Bayesian Procedures for File Linking with Application to Health Services Research

Analysis of partially linked datasets is increasingly important as researchers and policy analysts seek to integrate administrative datasets and registries while adapting to privacy regulations that limit access to unique identifiers. Record-linkage tools have been developed to identify records that represent the same entity across multiple datasets in the absence of unique identifiers. Past research mainly focused on the computational efficiency of record-linkage tools. Less attention has been given to features of the data that can improve the linkage and to statistical inferences with linked records.

To address these limitations, we view record linkage as a missing data problem and develop Bayesian procedures that utilize data features that are frequently encountered in public health applications. These procedures improve the linkage, and result in more accurate and precise estimates of scientifically important associations. One set of procedures incorporates associations between variables exclusive to one of the datasets in the linkage process. Another procedure ensures that individuals receiving care from the same provider in one dataset are linked to individuals receiving care from a similar provider in the other dataset. This latter procedure is implemented even when providers cannot be uniquely linked across the two datasets. Both procedures generate M datasets in which the links between the two datasets are imputed. The datasets can be analyzed independently and combined using standard multiple imputation rules. This approach makes fewer demands on the time and expertise of analysts while giving them flexibility in the downstream analysis of the linked data.

Two applications will be described: one combines Medicare claims records and Vital Statistics Mortality records to study the association between end-of-life medical expenses and causes of death. A second application combines records from the National Trauma Databank with Medicare claims data to study the relationship between injury characteristics and successful discharge to the community among patients with traumatic brain injury.