

An assessment of using frequency weights for record linkage

Roemer, M^{1*} and Campbell, S²

¹NCHS

²NORC at the University of Chicago

Many different techniques can be used to perform entity-to-entity record linkage. The optimal approach may relate to the entity type, such as individual or establishment, and source, such as sample survey, enumeration, program administration, or register. Within the Fellegi-Sunter record linkage framework, the frequency of occurrence of match variables' values can be either employed or ignored when estimating the probabilities from which the match weights are derived. Namely, the U-probability that a variable agrees in a pair of non-matched records, is estimated for each value of a match variable in the frequency-based approach, or in general in the non-frequency-based approach. The aim of this talk is to compare the quality of results produced by the frequency-based and non-frequency-based approaches when linking a household survey and an establishment survey to vital records. A household survey, the National Health Interview Survey (NHIS) and an establishment survey, the National Hospital Care Survey (NHCS) are each linked to the National Death Index (NDI) using the Fellegi-Sunter record linkage framework. We perform the linkage on each survey twice; first, employing frequency-based weights for all match variables, and second, simple agree/disagree weights for all match variables. We then examine any differences in quality within each survey, and assess whether any differences in the quality of the two approaches are attributable to the type of survey, household versus establishment.

*Corresponding Author:

Email Address: nlk5@cdc.gov (M Roemer)

