Link the death and CHARS files

Eric Ossiander
July 17, 2014

I created linked death-hospitalization files for deaths occurring in the years 2010, 2011, and 2012. I linked the deaths occurring in each year to the hospitalizations in that year and the previous year.

I used the following fields in the linking process:

birth date
name
last 4 digits of SSN
sex
zipcode of residence
county of residence
hospital code
death date
Hispanic ethnicity
race
state of residence

I used the RecordLinkage package in R for most of the linking. In all of the record linking that I did in R, I used birth date as a blocking field (i.e. I required that the birth date on the death certificate match the birth date on the hospitalization record). First, I computed a probabilistic linkage weight for each record pair. Second, I used a machine learning algorithm to predict which record pairs were links. (This required me to manually code a training set once, to create a statistical model for predicting links. Then I used the statistical model for each subsequent year of data.) Then I manually reviewed all of the record pairs which were predicted not to be a link by the machine learning algorithm, but which had a high probabilistic weight, and all record pairs which were predicted to be a link, but had a low weight. I also used a SAS program to compute a probabilistic linkage weight for all record pairs (i.e. not blocking on birth date), and manually reviewed all of the record pairs that had a high probabilistic weight in which the death certificate birth date did not match the hospitalization birth date. I combined the three linked sets (the machine-linked pairs, the manual review of the machine linking, and the manual coding of the non-birth date matching pairs). Then I checked for hospitalization records that linked to more than one death record, and manually adjudicated those links.