Getting to Know Your Data

Data Linkage Training Project for Vital Statistics and Medicaid Claims Data

February 6th and 7th, Washington D.C.

Russell Kirby and Craig A. Mason

Topics

- **Deterministic Matching**
  - Background, Planning and Preparation
- **Non-Deterministic Linkage**
  - Variation in Individual Variables
    - Phonetic Transformation, Textual Similarity
  - Inconsistencies Across Records
    - Theory, Weighted Cases, Probabilistic, Machine Learning
- **Evaluating Linkage Results**
  - Precedence, Creep, Metadata
  - Structuring the Results
    - ID Tables, Second-Order Linkage

In the Beginning

- Why link records?
  - One time research study?
  - Re-occurring programmatic strategy?
- What are the implications?
  - Do you need to link everyone?
  - Do some records or situations HAVE to be linked?
  - Must ALL records be linked?
  - To what degree can you tolerate error?

Research Versus Applied Service

- Linkage of records in research
  - One-time project
  - Small percentage of errors may be acceptable
  - May choose to ignore…
    - If random and small impact
    - Elimination requires disproportionate effort and/or may introduce additional error

Research Versus Applied Service

- Linkage in ***applied*** public health service
  - Existence of a known error is more likely to be unacceptable
  - Once an error or problem has been corrected, it may be important that it can not re-emerge
  - May be re-occurring project on a regular basis

Overview of Linkage Process

- Vital Statistics and Medicaid are two databases containing information on some of the same individuals
- There are many possible ways these individuals may be linked across databases:
  - Some babies in birth certificate records ***are*** enrolled in Medicaid
  - But many babies in birth certificate records are not enrolled in Medicaid
  - Similarly, some mothers who appear in birth certificate records ***are*** enrolled in Medicaid
  - But many mothers in birth certificate records are not enrolled in Medicaid

Overview of Linkage Process (continued)

- Medicaid may include enrollments for both a mother and a child…
- …or just the mother…
- …or just the child…
- Medicaid will also contain records for women and children who do not appear in the birth certificate data
  - Some women may have had pregnancies that did not result in a live birth.
  - They also may have given birth in a different state.
- …and a person in the Medicaid Enrollment data may appear a few times or many times in the Claims data
  - Theoretically, there are one-to-one and one-to-many relationships, assuming valid, accurate, correct identifiers. But in practice, this is not always so!

Getting to Know Your Data, or Preparing Datasets for Record Linkage Using SAS

Data Linkage Training Project for Vital Statistics and Medicaid Claims Data

February 6th and 7th, Washington D.C.

Russell Kirby and Craig A. Mason

Topics

- **What's in My Datasets?**
  - Data Dictionaries and MetaData
  - Variable Names, Field Formats/Informats, Field Lengths
  - Differences between Datasets
- **Working with Data**
  - Coding, Inconsistency, Field Length, Field Name, Field Formats
- **Working with Names and Dates**
  - A Brief Introduction
- **Keeping Track of Results**

What's in My Datasets?

- Importing data into SAS
  - DATA step with INPUT statement
  - Import data with defined database structure
  - Dynamically access dataset using PROC SQL
  - Other approaches
- Referencing external datasets in SAS
- Referencing SAS datasets in SAS

What's in My Datasets?

- Examining Database Structure
  - PROC CONTENTS
    - A versatile procedure, generates list of database contents including number of records, number of variables, variable list with field name, length, type, format, and label
  - Tools > Table Editor > File > Open > Select library > Select file

The Contents Procedure

Data Set Name	RL.MEDICAIDDATA2007	Observations	3238
Member Type	DATA	Variables	6
Engine	V9	Indexes	0
Created	Tuesday, April 24, 2007 03:46:46 AM	Observation Length	53
Last Modified	Tuesday, April 24, 2007 03:46:46 AM	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label	Written by SAS		
Data Representation	WINDOWS_32		
Encoding	Default		

Alphabetic List of Variables and Attributes				
#	Variable	Type	Len	Format
5	MEDI_Birth_Hosp	Num	8	
4	MEDI_Child_DOB	Num	8	MMDDYY10.
2	MEDI_MOM_FIRST	Char	9	
1	MEDI_MOM_LAST	Char	12	
6	MEDI_Mom_SS	Num	8	
3	MEDI_Mother_DOB	Num	8	MMDDYY10.

Default variable list is alphabetical, add keyword POSITION for sequential listing within the data table.

What to Look For

- For candidate linkage variables
  - Are variable names identical?
  - Are field types identical (generally, character vs. numeric)?
  - Are field lengths identical? ? This is most important for character string variables.
    - They don’t have to be, but you need to make sure the longer one is on the master dataset, otherwise you will lose information and potentially match cases incorrectly.

What to Look For

- For candidate linkage variables (continued)
  - If the variable is numeric, does it include decimals or percentages?
  - If the variable is a date field, how is it stored?
    - We could spend the entire training on this topic, but here we will only hit the highlights
    - Dates might be stored in separate fields for month, day, year. Or, they might be stored in any of a variety of date formats (MM/DD/YY, DD/MM/YY, MM/DD/YYYY, just to name a few)

What to Look For

- For candidate linkage variables (continued)
  - Similar attention must be paid to any other fields that might be used for linkage, including:
    - Social Security Numbers
    - Health care provider or financing IDs
    - Facility codes
    - Areal identifiers (place names, county codes, ZIP codes, etc.)

What to Look For

- For other variables
  - Make sure that any other variables on the master or transaction datasets have different names
    - Why? After you have created link ids for each record, you will come back and pull in the data for each record from the two datasets, and data in the master file will be overwritten by data in the transaction file if the fields in each have the same name
    - Of course, if they differ in field type or length, you might just get an error message in the SAS Log, which you then have to correct. But better to address this issue proactively

How to Fix It?

- The RENAME command can be used to change variable names without changing their structure
- Date fields should be converted into SASdate fields and stored with identical formats
- Where there is discord between character and numeric fields, choose the more useful and convert the variable in the other dataset

How to Fix It?

- Where coding differs, variables can be recoded a DATA step with nested IF-THEN-ELSE syntax, or directly reformatted using VALUE formats and a PUT statement
- We will explore examples of each . . . dataset