Data Management

Sasivimol Rattanasiri, Ph.D
Section for Clinical Epidemiology and Biostatistics
Ramathibodi Hospital, Mahidol University
E-mail: sasivimol.rat@mahidol.ac.th

Outline of talks

- Concept of data management
- Designing data collection form
- Creating databases
- Creating data quality control

Objective of Data Management

The objective is to prepare the data of the highest possible quality in a suitable form for statistical analysis

Data management processes

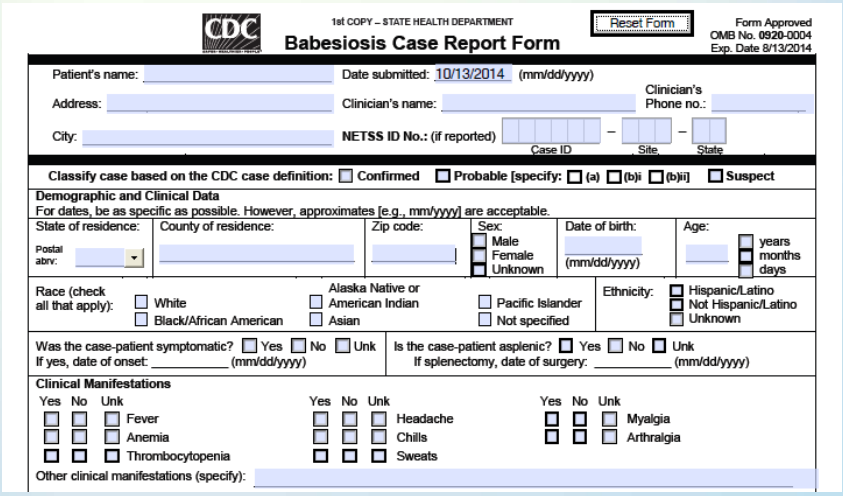
- Case report form (CRF) design
- Data collection
- Database management
- Data entry
- Data cleaning and checking

Case report form (CRF) design

Definition

Case report form (CRF) is a *paper*, or *electronic document* which is designed to collect all of the protocol required data.

Example of paper CRF



1st COPY – STATE HEALTH DEPARTMENT

Babesiosis Case Report Form

Reset Form

Form Approved
OMB No. 0920-0004
Exp. Date 8/13/2014

Patient's name: | Date submitted: 10/13/2014 (mm/dd/yyyy)

Address: | Clinician's name: | Clinician's Phone no.:

City: | NETSS ID No.: (if reported) Case ID – Site – State

Classify case based on the CDC case definition: ☐ Confirmed ☐ Probable [specify: ☐ (a) ☐ (b)i ☐ (b)ii] ☐ Suspect

Demographic and Clinical Data
For dates, be as specific as possible. However, approximates [e.g., mm/yyyy] are acceptable.

State of residence: Postal abrv: | County of residence: | Zip code: | Sex: ☐ Male ☐ Female ☐ Unknown | Date of birth: (mm/dd/yyyy) | Age: ☐ years ☐ months ☐ days

Race (check all that apply): ☐ White ☐ Black/African American ☐ Alaska Native or American Indian ☐ Asian ☐ Pacific Islander ☐ Not specified

Ethnicity: ☐ Hispanic/Latino ☐ Not Hispanic/Latino ☐ Unknown

Was the case-patient symptomatic? ☐ Yes ☐ No ☐ Unk
If yes, date of onset: _____ (mm/dd/yyyy)

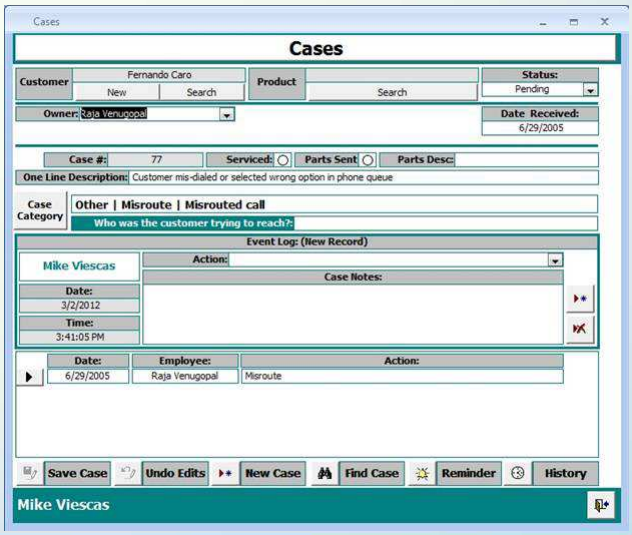
Is the case-patient asplenic? ☐ Yes ☐ No ☐ Unk
If splenectomy, date of surgery: _____ (mm/dd/yyyy)

Clinical Manifestations

Yes	No	Unk		Yes	No	Unk		Yes	No	Unk	
☐	☐	☐	Fever	☐	☐	☐	Headache	☐	☐	☐	Myalgia
☐	☐	☐	Anemia	☐	☐	☐	Chills	☐	☐	☐	Arthralgia
☐	☐	☐	Thrombocytopenia	☐	☐	☐	Sweats				

Other clinical manifestations (specify):

Example of electronic CRF



Objective of CRF design

The objective of CRF design is to collect complete, accurate, and unambiguous data which answer the research question specified by the protocol.

Principle for CRF design

1. Determine Basic questions

- What are the objectives of research?
- What is the type of study design?
- What variables will be involved?
- How variables will be collected?
- How often variables will be collected?

Principle for CRF design

1. Determine Basic questions

Example

For a retrospective cohort study of kidney transplantation, researchers would like to study the association between type of donor and risk of graft rejection.

1) What are the objectives of this research?

To study the association between type of donor and risk of graft rejection

2) What is the type of study design?

Retrospective cohort study

3) What variables will be involved?

- *Type of donor*
- *Graft status*

4) How variables will be collected?

- *Type of donor was classified as cadaveric or living relative.*
- *Graft status was classified as graft rejection or non-graft rejection.*

5) How often variables will be collected?

- *The data of type of donor was collected during enrollment period.*
- *The data of graft status was collected every 6 months during the follow up period.*

Principle for CRF design

2. Determine numbers of CRFs

- Decide how many different CRFs should be created to collect the data.

- Decide which data should be collected on which form.

2. Determine numbers of CRFs

Example

The data requirements for a cohort study of the kidney transplantation can be divided into four parts:

- General characteristics of recipients.
- General characteristics of donors.
- Details of kidney transplantation.
- Information after kidney transplantation .

2. Determine numbers of CRFs

1) General characteristics of recipients:
   - Age of recipients,
   - BMI of recipients,
   - Blood group of recipients,
   - Underlying diseases of recipients,
   - and so on.

These data were collected during the enrollment period.

2. Determine numbers of CRFs

2) General characteristics of donors:
   - Age of donors,
   - Type of donors,
   - Relationship between donors and recipients,
   - and so on.

These data were collected during the enrollment period.

2. Determine numbers of CRFs

3) Details of kidney transplantation:

- Cold ischemic time,
- Warm ischemic time,
- Initial immunosuppressive drugs,
- Immediate complications,
- and so on.

These data were collected during the enrollment period.

2. Determine numbers of CRFs

4) Information after transplantation:

- Changes of immunosuppressive drugs,
- Changes of creatinine level,
- Complications,
- Graft status,
- and so on.

These data were collected every 6 months during the follow up periods.

2. Determine numbers of CRFs

1. Enrollment form
ID number
Part I Recipient

Part II Donor

Part III Transplantation

---

2. Follow up form
ID number
Date of visit

Principle for CRF design

3. Rules for collecting continuous data

Example: Format for collecting continuous data

1. HN	_ _ _ _ _ _ _
2. Weight	_ _ _ . _ _
3. Height	_ _ _ . _ _
4. SBP	_ _ _
5. DBP	_ _ _

Principle for CRF design

3. Rules for collecting continuous data

The units to be used in recording the data should be specified.

1. HN	_ _ _ _ _ _ _	
2. Weight	_ _ _ . _ _	**kg**
3. Height	_ _ _ . _ _	**cm**
4. SBP	_ _ _	**mmHg**
5. DBP	_ _ _	**mmHg**

Principle for CRF design

3. Rules for collecting continuous data

Do not group continuous data at data collection time.

3a. Age at enrollment

1) 15-24
2) 25-35
3) 36-45
4) > 45

X

3b. Age at enrollment _ _ years

√

Principle for CRF design

3. Rules for collecting continuous data

• Do not make any calculations before data entry. Why?

- Since it may cause many errors and more time is consumed.

- You can obtain a value that you want later from the original input, e.g. a value BMI can be calculated from height and weight.

Principle for CRF design

4. Rules for collecting categorical data

- All possible categories of categorical variables should be displayed on the form.

Please circle the right answer What is your sex?
Male
Female

Principle for CRF design

4. Rules for collecting categorical data

- Numerical codes should be assigned for all possible categories.

What is your sex?
Male……..1
Female……2

Principle for CRF design

4. Rules for collecting categorical data

Underlying disease

- DM	1. yes	2. no
- HT	1. yes	2. no
- Stroke	1. yes	2. no
- CVD	1. yes	2. no

Principles of CRF design

5. Rules for dealing with missing data

- Assign special codes for missing values at the *data collection time*.

- It is bad practice to leave data collection field blank on the questionnaire because it can lead to confusion at data entry time.

- These special codes for missing values must be removed before statistical analysis.

Principles of CRF design

5. Rules for dealing with missing data

- The missing data codes should not be possible valid values.

- It is common practice to use 9, 99, 999 and so on to denote missing data.

Principles of CRF design

5. Rules for dealing with missing data

Example:

Age _ _ _ year (missing=999)

Height _ _ _ . _ _ cm (missing=999.99)

Sex

1. male 2. female 9. missing

Stage of cancer

1. I 2. II 3. III 9. missing

Principles of CRF design

6. Rules for collecting date

When dates are recorded, it is important to clearly identify the format to be used, for example, dates can be recorded as:

- Day, Month, Year (dd/mm/yyyy).
- Month, Day, Year (mm/dd/yyyy).

Principles of CRF design

6. Rules for collecting date

When years are recorded, it is important to clearly identify the standard to be used, for example, Western or Buddist standard:

- Western standard (dd/mm/20yy).
- Buddist standard (dd/mm/25yy)

Recommendations

- CRF should be designed along with the protocol to ensure collection of only the data the protocol specifies.

-It is better to have more forms, each with a small amount of data.

Recommendations

- The CRF should be simple and clear, and there should be no doubt how to fill it in.
- Collecting data without the CRFs is likely to result in incomplete and invalid data.

Example of weak CRF design

1. Have you ever been diagnosed with DM?
1. Yes 2. No 9. Missing
For female: if yes, answer the following questions
2. Did you have DM before pregnancy?
1. Yes 2. No 9. Missing
3. Did you have DM during pregnancy?
1. Yes 2. No 9. Missing
4. Have you ever taken drug for DM?
1. Yes 2. No 9. Missing

Example of strong CRF design

1. Have you ever been diagnosed with DM?
1. Yes 2. No 9. Missing
if yes, answer the question number 2.
2. Have you ever taken drug for DM?
1. Yes 2. No 9. Missing
If you are female, and have been pregnant, answer the questions number 3 and 4, otherwise go to question number 5.
3. Did you have DM before pregnancy?
1. Yes 2. No 9. Missing
4. Did you have DM during pregnancy?
1. Yes 2. No 9. Missing

Example of weak CRF design

Have you ever taken medications for osteoporosis?		
Calcium	☐	Start date _ _/_ _/_ _ _ _
Vitamin D	☐	Start date _ _/_ _/_ _ _ _
Calcitonin	☐	Start date _ _/_ _/_ _ _ _
Hormone	☐	Start date _ _/_ _/_ _ _ _

Example of strong CRF design

Have you ever taken medications for osteoporosis?
Calcium 1. Yes 2. No 9. Missing
If yes, specify start date _ _/_ _/25 _ _
Vitamin D 1. Yes 2. No 9. Missing
If yes, specify start date _ _/_ _/25 _ _
Calcitonin 1. Yes 2. No 9. Missing
If yes, specify start date _ _/_ _/25 _ _
Hormone 1. Yes 2. No 9. Missing
If yes, specify start date _ _/_ _/25 _ _

Data management processes

- Case report form (CRF) design ?
- Data collection
- Database management
- Data entry
- Data cleaning and checking

Database management

DBMS software package

There are many different DBMS software packages:

- dBase
- Paradox
- *EpiData*
- Access
- and so on

Introduction to EpiData

- EpiData is a program for data entry and documentation of data only.
- However, you can export the data to a number of data formats for statistical analysis.

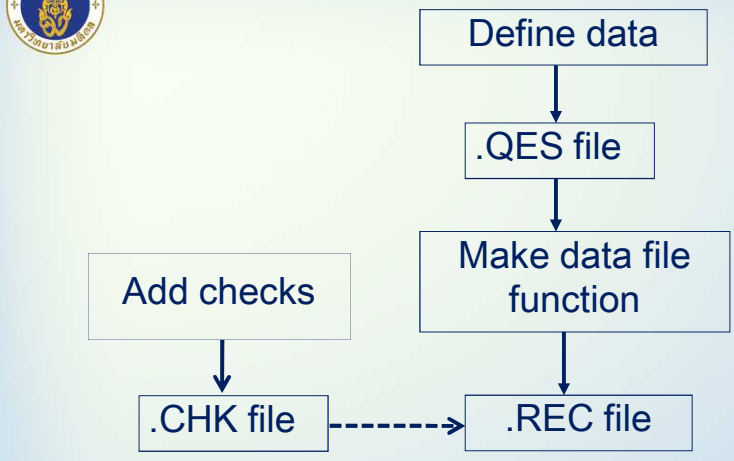
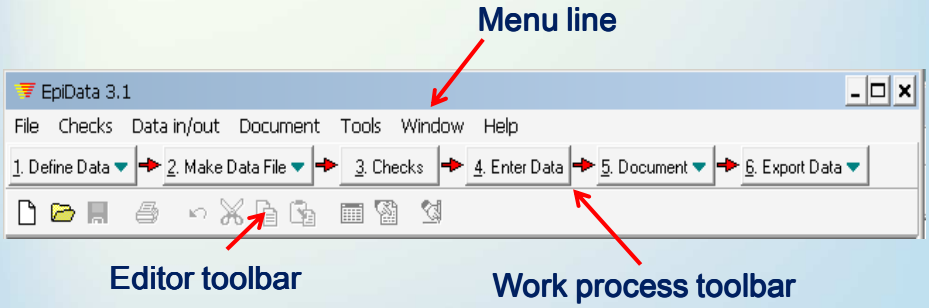


Figure 1 Flow charts of creating a database file in EpiData

Overview of EpiData

The EpiData screen has a standard windows layout with one menu line and two toolbars.



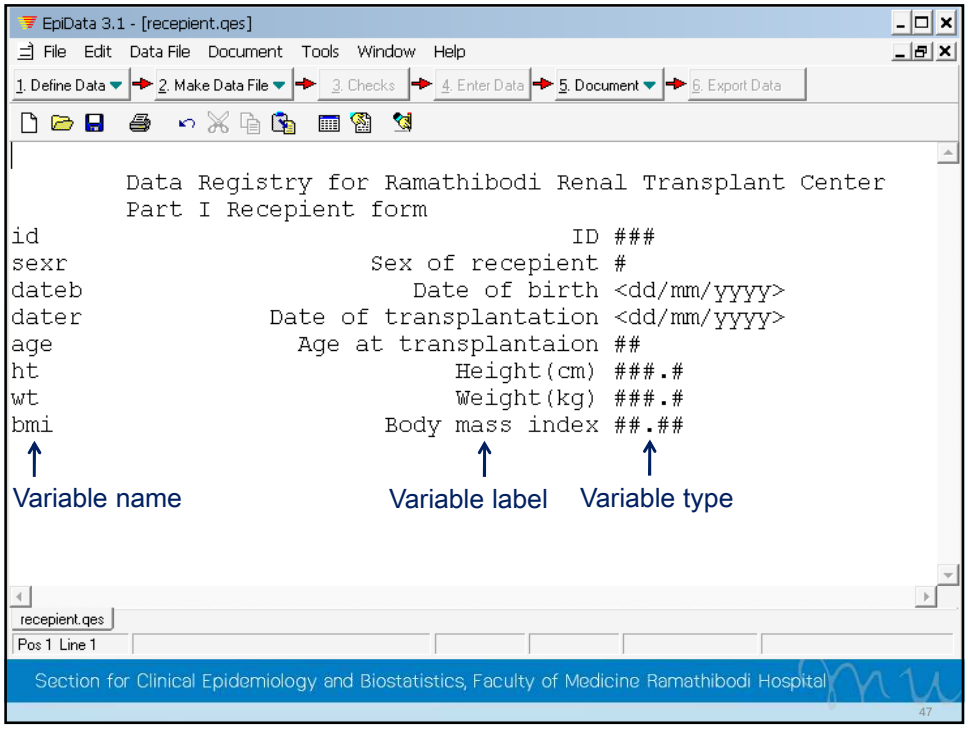
1. Define data

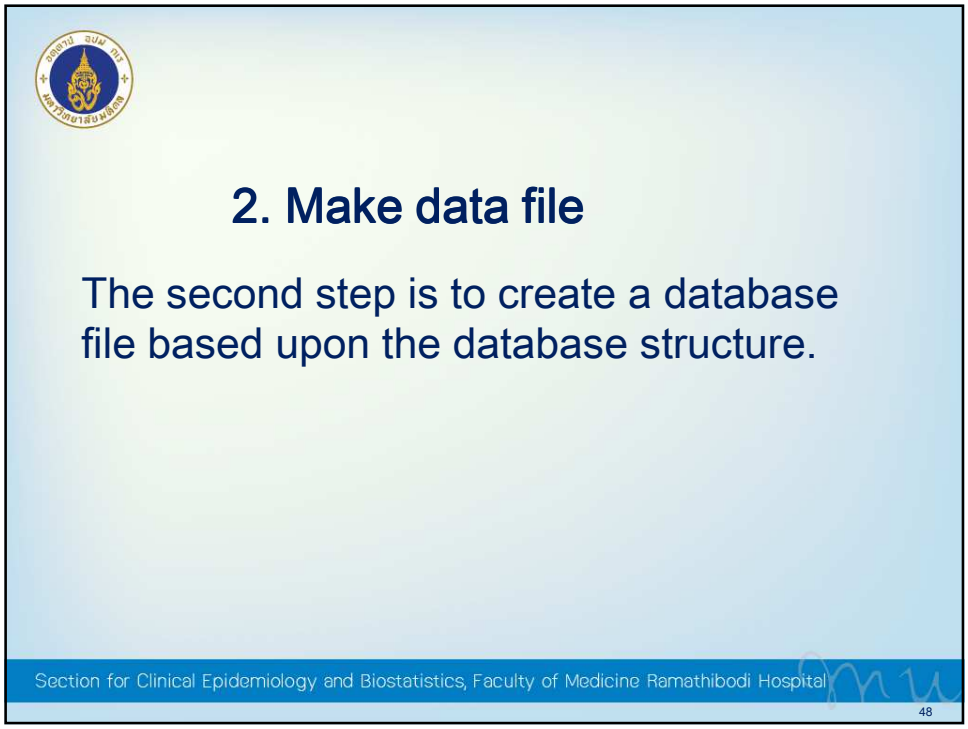
The first step is to define the structure of a database file by writing three types of information for each variable

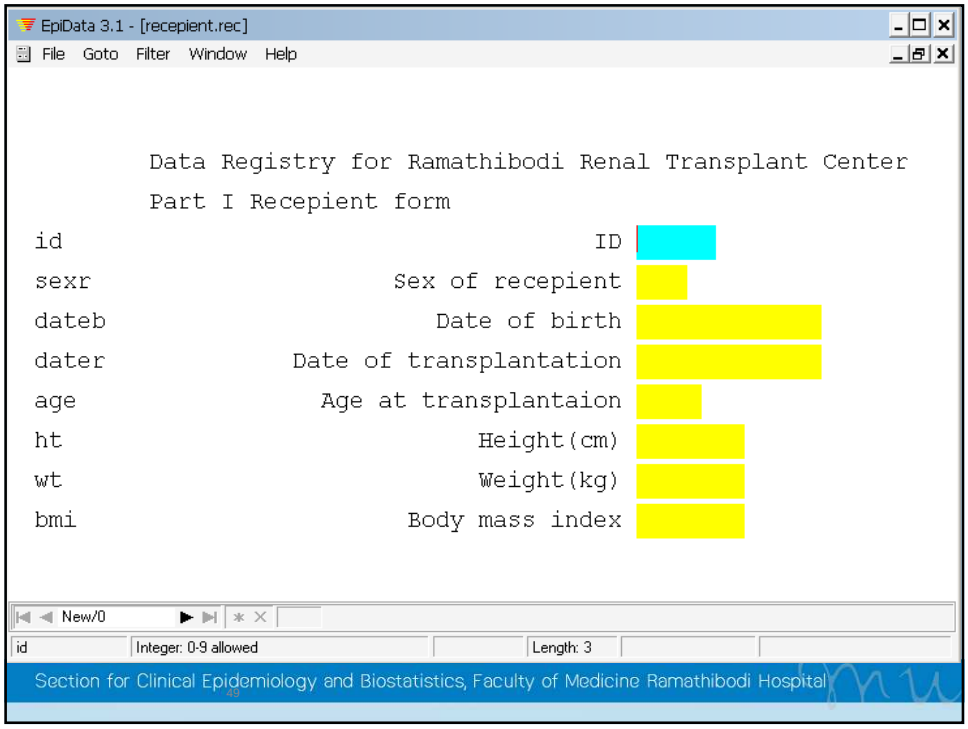
a) Variable name

b) Variable label

c) Variable type


EpiData 3.1 - [recepient.qes]
File Edit Data File Document Tools Window Help
1. Define Data
2. Make Data File
3. Checks
4. Enter Data
5. Document
6. Export Data
Data Registry for Ramathibodi Renal Transplant Center
Part I Recepient form
id ID ###
sexr Sex of recepient #
dateb Date of birth <dd/mm/yyyy>
dater Date of transplantation <dd/mm/yyyy>
age Age at transplantaion ##
ht Height(cm) ###.#
wt Weight(kg) ###.#
bmi Body mass index ##.##
Variable name
Variable label
Variable type
recepient.qes
Pos 1 Line 1
Section for Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital
47


2. Make data file
The second step is to create a database file based upon the database structure.
Section for Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital
48



3. Add/Revise Checks

- The third step is to specify edit checks and calculations during data entry.
- The major role of this step is to reduce errors during data entry.

3. Add/Revise Checks

The basic checks in EpiData consists of these following functions:

- Range/Legal
- Add value label to variable
- Jumps
- Must enter
- Repeat

Legal values

- You can specify the legal values for a certain numeric variable.
- The input must match one of a specified list of values.

Range checks

- You can specify that a certain variable must lie between two limited values.
- You may mix range checks and legal values checks (e.g. for missing value codes).

Attach value labels to variables

- You can specify a valid value label set for categorical variables such as sex, blood group.
- When you specify a value label set, you specify both the legal values for a variable and the meaning of each of the legal values.

Conditional jumps

- You can specify a value for a variable that will cause the entry to jump to a target variable.
- If the tests fail the entry moves to the next variable.

Must-enter variables

- You can specify that a certain variable must be filled with a value other than leave it blank.
- The variable can be left blank unless it is defined as a must-enter variable.

Repeat variables

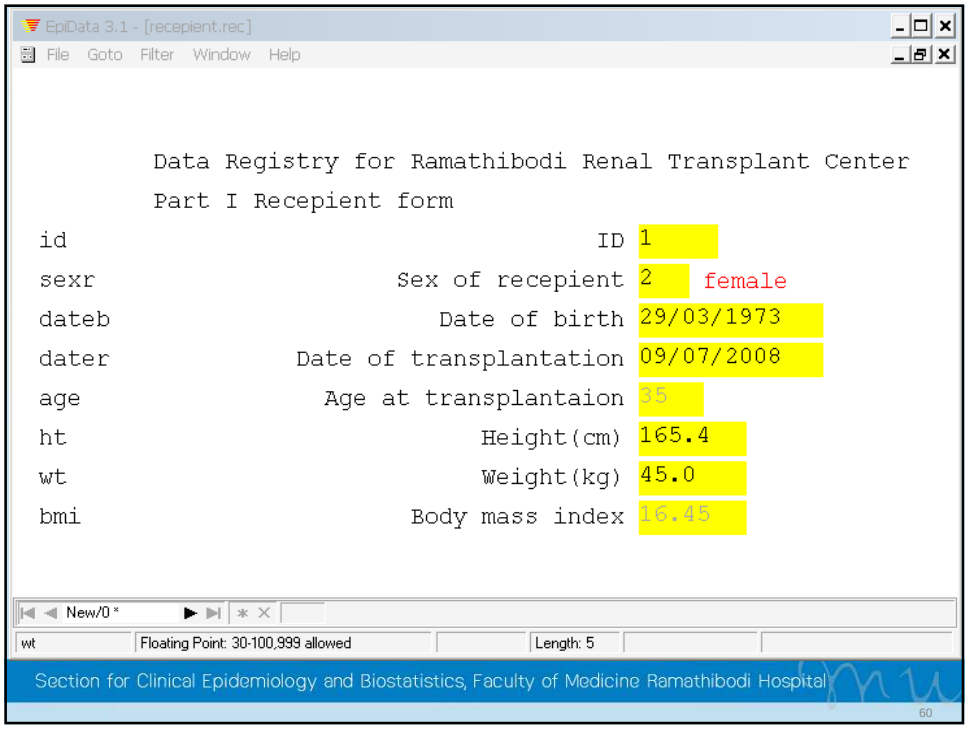
- You can specify that a certain variable on a new record will automatically keep the value from the previous case.
- This is useful for data that seldom changes.

Examples of edit checks

- Specify sequence of data entry, e.g. fill out certain questions about menstruation for females only.
- Restrict data entry to possible values, e.g. the SBP of patients must lie between 70 to 200 mmHg.

Examples of calculations

- Calculate age at visit based on date of visit and date of birth.
- Calculation BMI based on weight and height.

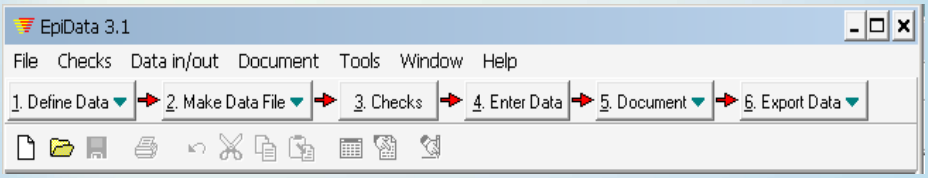


Summary of the work process with EpiData

Define data -> Questionnaire file (.QES)

Make data file -> Record file (.REC)

Add/revise checks -> Check file (.CHK)



Before data entry, you can document file structure

```
DATA FILE:              C:\data\recipient.rec
File label:             recipient form

File size:              9620 bytes
Last revision:          18. ส.ค. 2010 13:25
Number of fields:       9
Number of records:      352
Checks applied:         Yes (Last revision 18. ส.ค. 2010 13:24)


Fields in data file:

No. Name        Variable label          Field type      Width  Checks          Value labels
-------------------------------------------------------------------------------------------
  1 id          ID                      Number          3

  2 sexr        Sex of recipient        Number          1                      label_sexr
                                                                               1: male
                                                                               2: female
                                                                               9: missing

  3 datebirt    Date of birth           Date (dmy)      10

  4 ager        Age of recipient        Number          3
```

After data entry, you can list values for some or all records.

```
Observation 1  (record # 1)

        id              348          sexr          female       datebirt       20/04/1943
      ager               59      bloodgrr               0          hbsr         negative
      hber          missing        athbsr        positive        athcvr         negative


Observation 2  (record # 2)

        id              154          sexr            male       datebirt       01/07/1931
      ager               65      bloodgrr              AB          hbsr         negative
      hber          missing        athbsr        negative        athcvr         negative
```


After data entry, you can summary the data

```
id ------------------------------------------------------------------------------ ID
                      type:  Number

                   missing:  0/352
                     range:  [1 ; 384]
             unique values:  352

                      mean:  196.8438
                 std. dev.:  112.8792

sexr ---------------------------------------------------------- Sex of recipient
                      type:  Number
              value labels:  label_sexr

                   missing:  0/352
                     range:  [1 ; 2]
             unique values:  2

                tabulation:   Freq.    Pct.  Value   Label
                               217    61.6       1  male
                               135    38.4       2  female
```

Export data

-The last step is to export data into a format that can be read into your preferred statistical analysis package.

- EpiData provides a set of functions which allows you to export data in a variety of common file formats such as text, dBase, Excel, Stata, SPSS, etc.

Thank you for your attention