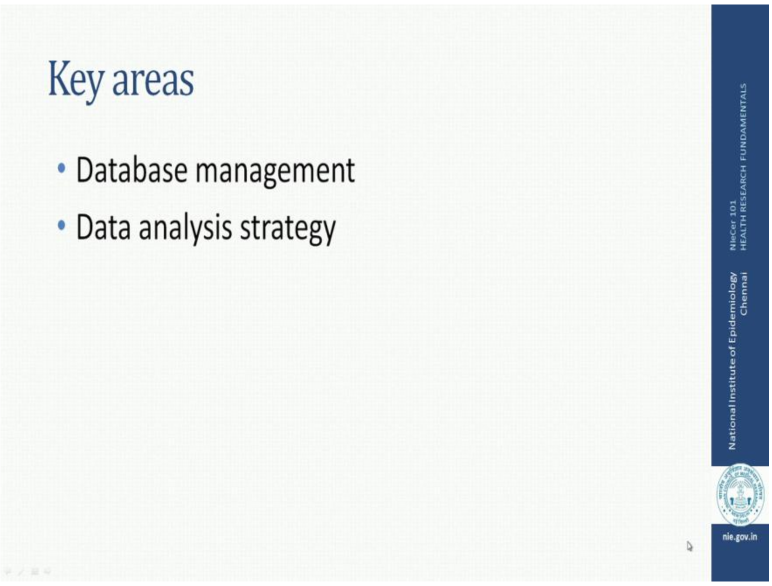
Health Research Fundamentals
Dr. Manickam Ponnaiah
ICMR-National Institute of Epidemiology, Chennai

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)



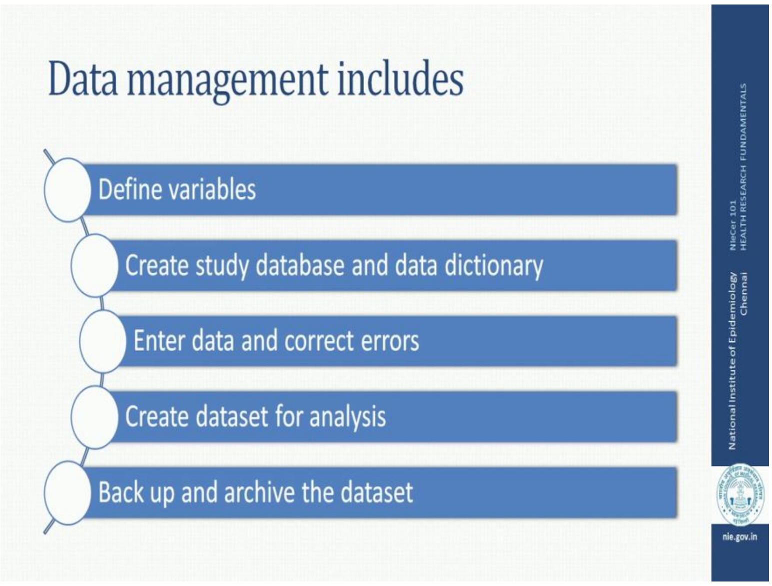
Key areas

- Database management
- Data analysis strategy

National Institute of Epidemiology
Chennai
NIeCer 101
HEALTH RESEARCH FUNDAMENTALS
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)



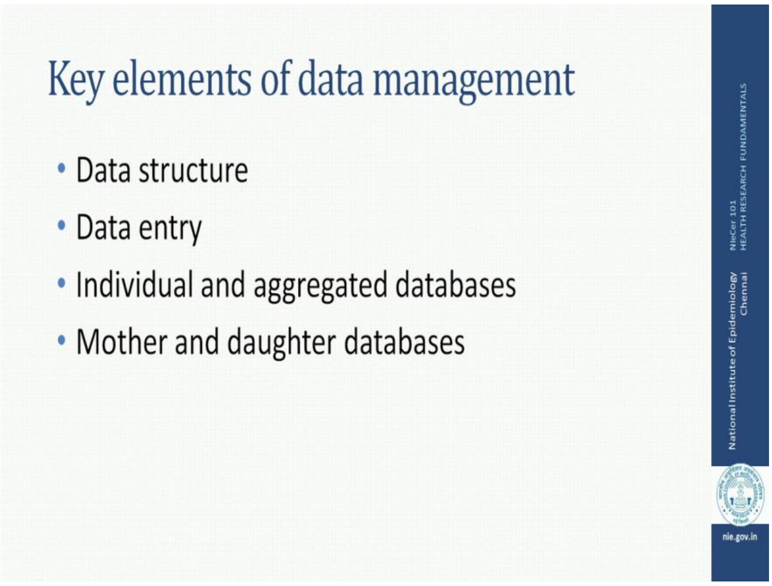
Data management includes

- Define variables
- Create study database and data dictionary
- Enter data and correct errors
- Create dataset for analysis
- Back up and archive the dataset

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

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)



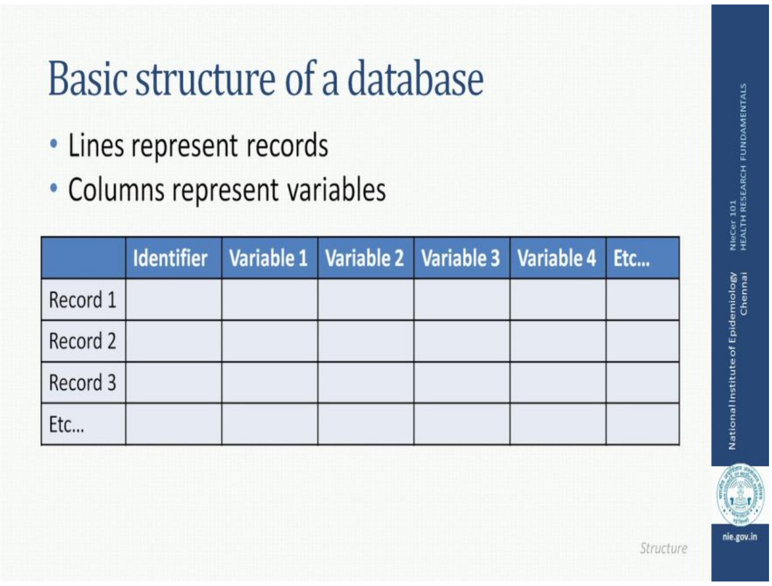
Key elements of data management

- Data structure
- Data entry
- Individual and aggregated databases
- Mother and daughter databases

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

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

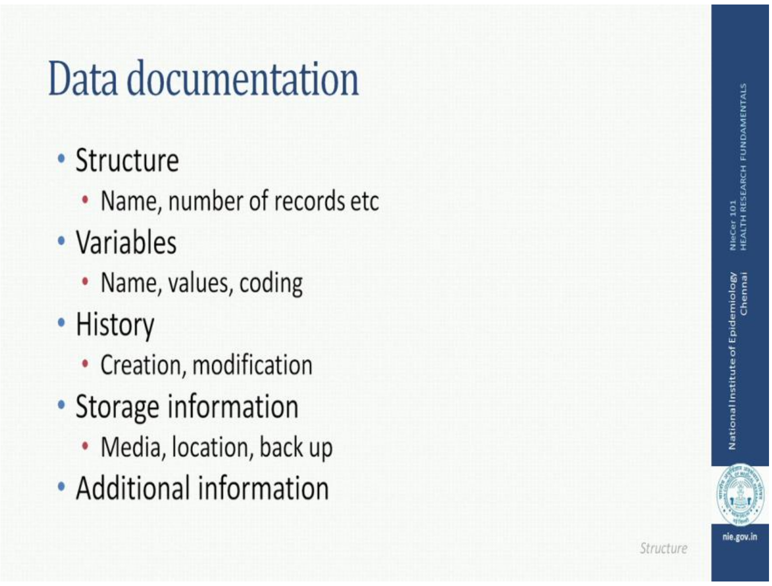
- Lines represent records
- Columns represent variables

	Identifier	Variable 1	Variable 2	Variable 3	Variable 4	Etc...
Record 1						
Record 2						
Record 3						
Etc...						

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

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, to one particular individual, each of these columns represent a variables, the information that is collected on certain variables based on the study question (Refer Time: 01:51).

(Refer Slide Time: 01:53)



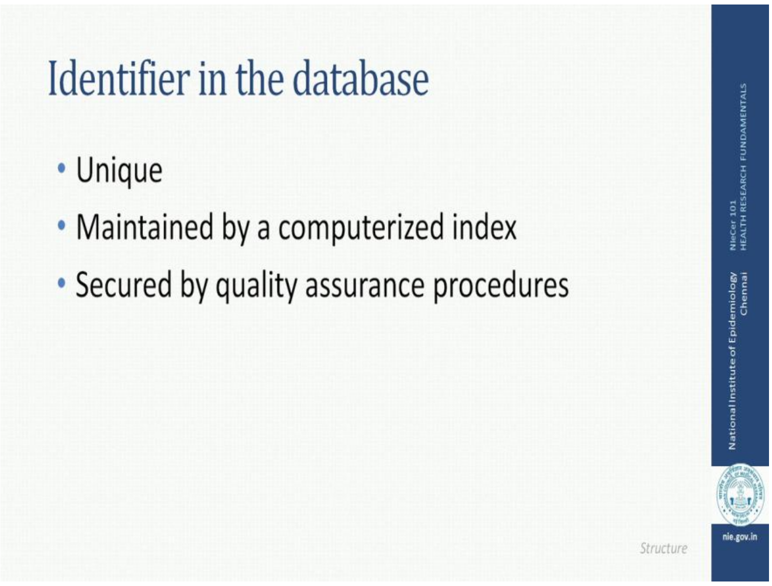
Data documentation

- Structure
  - Name, number of records etc
- Variables
  - Name, values, coding
- History
  - Creation, modification
- Storage information
  - Media, location, back up
- Additional information

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

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

Structure

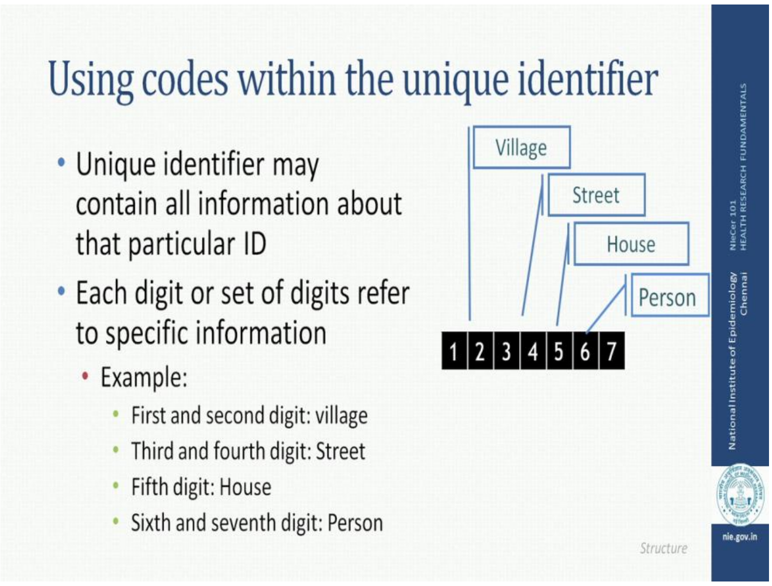
National Institute of Epidemiology Chennai

NieCer 101 HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

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)



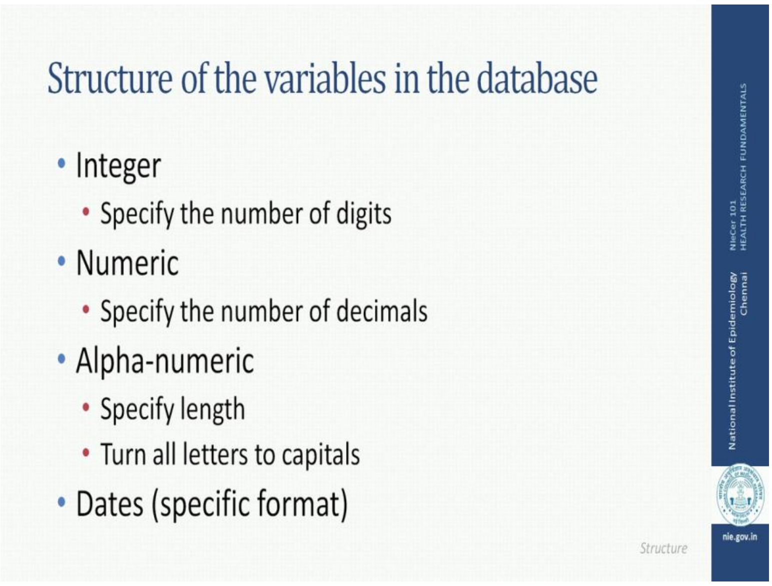
Using codes within the unique identifier

- 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

Village
Street
House
Person
1 2 3 4 5 6 7
Structure
National Institute of Epidemiology Chennai
NIeCer 101 HEALTH RESEARCH FUNDAMENTALS
nie.gov.in

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)



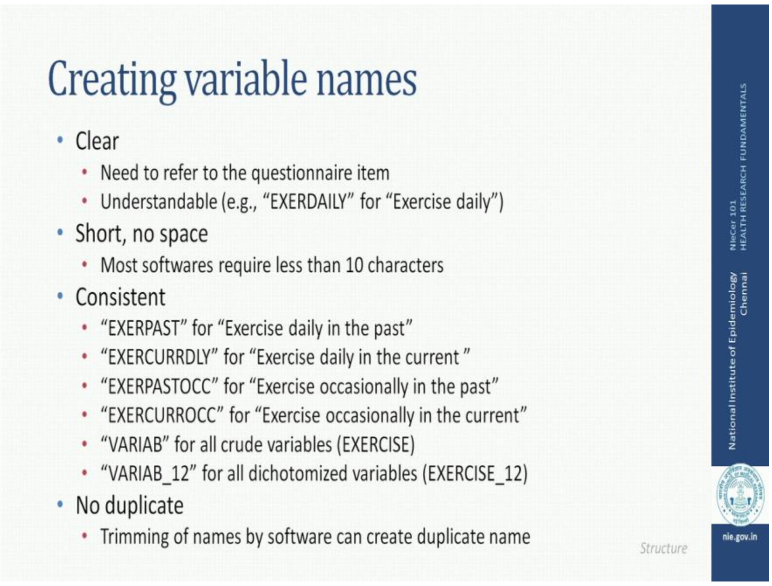
Structure of the variables in the database

- Integer
  - Specify the number of digits
- Numeric
  - Specify the number of decimals
- Alpha-numeric
  - Specify length
  - Turn all letters to capitals
- Dates (specific format)

Structure
National Institute of Epidemiology Chennai
NIeCer 101 HEALTH RESEARCH FUNDAMENTALS
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

Structure

National Institute of Epidemiology
Chennai

NIeCer 101
HEALTH RESEARCH FUNDAMENTALS

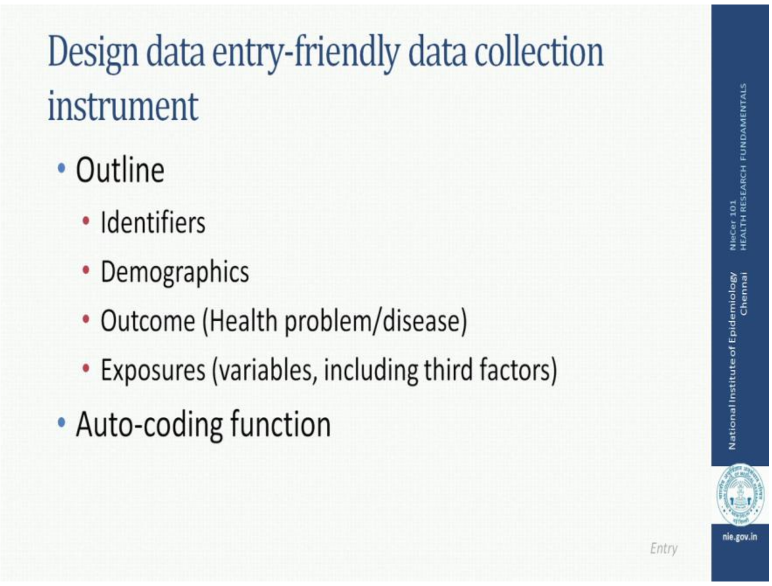
nie.gov.in

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 is 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 EXERPAST, 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

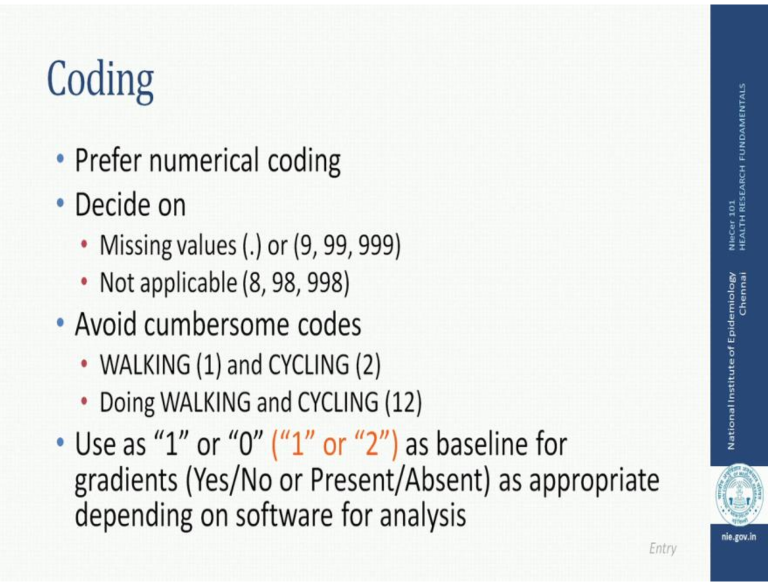
Entry

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

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 is 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

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

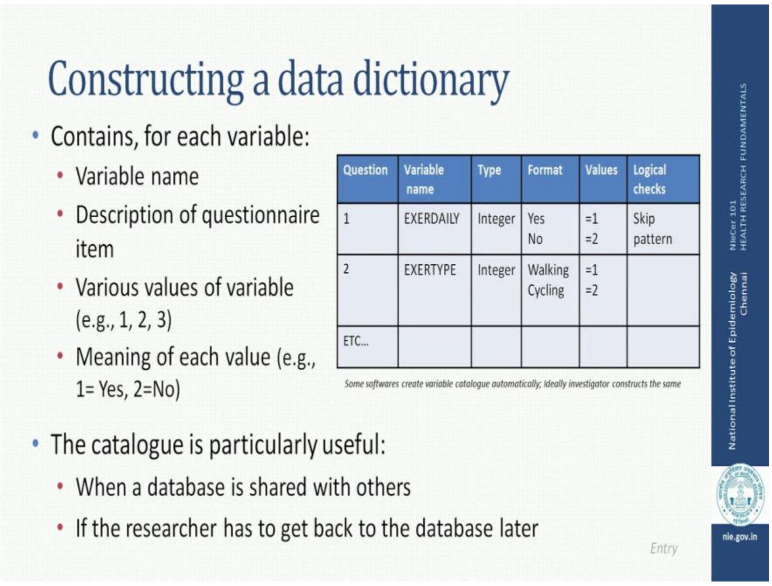
An important aspect of data entry is all about coding. It is always preferable to have numerical coding. Of course, you would have seen, if it is textual information in the form of qualitative there is a different way of dealing with it but with reference to this section we are talking about quantitative 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 consistently 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 researches.

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)

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

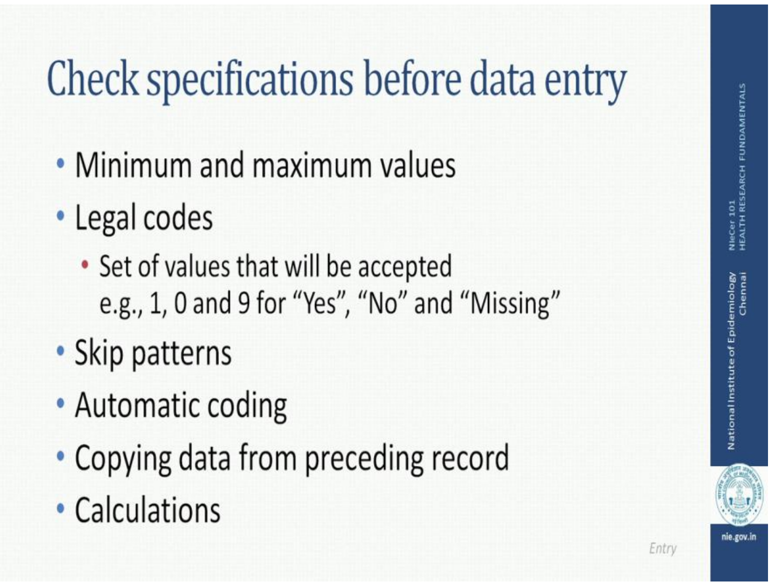
- The catalogue is particularly useful:
  - When a database is shared with others
  - If the researcher has to get back to the database later

Entry
National Institute of Epidemiology
Chennai
NIeCer 101
HEALTH RESEARCH FUNDAMENTALS
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

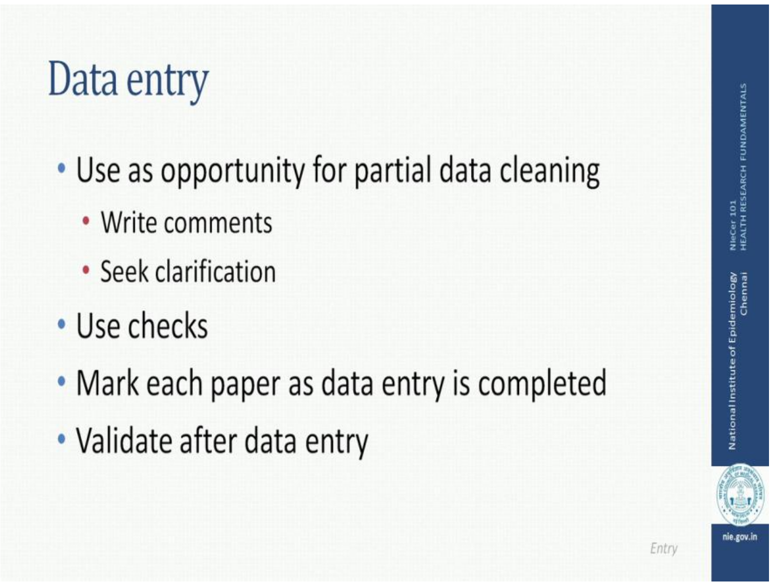
- Minimum and maximum values
- Legal codes
  - Set of values that will be accepted e.g., 1, 0 and 9 for "Yes", "No" and "Missing"
- Skip patterns
- Automatic coding
- Copying data from preceding record
- Calculations

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

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 your 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 can 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 refered 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 to 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

Entry

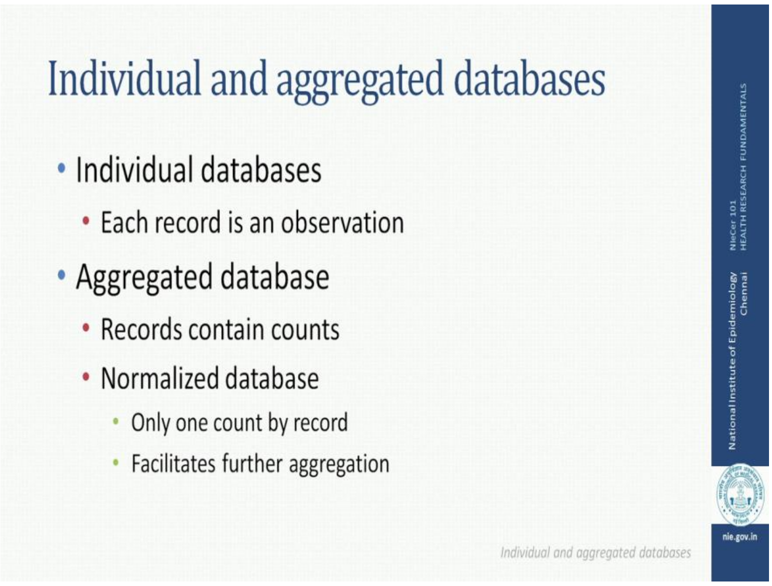
National Institute of Epidemiology Chennai

NIeCer 101 HEALTH RESEARCH FUNDAMENTALS

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

Individual and aggregated databases

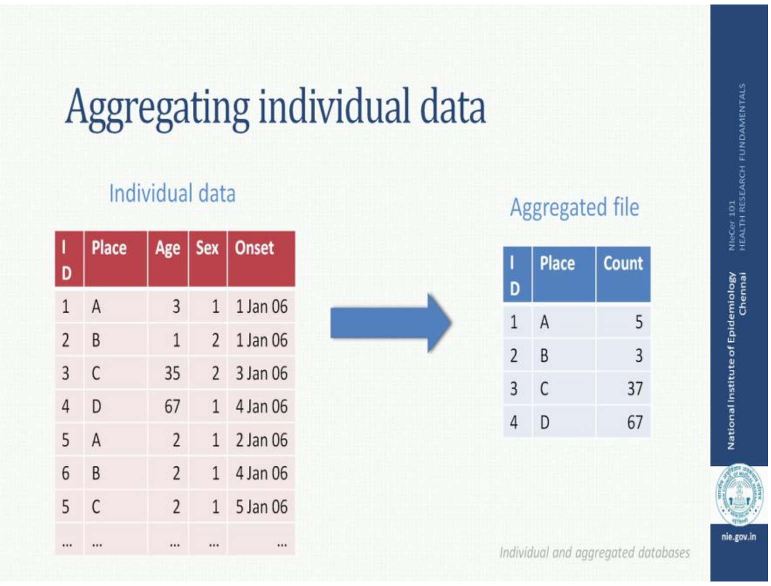
NeCer 101
HEALTH RESEARCH FUNDAMENTALS

National Institute of Epidemiology
Chennai

nie.gov.in

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 rows 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)



Aggregating individual data

Individual data

ID	Place	Age	Sex	Onset
1	A	3	1	1 Jan 06
2	B	1	2	1 Jan 06
3	C	35	2	3 Jan 06
4	D	67	1	4 Jan 06
5	A	2	1	2 Jan 06
6	B	2	1	4 Jan 06
5	C	2	1	5 Jan 06
...	...	...	...	...

Aggregated file

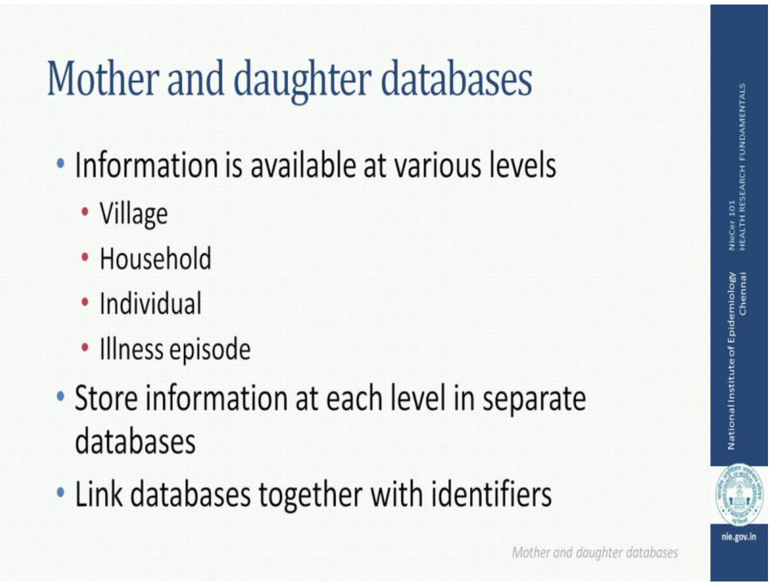
ID	Place	Count
1	A	5
2	B	3
3	C	37
4	D	67

Individual and aggregated databases

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

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 can 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)



Mother and daughter databases

- Information is available at various levels
  - Village
  - Household
  - Individual
  - Illness episode
- Store information at each level in separate databases
- Link databases together with identifiers

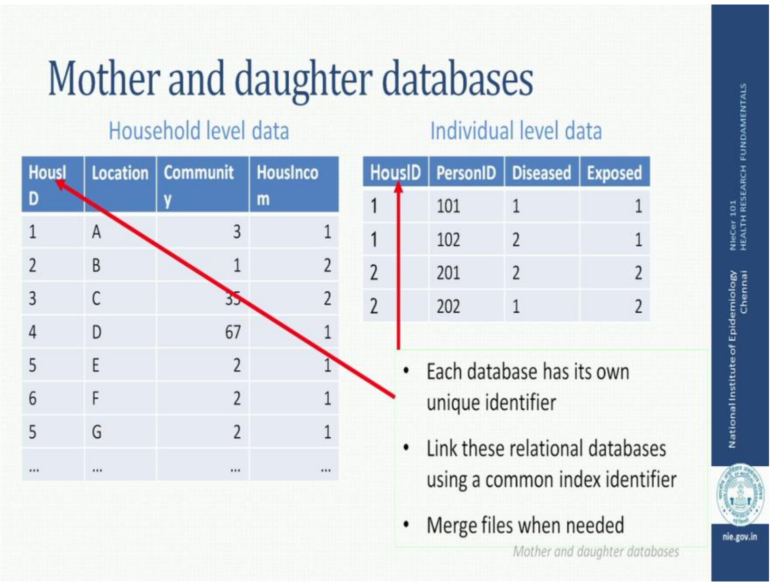
Mother and daughter databases

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

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	HousIncom
1	A	3	1
2	B	1	2
3	C	35	2
4	D	67	1
5	E	2	1
6	F	2	1
5	G	2	1
…	…	…	…

Individual level data

HousID	PersonID	Diseased	Exposed
1	101	1	1
1	102	2	1
2	201	2	2
2	202	1	2

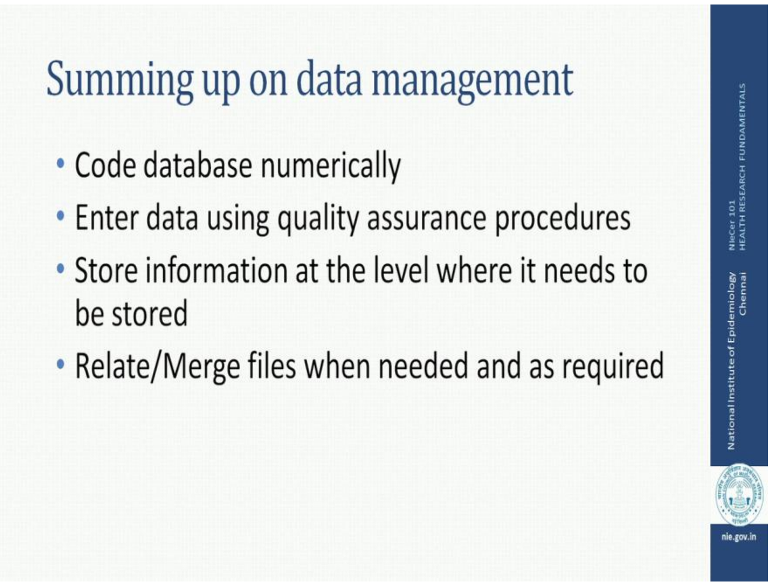
- Each database has its own unique identifier
- Link these relational databases using a common index identifier
- Merge files when needed

Mother and daughter databases

National Institute of Epidemiology
Chennai
NIeCer 101
HEALTH RESEARCH FUNDAMENTALS
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 diseased 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 Epidemiology
Chennai

NieCer 101
HEALTH RESEARCH FUNDAMENTALS

nie.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.