ICPSR's Disclosure Risk Guide for Data Depositors

August 30, 2022

Disclosure Risk Guide Disclaimer

This guide is not intended to provide comprehensive rules for dealing with disclosure risk of all data deposited to the Inter-university Consortium for Political and Social Research (ICPSR). All guidance and resources provided below should always be considered with respect to study context and its particular content. This guide is not representative of all applications of disclosure risk that may be necessary when curating any given study.

What is Disclosure Risk?

Disclosure risk is the degree of risk that data from a study can be linked to a specific person or organization, thereby revealing information that otherwise would not be known or known with as much certainty.

Data that tend to have high disclosure risk include data with information on third party subjects who have not given explicit consent to be in a study. Generally this happens when the respondent who has consented provides information about another person in the course of participating in the study. ICPSR considers all data about the third party subjects sensitive (i.e. potentially harmful) and requiring attention before it can be shared with secondary-users.

Other data with potentially high disclosure risk include data bound by privacy laws (e.g., HIPAA, FERPA, GDRP) and data focused on special populations. Attention must also be paid to any specific guarantees made to the respondents in a consent agreement.

When does ICPSR Conduct a Disclosure Risk Review?

ICPSR implements a disclosure risk review (DRR) of every data and documentation file deposited with us regardless of whether the data is public or restricted, an update of data that were previously released, or data that are publicly available elsewhere. On some occasions, ICPSR may bypass a DRR of a study when a trusted organization previously released the public data and confirmed that experts performed a thorough disclosure review.

The purpose of ICPSR's DRR is to ensure that sensitive information about a specific person or organization is not revealed and to determine the appropriate level of access necessary to protect sensitive information.

What does ICPSR Consider During a Disclosure Risk Review?

ICPSR's DRR evaluates the confidentiality and sensitivity of the data and documentation on two aspects: re-identification (data specifying a respondent or an organization) and possible harm (what might happen to a respondent or an organization if identified). Following DRR, ICPSR engages in remediation to protect study respondents (i.e. top-coding, variable masking, etc.) before disseminating the data at the requested access level.

The amount and type of disclosure risk remediation necessary before ICPSR releases a study depends on the intended access level (public vs restricted), the type of identifiers in the data, and potential data linkages.

Access Level

Access level is determined by ICPSR staff, but data depositors can make recommendations and are encouraged to make note of any particularly sensitive data or concerns they may have regarding the access level.

Public Access: Data with minimal risk of re-identification of research participants and disclosure of sensitive information are categorized as public-use and are publicly available via download and/or online data analysis. Note that ICPSR requires user authentication (e.g., ICPSR MyData) and agreement to terms and conditions before they are granted access and can download data from the ICPSR website, even if data is designated for public access.

Restricted Access: Sometimes data cannot be modified to protect confidentiality without significantly compromising the research potential of the data. For these data, ICPSR restricts access to impose further confidentiality safeguards. ICPSR distributes restricted-use data through several mechanisms:

- Secure Download: Upon approval, researchers will receive an encrypted file via email which they may download and use in a secure location specified in their initial application.
- Virtual Data Enclave (VDE): The VDE is a secure, online environment in which approved users analyze restricted data via a remote desktop environment using several available software options, including SAS, Stata, and SPSS. Researchers do not receive a copy of the data, but rather analyze the data stored on ICPSR's servers. Final analysis output is vetted and, if approved, released to the researcher.

- Physical Data Enclave (PDE): For highly restricted data, ICPSR has a PDE which requires that approved users be on site at ICPSR to use the data. Data use in the physical data enclave is monitored by ICPSR staff. Final analysis output is vetted and, if approved, released to the researcher.
- Secure Online Analysis: This option provides analysis of restricted-use data behind an interface with programmable disclosure protection for selected users. Users submit an application to access the data with this option.

Identifiers

Two kinds of variables often found in social science data present problems that could endanger research subjects' confidentiality: direct identifiers and indirect identifiers. Below are examples of each identifier and the remediation ICPSR implements. Details on each identifier and its remediations are available upon request.

	Definition	**Remediation**	**Examples**
Direct Identifiers	Variables explicitly pointing to particular individuals or groups	● Always masked in public releases. ● Masked in restricted releases with little exception, depending on the level of restriction and utilization of data	● Names ● Personal addresses ● Full dates ● Unique identifying numbers (SSN, account numbers, medical record, vehicle identifiers, institution ID) ● Telephone/email ● URLs or IP address ● Biometric identifiers ● Full-face photographic images ● GPS coordinates
Indirect Identifiers	Variables that can be problematic when used together or in conjunction with other information to identify individuals or groups	● Recoding values to combine responses ● Combinations of recodes and (possibly) masking ● Adjusting the level of restriction	● Race, ethnicity ● Other sociodemographic variables (e.g., exact age, exact income, occupation, institutional affiliation) ● Physical features/characteristics (exact height, weight, BMI, disability) ● Medical history (illness or diagnoses) ● Family characteristics (number of children, household size), ● Sensitive behaviors (drug use, sexual history, crime) ● Geographic information

			(state and county of residence)

Linkages

Even if data is not disclosive at the moment, it may link to previous studies or outside information that is disclosive, or vice versa. For example, if a new wave of a study contains sensitive information that previous waves did not contain, respondents could be identified from earlier parts of the series. Depositors should notify ICPSR of any known linkages to publicly available datasets upon deposit. ICPSR can assist in review and may adjust remediation measures as needed.

Helpful Resources

- ICPSR's Approach to Confidentiality
- Restricted-Use Data Management at ICPSR
- Qualitative Data Confidentiality
- FAQ for Depositing Data
- Guide to Social Science Data Preparation and Archiving

For questions about disclosure risk or preparing your deposit, please contact icpsr-help@umich.edu.