Linking data deterministically
Which variables are common to both datasets??

Do a PROC Contents
Mother’s information

Birth_mom_legal = Screen_mom_legal_last
Birth_mom_mid = Screen_mom_mid
Birth_mom_first = Screen_mom_first
Birth_mother_dob = Screen_mother_dob
Infant’s information

Birth_child_last = Screen_child_last
Birth_child_mid = Screen_child_mid
Birth_child_first = Screen_child_first
_Birth_gender=_Screen_gender
Birth_child_dob=Screen_date
Other information

Birth_zip_code=Screen_zip_code
Birth_hosp=Screen_hosp
Missing data

Look for missing data in linkage variables
Ranking of linkage variables

Which variables are the “best” variables?

• How much missing data in each variable?
• What do you know about the variables?
Our ranking

*Fill in here*
The art of creating a linkage algorithm

- Most discriminating combination of variables first
- Loosen criteria as you go along
The art of creating a linkage algorithm

Most strict criteria

Linkage step 1

Linkage step 2

Linkage step 3

Least strict criteria

Linkage step…
Create id in data set

- Allows you to easily merge back with original data
- Easy as:
  data new;
  set old;
  id=_n_;
  run;
Sort by chosen linkage variables

• What happens when you don’t use by variables??

• Let’s take a look . . .
Merge by chosen linkage variables

- Create data set with only linked records
- Keep track of the “link level” – level of linkage where records matched
Re-merge to get unlinked datasets

- Unlinked data sets contain only variables from that data set
- Unlinked records sent to next level of linkage algorithm
Last step

• Combine all linked data sets
• Investigate unlinked records
  – Look for systematic errors responsible for non-linking
  – Look for biases
• Evaluate quality of links in linked records