Data Deduplication Rules

Before starting to perform data deduplication, change all lower case to upper case to ensure better matching. Also set invalid names—such as BABYBOY, BABYGIRL, BB, BG—to missing values.

Step 1: Sort all records by five variables (fields) as matching criteria. Eliminate the variable Infant Gender because it is not a very strong filter and the potential for it to eliminate a potential match is greater than its usefulness to confirm a possible match. The five variables are (1) Infant First Name, (2) Infant Last Name, (3) Infant DOB, (4) Mother First Name, and (5) Mother Last Name. If several records match exactly, then one record could be eliminated after necessary documentation.

Step 2: Select four of the five variables listed in step 1 as a new criteria, and repeat step 1.

- Infant First Name, Infant Last Name, Mother First Name, Mother Last Name.
- Infant First Name, Infant Last Name, Infant DOB, Mother Last Name.
- Infant Last Name, Infant DOB, Mother First Name, Mother Last Name.
- Infant First Name, Infant DOB, Mother First Name, Mother Last Name.
- Infant First Name, Infant Last Name, Infant DOB, Mother First Name.

As fewer matching criteria are used, the likelihood of an accurate match decreases. Therefore, more human decision-making is involved. When making a decision regarding whether or not the match is true, place more weight on those criteria that are least likely to change and least likely to have data entry errors, such as date of birth or gender. Typographical errors in the infant's or mother's name will stop a likely automated match.

The most likely criteria to not match will be Mother Last Name (which could change with marriage or divorce) and Infant Last Name.

Step 3: Select three of the five variables listed in step 1, and repeat step 1.

Because the number of potential combinations of three criteria is very large, only certain matches will be performed. Perform the following matches:

- Infant First Name, Infant Last Name, Infant DOB.
- Infant First Name, Infant DOB, Mother First Name.

5.1 Review each of these lists and determine if there is a match. As fewer and fewer criteria are selected, the number of potential duplicates decreases so the burden of numbers should decrease dramatically with each stage of deduplication.