# USAID Data Risk Assessment Model

## An Additional Help for ADS Chapter 579

New Edition Date: 05/09/2024
Responsible Office: M/CIO/ITSD/KM
File Name: 579sad_050924

# Table of Contents

# 1. Purpose

The USAID Data Risk Management Program maintains and implements a Data Risk Assessment Model, which enables the Agency to perform risk analyses on datasets. The purpose of the Data Risk Assessment Model is to define the risk associated with disclosing any given dataset, interpret the risk alongside **USAID's Risk Appetite Statement**, and make that risk information available to inform decisions during the Data Publication Process and Data Sharing Process. This additional help to **ADS 579** provides technical details that show how the Bureau for Management, Office of the Chief Information Officer's (M/CIO's) Data Services Statisticians implement the USAID Data Risk Assessment Model.

USAID's Tiered Access System outlines three access levels for publishing data: 1) Public, 2) Restricted Public, and 3) Nonpublic. These are defined in **ADS 579.3.3.4**.

The USAID Data Risk Assessment Model consists of the following steps, which are described in more detail below:

- Risk Identification,
- Risk Evaluation,
- Risk Mitigation, and
- Risk Acceptance.

# 2. Risk Identification

Risk Identification is the process of assessing a dataset for risk factors. Risk Identification quantifies and qualifies the risk that the analyst will evaluate, mitigate, and accept using the following steps:

- **Step 1:** Consider related datasets both previously published by USAID and other public sources for the impact they could have on the disclosure risks associated with the dataset under review. For example, if a study is recurring, such as a baseline/midline/endline time series analysis, review the information already published in the related datasets and evaluate potential impacts on the dataset under review.
- **Step 2:** Analyze the variables in the dataset under review to determine their impact on the risk (directly identifying, indirectly identifying, sensitive).

- **Step 3:** Remove all direct identifiers[1] about individuals from the dataset under review.

- **Step 4:** Select the key variables to use for risk calculations. Key variables are the indirectly identifying or sensitive variables. Selecting these variables is a subjective process that should involve: 1) discussions with subject matter experts; 2) interpreting relevant current events and other contextual information; and 3) accounting for variables in related datasets previously published (see step 1 above).

# 3. Risk Evaluation

The Risk Evaluation process involves estimating initial disclosure risk using the following steps:

- **Step 1:** Analyze sample frequency counts and apply probability techniques to evaluate probability of re-identification. The probability of the re-identification and risk (see **section 7**) are taken into account together (see **section 8.2**).

- **Step 2:** Consider USAID's Tiered Access System, along with initial disclosure risk estimates, to determine the appropriate access level for the data. USAID's ability to mitigate sensitive data and indirect identifiers to an acceptable risk level has significant influence on the access level of the dataset.

There are other contextual factors which influence the access level and can cause mitigated data to be released at a more restrictive level. This is done to maintain data utility and reduce data disclosure risk. Examples of this include:

- Language of the informed consent for respondents, such as "your data will not be shared with anyone," "your data will be aggregated and shared publicly only in a report," etc.;

- Data collected in highly sensitive and restrictive environments such as countries with political, socio-economic, and religious conflicts; humanitarian crises; pandemics, etc.;

- Sensitive personal information, such as HIV status and sexual orientation;

- Sensitive populations; such as children (education and displaced youth), individuals with disabilities, and individuals who have a history of drug use, crime, assault, or trauma; and

- Data restricted by policy.

[1] Direct identifiers are information that relates *specifically* to an individual, including but not limited to names, addresses, and National ID or Social Security Numbers.

Finally, circumstances do change and in the event that there is a need to further restrict data beyond its original release, USAID Data Risk Management Program Statisticians and other USAID personnel will work together to implement *Post-Publication Emergency Redactions* (see **ADS 579mad**).

# 4. Risk Mitigation

The process of Risk Mitigation seeks to mathematically minimize a data asset's disclosure risk by applying precise and technical changes to the dataset. This process tries to balance the tradeoff between disclosure risk and data utility.

Analysts will apply a variety of risk mitigation techniques to the data to achieve an access level of Public, when possible. Sensitive and indirectly identifying variables need to be manipulated until re-identification risk is acceptable (see **Section 5**). Analysts do this by applying appropriate methods to reduce the re-identification risk. The choice of one or a combination of methods will depend on the data itself and the level of risk.

- **Method 1:** Apply **recoding** or **local suppression** to **categorical** key variables (race/ethnicity, marital status, city, etc).
    - **Recoding** is changing one value to another. Examples include combining smaller groups into one large group (divorced, widowed, and separated can all be changed to "not married") or creating intervals for numeric data if certain values have few observations (ages 24, 25, 26, 27 can all be changed to 24-27).
    - **Local Suppression** is a k-anonymity technique that removes unique values and replaces them with "missing" or N/A.
- **Method 2:** Apply **Microaggregation, noise addition**, or **top/bottom coding** to **numeric** key variables (age, income, years of education, etc).
    - Similar to **recoding**, **microaggregation** groups values and replaces them with another value, instead of a range, to keep the variable numeric. Typical replacement values are the mean/median of the group. For example, **microaggregation** could change an age group of 20-30 to the mean, 25.
    - **Noise Addition** adds or subtracts small values to data within a variable. Commonly, noise will be added so that the changes have a mean of zero. For example, if the ages in our dataset were 10 and 20, they could be changed to 12 (10+2) and 18 (20-2), where for statistical purposes the average age would still be 15, which helps to minimize information loss.
    - **Top/Bottom Coding** applies **recoding** to the extremes of the dataset (recoding the highest and/or lowest values). If the only outliers are older ages, for example, you can **top code** ages 85, 89, and 94 to 85+.

Choice of appropriate mitigation techniques depends on the structure of the key variables. Analysts will use recoding together with local suppression when the number of unique combinations of key variables is low. Microaggregation is applied when it is beneficial to researchers to maintain numeric variables.

One of the challenges of risk mitigation is that with higher risk protection, data utility decreases. Following mitigation, re-evaluate disclosure risk measures and utility (information) loss in the mitigated data to estimate the level of privacy protection and the utility of the data for future analysis. If the information loss or the disclosure risk measures are unsatisfactory, the analyst will repeat the risk mitigation process from the beginning, altering the techniques to produce a different outcome. If all attempts to mitigate the data lead to significant loss of utility, consider a more restrictive access level. A restricted but usable dataset is usually more valuable than a public dataset stripped of any utility.

# 5. Risk Acceptance

Risk Acceptance is the acknowledgement of an estimated risk and the delivery of release recommendations to appropriate officials. For example, with Development Data Library (DDL) data assets, clearance offices for risk acceptance include the Bureau for Management, Office of Management Services, Information and Records Division (M/MS/IRD) (Freedom of Information Act [FOIA]), the Office of Security's (SEC) Information and Industrial Security (IIS), M/CIO's Information Assurance Division (M/CIO/IA) (Privacy), and the originating Operating Unit (OU).

The analyst will share the mitigated dataset along with the estimated disclosure risk, information loss, suggested access level, and other supporting documentation with the officials for approval prior to release. Combined, these make up the data risk assessment package.

Once the relevant officials have approved the data risk assessment package, the mitigation workflow is complete. The remainder of this document expands upon the core components of the risk assessment model: **risk identification, evaluation,** and **mitigation**. The following sections provide examples to clarify what makes a variable directly identifying, indirectly identifying, and/or sensitive. They also include granular information about how an analyst can handle specific data characteristics and common variables often present in datasets.

# 6. Risk Identification Deep Dive

## 6.1 Common Data Characteristics

The table below displays common dataset characteristics and recommendations for addressing them. Note that this is general guidance. All characteristics should be considered on a case-by-case basis.

| Dataset characteristics | Considerations | Recommendations |
|---|---|---|
| Small sample size | Disclosure should be based on the sensitivity of data. | May be appropriate to release at the restricted public or nonpublic access level |
| Outlier cells | Suppression may be required. | If retained, restricted public or nonpublic access level. |
| Aggregated data | Sex disaggregation is required. Location, age, urban/rural disaggregation is recommended. | Ensure that USAID disaggregation requirements are met. (See **ADS 205** and **ADS 201** for more information) |
| Small values in aggregate data | Suppress values below acceptable thresholds to prevent reverse engineering. | If aggregation fails or USAID disaggregation requirements cannot be met, consider suppressing values or collapsing categories to maintain public access. |
| Small geographical area | Small geographic areas, when combined with other variables, increase the risk of re-identification. | This should be decided on a case-by-case basis, please see **ADS 579mab Activity Location Data** and **ADS 579saa Geographic Data Collection and Submission Standards** |
| Longitudinal data (data with multiple observations per participant over time) | Mitigations are complex to implement without accepting disclosure risk or reducing data utility. | Apply additional mitigations if the nature of the data will compromise protection strategy over time. Consider restricted public or nonpublic access levels. |
| Free-form text | It is resource intensive and challenging to identify disclosure risks in free-form text. | Remove non-recoded free form text from the dataset prior to release. |
| Non-English language | Data that cannot be fully and reliably evaluated for disclosure and security risks cannot be released. | Request a translated version of the asset or remove data and documentation containing non-English text prior to release. |

## 6.2 Direct Identifiers

Prior to disclosure, remove information that directly identifies individuals or respondents, unless they consented to the disclosure. Identifiers of institutions may also be removed, in accordance with Agency policy, for reasons such as safety and security of associated individuals. Analysis of consent requires a separate examination process, outside the scope of this Additional Help document.

| Common Direct Identifiers | | |
|---|---|---|
| ● Name<br>● Biospecimens/DNA<br>● Dates of birth<br>● Email addresses | ● IP Addresses (home or personal computer)<br>● Link to personal social media accounts<br>● National ID number (e.g. social security number, passport, drivers license)<br>● Phone/fax number | ● Photographs of individuals<br>● Position of employment<br>● Precise location data (coordinates, addresses)<br>● Student,employer, or similar ID numbers |

When direct identifiers[2] are relevant to secondary analysis, consider transforming them into an indirect identifier. For example, an individual's date of birth can be recoded to their age.

### 6.3 Indirect Identifiers

Indirect identifiers do not directly identify, but still provide some information about an individual. Indirect identifiers may not appear problematic on their own but, when used in combination, they can provide enough information to re-identify individuals.

Remove or modify indirect identifiers based on their level of risk. This ensures the data retain the greatest value and utility for secondary research and analysis while still protecting the privacy and confidentiality of data subjects.

## 7. Risk Evaluation Deep Dive

### 7.1 Impact Risk of Different Variables

This step focuses on identifying and categorizing variables within the data as sensitive and/or identifiable. Remove or mitigate sensitive and identifiable information that violates confidentiality expectations or that can be used to determine a data subject's identity. Note that the following list is not exhaustive, and there can be overlap between what is sensitive and what is identifiable.

[2] Federal de-identification standards are often informed by the HIPAA Privacy Rule. Synthetic UIDs generated by the researcher should not be considered direct identifiers.

| High Sensitivity |
| --- |
| ● Information marked as “Sensitive but Unclassified,” “For Official Use Only,” “Controlled Unclassified Information”<br>● Information collected under a promise of confidentiality<br>● Information related to deliberative processes<br>● Information related to criminal investigations or law enforcement practices |

| Sensitive | Identifiable |
| --- | --- |
| **Definitions** | |
| Data are sensitive if disclosure would reveal locations, resources, associations, behavior, or other information that may cause physical, financial, psychological, or reputational harm. | Data are identifiable if they can be used to help determine the identity of an individual data subject, or to connect or associate an individual data subject with a vulnerable population, institution, or household represented in the data. |
| **Location** | |
| ● Refugee/migrant camps<br>● Stock facilities/warehouses<br>● Duty station in conflict-affected area<br>● Port of entry<br>● HIV clinic or other related service delivery location | ● Place of birth<br>● School address or specific location<br>● Building or dwelling characteristics<br>● Address (residential or employment)<br>● Geographical subdivision with populations under 100,000 |
| **Personal** | |
| ● Political association(s)/activism<br>● Sexual preference(s)<br>● Signature<br>● Fingerprint<br>● Bios (e.g. blood or DNA samples)<br>● Photograph<br>● Video/audio recording<br>● AIDS/HIV status<br>● Physical and psychological impairments<br>● Status of displaced populations<br>● Criminal background<br>● Alcohol/drug addiction<br>● Sex<br>● Income<br>● Religion/ethnicity | ● Number of children<br>● Household size<br>● Family or household associations<br>● Height/weight<br>● Dwelling<br>● Sex<br>● Date of birth<br>● Age<br>● Language spoken<br>● Marital status |

| Employee | |
|---|---|
| • Duty station in conflict-affected areas<br>• Performance reviews/ratings<br>• Employee and labor relations complaints/grievances<br>• Information protected by statute (e.g. whistleblower identity)<br>• Disability/reasonable accommodations<br>• Responses to opinion surveys | • Salaries<br>• Name<br>• Job title<br>• Position<br>• Grade level |
| **Business/Proprietary** | |
| • Management plan<br>• Detailed financial information<br>• Accounting method<br>• Proprietary source codes<br>• Monitoring information and reports<br>• Research/development information<br>• Analytical reports<br>• Trademarked or confidential operating procedures<br>• Tools and resources used to manage programs | • Area of specialization<br>• Number of employees<br>• Market share<br>• Budget information |

# 8. Risk Mitigation Deep Dive

## 8.1 Considerations for Data Mitigation based on Common Variables

Individual level data, or data where each row pertains to one individual, contained in a restricted public data asset may be released to qualifying research entities based on documented need and a data sharing agreement. Individual level data which has been mitigated and approved by appropriate clearance officials can also be released publicly with proper risk mitigation. The following table contains variables that are often deemed sensitive or indirectly identifying and highlights the considerations and mitigation strategies used to reduce risk. Note that these recommendations should always be considered on a case-by-case basis and take into account any relevant context.

| Variable | Considerations | Mitigation Recommendations |
|---|---|---|
| Location | Different countries use different naming conventions for location granularity. Understand the particular country's definitions before performing | **Recode** any locations with small sample size and consider **suppressing** any granularity levels that may be too refined. |

| Variable | Considerations | Mitigation Recommendations |
|---|---|---|
| | mitigations. | |
| Age | Exact dates (1/1/2000), individual ages, or month and year (09/2000) should not be disclosed except when disclosure does not increase risk of re-identification. | **Microaggregation** or **recoding** |
| Sex | Sex must be disaggregated whenever possible[3]. | Male/Female |
| Marital Status | Ensure no unique or near-unique entries. Aggregate up to ensure k-anonymity met. | **Recoding** can be used for low-frequency entries (for example divorced, separated, or widowed changed to not married). |
| Employment Status/Income | Treat outliers. | Collapse income into intervals through **recoding** or **microaggregation.** |
| Education | Treat outliers. | If needed, can use **recoding** to group into intervals or common thresholds (primary, secondary, university, graduate). |
| Household Size and Number of Children | Context-specific: In some regions, small household sizes will be the outliers. | **Top/Bottom Code** extreme household sizes/number of children |
| Race/Ethnicity | Treat outliers. | **Recode.** Underrepresented groups can be combined and reported as “other”. |
| Language Spoken | Common groupings could be: National Primary, Secondary and Other. | **Recode.** Underrepresented groups can be combined and reported as “other”. |
| Country of Birth | Decisions should be based on frequency, or tailored to the specific research question. | Can be **recoded** regionally if appropriate to keep within the data. |
| Religion | Suppress outliers and consider context. Religious minority groups may be considered vulnerable populations. | **Local Suppression** to remove outliers. |

[3] ADS 205.3.7

| Variable | Considerations | Mitigation Recommendations |
|---|---|---|
| Provider/Household/ Institutional ID | If a variable is not essential to secondary analysis, the variable should be redacted. | In cases where multiple datasets are presented, Provider/Household/ Institutional ID Data may be substituted by replacing the identifying ID with a random number to preserve any relevant links in the datasets. |
| Site ID | If a variable is not essential to secondary analysis, the variable should be redacted. | Site ID or Information Data may be substituted by replacing an ID with a random number. |
| Investigator ID | If a variable is not essential to secondary analysis, the variable should be redacted. | Synthetic data may be substituted by replacing an ID with a random number to preserve any relevant links to other datasets. |
| Medical Diagnosis | Treat outliers. | **Recode** |
| Treatment Regime | Treat outliers. | **Recode** |
| Household characteristics (Roof material, household size, land dimensions) | Treat outliers. | **Recode** |
| Water source | Can be a highly sensitive variable dependent on context. If recoding is not practical, the variable could be suppressed. | **Recode** |
| Occupation | Treat outliers. | **Recode** |

After application of various mitigation techniques, k-anonymity is typically the final risk mitigation technique for individual-level data.

## 8.2 Applying K-Anonymity

With k-anonymity, sensitive information and indirect identifiers are retained or mitigated. To apply a k-anonymity standard, continue previously mentioned mitigations on indirect identifiers until the k-anonymity threshold is satisfied.

To achieve a k-anonymity threshold, there must be at least *k* individuals in the dataset sharing any given combination of determined key variables. As an example, if in a dataset there was an instance where only four individuals shared a common age, sex, and religion, this scenario would not satisfy the threshold

level of k=5, and the dataset would require more mitigations prior to publication. The k-anonymity thresholds below reflect the sensitivity of the data; these are recommendations that can change depending on context.

| Asset Properties | K-Anonymity Threshold |
|---|---|
| Agricultural, employment, education level, and non-sensitive opinion surveys | k = 3 |
| Behavioral, substance abuse, health surveys, or information regarding vulnerable populations, e.g. high-risk ethnic groups, individuals with physical or psychological disabilities, children, pregnant women, sex workers, prisoners, displaced populations, non-conforming, or other groups that may be targeted for discrimination, stigmatization or exploitation) | k = 5 |
| Sensitive health or medical information, e.g. HIV status, political opinion or engagement surveys that may be targeted by local governments, state actors, and Information about criminal activity | k = 10 |

579sad_050924