Data Linkages & Disclosure Risk

Data Sharing for Demographic Research

John Marcotte, PhD, and Sarah Rush, MSW
ICPSR, University of Michigan

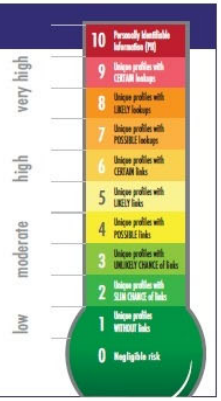
Background

A great variety of data sources are now available to researchers. Analysts may wish to study outcomes from one dataset with predictors from another data source. Combining data from multiple sources can enrich research and increase analytic potential. At the same time, linking data can increase the risk of re-identification and disclosure.

Measuring Disclosure Risk

Degree of disclosure risk is related to presence of:

- Unique Profiles: Set of variables that when combined together can be used to link data to other sources
- Links: Pieces of information that can be used to connect data (e.g., external ID)
- Lookups: Information that translates profiles into identities



Ways to Reduce Disclosure Risk

- Linked data can be made available as restricted-use, requiring an application and approval process for access
  - Linked data with high or very-high risk can be made available via enclave only
- Strategies can be implemented to mitigate risk
  - Swapping records
  - Minimum cell and sub-sample sizes
  - Suppressing link variables
- Tighter security and enhanced monitoring of researchers performing own linkages
  - Masking of internal IDs when possible

Types of Linkages

Data files are linked using identifiers (linking variables). Identifiers differ based on the type of linkage and matching used.

TYPE OF LINKAGE	LINKS OCCUR BY	EXAMPLES	MATCH TYPE
Between waves	Internal IDs	Case ID, unique to study	Exact
Overlapping cross-sections	Internal IDs	Case ID, unique to study	Exact
Separate sources	External IDs	PII: SSN, Health ID	Exact
Contextual linking (incl. geography)	Geocodes, Org codes	Zipcode, Tract Number, County Code, School Name, Hospital Name	Exact
Matching	Common variables	Age, Sex, Ethnicity, Education	Probabilistic

Sources of Disclosure Risk in Linked Data

Disclosure Risk varies based on the type of data linked and the combined information contained in the linked files.

SOURCE OF DISCLOSURE RISK	TYPE OF LINKAGE: MATCHING (Common Variables)	PANEL WAVES (Internal IDs)	OVERLAPPING CROSS-SECTIONS (Internal IDs)	CONTEXTUAL LINKING (incl. geography) (Geocodes, Org Codes)	SEPARATE SOURCES (External IDs)
LINKAGE VARIABLE — Examples: Case IDs, PII, PHI, FIPS	6	7	8	9	10
RESPONDENT'S HISTORY — Examples: Employment History, Marital History	5	6	4	2	3
OTHER RESPONDENT INFORMATION — Examples: Household Structure, Employment Status, Area Characteristics	3	4	6	8	7

Summary

Linkages provide richer data for researchers, but disclosure risk may increase substantially. While access to linked data can be provided through restricted-use data agreements and security plans, results from these combined datasets must still be reviewed for disclosure risk. Even if IDs and other linking variables are removed after data are combined, histories and additional information still increase disclosure risk.

dsdr.icpsr.umich.edu