INDIAN COUNCIL OF MEDICAL RESEARCH | NATIONAL INSTITUTE OF EPIDEMIOLOGY

Data management

Dr. Manickam Ponnaiah
BSMS, MSc, PhD

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

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

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

Key elements of data management

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

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…						

Structure

Data documentation

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

Structure

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

Identifier in the database

- Unique
- Maintained by a computerized index
- Secured by quality assurance procedures

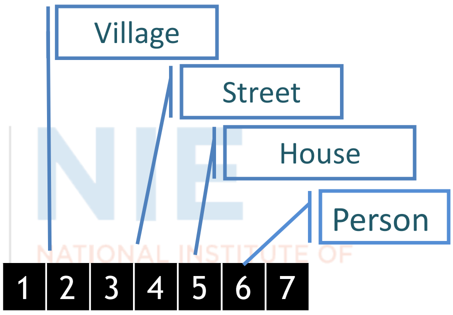
Structure

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

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



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)

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

Design data entry-friendly data collection instrument

- Outline
  - Identifiers
  - Demographics
  - Outcome (Health problem/disease)
  - Exposures (variables, including third factors)
- Auto-coding function

Entry

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

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

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

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

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

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

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

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
…	…	…	…	…

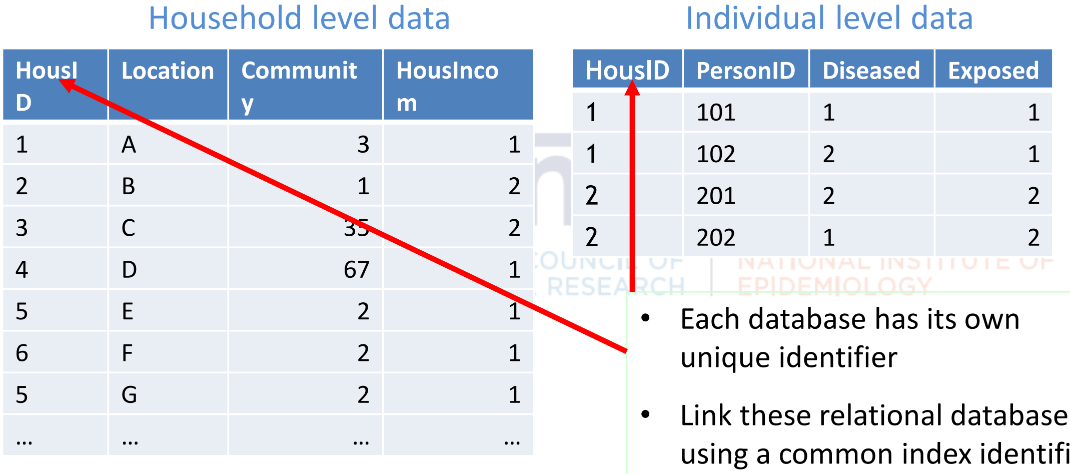
Aggregated file

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

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



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

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

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in

INDIAN COUNCIL OF MEDICAL RESEARCH

NATIONAL INSTITUTE OF EPIDEMIOLOGY

Thank you

HEALTH RESEARCH FUNDAMENTALS

nie.gov.in