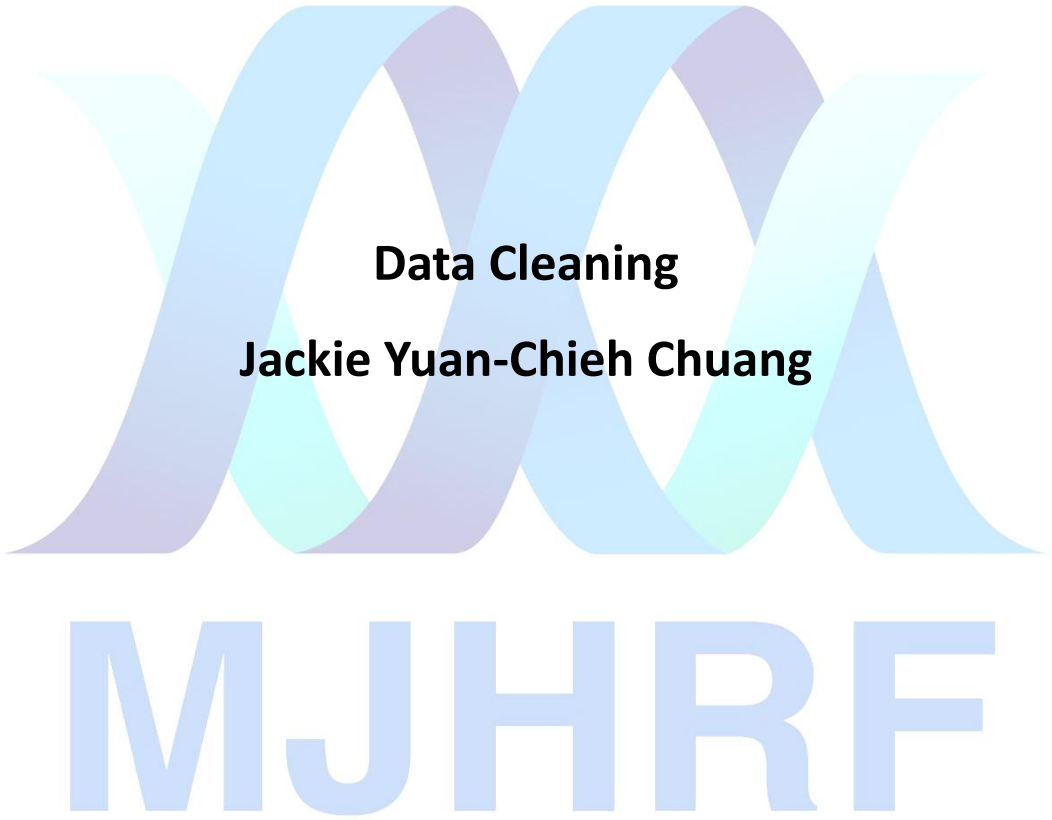
MJ Health Research Foundation

MJ Health Resource Center

Technical Report



Data Cleaning

Jackie Yuan-Chieh Chuang

MJHRF-TR-04

2016/03/18

Data Cleaning
Jackie Yuan-Chieh Chuang

I. Foreword

Since its inception in 1994, the MJ Health Database has accumulated data from health questionnaires and physical examination for over 20 years. However, due to the long period of time collecting the data, question changes and unavoidable omission in the recordings inevitably lower the quality of the data and information, thus affecting the research analysis using the data.

In order to prevent research errors caused by the quality of the data, MJ Health Resource Center (hereafter the Resource Center) has conducted data cleaning through a series of error inspection and organization. Not only could we lower the rates of error and increase the quality of the data, but also increase the data's usability. The following paragraphs will illustrate the process in which our Resource Center conduct the organization and inspection of our data.

II. Establishing Basic Personal Information File

1. Basic Characteristics

   One of the main characteristics of MJ Health Screening is that it is based on a membership system. Every member periodically receives health examinations at a MJ Clinic, therefore all data collected in the MJ database includes all the basic personal information, physical examination data, and health questionnaires from previous health examinations. For future convenience, we cross-referenced all past information to confirm each member's personal information (national identification number, gender, date of birth, etc.) and create one set of correct and uniform basic personal information file for each member. This way we are able to avoid data corruption due to errors or missing data generated by members themselves during a single physical examination visit.

2. Matching National Identification Number with Name and Date of Birth

   Often times, when organizing basic information, multiple people with different names and date of births may have the same national identification number

(foreigners use date of birth plus first three letters of their English surname). The main reason for such occurrences is because of either a change of names, error in data entry, or the subject is not the person receiving the physical examination. When this occurs, we use their address, telephone number, or email, etc. to determine if this is the same person. If it is indeed the same person, we correct the errors in our record (name or date of birth); if not, then we inspect all the files to find if there is a subject who has the same name and date of birth. Consequently, if there is another subject with the same name and date of birth, then we correct the national identification number; if not, due to us being unable to determine which person as the legitimate owner of the national identification number, we will assign such a person a new identification code.

3. Matching National Identification Number with Gender

Due to the second number of Taiwan's national identification number being an indicator of a person's gender, we use one's national identification number to correlate with the data in the subject's files. Procedurally, we inspect the subject's national identification number first to determine if it is from the Republic of China; if such is the case, we correlate the subject's gender with the national identification number. When contradiction appears, we inspect gender-specific physical examination items such as breast examination or prostate examination, and correct the gender accordingly. If no gender-specific examination items exist, we refer to the health questionnaire, and use gender specific questions to determine the subject's gender.

4. Storage and Release of Basic Personal Information

We refer to the information security regulations (see MJ Health Resource Center Technical Report on Information Security) when handling and storing our member's basic personal information. We ensure the security of these personal information through physical isolation, access management, auditing and other methods.

The main purpose of storing personal information is for identity confirmation during internal use of the data. Except age and gender information that may be released to data users outside the foundation, all other personal information are for internal-use only.

III. Health Questionnaire Data Cleaning

1. Illegal Value Treatment

All of the questions in the health questionnaires are close-ended, meaning each question has a reasonable confine of answers. Each question has a reasonable confine of answers. For example, for the question; what is your blood type? 1) A, 2) B, 3) O, 4) AB, 5) do not know. The answer could only be of number 1-5. However, due to data entry errors, or data transfer errors, there may be minuscule amount of illegal values. These values will be changed to missing values.

2. Skip Logic Questions

MJ Health Questionnaires contain skip logic questions. However, due to the questionnaires being answered on scantrons, the skipping of certain questions could not be enforced, thus causing logic errors. For example:

1) Questions:

[Smoking]

87. Do you smoke? (For those who answer "No", please skip to question No.91)
☐ No ☐ No, but often inhale second-hand smoke
☐ Have quit smoking ☐ Occasional smoker ☐ Smoke daily

88. How many years have you been smoking? (For those who have quit, please refer to past experience.)
☐ <1 ☐ 1≤years<3 ☐ 3≤years<5
☐ 5≤years<10 ☐ 10≤years<20 ☐ ≥20

89. If you no longer smoke, how many years has it been since you quit smoking?
☐ <1 ☐ 1≤years<3 ☐ 3≤years<10
☐ 10≤years<20 ☐ ≥20

90. How many cigarettes do you smoke every day on average? (For those who have quit, please refer to past experience.)
☐ <½ pack ☐ ½ -1 pack ☐ >1 pack

2) Inspection Steps:

- ❖ Those who answered "No" on question 87 should not have answered question 88-90.
- ❖ Only those who answered "Have quit smoking" on question 87 have to answer question 89.

Some subjects, due to not seeing the instructions, tend to answer questions that do not match their own circumstances with the next best answers. For example, those that do not smoke will answer “less than a year” for the question “How many years have you smoked?” However, in order to maintain data consistency, we will change such answers to missing value.

3. Multiple Answer Questions

Often times, multiple-answer questions also have mutually exclusive answers in the selection pool, for example:

44. Are you currently on any medication? (You can choose more than one from the following list.) (taken at least once a day on average)
☐ None ☐ Sedatives or sleeping pills ☐ Uric acid drugs
☐ Steroids ☐ Drugs for cardiovascular diseases ☐ Hormones
☐ Drugs for hypertension ☐ Painkillers ☐ Drugs for diabetes
☐ Gastrointestinal drugs ☐ Thyroid inhibitor
☐ Chinese herbal medicines ☐ Drugs for high blood lipids
☐ Psychiatric medications ☐ Medicine for asthma
☐ Over-the-counter drugs ☐ Others

If one selects “None”, then he or she should not select any of the following medications. If contradiction occurs, we will use the selected medications, and change the selection “None” to missing value.

4. Logic Check between Questions

Although some questions do not have reminders for skipping questions, there are logical connections between the questions. Such datas will also be examined to ensure its consistency. For example:

(Questions below are for female only)
37. Have you ever been pregnant or given birth? ☐ No ☐ Yes
38. Number of pregnancies you have had:
☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ ≥5 times
39. Number of childbirths you have had:
☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ ≥5 times
40. At what age did you give birth to your first child?
☐ ≤ 19 ☐ 20-24 ☐ 25-29 ☐ 30-34 ☐ ≥ 35 ☐ N/A

The four questions above have direct logical correlations between each other. For example, if you have never given birth, then the amount of times you have given birth would be 0, and thus the answer for the age of your first born should be “N/A”. On the contrary, if you have given birth before, then the answers for the amount of times pregnant and given birth should not be 0, and thus the answer to the age of your first born should also not be “N/A”. Also, the amount of times one has gotten pregnant should not be lower than the amount of times given birth. If there are inconsistencies in such circumstances, we will examine answers to all four questions, and fix the inconsistencies.

5. Organizing Questionnaire Revisions.

MJ Health Resource Center has been collecting data using health questionnaires for over 20 years. During the past 20 years, there has been many revisions to the questionnaire contents, whether it be adding, subtracting questions or changing answer choices. Such revisions could cause inconvenience for data users. To make our data more user-friendly, data has been reorganized and code book written to aid users in better understanding the data.

Using the marital status question as an example (see table 1), it has been modified both 1998 and 2014. In the raw data, marital status is placed under one variable, therefore, if choice #3 was selected for the question, the meaning is different in 1997 and 2013 versions. Because of this, we split the marital status into three separate variables, each with its own set of answer choices to reflect question revisions over the years. We have also prepared code book with information as shown below to help data users understand differences between marital status questions from different questionnaire versions:

Table 1. Marriage Situation Question Code Book

Variable Name	Question	Options	Year
marriage_96	marital status	【1996.02-1997】 (1)Single (2)Married (3)Remarried (4)Widowed (5)Divorced (6)Separated	96-97
marriage_98	marital status	【1998-2013】 (1)Single (2)Married (3)Divorced (4)Widowed	98-13
marriage_14	marital status	【2014-】 (1)Single (2)Married, Remarried, Cohabitation (3)Divorced (4)Widowed	14-

IV. Physical Examination Data Cleaning

1. Illegal Value Treatment

   Out of more than one hundred physical examination items, few data such as early height, weight measurements and waistlines were entered into the system by hand. There were inadvertent errors as a result of manual entry. We cross-reference suspected illegal values such as height over 250cm, or grown adults lighter than 30kg with data entries from subject's other physical examination appointments. If there are no data available, then the illegal value will be changed to missing value.

2. Extreme Value Treatment

   When checking the physical examination data, we listed all the extreme values in our variables such as systolic blood pressure exceeding 200mmHg and reviewed them with laboratory and nursing department colleagues from the clinic. For the questionable values, we will confirm their accuracy by retrieving the original physical examination data from the clinics. Cases such as these are very rare.

3. Inspection of Gender Related Items

   Some items (such as pap smears) in the physical examination are specific only to the female gender. We use this information to cross check the subject's gender variable to see if the gender is recorded correctly. This step also ensures that data were not displaced en masse.