Data Sharing for Demographic Research

Specifying Disclosure Protection Rules for Reporting Results from Restricted-use Data

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

Background

- Restricted-use data analyses require vetting of output for disclosure risk prior to presenting or publishing results.

- A framework for setting disclosure protection rules can guide data providers in determining thresholds and communicating rules to data users.

- Providing a clear set of rules in an easy-to-use format can help secondary users of research data ensure that summary results comply with data protection rules for the study.

- Applying study-specific disclosure protection rules to analysis results can protect against indirect reidentification of respondents and other vulnerable groups.

- **Data collectors** and **data users** have different perspectives and goals that a framework must address.
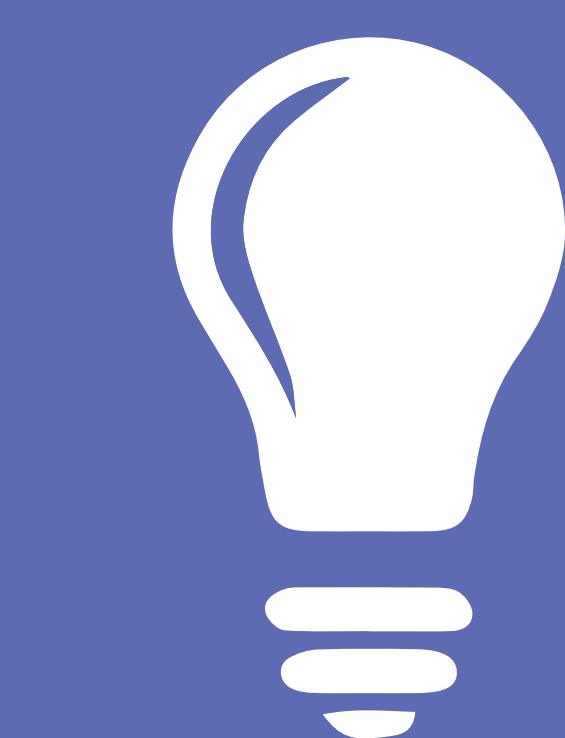
Data Collectors
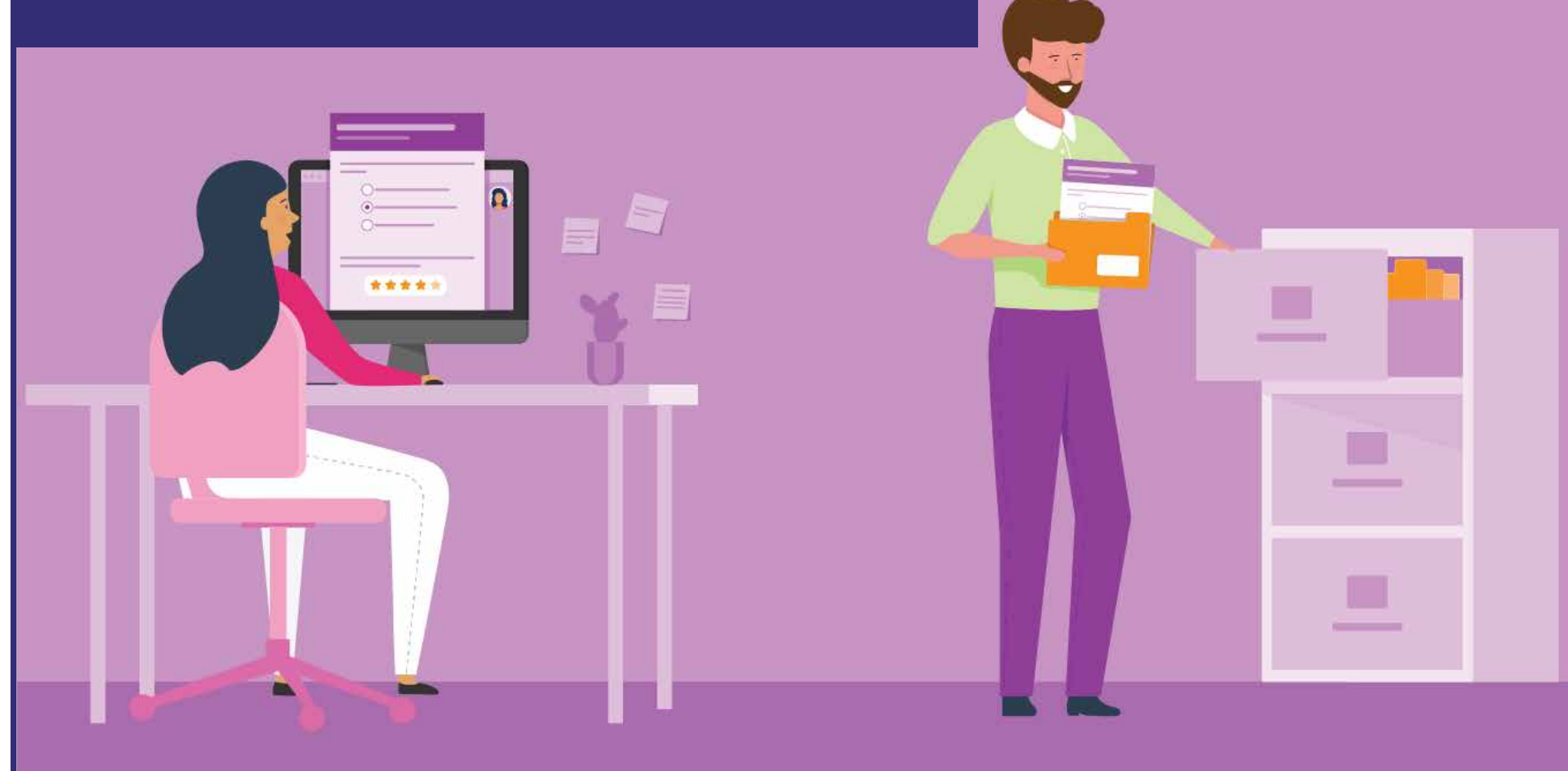
- Define clear rules
- Encourage compliance
- Promote data use
- Protect future data collection efforts
- Protect Human Subjects

Data Users

- Conduct research
- Refer to rules easily
- Publish results
- Protect future research
- Protect Human Subjects

Takeaways

- Facilitate secondary research by helping data users comply with human subjects protections
- Prevent data breaches by specifying clear disclosure protection rules

Disclosure Protection Rule	Recommended Threshold and Example Value
Personally Identifiable Information (PII) or Protected Health Information (PHI) PII and PHI includes Information such as name, address, and respondent ID	*Names, SSN, Email, Insurance Number, etc. cannot be reported*
Suppressed Variables While these variables can be included in an analysis, the resulting coefficients and tables cannot be reported	*Sampling Clusters, Geocodes*
Suppressed Combinations of Variables While these variables can be reported separately, they may not be used together in tables or interactions	*Some 4-way and higher tables*
Minimum Cell Sizes In tables, cells below this value require rows or columns to be combined. Redaction of the individual cell is insufficient.	*10*
Minimum Sample and Sub-Sample Size Minimum number of valid observations (excluding missing data) for regression analysis	*30*
Disallowed Sub-Samples Sub-samples that are not allowed even if the sub-sample meets minimum sample size requirements	*Sub-groups for which comparisons are not allowed. Data may not be used to rank service providers.*
Dummy Variables Dummy variables for which coefficients cannot be reported	*Fixed effects for disallowed sub-samples*
Organizations and Groups Organizations and groups for which results cannot be presented separately	*School districts, Hospitals and Health Facilities*
Nested Tables Tables that can be combined into one table	*Should be presented as single table*
Saturated or Near Saturated Models Models that reproduce the data exactly	*Maximum R-squared: 0.4 unless lagged variable Minimum df remaining: 25*
List cases including predicted values An individual case or roster of cases cannot be reported	*Extracts are not allowed*
Weights Do results have to be weighted? When are unweighted counts allowed?	*Unweighted counts for table totals only*
Visualizations Includes illustrations, maps and other visualizations	*Maps must obscure exact locations*

Find out more! dsdr.icpsr.umich.edu