WORKSHOP #5: DATA MANAGEMENT FOR DIETARY INTERVENTIONS

Biling Hong and Donna Spiegelman
Harvard School of Public Health
<table>
<thead>
<tr>
<th>GNET Study</th>
<th>Software used in data entry</th>
<th>Why chose this software?</th>
<th>What software would be recommended?</th>
</tr>
</thead>
</table>
| Shanghai, China     | EpiData                    | - Small, simple, and free software specially designed for data entry  
- It can run without installation by simple copying  
- It can make flexible change in questionnaire style  
- It has error detection features (such as double entry verification), list of ID numbers in several files, codebook overview of data, date added to backup and encryption procedures.  
- It’s more suitable to be applied in large data management | Epidata                            |
| Chennai, India      | MS Access (before)         | - Does not get corrupted/lost and user friendly to enter.                                                                                                                                                                | Online data entry method with username and password protected for each site and can be monitored centrally |
|                     | ASP.net for entry and back end SQL (now)  
EpiNu (for all dietary information like 24 hr recall or FFQ) |                                                                                                                                                                                                                      |                                    |
### Database used in the current GNET groups

<table>
<thead>
<tr>
<th>GNET Study</th>
<th>Software used in data entry</th>
<th>Why chose this software?</th>
<th>What software would be recommended?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dar es Salaam, Tanzania</td>
<td>MS Excel</td>
<td>- Easy to use and its compatibility with other statistical software</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- But it does not give much room to play around with the data</td>
<td></td>
</tr>
<tr>
<td>Abuja, Nigeria</td>
<td>EpiData (for Fufu study)</td>
<td>- EpiData was to allow for double data entry</td>
<td></td>
</tr>
<tr>
<td>Mombasa, Kenya</td>
<td>SPSS or MS Excel</td>
<td>- Because they had used SPSS several times in the past.</td>
<td></td>
</tr>
<tr>
<td>Kuwait City, Kuwait</td>
<td>Not data collected yet. REDCap (used in the research center in Kuwait)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mexico City, Mexico</td>
<td>Not data collected yet. Access (may use it when collecting the data)</td>
<td>- The choice has to do with the infrastructure and expertise being already in place.</td>
<td>Access</td>
</tr>
<tr>
<td>Puerto Rico</td>
<td>Access</td>
<td></td>
<td>Access</td>
</tr>
<tr>
<td>San Jose, Costa Rica</td>
<td>Excel</td>
<td>- Excel is very friendly and allows export data to other software for further analysis</td>
<td>Excel</td>
</tr>
</tbody>
</table>
- **EpiData** was initially released in 2000 by a non-profit organization in Denmark (http://www.epidata.dk/).
- This software was developed specifically to translate **Epi Info DOS** into Windows software.

**Background information for Epi Info –**

- **Epi Info DOS** was created by the CDC in 1985 which were designed to support field epidemiological investigations of outbreaks. In 2001, the CDC released the final version of Epi Info DOS (Epi Info 6.04d).
- Epi Info 2000 (Epi Info 3) for windows was released in 2000.
- Epi Info allows the user to
  - develop a questionnaire
  - customize the data entry process
  - enter data and analyze the data
  - produce statistics, graphs, tables, and maps by using simple commands
- EpiData is **free** and widely used by organizations and individuals to create and analyze large amounts of data.
- The WHO uses EpiData in its STEPS method of collecting epidemiological, medical, and public health data, for biostatistics, and for other quantitative-based projects.
- Data can be exported to Stata, SPSS & SAS with labels and missing value definitions, CSV and can be imported from CSV, Stata with labels and missing value definitions.
- The installation can be as simple as copying the program files. For example, it can be run from a USB drive (memory stick) and is small (<2.5MB).
“controlled data entry” - EpiData will only allow the user to enter data which meets certain criteria. For example:

- specified legal values with attached text labels (1 = No 2 = Yes)
- range check (only ages 20-100 allowed)
- legal values (e.g. 1, 2, 3 and 9)
- legal dates (e.g. 29febr1999 is not accepted)

EpiData has two parts:

- **EpiData Entry**: double entry verification, list of ID numbers in several files, codebook overview of data, date added to backup and encryption procedures.

- **EpiData Analysis**: basic statistical analysis, graphs, and comprehensive data management (recoding data, label values and variables, defining missing values).
Microsoft Excel

- Even though Excel is not a database, it is widely used to store data, and it is often used to solve simple database problems. But, Excel is a "flat file database", not a relational database.

- When simple tables need to evolve into multiple tables of related data, Access is the first choice for information workers to quickly create a database application.
Microsoft Access

- Access is the most widely used desktop database system in the world, and has more support and development consultants than any other desktop database system.
- It is available on any computer that has installed the Microsoft Office (no additional cost).
- It is a relational database (multiple tables) to store data.
- It can handle a large amount of data.
- It’s easy to share the Access data with others since Access is a local based program.
- When designed correctly, Access database can be ported to SQL Server or Oracle (the larger database management system).
MySQL or Microsoft Access

MySQL is a relational database management system (RDBMS) that runs as a server providing multi-user access to a number of databases.

(MySQL was owned and sponsored by the Swedish company MySQL AB, now owned by Oracle Corp.)

MySQL or ACCESS?

- With MySQL, people can use client program or other administrative tools to get access to your database by authentication, and can be integrated with a web server which provide a more flexible choice than Access alone.
- MySQL requires authentication before opening a connection to a database. This enhances the security issue.
- MySQL is capable to manage much more data (>=1 Million records) than ACCESS
- MySQL is cross-platform (Linux, Mac OS X, UNIX, and Microsoft Windows).
- MySQL can be obtained for free while Access cannot.
The EpiNu [Nutritional Epidemiology] software was developed intentionally for Epidemiological studies to assess the dietary habits of the population.

The software is supported by a reference database of ingredients, macro and micronutrients, the ideal recipes, their nutritive value, recommended dietary allowances.

The EpiNu database is the major source of dietary assessment database in India and provides foundation for most public and private sector databases.

More information can be found from the website of Madras Diabetes Research Foundation (MDRF) [http://mdrf.in/department/food&Nutrition2.html]
REDCap (Research Electronic Data Capture) is a secure, web-based application for building and managing online surveys and databases.

Using REDCap's stream-lined process for rapidly developing projects, you may create and design projects using
1) the online method from your web browser using the Online Designer; and/or
2) the offline method by constructing a 'data dictionary' template file in Microsoft Excel, which can be later uploaded into REDCap.

REDCap provides automated export procedures for seamless data downloads to Excel and common statistical packages (SPSS, SAS, Stata, R).

More information can be found from http://project-redcap.org/
Donna suggests me to talk to Amara Ezeamama who works for a project in Uganda study (the clinical trial, Multivitamins, HAART and HIV/AIDS in Uganda, which has an accrual goal of 400 HIV positive patients and a follow up duration of 18 months) about the database software (DataFax) used in their project.

**The DataFax clinical trial management system** [http://www.datafax.com/]

- Since its introduction in 1991, DataFax has been used in clinical trials conducted by pharmaceutical companies, contract research organizations, universities, and NGOs; in the United States, Canada, South America, Europe, Africa, and Asia.

- To assist new clients in their first DataFax study, a consultant team would assist client for the case report form design and would do the on-site software installation and integration testing. The consultant team would manage the database throughout the study.

- It supports data collection by both Fax and Electronic Data Capture over the internet in the same study with complete integration designed.
Open discussion- Should all studies in GNET use the same data management software?

- Use the same data management software would be good to share the data across studies and easy for later statistical data analysis.
- Should we do that? - We can think about it and discuss it later ...
- Among all softwares covered today, I would recommend EpiData or Access, since
  - EpiData is a free software and Access is most likely to be available for every computer.
  - Both are widely used by organizations and individuals to create and analyze large amounts of data.
  - Both can be imported/exported to different data formats that can work with other database software and statistical software (such as SAS).
  - Defining a dataset and its associated metadata (i.e. variable and value labels, edit checks etc.) is easier in EpiData than in, for example, Microsoft Access. The trade off is that EpiData does not have the facilities for designing a user interface or the reporting capabilities that Access offers.
After data collection, set up the standardized database and a clear codebook

I used Shanghai study as an example.

**Study introduction:**

- Our Shanghai study aimed to assess the effects of brown and white rice on metabolic risk factors in Chinese adults. A total of 202 middle-aged people with metabolic syndrome (MetS) were randomly assigned to a white rice group (n=101) or brown rice group (n=101) and consumed the rice *ad libitum* for 16 weeks. Metabolic risk markers, including BMI, fasting serum lipid profile, glucose, Hemoglobin A$_{1c}$ (HbA$_{1c}$) and insulin were measured before and after the 16-week intervention.

**How to store the analysis data set, long format vs. wide format?**

- **Wide format data:** one record per person (see example 1)
- **Long format data:** one record per person-time (see example 2).
**Wide format data set**

**Example 1**

<table>
<thead>
<tr>
<th>code</th>
<th>group</th>
<th>Sex</th>
<th>AGE</th>
<th>a5</th>
<th>a6</th>
<th>GLU_1</th>
<th>GLU_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>49</td>
<td>1</td>
<td>4</td>
<td>5.85</td>
<td>5.41</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>47</td>
<td>1</td>
<td>3</td>
<td>5.63</td>
<td>6.02</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>49</td>
<td>1</td>
<td>2</td>
<td>6.05</td>
<td>6.12</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>54</td>
<td>1</td>
<td>3</td>
<td>6.24</td>
<td>5.26</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>variable name</th>
<th>format</th>
<th>English description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code</td>
<td>numeric</td>
<td>The id number of the subject</td>
</tr>
<tr>
<td>Group</td>
<td>numeric</td>
<td>The group of the intervention (1: brown rice group; 0: white rice group)</td>
</tr>
<tr>
<td>Sex</td>
<td>numeric</td>
<td>1: Male; 2: Female</td>
</tr>
<tr>
<td>Age</td>
<td>numeric</td>
<td>Marriage (1: Married; 2: Unmarried; 3: Divorced; 4: Widowed)</td>
</tr>
<tr>
<td>a5</td>
<td>numeric</td>
<td>Education level (1: Below senior high school level; 2: Senior high school level; 3 Bachelor; 4: Master or higher level)</td>
</tr>
<tr>
<td>a6</td>
<td>numeric</td>
<td>fasting glucose (mmol/L) at baseline</td>
</tr>
<tr>
<td>glu_1</td>
<td>numeric</td>
<td>fasting glucose (mmol/L) at endpoint</td>
</tr>
</tbody>
</table>
### Example 2

<table>
<thead>
<tr>
<th>Code</th>
<th>group</th>
<th>timeindex</th>
<th>sex</th>
<th>AGE</th>
<th>a5</th>
<th>a6</th>
<th>GLU</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>49</td>
<td>1</td>
<td>4</td>
<td>5.85</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>49</td>
<td>1</td>
<td>4</td>
<td>5.41</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>47</td>
<td>1</td>
<td>3</td>
<td>5.63</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>47</td>
<td>1</td>
<td>3</td>
<td>6.02</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>49</td>
<td>1</td>
<td>2</td>
<td>6.05</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>49</td>
<td>1</td>
<td>2</td>
<td>6.12</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>54</td>
<td>1</td>
<td>3</td>
<td>6.24</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>54</td>
<td>1</td>
<td>3</td>
<td>5.26</td>
</tr>
</tbody>
</table>

Actually it’s easy to transfer long to wide or wide to long by using SAS ([http://www.ats.ucla.edu/stat/sas/modules/towide.htm](http://www.ats.ucla.edu/stat/sas/modules/towide.htm), [http://www.ats.ucla.edu/stat/sas/modules/tolong.htm](http://www.ats.ucla.edu/stat/sas/modules/tolong.htm))
How to standardize the analysis data set across sites?

**Example 3: Reformat data in Example 1**

<table>
<thead>
<tr>
<th>gnetid</th>
<th>studyid</th>
<th>ricegroup</th>
<th>female</th>
<th>age</th>
<th>marriage</th>
<th>education</th>
<th>glu_1</th>
<th>glu_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sh001</td>
<td>1</td>
<td>0</td>
<td>49</td>
<td>1</td>
<td>4</td>
<td>5.85</td>
<td>5.41</td>
</tr>
<tr>
<td>2</td>
<td>sh002</td>
<td>1</td>
<td>1</td>
<td>47</td>
<td>1</td>
<td>3</td>
<td>5.63</td>
<td>6.02</td>
</tr>
<tr>
<td>3</td>
<td>sh003</td>
<td>1</td>
<td>1</td>
<td>49</td>
<td>1</td>
<td>2</td>
<td>6.05</td>
<td>6.12</td>
</tr>
<tr>
<td>4</td>
<td>sh004</td>
<td>1</td>
<td>0</td>
<td>54</td>
<td>1</td>
<td>3</td>
<td>6.24</td>
<td>5.26</td>
</tr>
<tr>
<td>5</td>
<td>sh005</td>
<td>2</td>
<td>0</td>
<td>56</td>
<td>1</td>
<td>1</td>
<td>5.84</td>
<td>6.81</td>
</tr>
<tr>
<td>6</td>
<td>sh006</td>
<td>1</td>
<td>0</td>
<td>36</td>
<td>2</td>
<td>3</td>
<td>7.13</td>
<td>7.6</td>
</tr>
</tbody>
</table>

**variable name** | **format** | **label**
---|---|---
gnetid | numeric | The unique index id for each subject
studyid | numeric or character | The unique id for each subject in study
ricegroup | numeric | The group of the intervention. It’s better to use the categorical form (1/2/3/…), setting the largest number for the reference rice group (such as white rice group)
female | numeric | 1=female, 0=male
age | numeric | integer
marriage | numeric | Question a5. Marriage status (categorical variable): 1=Married; 2=Unmarried; 3=Divorced; 4=Widowed.
education | numeric | Question a6. Education level (categorical variable): 1=Below senior high school level; 2=Senior high school level; 3=Bachelor; 4=Master or higher level.
glu_1 | Numeric | Fasting glucose (mmol/L) at baseline
glu_2 | Numeric | Fasting glucose (mmol/L) at endpoint
Some basic guidelines for setting up analysis data set (1)

- A unique index id (can be names as **gnetid**) for each subject from 1 to n. So it’s easy to (1) be merged with other data set; (2) to be sorted by index id; and (3) to be used in SAS procedure, such as Proc Mixed.

- An additional **“studyid”** for the unique subject id in different studies, which allows for numeric or character format.

- Name all variables using **lower case** to eliminate mistakes if software programs are case sensitive.

- Define **“ricegroup”** variable to be categorical variable to indicate different rice groups (for example, 1=white rice, 2=brown rice, 3=whole grain, etc.).

- For **categorical variable** ( >=3 levels), it’s better to set it as numeric variable rather than character variable and has the detailed label for the meaning of each level (the values of “Married”, “married” and “marrie” would be treated as different categorical levels in SAS).
Some basic guidelines for setting up analysis data set (2)

- For **binary variable**, sometimes set it as indicator variable (0 / 1) rather than categorical variable with 1 / 2 values would be easier to use in the statistical analysis. For example:
  - “gender” variable can be set as 1=Male and 2=Female, but it’s more meaningful if we define it as “female” and 1=female and 0=male.
  - “diabetes” variable can be set as 1=yes and 2=no, but we could also set it as 1=yes and 0=no.

- Some raw data set like to use the question number as the variable name, for example “a5”. It may be good for data entry, but it’s hard for later data management and analysis. We should rename it with meaningful variable name (for example, rename “a5” into “marriage”).
Some basic guidelines for setting up analysis data set (3)

- Include **unit** in the codebook, and detailed definition. For example:
  - “central_obesity” - in codebook “waistline_1>90cm”
  - “glu_1” – in codebook “fasting glucose (mmol/L) at baseline”

- It’s very important to record the **missing values** correctly with detailed codebook or label. Some people would use “999” or similar large number that out of the possible range of the variable values for missing values for numerical variables. Sometimes this would generate mistakes in the later data analysis if there is not clear codebook to explain “999” means missing values.
Some open discussion topics

- **Generally who is allowed access to the data**
  - Research team for each study can access their own data
  - Harvard Statistician Team can access each study’s data?
  - Would we allow all data are accessible for the whole GNET group?

- **How should data be transferred?**
  - By email?
  - By CD?
  - Or save the database directly in Harvard Channing Unix System?
Acknowledgements

- Donna Spiegelman
- Nan Li
- Amara Ezeamama
- Akum Aveika Awasana
- Tao Hou, Mathew Pazaris, Jiatao Ye
- GNET each site’s PI and data analyst
**EpiData Entry**

<table>
<thead>
<tr>
<th>Define data in simple text</th>
<th>EpiData creates this data entry form</th>
</tr>
</thead>
<tbody>
<tr>
<td>My first DataEntry Form</td>
<td>My first DataEntry Form</td>
</tr>
<tr>
<td>Id  <code>&lt;idnum&gt;</code>  Encrypted id <code>&lt;e&gt;</code></td>
<td>Id</td>
</tr>
<tr>
<td>Height <code>###</code>  Weight <code>####</code></td>
<td>Height <code>1.90</code>  Weight <code>75.0</code></td>
</tr>
<tr>
<td>Born <code>&lt;dd/mm/yyyy&gt;</code>  age <code>###</code></td>
<td>Born <code>12/12/2001</code>  age <code>###</code></td>
</tr>
</tbody>
</table>

Several field types are available (e.g. date, numerical, string, today, auto id). Color setting defined by user.
EpiData Entry

Flowsheet of how you work with EpiData Entry

The work process is as this (optional parts are dotted):

1. Define datastructure and layout of DataEntry
2. Change structure or layout when necessary refine structure
3. Define checks and jumps
   - attach labels to variables
   - range checks
   - define conditional jumps (filters)
   - consistency checks across variables
4. Attach labels to values
   - Reuse from collection
   - Define new
5. Define values as missing value
6. Preview DataForm and simulate dataentry
7. Create datafile
8. Enter all the data.
   - Enter data twice and compare directly at entry or enter separately and compare afterwards.
9. Correct errors based on original paper forms
10. Revise structure without losing data
11. Generate documentation:
    - List of data, codebook and variable overview including defined checks and labels
12. Dataset is ready.
    - archive copy with documentation.
    - export data for special analysis
    - analyse with

Version 25® August 2005
EpiData Analysis

Flow Sheet of working with Analysis

1. Read a data file (REC, DBF, CSV)
   - For REC files: Read variable labels in Rec File and category label and missing value definitions in associated CHK file

2. Keep a COPY of data in memory
   - Save data to disk
   - Command savedata
   - Modify data in memory
   - Add new variables
     - generate or define
     - commands
     - Change contents of variables
     - let, recode
     - if ... then ...
     - Change sorting
     - sort command

3. "Next" Analysis command

4. Show results on the Screen

5. Copy of results written to disk
   - Default name EAOOutput.htm
   - Close: logclose

6. Stop Analysis
   - quit or exit
   - A copy of all commands are saved as temp.pgm