FACILITATOR GUIDE

Managing Data

Created: 2013

Managing Data. Atlanta, GA: Centers for Disease Control and Prevention (CDC), 2013.

Managing Data

Table Of Contents

Introduction

LEARNING OBJECTIVES

At the end of the training, participants will be able to:

1. Create a data dictionary that includes, at a minimum:
    a. Variable names
    b. Variable descriptions or labels
    c. Variable types
    d. Response options and allowable values

2. Clean the data
    a. Identify errors, including duplications, missing data, miscodes, and outliers
    b. Use statistical software to correct the errors

ESTIMATED COMPLETION TIME

The workbook should take between 6 and 7 hours to complete.

TARGET AUDIENCE

The workbook is designed for FETP fellows who specialize in NCDs; however, participants can also complete the module if they are working in infectious disease.

PREWORK AND PREREQUISITES

Before participating in this training module, participants must complete training in:

- Basic epidemiology and surveillance
- Statistical software program your FETP is using (e.g., SPSS, Epi Info)

OPTIONS FOR FACILITATING THIS TRAINING

There are two options for facilitating this training:

1. Individual mentor-directed: A mentor helps the participant complete the training. The mentor's main responsibility will be to review the participant's work and provide feedback.

    A mentor will meet with the participant a minimum of two times. At the first meeting, the mentor should orient the participant to the training, provide examples and direction indicated, answer questions, and set future modes of contact and meeting time(s). Very small groups (less

than 5 individuals) may choose to work on the training together and find individual or collective mentor(s).

2. Classroom: There are two options for classroom training. For option a, participants read the training material *prior* to attending class and then review what they read in class. For option b, participants read the training material *during* class.

    a. **Participants read training material *prior* to attending class**. At the start of each module section, the facilitator reviews key points. (**Note:** The facilitator may prepare PowerPoint slides for a brief presentation of key points, lead an informal discussion about the reading, or ask participants to answer questions individually or in small groups about what they read. Appendix A contains sample questions.) After each review, participants will complete practice exercises and skill assessments as directed.

    b. **Participants read training material during class:** The facilitator directs students to read the training material and complete the exercises as indicated in the workbook. The facilitator leads group discussions to review what participants have read and reviews participants' answers to the exercises and skill assessments.

MATERIALS AND EQUIPMENT

For the Facilitator or Mentor:

- Facilitator/Mentor Guide
- Computer and projection screen to show either screen shots or the statistical software program
- Optional Powerpoint slides for:
    - Introduction to the module
    - Discussions on how to clean datasets and working with datasets that have multiple errors

For the Participant:

- Participant Workbook
- Flip chart and markers
- Dataset and questionnaire for hypertension case study
- Dirty dataset and (sampling of) questionnaires from their own country's NCD study
- Laptops with statistical software (e.g., SPSS)

CONFIGURATION OF THE TRAINING ROOM

If this training will be implemented in a facilitator-led setting, please note the following recommendations:

1. Use a room large enough to host breakout groups of 6–8 participants.
2. Each breakout group should have one rectangular or round table for completing small group work.
3. An ideal training room will have enough space between tables to have flip charts for each group and enough space between tables so that groups will not be too distracted by each other.

ICON GLOSSARY

The following icons are used in this guide:

Image Type	Image Meaning
	**Book** – participants read a section in the participant workbook
	**Activity** – an activity, exercise, assessment or case study that participants complete
	**Group** – a group discussion that you will lead, either to review key points or answers to an activity
	**Flipchart –** write responses during facilitator-led discussions or debriefs
	**Prepare** – an activity in the module for which you need to prepare (e.g., making handouts of a report, identifying a local example)

PREPARATION

Be sure to review this facilitator guide and read the descriptions of how to prepare for the following discussions and activities:

- Sample questionnaires for the NCD study in their country and a partially completed data dictionary

- Overview of data errors and how to detect/correct duplicate records
- Detecting and correcting missing, miscoded, and out-of-range values
- Preparing a dirty dataset with multiple errors
- Questionnaires and a "dirty" dataset (from your country) with the following errors:
  - One or two duplicate records
  - One or two miscodes
  - One or two missing values
  - One or two outliers

As noted in the "Materials and Equipment" section above, you may prepare slides for these discussions/presentations.

ACKNOWLEDGEMENTS

Many thanks to the following people from the Centers for Disease Control and Prevention (CDC) who contributed to this module:

- Fleetwood Loustalot, PhD, FNP, Andrea Neiman, MPH, PhD (Division for Heart Disease and Stroke Prevention), and Edward Gregg, PhD (Division of Diabetes Translation), for creating the hypertension case study.

- Indu Ahluwalia, (Senior Scientist, Division of Reproductive Health, National Centers for Chronic Disease Prevention and Health Promotion), and Richard Dicker, MD, MS, from the Centers for Global Health, Division of Global Health Protection for their subject matter expertise and for reviewing the training module.

Some of the content of this module was taken from a training manual developed by CDC's Division of Epidemiology and Surveillance Capacity Development: *Advanced Management and Analysis of Data Using Epi Info for Windows: Risk Factors for Sexually Transmitted Infections in Kuwadzana, Zimbabwe; 2006.*

How to Facilitate This Module

FACILITATOR/MENTOR RESPONSIBILITIES

This training module is self-paced. Participants learn the content by reading their workbook and participating in group discussions. They apply what they learn by completing practice exercises and skill assessments. Participants use a hypertension case study for the practice exercises; for the skill assessments, they use information about an NCD study from their own country.

As a *facilitator*, you will *facilitate* or assist in the participants' learning. Your main roles will be as follows:

- **Introduce** the module topic.
- **Lead** group discussions to review or elaborate on what participants read.
- **Answe**r questions that participants may have during the training.
- **Review** participants' work and provide feedback.
- **Be a timekeeper**, ensuring participants stay within a general schedule.

As a *mentor,* you will perform the same tasks and play a more active role in supporting the learner *after* the training with his or her field work.

Overview of Modules

SECTIONS 1 AND 2: INTRODUCTION AND OVERVIEW OF DATA MANAGEMENT

Total Estimated Time: 25 minutes

Introduction and brief overview: 10 minutes

Readings: up to 15 minutes

Duration/Session Type	What to Do/What to Say
 **Group Discussion**	**Introduction and Brief Overview** • Introduce yourself if you are a new facilitator. • Briefly explain the two components this module will cover: 1) creating a data dictionary, and 2) cleaning data. • Distribute all the materials participants will need to complete the

Duration/Session Type	What to Do/What to Say
 **Clipboard** **10 Minutes**	Practice Exercises (e.g., CD-ROM, questionnaires).
 **Reading** **15 Minutes**	**Readings** Participants read the module introduction and overview of data management. Skip this step if participants have read the material prior to class.

SECTION 3: DATA DICTIONARY

Total estimated time: 110 – 120 minutes
Readings: up to 10 minutes
Group discussion: 10minutes
Practice Exercise #1: 30 minutes
Debrief Practice Exercise #1: 10 minutes
Skill Assessment #1: 40 minutes
Debrief Skill Assessment #1: 20 minutes

Duration/Session Type	What to Do/What to Say
 **Reading** **10 Minutes**	**Readings** Participants read 2 pages about data dictionaries (section 3). Skip this step if participants have read the material prior to class.
 **Group Discussion** **10 Minutes**	**Group Discussion** • You may use the sample questions in Appendix A as a guideline for the discussion. Participants can answer questions orally or you can provide them with the written questions and ask them to record their answers individually or in a group. Participants can use the "Key Points to Remember" section in their participant workbook to record notes or answers.

<table>
<tr><th>Duration/Session Type</th><th>What to Do/What to Say</th></tr>
<tr><td></td><td>

- Provide any specific instructions participants will need to know about using their statistical software to create or locate the data dictionary.

</td></tr>
<tr><td> 
Clipboard

 
Activity

30 Minutes</td><td>

Practice Exercise #1

- Divide participants into small groups or pairs.
- Distribute the dataset and questionnaire and measurements handout for the hypertension case study.
- Ask them to spend approximately 30 minutes completing the exercise. The first 12 lines of the data dictionary have been provided to the participants.
- Make sure you are available to answer any questions during the exercise.
- Provide participants with a 5-minute warning.

</td></tr>
<tr><td> 
Reading

10 Minutes</td><td>

Debrief Practice Exercise #1

- Ask each group to provide one line of data from the data dictionary, and answer any questions about the practice exercise.
- Possible answers are in the table below.

<table>
<tr><th>#’s</th><th>Variable Name</th><th>Type</th><th>Label</th><th>Value</th><th>Measure</th></tr>
<tr><td>1-12</td><td>SEQN</td><td>Numeric</td><td>ID number</td><td></td><td>Scale</td></tr>
<tr><td>13</td><td>HTN</td><td>Numeric</td><td>BP >=140/90 or taking blood pressure medication</td><td>1 = Yes
2 = No
6 = Missing</td><td>Nominal</td></tr>
</table>

</td></tr>
</table>

Duration/Session Type	What to Do/What to Say					
	14	BP3C	Numeric	Blood pressure classification in 3 categories	1 = Normotensive 2 = Prehypertension 3 = Hypertension 4 = Missing	Nominal
	15	WT_KG	Numeric	Weight in kilograms	6666 = Missing 9999 = Refused	Scale
	16	HT_CM	Numeric	Height in centimeters	6666 = Missing 9999 = Refused	Scale
	17	BMI	Numeric	Body Mass Index	6666 = Missing	Scale
	18	BMI3C	Numeric	Body Mass Index 3 categories	1 = Underweight/Normal 2 = Overweight 3 = Obese 4 = Missing	Nominal
 **Clipboard**	**Skill Assessment #1** • Divide participants into small groups. If they are from the same country, then you may keep them in the same groups as the Practice Exercise. Or, if participants have come to class without an NCD study, pair them with a participant who does have a study.					

Duration/Session Type	What to Do/What to Say
 **Activity** **40 Minutes**	• Distribute the sample questionnaires for the NCD study in their country and a partially completed data dictionary. (Ask them to add approximately **5** variables to the existing data dictionary.) • Ask participants to turn to the appropriate page in their Activity Workbook. (Participants can record their answers in the space provided in the workbook or they can enter the variables directly onto the data dictionary.) • Ask them to spend approximately **40 minutes** completing the exercise. • Make sure you are available to **answer** any questions during the exercise. • Provide participants with a 5-minute warning.
 **Group Discussion** **20 Minutes**	**Debrief** Review participants' answers by going around the room and asking each group or pair to provide you with one or two lines from the data dictionary.

SECTION 4: CLEANING DATA

Total Estimated Time: 3 ½ – 4 ½ hours

Part 1 Total estimated time: 55 – 80 minutes

Readings: up to 20 minutes
Group discussion: 15–20 minutes
Practice Exercise #2: 30 minutes
Debrief Exercise #2: 10 minutes

Part 2 Total estimated time: 2 ½ – 3 hours

Readings: up to 30 minutes
Group discussion: 25 – 30 minutes
Practice Exercise #3: 45 minutes
Debrief Exercise #3: 15 minutes
Skill Assessment #2: 45 minutes
Debrief Skill Assessment #2: 15 minutes

Duration/Session Type	What to Do/What to Say
 **Reading** **20 Minutes**	**Readings** Tell participants to read 4 pages about an overview of data errors and how to detect/correct duplicate records. Skip this step if participants have read the material prior to class.
 **Group Discussion**  **Clipboard** **15-20 Minutes**	**Group Discussion** • You may use the sample questions in Appendix A as a guideline for the discussion. Participants can answer questions orally or you can provide them with the written questions and ask them to record their answers individually or in a group. Participants can use the "Key Points to Remember" section in their participant workbook to record notes or answers. • Provide any specific instructions participants will need to know about using their statistical software to identify and correct duplicate records. • The following explains how to check and identify duplicate records from *IBM SPSS Statistics 19*: o Select Data > Identify Duplicate Cases. o Define matching cases by: ID number (primary ID value). o Keep default settings: Moves duplicate to top of dataset and identifies duplicates with new variable (Labeled:

<table>
<tr><th>Duration/Session Type</th><th>What to Do/What to Say</th></tr>
<tr><td></td><td>PrimaryLast) as 0.
o Review Data View and check for duplicate records (by Primary ID value).
o Delete duplicate records and record corrections.</td></tr>
<tr><td>Activity

30 Minutes</td><td>Practice Exercise #2
Keep participants in the same small groups or pairs.
Ask them to spend approximately 30 minutes completing the exercise.
Make sure you are available to answer any questions during the exercise.
Provide participants with a 5-minute warning.</td></tr>
<tr><td>Group Discussion

10 Minutes</td><td>Debrief Exercise #2
Review participants’ answers by going around the room and asking each group or pair what duplicate records they found and how they corrected them.
Possible answers include:
<table>
<tr><th>Number</th><th>Primary ID</th><th>Secondary ID</th><th>Problem</th><th>Record ID's Affected and Resolution</th></tr>
<tr><td>1</td><td>SEQN</td><td></td><td>Duplicate</td><td>51657 - Removed</td></tr>
<tr><td>2</td><td>SEQN</td><td></td><td>Duplicate</td><td>51830 - Removed</td></tr>
<tr><td>3</td><td>SEQN</td><td></td><td>Duplicate</td><td>52152 - Removed</td></tr>
<tr><td>4</td><td>SEQN</td><td></td><td>Duplicate</td><td>52566 - Removed</td></tr>
<tr><td>5</td><td>SEQN</td><td></td><td>Duplicate</td><td>53117 - Removed</td></tr>
</table></td></tr>
</table>

Duration/Session Type	What to Do/What to Say
 **Reading** **30 Minutes**	**Readings** Tell participants to read 6 pages about detecting and correcting missing, miscoded, and out-of-range values. Skip this step if participants have read the material before class.
 **Group Discussion** **25-30 Minutes**	**Group Discussion** • You may use the sample questions in Appendix A as a guideline for the discussion. Participants can answer questions orally or you can provide them with the written questions and ask them to record their answers individually or in a group. Participants can use the "Key Points to Remember" section in their participant workbook to record notes or answers. • Provide any specific instructions participants will need to know about using their statistical software to identify and correct missing, miscoded, and out-of-range values. • The following explains commands for SPSS: 1. Check for Missing Data o Analyze > Descriptive Statistics > Frequencies. o Select variables to review. o Output includes Valid and Missing Data and a Frequency Table. o Frequency Table can be reviewed for outliers. 2. Addressing Missing Data using in the Variable View window to code Missing Data o In the Variable View window, use the Missing Column to identify missing data. o Use '999' or similar for missing data or 'NR' for string variables. 3. After adding information to the Missing Column, assign values to the missing data codes. Use SPSS to Identify Missing Values (potentially more advanced) o Analyze > Missing Value Analysis. o Select quantitative variables and categorical variables. o 'Click' Demographics and select t-test and cross-

Duration/Session Type	What to Do/What to Say
	tabulations. o Run the analysis. o SPSS Tutorial "Describing the Patterns of Missing Data" can be reviewed for additional instruction. 4. Miscoded variables o Review frequency table and compare with questionnaire results. 5. Out-of-range variables o Review frequency table and compare with questionnaire results. o Go to Graphs > Chart Builder > Scatter/Dot. o Double click on scatterplot to review. o Right click > Go to case (to review outliers). 6. Logic checks o Use Frequency command or Crosstabs. o Analyze > Descriptive Statistics > Crosstabs or Frequencies.
 **Activity** **45 Minutes**	**Practice Exercise #3** • Keep participants in the same small groups or pairs. • Ask them to spend approximately **45 minutes** completing the exercise. • Make sure you are available to answer any questions during the exercise. • Provide participants with a 5-minute warning.
 **Group Discussion** **15 Minutes**	**Debrief Practice Exercise #3** • Review participants' answers by going around the room and asking each group or pair what duplicate records they found and how they corrected them. • Possible answers include:

Duration/Session Type	What to Do/What to Say

Number	Variable name	Format	Problem	Record ID Affected and Resolution
1	HCP	Numeric	Large amount of missing data	Variable deleted from analytic dataset
2	SEX	Numeric	Miscoded (response = 3)	51929 – changed to missing
3	SEX	Numeric	Miscoded (response = 3)	51930 – changed to missing
4	AGE	Numeric	Out of range (response = 120)	52583 – changed to missing
5	AGE4	Numeric	Derived variable of age (yrs), changed due to out of range value	52583 – changed to missing
6	BMI	Numeric	Out of range value (response 182)	52658 – corrected using BMI calculation (BMI = 30)
7	BMI3C	Numeric	Derived variable of BMI, changed due to out of range value	52658 – corrected based on new BMI calculation (category = 2)

<table>
<tr><th>Duration/Session Type</th><th>What to Do/What to Say</th></tr>
<tr><td></td><td>
<table>
<tr><td>8</td><td>BPSYS, BPDIA</td><td>Numeric</td><td>Missing all blood pressure measurements</td><td>52814 – Delete case from analytic dataset, or label as missing</td></tr>
<tr><td>9</td><td>BPHI, HTN</td><td>Numeric</td><td>Incorrect coding – no blood pressure measurements for classification</td><td>52814 – Delete case from analytic dataset, or label as missing</td></tr>
</table>
</td></tr>
<tr><td> 
Clipboard

 
Activity

45 Minutes</td><td>Skill Assessment #2
Provide participants with questionnaires and a “dirty” dataset with the following errors:
One or two duplicate records
One or two miscodes
One or two missing values
One or two outliers
Tell them to open their Activity Workbooks and complete Skill Assessment #2. Participants have 45 minutes to complete the skill assessment.
After 40 minutes, provide them with a 5-minute warning.</td></tr>
<tr><td> 
Group Discussion

15 Minutes</td><td>Debrief over Skill Assessment #2
Review participants’ responses by asking participants to share one of the errors they found and explain how they would resolve it.</td></tr>
</table>

CONCLUSION

Total Estimated Time: 15 minutes

Duration/Session Type	What to Do/What to Say
 **Flip Chart**  **Group Discussion** **15 Minutes**	**Conclusion** • Ask participants for some main points they learned in the module. • Ask participants for their reactions to what they learned in the training and how they will apply the skills when they return to their job.

Appendix

Sample Review Questions for Section 3: Data Dictionary

1. What are the four types of information a data dictionary should include at a minimum?

   Possible answers:
   1. *Variable names*
   2. *Variable descriptions or labels*
   3. *Field type*
   4. *Response options and codes used to represent options.*

2. Name at least five features of a variable name:

   Possible answers:
   - *Easily identifies the question on the data collection form (if one is used) or type of information collected.*
   - *Begins with a letter.*
   - *Cannot end with a period.*
   - *Can have special symbols or characters.*
   - *Should be short with a maximum length of 64 characters.*
   - *Limits the use of symbols.*

3. What is an example of a variable name and variable description or label?

   Possible answer:
   DOB for "date of birth"
   EDU for "educational level"

4. What would you record for response options (or values) of open-ended text fields?

   Possible answer: Indicate that the response can contain up to a certain number of characters; for example: {up to 60 characters of text}.

5. What is an example of response options for marital status?

   Possible answer:
   1 = single

2= married
3 = divorced
4 = widowed
99 = missing

Sample Review Questions for Section 4: Cleaning Data – Overview through Practice #2

1. What is a key principle of data management?

   Possible answer: To document everything.

2. True or False: You should fix an error as soon as you become aware of it.
   Answer: True

3. What should you do with the original dataset before you make edits that permanently change it?
   Answer: Make a working copy

4. What are the most common types of data errors?
   Possible answers: Entering duplicate information, miscoding, assignment of missing values, and inclusion of out-of-range values.

5. What steps should you take to check for duplicate records?
   Possible answers:
   1. *Use your statistical software to check the record count.*
   2. *Determine if the record count equals the number of questionnaires.*
   3. *If the record count is greater than the number of questionnaires, run a* ***frequency*** *listing to look for multiple records with the same identifying information.*
   4. *If there are two records with the same ID number of name, select the records and examine them to determine if they are identical or whether an ID number of name was entered incorrectly.*

Sample Review Questions for Section 4: Cleaning Data – Detecting and Correcting Missing, Miscoded, and Out-of-Range Values

1. What is the first thing you should do to begin detecting errors?

   Possible answer: Conduct a frequency distribution

2. What should be the first step in correcting records with missing data?
   Answer: Find the questionnaires that are missing the value and determine if the data were missing from the questionnaire or were not entered.

3. What should you do if missing data are <u>not</u> caused by data entry errors?
   Possible answers:
   - *Correct the error by contacting the study participants, if possible.*
   - *Delete subjects with missing data from the analysis (if missing data is less than 10%).*
   - *Omit record from analysis of that particular variable.*
   - *Delete variables that have large amount of missing data.*
   - *Use imputation techniques. (Consult a statistician first.)*

4. How should you correct miscodes?
   Possible answer: Look at original data source (e.g., questionnaire) to determine true value and make the correction in the database.

5. How can you correct a miscode if you do not have access to the original data sources such as the questionnaire?
   Possible answer: Recontact the subject to confirm the information, if feasible. Or record the value as missing

6. What should you always do AFTER you make any change to the database?
   Answer: Document the changes made.

7. What is something you can make to visually compare two variables in order to check for out-of-range values?
   Answer: A scatterplot.

8. What should you do when you cannot resolve an outlier by looking at the questionnaire?
   Possible answer: You should decide whether to verify the data or leave it as entered. For a small dataset and a key variable you should verify the data.

9. What are some types of logic checks you may conduct?
   Possible answer: Looking for impossibilities; looking for inconsistencies; ensuring that skip patterns have been followed.