Data Linkages & Disclosure Risk

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.

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

<table>
<thead>
<tr>
<th>TYPE OF LINKAGE</th>
<th>EXAMPLES</th>
<th>MATCH TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between waves</td>
<td>Case IDs, unique to study</td>
<td>Exact</td>
</tr>
<tr>
<td>Overlapping cross-sections</td>
<td>Case IDs, unique to study</td>
<td>Exact</td>
</tr>
<tr>
<td>Separate sources</td>
<td>PHN, SSN, Health ID</td>
<td>Exact</td>
</tr>
<tr>
<td>Contextual linking (incl. geography)</td>
<td>Geocodes, Org codes</td>
<td>Exact</td>
</tr>
<tr>
<td>Matching</td>
<td>Age, Sex, Ethnicity, Ed</td>
<td>Probabilistic</td>
</tr>
</tbody>
</table>

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

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.

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