vaccine or product imports, they have to be assessed and this is important if there are certain trials or experiments which are being done using the products that have been developed abroad. There could be certain special situations, where genetically engineered products are being used or there is a stem cell research, research on the reproductive technology, organ transplantation etcetera and so the concerned regulatory agencies have to ensure that all the necessary guidelines have been followed here.

The issues around sample shipment and transfer as well as transfer of raw data is looked at very seriously by the Government of India because there are any issues around the intellectual property rights in this area and the government is very protective about those. So one, the researchers have to know these issues fairly well and the regulation in this regard as well. There are certain sorts of caveats; there are certain kind of restrictions regarding exchange of scientists and visitors. The funding particularly, if there is foreign funding coming in for a project to be done in India and during such situations, where the research is to be done in the border or high security areas and the researchers have to be

aware of the regulatory requirements in such situations. Again, like scientific review, careful regulatory review also results in answering some of the ethical concerns.

(Refer Slide Time: 22:00)

## Range of ethical issues that need to be addressed in health research

- Competence of the researchers and the research team
- Provisions for protection of human rights and ethical issues: vulnerable populations, women, children
- Measures for protecting confidentiality and non-discriminatory practices
- Appropriateness of Informed consent and study specific educational material
- Mechanisms for reporting and management of adverse events and serious adverse events
- Care and support for research participants: standard of care, long-term care, post-trial access to care and product
- Reimbursement and compensation
- Continuing review of progress of the study



Now, coming to the various ethical issues that need to be addressed; well, it is important that the competence of researchers and a research team has to be accessed appropriately. Then whatever are the provisions made for protection of human rights and vulnerable populations, in particular they have to be accessed. Then measures for protecting confidentiality, they have to be seen. Appropriateness of the informed consent form the correct completeness of the informed consent form has to be assessed. Mechanism for reporting and management of adverse events and serious adverse events, this is particularly important in case of drug trials. Then care and support mechanism for participants, is the support going to be extended to the participants after the trial? Will the post trial benefits be given to the community after the trial is finished and is proved to be useful? These aspects are also; have to be looked at from the ethical angle.

The reimbursement and compensation are important issues, which the ethics review looks at. Because one has to ensure that reimbursement is for the time lost and also the expenses paid for traveling to the clinical research site. But the compensation or the incentive given or the money that is given should not be as much as to course the

participant to participate in the research trial. So, this decision is also made by the ethics committee. And it is also important to continue the review of the progress of the study till it is completed. So, the committee which really looks at all this thing is what is called as Institutional Ethics Committee in India.

(Refer Slide Time: 23:41)

## Main responsibility of institutional ethics committees or institutional review board

- Does the study have real/ potential individual/ community benefit?
- Are the rights of research participants adequately protected?
- Does the potential benefit far outweigh the risks associated with research participation?
- Will the participants and communities have access to study findings and benefits of research?
- What is the mechanism for provision of safety, care and support to research participants?



It is called as Institutional Review Board abroad. Basically, it looks at whether the study has potential benefits or community benefits? Whether the rights of the participants are adequately protected? Whether the benefits of this particular study out with the risks that are involved here? Whether the participants and communities have access to study findings? Whether they are eventually going to benefit from the research participation? And what kinds of mechanisms are built-in to provide the safety and care and support to research participations even during the study and after the study?

(Refer Slide Time: 24:29)

## Ethics influencing health research and practice of medicine ....

- Growing expectation about accountability:
  - Questioning of Government responsibility [local, state and national] and investigators' responsibility
  - Growing public awareness due to advocacy movement
- Collective demand for health benefits - Universal right to health care (health for all)
- Place for self responsibility (lifestyle) – should it always be researchers to be blamed for mishaps
- Need for including bioethics in medical curriculum being increasingly stressed



So, there are growing expectations about accountability from the researchers now. Various researchers are being questioned about their responsibility and governments' responsibility as well in fairly conducting research. There is a growing public awareness also; all this is eventually going to improve the quality of research in our country and also, the practice of medicine in our country over a period of time. There is a collective demand for health benefits, people are demanding. So, universal right to health care which emanates from the principle of health for all. So, this which demands that more and more research will have to be conducted in making more and more benefits available to the common man of this particular country.

It is also important to understand that it is not only the researchers who have the complete responsibility to follow ethics. It is also important that the research participants follow certain or fulfill certain expectations. So, whatever is has been described in research they have to follow appropriately. It is explained in the informed consent form and so there is a great need for including bioethics in the medical curriculum which is being stressed.

(Refer Slide Time: 25:47)

Ethics in practice of public health and health research	
There is hope	Ethics in practice of public health and health research is being increasingly addressed.
There are challenges	Public expectations and demands will continue to increase.
The search for solution should be an ongoing process	Public health system, policy makers, researchers and program managers should show enough sensitivity and realize that there is a scope for further improvement

NIHERF IDI  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai  
  
nie.gov.in

There is certainly a hope because ethics in the practice of public health and in health research is being increasingly addressed. We know that there are challenges as well because the public expectations and demands are continuously increasing, but we will continue to find solutions and this has to be an ongoing process. For this, various public health stakeholders like public health system, policy makers, researchers and program managers, they should show adequate sensitivity and realize that we always can improve and a practice always can improve.

Thank you very much for your attention.

**Health Research Fundamentals**  
**Dr. Sanjay Mehendale**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**Lecture - 20**  
**Conducting clinical trials**

Hello. In the course of Health Research Fundamentals, today I am going to discuss the experimental study designs or clinical trials.

(Refer Slide Time: 00:16)

### Scenario of clinical trials in India

- The clinical trial industry rapidly expanded in the first decade of 21<sup>st</sup> century, but has faced some challenges due to regulatory reforms in 2012-13
- The main challenges perceived by international investigators and sponsors in undertaking high quality clinical trials in India include
  - Delayed approvals
  - Concerns about quality of ethics review
  - Shipment of samples and transfer of data due to Governmental restrictions
  - Overall lack of duly trained investigators and centers
  - Clause of compensation for clinical trial participants
  - Recent requirement of audio visual recording of consent process for IND [investigational new drug] trials

Ruchi 101  
HEALTH RESEARCH FUNDAMENTALS



nie.gov.in

The scenario of clinical trials in India has changed over the last 25 to 30 years. In the late 20th century, what was happening was, mostly the clinical trials that were undertaken were by the research institute and also by the pharmaceutical industry, which were limited in nature. But, towards the beginning of 21st century, in the first decade in particular, the clinical trials industry in India grew and expanded very, very rapidly. In the present decade, it has been facing some kind of challenges because of various regulatory reforms that have happened recently. But overall, the international community particularly the international investigators and sponsors, who want to undertake high quality clinical trials perceive that there are certain challenges for conducting clinical

trials in India.

They feel the important ones among them include, the approvals are often delayed; approvals in terms of the ethical approvals, by the ethics committees as well as the regulatory approvals by the regulatory authorities. There are also concerns about the quality of ethical review that takes place because it has been only in the recent times, a very systematic efforts is being taken to improve the performance of various ethics committees of research organizations in particular.

The shipment of samples and transfer of data has always been an issue for international studies in particular because government has some specific restrictions and regulations about transfer of the data as well as samples. It is also felt that, there is one overall lack of duly certified and trained investigators as well as centers in India and again a very systematic effort in this direction is ongoing in the country. What has happened as a part of the regulatory reforms are two important changes suggested now and there is a specific clause of compensation, which has now come up for clinical trial participants and the office of the Drug Controller General of India has given very specific recommendations about, how this compensation has to be calculated and up to what period these laws are applicable, that has put in some kind of pressure on the people, who are conducting the trial and there is some kind of haziness around this particular area.

Also, in the area of audio visual recording of informed consent process, all of us will certainly agree that to make ourselves very accountable, to make ourselves very transparent, probably this is the best procedure. If we protect the confidentiality of the trial participant adequately, but carry out audio visual recording that is possibly the best proof of how the consent procedure was actually undertaken here. It is currently applicable to mostly the IND trials, which are Investigational New Drug trials, but again people are confusing this. They are trying to do it in all different kinds of scenario and some kind of thinking, rethinking on this particular matter is happening at the center on this particular thing in India.

(Refer Slide Time: 03:35)

Scientific, ethical and regulatory reviews of clinical trials are critical		
Concerns in implementation of research protocols	Type of review	Some examples of available agencies/ mechanisms
Is the research question sound?	Scientific review	Institutional Scientific Advisory Committee
Is the safety and welfare of the research participants adequately protected?	Ethical review	Institutional Ethics Committee Central/ National Ethics Committee
Are the research methods appropriate?	Regulatory review	Health Ministry Screening Committee Drug Controller General of India Genetic Engineering Approval Committee

But, for any clinical trial to happen or to start, certain approvals are mandated and they have to be taken in. It all starts with the scientific review; it is a very systematically planned experiment, very meticulously planned. Every step is outlined in detail and basically, what we ensure through the scientific review is whether the research question is sound and usually it is the Institutional Scientific Advisory Committee that decides, whether the study designs that is being proposed here is appropriate. All the methods that are being described under that are accurate and correct.

Then, it goes to the next stage of what is called as the Ethical review. It is done by the Institutional Ethics Committee; sometimes at the national level also it can be done by National Ethics Committee as well. We do it, in our set up of Indian Council of Medical Research and what is really looked at here, is whether the safety and welfare of the research participants is adequately protected or not? This is a very significant aspect of a clinical trial. This review is considered very critical, why? Because, the basic principle is we should do no harm to research participants by making them participate in a research study and that is a very critical component, which we have to keep in mind.

Another kind of review that happens is that of regulate, which we call it as a Regulatory review. Sometimes it is new drugs, so the setup in our country is the previous Drug

Controller General of India, which is now the Standards Control Organization, that we have CDSCO. Sometimes, if the project is getting funded internationally, then there is a health ministry screening committee, which looks at various regulations around that particular issue. If there are genetically modified or engineered products being used then there is a Genetic Engineering Approval Committee, which looks at that. It is the responsibility of the sponsors and investigators to find out, what kind of regulatory approvals are required for the kind of clinical trial that is being undertaken and all those clinical trial approvals must be taken before initiating the study. Appropriate certification has to be kept on file for ready review by the external monitors or by any other agencies that are authorized.

(Refer Slide Time: 05:57)

The slide has a light blue header and a white main content area. The title 'Addressing ethical issues in clinical trials' is centered in a large, dark blue font. Below the title is a bulleted list of nine items, each preceded by a small blue circle. The footer on the right side contains the text 'NIEIR, IITI, Chennai' above a circular logo, and 'nieir.gov.in' below it.

## Addressing ethical issues in clinical trials

- Is there a mechanism for independent ethical review? [Approvals from Ethics Committee and in country Regulatory Authority]
- Which mechanisms exist to ensure protection of human subjects throughout trial participation?
- Is there adequate community engagement and support?
- Informed consent
- Standard of care and post-trial support
- Use of placebos
- Confidentiality

NIEIR, IITI, Chennai  
nieir.gov.in

But one of the most important things that has to be kept in mind is because it is an experimental study, because we are manipulating the environment, we have to take care of a lot of ethical issues that can arise in this particular scenario. Primarily, the most important thing is have all the approvals been sort for this particular thing, which include the scientific, the ethical as well as the regulatory approval, which is currently the norm in the country at any point of time. Then what? Just getting the initial approvals is not enough; to ensure that the patients or the participants are protected all throughout the trial is also critically important and so some role of ethics committee to monitor this

particular process is also important.

Is there an adequate community engagement and support for this? Somehow, it has so happened that sometime when the people were not made aware of some of the interventions that were being tried at the national level, there were backlashes because there was something which was believed to be culturally not acceptable and hence it is always important to engage the community and face of the community is visible through many groups. They may be the private practitioners and doctors, they may be the program managers of the country, they may be the doctor attending to the patients, they may be the political people who are taking say, sort of care of the interest of the people and so on and so forth.

One of the important things is to be remembered is every single participant has to sign and informed consent form before participating in a clinical trial. Any trial without an informed consent form is a problem and we have to ensure that this particular process is completed. Also, after the trial gets completed, probably the responsibility of the investigator does not necessarily end. There has to be some kind of a mechanism for providing follow-up care to or post-trial support to the trial participants and that maybe all well thought of right ahead of time and planned in the research protocol.

Use of placebos has raised some questions in the past and it is a topic in itself, but wherever there is a vaccine which we are talking about, where there is no comparable vaccine available earlier of that particular disease, doing use of placebos is considered as justifiable. One of the important things which is an ethical issue, is the confidentiality of the study participants. The study participants sometimes are worried that nobody other than them, even in their own family should know about their research participation and it is our duty as investigators or researchers to protect this interest of the participant and to actually keep all the information totally confidential.

(Refer Slide Time: 08:50)

## Critical issues in trial implementation -1

- Informed consent procedure
- Screening and enrollment: Strict adherence to inclusion and exclusion criteria
- Good clinical and laboratory practice, quality control and quality assurance
- Adherence to intervention and follow-up is very important in the context of study outcome assessment
- Multi-centric trials: standardization of study protocols

As I said, the whole process of clinical trials starts with the informed consent process procedure. I have deliberately used this word, it is a procedure and it can be multiple step procedure also. Some patients and some participants may not be able to finish this particular process in one single sitting. You may require multiple sittings to explain the things. The word informed consent here is to be critically looked at, informed is where the duty of the investigator is to explain all the study procedures to the study participant very carefully and also ensure that any question or doubt the person might have is appropriately answered. This has to be now documented, all of us know this. And now, there is also an instances that for clinical trials it should be, a duplicate copy should be signed and because it is a document, which give some information about the study also. One copy should be kept by the patient and the other should be kept with the investigators team.

The process of clinical trial participation is generally two step. There is a screening protocol first, wherein those who are interested in participating in the trial undergo some level of screening that is the interview, followed by medical examination, followed by some samples collection and it is ensured that they fit into the eligibility criteria and because as investigators, we have to adder to the inclusion and exclusion criteria very specifically. It is important that the investigators follow good clinical and laboratory

practice. This is absolutely important, this has to be followed all throughout the trial and what this means? This is, although this looks like something like a common sense. Do the enrollments as exactly as have been defined in the protocol. Do the study procedures as they have been defined in the protocol. Collect the volume of blood as, has been defined in the study protocol, ensure that all the visits are made. So, these are the basic procedures, there are responsibilities for the sponsors, there are responsibilities for the investigators. All of them have to follow this.

In clinical trials, one of the most important things, which we have to ensure, is adherence to the intervention. If it is a drug trial, we have to ensure somehow that the patient is taking the drug regularly. If it adherence with respect to follow-up evaluation, also we have to ensure that the person comes for follow up at the define time intervals. Say for example, we are asking a particular participant to participate in a vaccine trial and if we want to study, periodically up to 2 years from giving that particular dose of vaccine. Till what length of time, the vaccine immunogenicity or the antibody levels to that particular vaccine candidate are maintained and if it is decided that once in every 3 months, this evaluation will be done. If patients miss some of these visits, it is very likely that we will miss some of the important data because we may not be able to decide up to what level the antibodies where maintained and beyond which they were not maintained.

There can be a lot of issues which can arise in case of multi-centric trials. Sometimes, trials are conducted in 5, 6, 7, 10, 50 centers in the country and here is where, this is done with the basic purpose is sometimes, it becomes very difficult for a few sites to gather the required number of study participants. So, multiple large number of centers have to be used, but the compromise that we make here is the quality can come down particularly, variations can occur within the procedures that are adopted in various centers and hence, one of the key elements here is the standardization of the study protocols. Training of the people, right in the beginning, periodic training of the people and also then using very standard well defined protocols is something which is absolutely important.

(Refer Slide Time: 12:56)

## Critical issues in trial implementation -2

- Independent monitoring
- Safety assessment: Reporting and management of adverse and serious adverse events [clinical, laboratory and social/ familial]
- Reimbursements, compensation and grievance redressal
- Trial stoppage rules
- Documentation archival

NATIONAL INSTITUTE OF  
ENVIRONMENTAL  
HEALTH SCIENCES  
NIH  
National Institute of Epidemiology  
Chennai  
nie.gov.in

Also in a clinical trial, one of the important prerequisites is to meet some kind of an arrangement for an independent monitoring mechanism. This is an agency, which you generally employed by the sponsor and that carries out independent monitoring, looks at for the adherence to GCP, dispensing of study products, how the records are maintained, how informed consent procedure is happening and so on and so forth.

This particular agency has to have a total third party view in doing that particular assessment and should not be influenced either by the sponsors or by the investigators. As we know, as I have been mentioning right from the beginning, safety is one of the most important thing as far as clinical trials are concerned and maintaining or ensuring safety of research participants is of paramount importance and hence, the safety assessments are built in in all clinical trials. What is also important is, there are very well defined reporting mechanisms for management of various kinds of adverse and serious adverse events as they happen in, among the clinical trial participants and they have to be timely reported to the regulatory authorities as well as sponsors as well as ethical committees.

We have to remember that all the adverse events may not necessarily be clinical. Sometimes, they can be odd laboratory values or they could be even social or familial

problems that are happening as a part of clinical trial participation. All these pieces of information have to be appropriately trapped. Generally, there is an agreement that some kind of reimbursement should be made to the trial participants. One is for the kind of time they spent in coming over there and the loss of daily wages that might happen as a result of that and also for the travel cost involved, plus maybe some kind of food expenses, which are made particularly, in case of trials which require long term participation of the study participants.

So, but here is where a clear understanding has to be there as to we should know that we are doing reimbursements and we are not doing incentives because sometimes incentives can become coercive. In the sense, the incentives can persuade patients to participate in the trials, even if they are not interested, if the amounts are really large. Recently, some kind of laws and regulations has come in with respect to compensation. We have to be all aware of what is prevalent regulation in the country with respect to the clause of compensation.

It is also important that we have a mechanism for grievance redressal, some of the research participants because it is an experiment and some kind of a experimental intervention is being tried and the some of the people might come up with some kind of complaints, which the investigators may feel are not related to the intervention that is being tried, but the patients keep insisting that it is essentially related to that and hence, there has to be a kind of third party body and it is always believed that grievance redressal team should be there, which is easily available to address these issues in real time.

Every single trial has to have a defined trial stoppage rule right in the beginning itself. Generally, 3 serious adverse events which are defined as maybe, deaths or very serious complications on the previous clinical condition that is what happening, hospitalization due to any cause, these are considered as serious adverse events. So, after any 3 serious adverse events, it is normally decided whether they are related or not related to the trial by a third party body called as Data Safety Monitoring Board, which is an independent entity, which is also predefined earlier and when they allow us to move forward, we move forward in the clinical trial, but these rules have to be defined up front.

Another important thing in a clinical trial is documentation archival. Various sponsors have various expectations with respect to this; some of the trials sponsors require this. All the trial related documents to be stored for a period of 5 years, sometimes 10 years and sometimes 15 years, whatever are the regulations we have to comply with that particular expectation.

(Refer Slide Time: 17:30)

## Impediments in clinical trial participation

### At the level of patients

- Don't know about clinical trials
- Don't have access to clinical trials
- May be afraid or suspicious of research
- Can't afford to participate

### At the level of health care providers

- Lack awareness of appropriate clinical trials
- Be unwilling to "lose control" of a person's care
- Believe that standard therapy is best
- Be concerned that clinical trials add administrative burdens

There could be some impediments in clinical trial participations. At the patients' level, many times patients do not know about the clinical trials. Sometimes, they do not have access to clinical trials. There is mistrust, suspicion or people are afraid of being participating in clinical trials. Sometimes, they feel there are fees attached to this particular thing. If clinical trials have to succeed, patients have to agree or volunteers have to agree to participate and hence, disseminating this particular information becomes extremely important.

Sometimes, there are some issues at the level of health care providers also. There is a lack of awareness of appropriate clinical trials that sometimes they are unwilling to lose controls for persons care, say for example, a trial is going on in cancer patients, there could be some kind of a fear among health care providers that they would lose their patients to this particular trial once it is referred. Sometimes, they believe that the

standard therapy is best and could be also concerned about the administrative burdens that this may add. So, even if there are impediments here the clinical trials are the only way for making a progress in medical science.

(Refer Slide Time: 18:38)

## Advantages & disadvantages of RCTs

**Advantages**

- The only effective method known to control selection bias
- Controls confounding bias without adjustment
- Facilitates effective blinding in some trials
- Maintains advantages of cohort studies

**Disadvantages**

- May be complex and expensive
- Lack representativeness - volunteers differ from population of interest
- Ethical challenges are immense



nie.gov.in

Because, if there are no clinical trials, no new drugs will come, if there are no clinical trials, no new technologies will be tested, no new vaccines will be tested and so they must be supported and the adequate information about clinical trials must be disseminated. Randomized controlled clinical trials have to be carried out in the best possible manner, with the highest possible quality adherence.

Thank you for your attention.

**Health Research Fundamentals**  
**Dr. Manickam Ponnaiah**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**Lecture - 21**  
**Preparing a concept paper for research projects**

Dear friends, welcome to this session on Health Research Fundamentals. All of you by now, have gone through the basics of research in terms of conceptualizing the ideas, study designs, ethical and scientific contact of research. I am sure now you are ready, to give life to your research ideas in the form of concept paper. I am going to introduce to you, how to prepare a concept paper for research projects.

(Refer Slide Time: 00:43)

**Competency to be gained from this lecture**

- Write a concept paper for a research project



At the end of this session, you will be in a position to write a concept paper for research project.

(Refer Slide Time: 00:51)

## The seven steps of a successful protocol

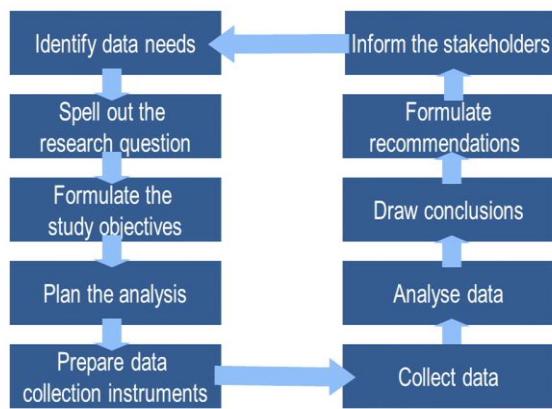


Now, let us revisit the seven steps of a successful protocol. These are logical, sequential and essential steps. The steps include, identifying research topic, framing research question and objectives and then moving on to outlining a one-page concept paper, preparing dummy tables as per the analysis plan, writing a detailed drafted protocol, preparing instruments and annexes including information sheet, consent form and other things related the study protocol; submit this to your competent peer review committee and finally, seek review by an ethics committee.

Let us look at the life cycle of research and how the steps are important relates to the life cycle of research.

(Refer Slide Time: 01:44)

## The life cycle of research



We will see that, we started with identifying data needs and we end with requiring more information that necessitates for the research. So, therefore, in the concept paper has to capture all of these elements in a miniature form of a protocol.

(Refer Slide Time: 02:06)

## The seven steps of a successful protocol



We are today in the second step, how do we outline a one-page concept paper? You may

wonder, why one-page and there is a reason for it.

(Refer Slide Time: 02:19)

## Rationale for using one-page concept paper

- Time is precious
  - For you
  - For your faculty / guide / reviewer
  - For funding agencies
- Brevity forces focus
- Many concept papers are not developed
  - Save time for an idea that may abort



The time, for everybody is very precious for you, for your faculty, for your guides, for your reviewers', time is precious. Equally, the funding agencies are also hot pressed for time. So, one-page in a shorter version forces you to be focused. Many of our ideas just are not born, they remain as ideas. One page, to write a one page as compared to a long detailed protocol, may help us in overcoming the inhibition that we have in writing the detail protocol. Therefore, the ideas may be born rather than getting aborted, that is the reason for one-page. I am going to now outline this one-page, what it requires in bullet styles in each of the sections of the concept paper.

(Refer Slide Time: 03:11)

## Outline of the one-page, bullet-style, concept paper

Background and justification

Objectives

Methods

Expected benefits

Key references

Budget

NICER FOR  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

It has these sections, which is essentially, as I said earlier, miniature of the protocol. It has sectional background and justification, statement of objectives as a separate section, method section, expected benefits, key references and budget. I am going to elaborate on each of this section. Let us begin with looking at background and justification.

(Refer Slide Time: 03:42)

## Outline of the one-page, bullet-style, concept paper

Background and justification

- Importance of the problem
- What is known and unknown about the problem
- The information that is missing to address the problem effectively

NICER FOR  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai



nie.gov.in

The background justification, essentially will have 3 bullets. One on the importance of the study problem, second what is known and unknown about the problem in relation to the literature and in relation to the local contexts, lastly the information that is needed to address the problem effectively.

(Refer Slide Time: 04:06)

## Outline of the one-page, bullet-style, concept paper

### Objectives

- 2-3 objectives
- Can be general and specific
- Can be primary and secondary



Now, let us look at the statement of objectives. In the concept paper, we want you to state not more than 2 or 3 objectives. If needed you may have to split them has general and specific. As we have discussed, in the earlier session on research, questions and objectives, you need to indicate which is your primary objective and which are your secondary objectives. This is very critical because the objectives give clarity to the reviewers about your research process.

(Refer Slide Time: 04:38)

## Outline of the one-page, bullet-style, concept paper

### Methods

- Outline of the methods
  - Study design
  - Study population
  - Operational definitions
  - Sampling procedure
  - Sample size
  - Data collection
  - Analysis plan
  - Human participants protection
- One bullet per point

NICER FOR  
HEALTH RESEARCH FUNDAMENTALS



nie.gov.in

The next section is on methods. It has to give by bullets an outline of all the aspects of methods covering, for example, study design, study population, key operational definitions, the sampling methodology by which you will select your study participants, the calculation of sample size and key considerations for sample size. The data collection procedures, who will collect? What will they collect? And an analysis plan around the key objectives. Finally, this section will end up by saying, what are the ethical issues and which ethics committee will review this protocol, if developed.

And next section is on expected benefits. As you have listened from the earlier session, the research question is to be answered. There are two aspects of benefits that we look forward in answering the research question. One, what action will be taken following the results? Second, what is a future research or planning or policy agenda as a result of this finding? This is very important from the concept paper point of view, you have to set out this section and finally, you need to show that you have mastered the literature.

(Refer Slide Time: 05:55)

## Outline of the one-page, bullet-style, concept paper

### Key references

- Not more than 5
- As per standard guidelines
  - e.g., International Committee of Medical Journal Editors- [icmje.org](http://icmje.org)



NIEIR IIT  
HEALTH RESEARCH FUNDAMENTALS



nie.gov.in

These key references may be referred to in the introduction section and may be needed to refer to some operational definitions in the method section, citation of a reference. We suggest that you need not put in more than 5 references. However, we recommend strongly, you follow internationally acceptable standard guidelines for writing these references. We recommend, you to adopt and use International Committee of Medical Journal Editors guidelines, which is accepted globally by many bio medical journals.

(Refer Slide Time: 06:39)

## Outline of the one-page, bullet-style, concept paper

### Budget

- 4-5 lines
- No detailed justification
- Divided in salaries/per diem, travel, equipments & supplies and miscellaneous



NIEIR IIT  
IIT MADRAS  
HEALTH RESEARCH FUNDAMENTALS



nie.gov.in

The last section is on budget. You may wonder, why budget is important in this concept paper, but it is equally important to technical and other aspects of preparing a concept paper. This is an indicative budget; it need not have detailed justification. It can cover key items like salaries-per diem, travel, equipments and supplies and miscellaneous, whatever is applicable to the context of the research that you are proposing. Many research agencies insist on having an indicative budget. So, this is also equally important.

(Refer Slide Time: 07:15)

## Indian Council of Medical Research (ICMR), Department of Health Research, Govt. of India

The screenshot shows the homepage of the Indian Council of Medical Research (ICMR). At the top, there is a banner for 'Web based Submission, Processing & Management for Extramural Proposals'. Below the banner, there is a logo of the Indian Council of Medical Research and the text 'भारतीय आयुर्विज्ञान अनुसंधान परिषद' (Bharatiya Ayurvedic Vigyan Anusandhan Parishad). To the right of the banner is the Indian National Emblem. The main content area features a 'Welcome to e-PPMS' message with a DNA helix graphic. On the left, there is a 'Quick Links' sidebar with several hyperlinks. In the center, there is a 'Member Login' form with fields for 'Username' and 'Password', and links for 'Forgot Password?' and 'Log In'. To the right of the login form is a 'Broad Areas' section listing various research categories. A blue arrow points from the 'Quick Links' sidebar towards the 'Pre-Proposal Submission Form' link.

Let us look at some of the aspects of application of this concept. If you look at Indian Council of Medical Research, the premier medical research agency in the country, ICMR on its home page offers funding for researchers, like you and me. There is a provision called extramural funding on its home page and you will see on the home page there is a provision to submit, what they call as pre-proposal format, this is similar to concept paper. This pre-proposal format covers the following, a title, an introduction, a novelty, applicability and description of the project.

(Refer Slide Time: 07:48)

## ICMR's pre-proposal format

- Title of the project (*25 words*)
- Introduction (*250 words*)
- Novelty (*100 words*)
- Applicability (*100 words*)
- Description of the project (*700 words*)
  - Methodology, Feasibility, Outcome, Budget, etc



You will see that they have specified word count for each of the section. Now, if you see the outline that be provided, most of this can be taken from the concept paper that we just now, generic concept paper that we just now provided you, except may be, novelty you may have to write it as, in a fresh section.

(Refer Slide Time: 08:21)

## ICMR's Short-term studentship (STS) for medical undergraduates: Format for STS proposals / project

- Title (*25 words*)
- Introduction (*300 words*)
- Objectives (*100 words*)
- Methodology (*800 words*)
- Implications (*100 words*)
- References (*300 words*)

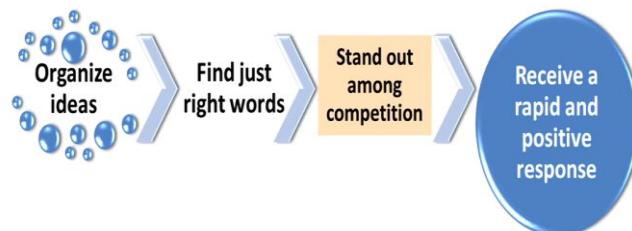


The ICMR is also promoting research culture among medical undergraduates. They are offering, what is called Short-term studentship for medical undergraduates in India, they call it STS proposals. The format for STS proposal, the preliminary one goes like this. It has a title, introduction, objectives, methodology, implications and references. This looks exactly similar to the generic format that we just now recommended. The implication is something similar to the expected benefits that we just now discussed. So, what we are recommending is that, many Indian, other Indian and international agencies accept this pre-proposal or concept note or concept paper as a first step before awarding funding to the researchers. This first step helps them in screening the proposal for its worthiness, merit to award the funds.

In fact, some of the research funding agencies even fund and support developing a concept paper, if it is meritorious to full pledged protocol through their own funding. So, ready, handy concept paper is very useful to submit to such research funding opportunities.

(Refer Slide Time: 09:50)

## What can you achieve with one-page concept paper?



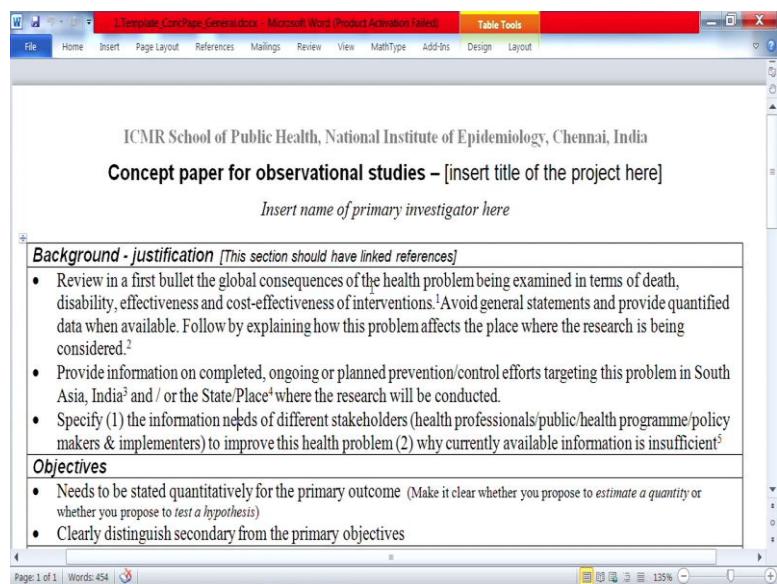
Patrick G Riley. *The one-page proposal*. 2002



Therefore, what can you achieve with one-page concept paper? It helps you to organize your ideas, find just right words in a brief, but succinct manner that helps you to stand out in the competition; you may receive rapid and positive response. For your benefit in

the reading section, you will find a template concept paper, one pager, a word document which you may use for writing your concept paper. Second, we have also provided a sample concept paper.

(Refer Slide Time: 10:25)



Now, here we look at two templates of concept papers, one for writing observational studies another for intervention studies. I am going to show them and go through the parts of these two templates. The first one is an observational studies, as you can see you can type your title of the project, you can type your name, the first section is about background and justification and we have provide tips. You can actually, overwrite on these 3 bullets; the first bullet is about the contexts of the study problem in a quantified manner with linked references. The second bullet is about the local contexts, which what we know and what we do not know is stated. The third bullet, in terms of what information we need to manage the problem effectively, you can just overwrite on these 3 bullets.

The next section is about statement of objectives. We have some tips here, you need to state them qualitatively for the primary objective and you may make it clear, whether you want to propose to estimate a quantity or test hypothesis depending on the nature of the research question and statement of the objectives, which we discussed earlier. And last,

but very important, you need to clearly distinguish the secondary objectives from the primary objectives.

(Refer Slide Time: 12:00)

The screenshot shows a Microsoft Word document window with the title bar 'Template\_ConcPage\_General.docx - Microsoft Word [Product Activation Failed]'. The ribbon menu is visible at the top. A table of contents is displayed on the page, listing various study components with bullet points:

- Clearly distinguish secondary from the primary objectives
- Methods** /Refer to [www.equator-network.org](http://www.equator-network.org) for specific requirements for different studies
- Study population**
  - Specify the population in which you will undertake the study.
- Study design**
  - Describe the type of study (e.g., survey, case-control, cohort studies) in one short bullet.
- Operational definitions**
  - Provide information regarding the key definitions, criteria and / or control recruitment strategy that you will be using.
- Sampling procedure**
  - Describe the type of sampling you will be using.
- Sample size**
  - Briefly mention your sample size and the main assumptions you used to calculate it. This should contain enough information for the reader to redo the calculations to check the estimate.
- Data collection**
  - Explain shortly who will collect what kind of data, what the timeline is and what quality assurance mechanism will be used.
- Analysis plan**
  - Summarize the type of analysis (e.g., descriptive, analytical, stratified, multivariate) that you plan to carry out. Mention laboratory analysis if they will be part of the study.
- Human participant protection**

At the bottom of the screen, the status bar shows 'Page: 1 of 1 | Words: 454' and a zoom level of '135%'. The Microsoft Word ribbon tabs (File, Home, Insert, Page Layout, References, Mailings, Review, View, MathType, Add-Ins, Design, Layout) are visible above the table of contents.

The next section is elaborate section on methods. As we discussed, it should be in bullets and you can derive from tips for the type of study design from this widely accepted guidelines available in this website, equator hyphen network dot org. You can state the study population; you can state the study design in one bullet. It could be in terms of observational study design, it could be a survey, it could be case control cohort studies or ecological studies or sometimes, it could be you know, even case report or case series depending on the study design that you choose, based on the objectives.

Then the operation definitions, you need to provide key definitions only and here is where, as I discussed earlier, you can make a reference to the standard definition for the literature criteria. You can even state control recruitment strategy for the case control study. Sampling procedure, you can state in one bullet, sample size you can briefly mention the sample size and the assumptions. Data collection you need, as I said, you need to spell out, who will collect the data? What kind of data they will collect and within what time frame? And briefly, the quality assurance that is going to be used, finally, the analysis plan, we need to summarize the type of analysis based on the

primary and secondary objectives.

You may mention laboratory analysis, if they are part your study. Human subject production, you may have to mention about key measures that are taken to ensure the production of the human participants and which ethics committee will review this proposal and last, but important section is on expected benefits. You need to describe the report that will be generated and the outcome of the study in order to manage the problem, study problem effectively in the area where the studies conducted.

(Refer Slide Time: 13:56)

The screenshot shows a Microsoft Word document with a table structure. The table has four main sections:

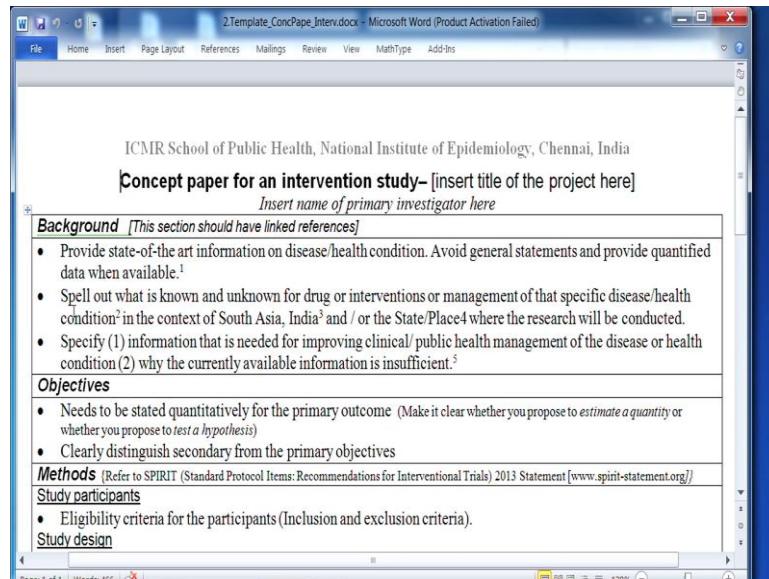
- Human participant protection**: Contains a bullet point: "Mention key measures taken to ensure the protection of human participants in your study and which ethics committee will review the proposal."
- Expected benefits**: Contains two bullet points: "Describe the expected output (e.g., reports) that this study will generate and the timeline." and "Describe the expected outcome: How this study will influence management of this problem in question in the area where the research will be conducted."
- References (As per ICMJE guidelines, not more than 5)**: A numbered list:
  - United Nations, Title, 2011
  - WHO, Title, Place 2011
  - X, Y, Z et al. Achieving the programme objectives. India International. 2011;12:22-26
  - Govt. of India, Title, Place 2010
  - Govt. of Tamil Nadu, Title, Place, 2011
- Budget**: Contains four bullet points: "Staff (Salary and per diem): Rs. XX,XXX", "Transport: Rs. XX,XXX", "Supplies (e.g., laboratory reagents, stationary and others): Rs. X,XXX", and "Miscellaneous: Rs XX,XXX".

**Total amount needed: Rs.X,XX,XXX**

At the bottom of the table, there is a note: "Page: 1 of 1 | Words: 454 | 135% |".

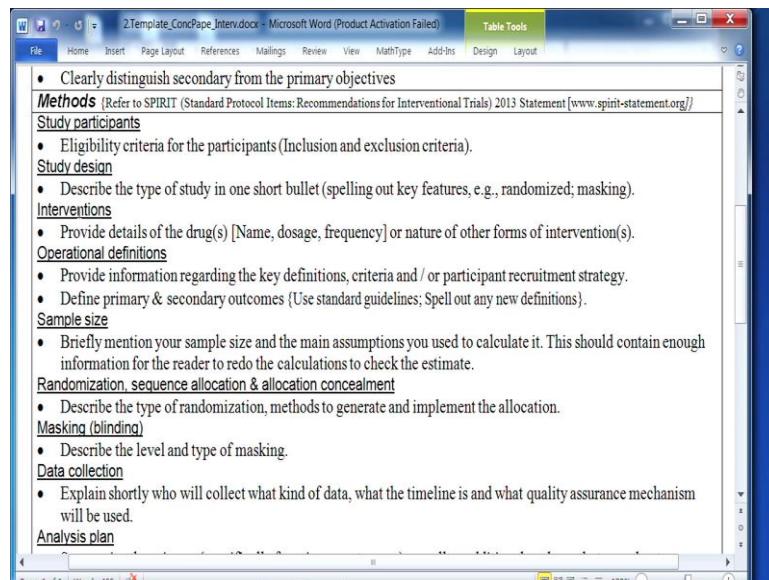
The references, the 5 references, you can even make it small font size, but adhere to the standard guidelines, which we recommend that you use ICMG guidelines. Finally, in the budget section, in a brief or 4 bullets. So, this is on observational studies.

(Refer Slide Time: 14:12)



So, we will now look at concept paper for intervention studies. For intervention studies, the background has to state, what is known and unknown for the drug or interventions or management of the specific study problem that you talking about and you need to say why the information currently available is insufficient. You need to state the objectives clearly here in terms of primary objective, which will look at the primary outcome.

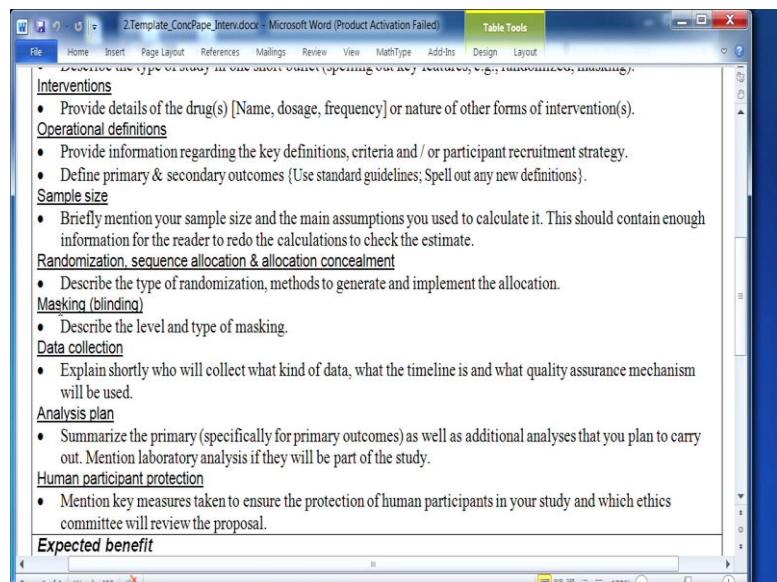
(Refer Slide Time: 14:47)



In terms of methods, you may benefit from using the guidelines that are available for writing a clinical trial protocol. So, you may use the elements from that guideline, which is called SPIRIT. The method section is pretty much similar except that you need to state in the clinical trial concept paper, specifically interventions, the drug or interventions with dosage, frequency, nature and all the other forms of interventions. Operational definitions, you may have to state the primary and secondary outcomes. You may have to state the participant recruitment strategy.

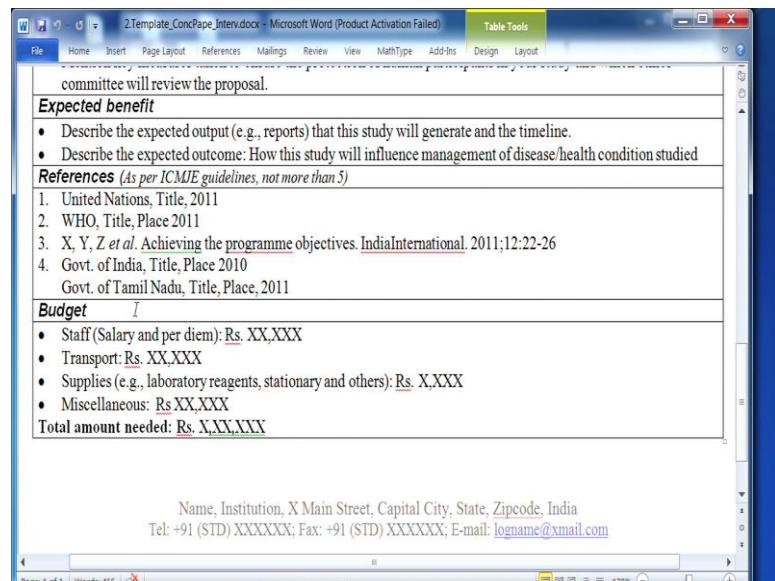
The next important section is on randomization sequence allocation and allocation concealment. You have to briefly mention the type of randomization, methods used to generate and implement the allocation, and masking has to mention as to the level and type of masking.

(Refer Slide Time: 15:37)



The other sections are pretty much similar to what we saw in the concept paper template for observational studies.

(Refer Slide Time: 15:44)



We have a sample out there in the reading section; please makes make use of it while writing your concept papers. We hope you will benefit by making use of it and make writing concept paper a habit and we wish you give birth to your ideas.

Wish you all the best.

Thank you.

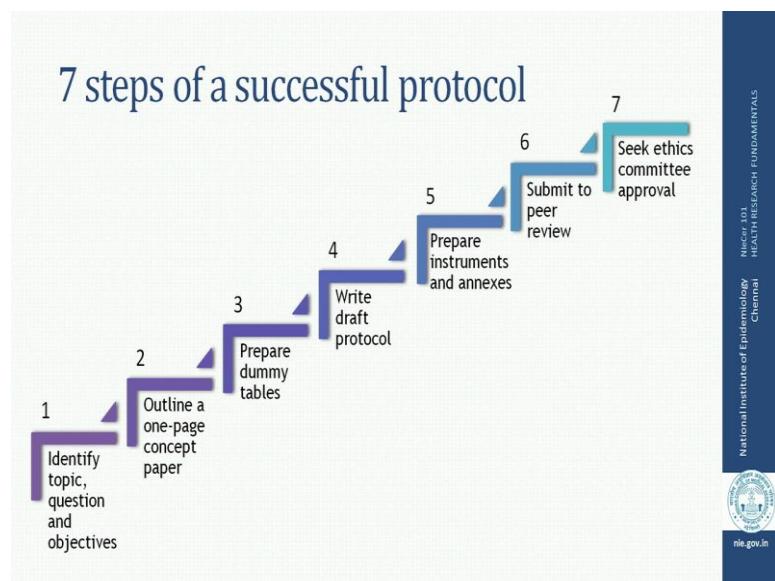
**Health Research Fundamentals**  
**Dr. Tarun Bhatnagar**  
**Department of Humanities and Social Science**  
**Indian Institute of Technology, Madras**

**Lecture - 22**  
**Elements of a Protocol for Research Studies**

Hello and welcome to this session of Health Research Fundamentals. Today, we are going to talk about the Elements of a Protocol that you would write for a Research Study. Remember, that we are in the last week of the online course and you have gone through all the aspects of how to start designing a research study, starting from thinking about a research question, doing sampling, selecting the research participants, doing measurements for exposures and outcomes, looking after the human subject protection and so forth.

Now, these are all the structures of a research project and then all these structures for a research project come together in a written format and that is what is known as a protocol, which is basically a written plan for a study.

(Refer Slide Time: 01:00)



Now, in the last session of writing a concept paper if you remember, Dr. Manickam talked you through the 7 steps of writing a successful protocol, beginning from identifying your research topic and the study objectives, outlining a one-page concept paper, preparing dummy tables and that is when you get on to the part of writing your protocol. We usually start off with writing a draft protocol and then it is refined adding various elements to it and ultimately, ending up with seeking ethics committee approval before you are able to actually do the study.

(Refer Slide Time: 01:40)

## The first draft of the protocol

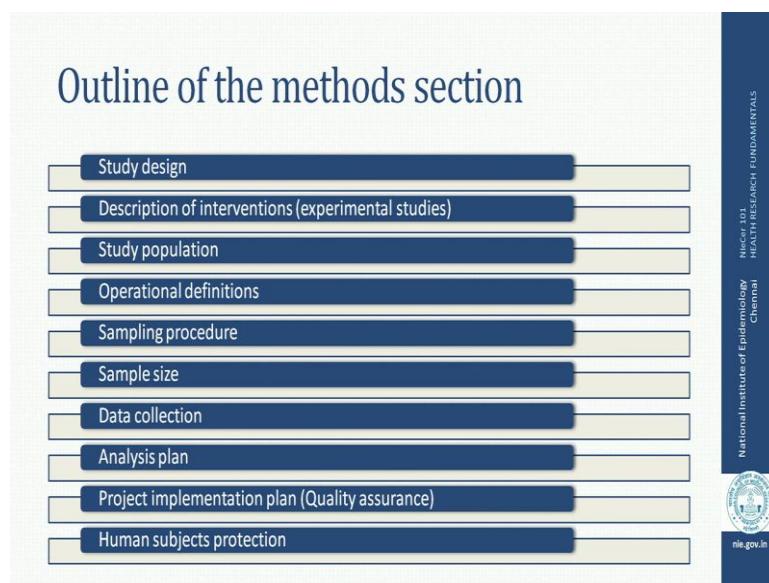
- Thought as it is written
- Keeps concept paper as summary
- Uses the concept paper outline
  - Background/justification
  - Objectives
  - Methods
  - Expected benefits
  - Budget
- Does not exceed 2000 words
  - Introduction < 20% of length
- Contains 5 – 10 key references

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

Remember that, when you first start out writing the protocol, you are basically writing it as you are thinking about it and if you already made a concept paper, you are in a good position because you can then just use a concept paper as an outline to actually write in more detail about various aspects that you have already mentioned as bullet points in your concept paper and developed in to a protocol. Usually, the objectives section of your protocol, we would take it, as you have stated out already in your concept paper, but the other elements like background and justification, the methods, the expected benefits. These are some of the sections that you would have to now give more elaborate details in your research protocol.

It is also generally a norm that usually, the first draft should not be too long, generally about 2500 to about may be 2000 words with a small introduction less than one- fifth of the total protocol, which will keep you focused. And of course, remember that, all you would give a few references in your concept paper and now you can increase a references, depending on how much detail you are adding to your protocol and so you have may be 5 or 10 or even more key references as is needed for your protocol.

(Refer Slide Time: 03:09)



So, as I said, we need much more details and the major part of the protocol would actually be your method section, wherein you give the whole detailed plan of how you are going to conduct the study, starting from the study design, describing the interventions, if you are doing an experimental study. Describing a study population, writing out your operational definitions, what was the sampling method that you used? How did you calculate the sample size? What would be the data collection methods? What would be the tools to do that? What is your analysis plan? And then, a detailed project implementation plan to maintain the quality and finally, the last section on human subject protection. We will walk through each one of these steps, little bit more in detail to see, how you can develop your protocol?

(Refer Slide Time: 04:03)

## Study design paragraph

- Explains how the objectives lead to indicators and to the study design
- Describes the type of study
  - Experimental
  - Cohort
  - Case control
  - Cross sectional
- Describes logistical arrangements
  - Prospective
  - Retrospective

NICER IRI  
HEALTH RESEARCH FUNDAMENTALS



nie.gov.in

So, in terms of the first paragraph would be the paragraph on study designs. Now, this is where you explain, how the objectives that you have stated at the end of the introduction section, how they would be measured? How you may frame indicators? And what design would be used in your study to fulfill these objectives? Remember, your study design could be experimental or observational; in term it could be a cohort study, a case control or cross sectional study. Again, depending on what your study objectives is you need to define again, a little bit more in detail in the protocol, whether your study would follow a prospective design or a retrospective design, in terms of how you are going to collect data and how your study is designed.

(Refer Slide Time: 04:57).

## Description of the interventions

- Applicable if an intervention is planned
  - Clinical trial
  - Community intervention
- Describes the “treatment” applied to the intervention and control group
  - Who?
  - What?
  - When?
  - How?

NICER IRI  
National Institute of Epidemiology  
Chennai  
  
nie.gov.in

In the next paragraph, you would now want to describe the interventions assuming that, if you are doing an experimental study, whether it is a clinical trial or a community based intervention. Accordingly, you will describe your intervention, who is going to be intervened? What exactly is your intervention, give it in more detail? What would be the time period over which this intervention is going to be applied to your intervention group? And how would, that be done? So, the whole procedure of whether, you are doing a clinical intervention or a community based intervention has to be described in detail in this paragraph, if you are doing an experimental study.

(Refer Slide Time: 05:41)

## Study population paragraph

- Use time, place and person:
  - Inclusion criteria
  - Exclusion criteria
    - May be added as a separate section but do not differ conceptually from the inclusion criteria
- Do not confuse the study population and the study sample
- Ensure that the study population is suitable to address the objectives

NICER IRI  
HEALTH RESEARCH FUNDAMENTALS  
Chennai



nie.gov.in

The next paragraph would be detailing about your study population. What is the inclusion and the exclusion criteria that you are going to be using to select study, to select the population either, which may be hospital based, community based or population based into your study. You would use over what time period you are going to select this people? What would be say the geographical location from where you are going to select this people? And what would be the individual characteristics of the people that you going to select into your study or keep them out of the study, which would be written in the form of an exclusion criteria.

Remember that, study population is different from the actual study sample that you are going to take from this population, or the study population is general population among which you are going to be doing the study and a part of them, you are going to be sampling based on your sample size and sampling strategy into your study.

(Refer Slide Time: 06:46)

## Operational definitions paragraph

- Spells out and justifies
  - Key outcomes
  - Key exposures
- Clarity and specificity essential
- References, if applicable

NICER IRI  
HEALTH RESEARCH FUNDAMENTALS



nie.gov.in

Next paragraph gives the details on the various operational definitions that you are going to be using in your study. You spell out, how you are going to define and measure your key outcome measures? What would be your key exposure measures? And again what would be the operational definitions of these measures? Remember that, you need to be very clear and specific in terms of these definitions. And many a times, there may be standard definitions, standard ways and means of measuring the outcomes or measuring certain exposures and then it is always good to use these standard definitions and again for which you would need to provide references, if you are using any of the standard methodologies to define your variables that you are going to be using in the study.

(Refer Slide Time: 07:40)

## Sampling procedure paragraph

- Describes and justifies
  - The type of sample used
    - Convenience sample (Avoid if possible)
    - Random sample
    - Systematic sample
    - Cluster sample
  - The way the sample will be selected in practice
- Provides references, if needed
- Explains randomization, if applicable

NICER IRI  
HEALTH RESEARCH FUNDAMENTALS



nie.gov.in

The next paragraph is about sampling. You need to give in detail, what type of sampling you are going to be using, whether it is a random sample or a non random sample, using systematic sampling, simple random sampling, cluster sampling, multi-state sampling or whatever the case may be, as is applicable to your study and how are you going to select this sample in actual practice. So, all the steps in doing this sampling have to be detailed out in this section of your protocol. Again, if you are using standard methods of sampling, it is also good idea to provide references of the standard ways in which this is already been applied in other settings.

If you are doing, randomized clinical trials then you would be using randomization and as was discussed earlier there are several ways of doing randomization. So, this would be the paragraph, where you would explain which kind of randomization technique you are going to be using and how are the study subjects going to be randomized into your control group and your intervention group.

(Refer Slide Time: 08:55)

## Sample size paragraph

- Details all parameters used to estimate the sample size
- Explains how the estimate was generated
  - Software used
  - Formula used
- Provides references, if needed

Of course, what follows is the paragraph on sample size, where you will explain, how you have calculated the sample size? What are all the parameters that you have used to estimate the sample size? Remember that, the sample size will depend on their study objective as well as the sampling methodology that you are going to employ for your study. You further need to give in detail, how this estimate is generated? What formulas were used? Whether you will be using any softwares for doing that? And again you need to provide references for these things.

(Refer Slide Time: 09:32)

## Data collection paragraph

- Lists the data that will be collected
- Specifies how the data will be collected
  - Who?
  - How?
    - Type of instrument to be used
    - Type of data collection method

NICER IRI  
National Institute of Epidemiology  
Chennai



nie.gov.in

Following the sample size is the paragraph on data collection. There would be 2 sections to this paragraph; one part is where, what is all that kind of data that you are going to collect? Whether it is socio demographic characteristics, the individual characteristics of the study participants, their clinical histories, the science and symptoms and so forth? Secondly, you would also need to specify, how you are going to collect this data? Who is going to do that? What data is going to be collected by the principle investigator or is it going to be some other study investigators, who are going to collect it? Is it going to be a staff nurse? Is it going to be an outreach worker and so forth?

And of course, how you are going to collect this data? What study instrument are you going to be using to collect this in data? Are you going to be doing interviews? Are they going to be face to face? Is it going to be computer based? Are you going to extract data using data abstraction forms? Or if you are doing qualitative methods, would you be doing focus group discussions or in-depth interviews? And what would be the guides for collecting this kind of data?

(Refer Slide Time: 10:48)

## The analysis plan paragraph

- Data entry
- Software used
- Recoding stage
- Descriptive stage
  - Prevalence, incidence
- Analytical stage
  - Univariate
  - Stratified
  - Multivariate analysis

NICER IRI  
HEALTH RESEARCH FUNDAMENTALS  
Chennai



nie.gov.in

Once, you have spelled out your data collection methods. The next paragraph will be explaining, how you are going to analyze this data? Once you have collected the data what would be your mode of data entry? Is it going to be manual? Is it going to be computer based? What software are you going to be using to analyze this data? You also need to give in detail various stages that you may go through in doing the data analysis, if you are going to be re coding certain variables that you are going to collect during your data collection. What are you going to estimate? What parameters? Whether it is prevalence, incidence or if you are doing analytical studies, what kind of analysis are you going to go through and the various steps that you are going to go through that.

So, all of these has to be mentioned in this analysis plan paragraph. It is not just enough to say that suitable analysis will be carried out for this study, that seems quite vague and it is not a good idea. It does not give a good impression; it shows that maybe you are not planned your study well enough. So, it is always a good idea to actually lay out what all analysis you are going to be carrying out to fulfill your study objectives.

(Refer Slide Time: 12:09)

## Project implementation plan paragraph

- Details the steps that will be used to ensure data quality at all stages
- Addresses
  - Data collection
  - Data entry and analysis
  - Reporting
  - Roles and responsibilities of investigators
  - Project governance
  - Coordination of project activities
  - Project timeline

NICER IRI  
National Institute of Epidemiology  
Chennai  
  
nie.gov.in

The next paragraph is quite important and it is actually quite a lengthy part of your protocol and which is basically, detailing the steps that you are going to use to actually implement the study. To ensure data quality at all stages of the study, in the way you are going to collect the data, in the way you are going to do data entry analysis. How you going to report these study results. This is also the place, we are going to give in detail the various roles and responsibilities that the investigators in the study have. It is also good place to provide, what would be the project governance? Who would be responsible for this project, both in terms of administrative responsibilities as well as the technical responsibilities? Who is the PI, the primary investigator? Who are the co-investigators? What their roles are and how various project activities are going to be coordinated?

Basically, all the logistics of actually implementing the data collection phase in the field and of course, you need to also provide, what is going to be the timeline for your study? What is going to be a study period? How are different phases of your study starting from say, designing the instruments, getting approvals from ethics committees, then doing pilot testing and then actually collection of data followed by data entry analysis and so forth? So, you need to give a timeline over the period of your study or for various activities that you are going to be doing as part of implementing your study.

(Refer Slide Time: 13:52)

## Human subjects protection paragraph

- Explains the steps that will be used to protect the study participants
- Addresses
  - Minimization of risks (Confidentiality)
  - Maximization of benefits
  - Compensations (without undue incentive)
  - Informed consent
  - Approval procedures (Ethics committee)



The last paragraph in the method section is the paragraph on human subject protection and this is where you explain in detail, how you are going to make sure that your study participants are protected from any harm and here, this would usually have sub-headings in terms of what are the potential risks that you foresee your participants may be going through, when you are collecting data? What are the procedures that you are going to minimize these risks? How you going to maximize the benefits? Are you going to give them any compensation?

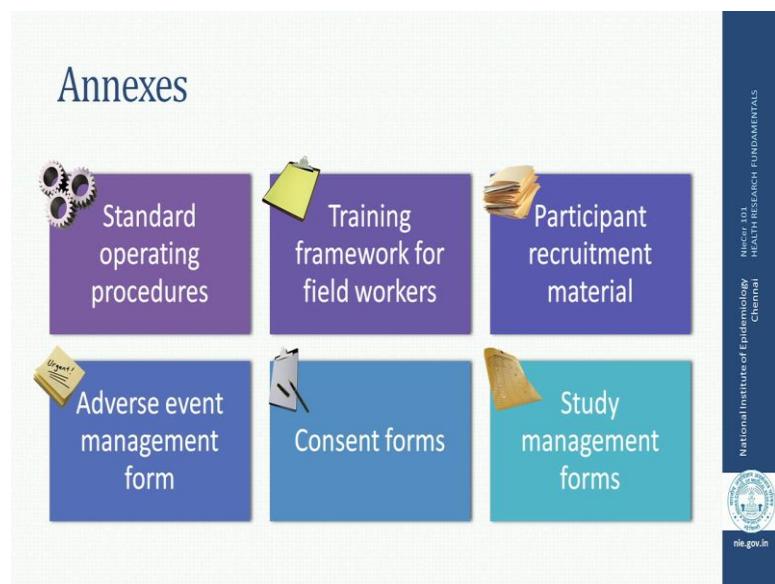
Again, this is the place, where you gave in detail of what kind of informed consent are you going to take from the study participants. Is it going to be a written consent, oral consent? Who is going to take the consent? How long will it take? What would be the various elements of this consent? And then what would be the various procedures that you would need to get approval for your study from the ethics committee of your institution?

(Refer Slide Time: 14:58)



Now, once you have put together all these elements in the protocol, you also need to append to it, all your data collection instruments depending on your, this study design, your study objectives and so forth. You would have decided what data collection instruments you are going to be using, whether these are questionnaires or abstraction forms, a structured observation guide or if you using qualitative methods and either an in-depth interview guide or a focus group discussion guide, whatever be it. This is the place where you are going to append all of these instruments to your protocol and it is always a good idea to have these instruments in local languages as well because that is how in the field you are going to be collecting data. So, you would have both the data collection instrument in English as well as in the local language.

(Refer Slide Time: 15:52)



And then at the end, what we need to do is to put annexes, which gives in more detail various different components of actual study implementation. These could include things like forms, which state the various standard operating procedures of how you are going to be collecting information, whether this is field base or laboratory based or clinic based and so forth. You may want to provide, how are you going to train your field workers or train your, if study investigators in collecting data. What would be the training module and the framework for the same? If you have participant recruitment material, that is something that gets appended into the annexes.

Other forms, such as adverse event management forms, study management forms again would form a part of the annexes and most importantly would be your consent forms, the participant information sheet as well as the consent statement form, wherein the study participant, who would be signing for agreeing to participate in the study. Again, remember that wherever participant are involved and wherever you have forms that involve participants, you may need to have forms both in English as well as in the local language.

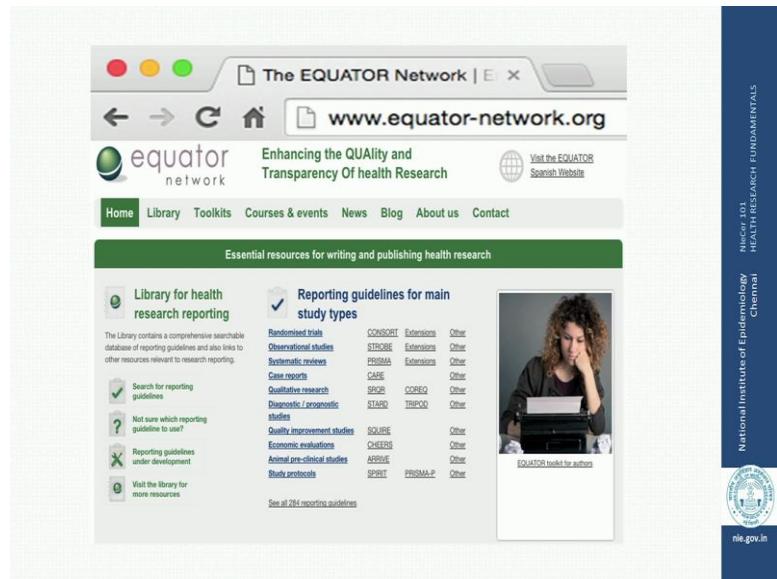
(Refer Slide Time: 17:14)



And of course, once you have put this altogether, you need to firm it up. Remember that, writing a protocol is usually not a one step process. It goes through several drafts and it is always a good idea to actually get it reviewed through your peers, which could who could be your colleagues or even subject matter experts and get feedback from them and then revise and revise your protocol to get your final stage, where it can be submitted for review to the ethics committee.

Now, the ethics committee may themselves have certain suggestions or they may want some things to be changed or revised in your protocol. So, you will have to be doing that and so again, there will lead to another draft of your protocol. By the time you are ready with the final draft, you would actually have a multiple drafts and it is always a good idea to archive all these drafts. So, that you know how your study is evolved and what are the different ways in which you have planned you study.

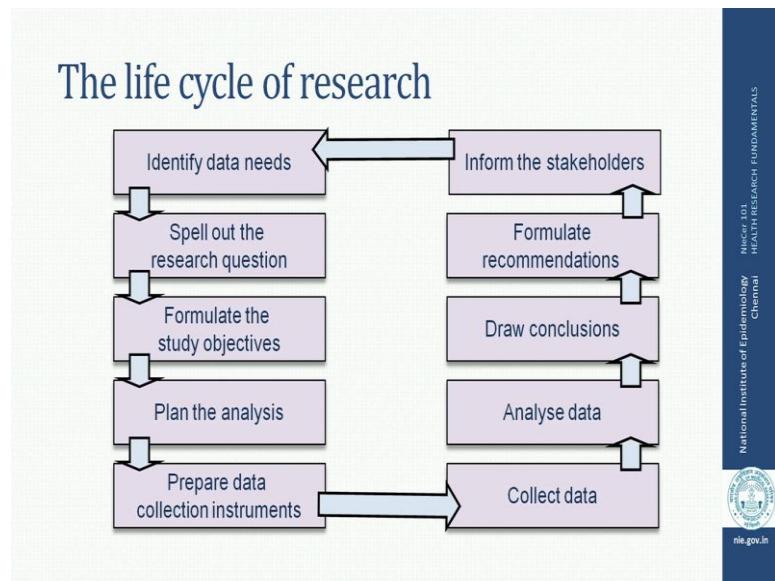
(Refer Slide Time: 18:24)



Now, as a resource material, I would like you to take a look at this website, which is called equator hyphen network dot org. And this is a website, which would be very useful for you to develop study protocols. It has lot of links and lot of resource material. One of the important one, which is right there on it is home page is what they called the reporting guidelines for main study types and if you look, you have guidelines for observational studies, for experiments, for clinical trials, for qualitative research diagnostic studies and so forth.

And you even have one section on study protocols, so there is the SPIRIT protocol; the SPIRIT guidelines, which basically gives you guidelines on how to write a protocol for clinical trials in particular. And then, the other reporting guidelines, whether it is the consort statement or the strobe for the observation studies, prisma for looking at meta analysis and the systematic reviews, although these are called reporting guidelines, but you could actually use these templates to even make your protocols because ultimately, what you are going to report is going to based on how you have written your protocol.

(Refer Slide Time: 19:47)



So, remember that, we have come to actually the last week of this course and you would have seen this life cycle in the week one and we have been following this life cycle of research, which takes you through various steps in which you would do a research study and please keep this in mind that following this life cycle is vital. So to have a study that is logical, that is focused and that is efficient in the way that you are going to conduct the study. So, whenever you are thinking of writing a study protocol, designing the study, make sure you do not jump across the steps of this life cycle and you follow the step one after the other and I ensure you that you are going to end up, with a very good protocol for your study design.

Thank you.

**THIS BOOK  
IS NOT FOR  
SALE  
NOR COMMERCIAL USE**



(044) 2257 5905/08



nptel.ac.in



swayam.gov.in