MCC Evaluation Microdata

Data Package

# Instructions

This template is informed by MCC’s Evaluation Microdata Documentation and De-Identification Guidelines. In addition to reviewing these Guidelines, MCC contractors responsible for preparation and documentation of evaluation-related microdata for public and/or restricted-access use should be familiar with the following US government guidelines for data de-identification and re-identification:

* [NIST 2015](http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf)
* [NIST 2016](http://csrc.nist.gov/publications/drafts/800-188/sp800_188_draft2.pdf)

 **MCC, the evaluator, and stakeholders should consider the following multi-stage process for data review and release**:

1. Evaluator and M&E PM should agree on expected DRB review date as early as possible to confirm. This should be scheduled at least one month before Evaluator’s contract expires.
2. Evaluator should submit full package to M&E PM. The package includes:
* One completed Section 1 of the DRB Data Package Worksheet for ALL data components (i.e. individual, household, and community data for one survey round are three data components with different risks)
* One completed Section 2 & 3 for EACH data component
* Datasets and code package(s)
* Informed consent(s)
* Questionnaire(s)
* Most recent Metadata file (for [Evaluation Catalog](https://data.mcc.gov/evaluations/index.php/catalog) entry)
1. M&E PM should review Metadata and [DRB](http://intranet.mcc.gov/department/DPE/Team/ME/Data%20Protection/2.%20Data%20Documentation%20and%20De-Identification/2_DRB%20Data%20Package%20-%20Cover%20and%20Worksheet.docx) Data Package Worksheet for clarity and completeness. This may require one round of revision based on the M&E PM requests for clarity and completeness.
2. Evaluator should submit full package to M&E PM. M&E PM and the M&E DRB members should establish a first-round review and feedback to the Evaluator on the proposed data de-identification process. This may require a second round of revision to the package.
3. Evaluator should submit full package to M&E PM for the confirmed MCC DRB review date at least 2 weeks prior to confirmed DRB review date.
4. If any feedback/revisions are required following MCC DRB review, Evaluator should revise and resubmit full package to M&E PM with documented responses to MCC DRB feedback to ensure timely virtual review and clearance of the full package. All final de-identification efforts and their impact on verification of analysis should be documented in the evaluator’s Transparency Statement available on the Evaluation Catalog.

All red font text are instructions in the Worksheet and must be replaced with standard black font with the contractor’s response.

**Unless otherwise agreed with MCC, the final document will be made public to complement/underlie the contractor’s Transparency Statement to document the data preparation and de-identification process required for the public and/or restricted-access microdata and any impact on the data for verifying evaluation analysis and broader data usability.**

# Section 1: Cover Sheet

**Overview of Data Package**

(*Instructions: Include a paragraph summarizing each data package component included in the package. For example, if the package includes household, individual, and community level data sets, please include a paragraph summarizing each of these three components, including information on the content and timing of the data collection*.)

This data package includes the following components:

* Data Package Component 1:
* Data Package Component 2:

**Complementary Data**

(*Instructions: Complementary data collection efforts are those efforts that complemented the data packages under review for de-identification, but do not necessarily require de-identification. The evaluator should list these data and provide a brief summary on how they connect to any data package components and affect the data package components’ de-identification. For example, if the geospatial data for the project infrastructure is collected and will be publicly released, it should be listed in the complementary data collection efforts*.)

This data package considers the following complementary data efforts:

* Complementary Component 1:
* Complementary Component 2:

**Data Package Folder Contents**

(*Instructions: Please list the Data Package Component File Name, and then include the File Names of each of the corresponding required documents [Metadata, Worksheet, Informed Consent, Questionnaire, Other docs]. Only one de-identification worksheet per survey is requested unless discussed.)*

Table 1: Data Package Components

|  |
| --- |
| **Data Package** |
| **Component** | **Worksheet** | **Informed Consent** | **Questionnaire** | **Other Documents** |
| Data Package Component #1 Name |  |  |  |  |
| Data Package Component #2 Name |  |  |  |  |
| Data Package Component # Name |  |  |  |  |

# Section 2: Data Component Preparation Overview

|  |  |  |
| --- | --- | --- |
|  | **Response** | **Discussion/Explanation** |
| Data + Code Completeness | Complete | Response | *To be considered Complete: The available data must allow new users to replicate evaluator analysis to the extent allowable by providing the full data set + analysis code. The constructed variables may also be included in a dataset, but if the dataset+code produces those variables, it is not necessary.**To be considered Incomplete: The available data only provides a sub-section of data as produced by the survey and/or the constructed variables only. Incomplete data files are limited in terms of full verification of analysis and/or broad usability of data and must be justified.* |
| Incomplete |
| Data Round(s): | Baseline only | Response | *MCC is willing to trade-off broad use of individual rounds for more consistent de-identification protocols across rounds of data. Therefore, unless there is specific demand for the baseline/interim only data, or contractual requirements, MCC prefers contractors to prepare all data rounds in one package.**If one stage only – please (i) confirm demand and/or contractual justification and (ii) discuss how preparation and release of this data as presented to the DRB may affect future data round releases.* *If combination, please discuss if this file replaces any previously published datasets.* |
| Interim only |
| Endline only |
| Combination of rounds |
| Informed Consent and IRB | High restriction | Response | *MCC assumes DIRECT identifiers are always removed from any public-use file. With this assumption: Please refer to the informed consent statement – does it require: High restriction: access to data that includes indirect identifiers is limited to the contractor only; Medium restriction: access to data that includes indirect identifiers is limited to the contractor and qualified researchers, including MCC; Low restriction: data with indirect identifiers may be made public.**Please discuss how the promises of confidentiality in the informed consent informed de-identification efforts. Please include any additional guidance provided by the IRB as applicable.* |
| Medium restriction |
| Low restriction |
| Geographic Identifiers | Highest (i.e. Province) | Avg. pop size | Identify/De-identify/Remove | *Please provide justification on the identification/de-identification/complete removal of specific geographic regions. De-identifying at a higher geographic level may support privacy protection, but it may also reduce data usability. Please provide justification for recommendation.* |
| --(i.e. District) | Avg. pop size | Identify/De-identify/Remove |
| --(i.e. State) | Avg. pop size | Identify/De-identify/Remove |
| --(i.e. Village) | Avg. pop size | Identify/De-identify/Remove |
| Lowest --(i.e. Census Blocks) | Avg. pop size | Identify/De-identify/Remove |
| Knowledge of Treatment  | High risk | Response | *In some cases, general knowledge of treatment areas and/or inclusion of a treatment variable can significantly increase re-identification risk depending on the population affected. Please provide assessment of this re-identification risk and recommendation if considered high/medium risk.* |
| Medium risk |
| Low risk |
| Publication Type | Public-use only | Response | *Please state for this data package: will there be public-use data only, restricted-use data only, or both and provide justification as this relates to enabling verification of evaluation results and/or broad usability of the data.* |
| Restricted-use only |
| Both |

# Section 3: Data Component Preparation Details

| **Specific Issues** | **Risk Analysis** | **Risk Mitigation** |
| --- | --- | --- |
| *Instructions* | *Response* | *Instructions* | *Response* |
| 1. | Who has significant financial, legal, cultural, or other incentives to re-identify survey respondents?  | *List all potential threats[[1]](#footnote-2)* |  |  |  |
| 2. | What is the potential value to these intruders?  | *List all uses (for example:* capture delinquent tax payments, or stigmatize the respondent) |  |  |  |
| 3. | What is the expected cost to these intruders to re-identify the data? | *Describe degree of difficulty for re-identification* |  |  |  |
| 4. | Assess availability of ‘linkage’ data that can be used to re-identify respondents. This includes other datasets or archives with information that can be used to re-identify individuals in the dataset. | *List all potential existing data* |  | *Describe how to mitigate link to existing data that enables re-identification* |  |
| 5. | **Identity Disclosures:** What are the DIRECT identifiers in the raw data?  | *List the DIRECT identifiers* (names, addresses, geographic information, government-issued ID numbers, etc.) |  | *List all DIRECT identifiers removed from the dataset.* |  |
| 6. | **Attribute Disclosures:** For GIS/GPS data, this distance data can be a direct identifier that is VERY useful analytically. Therefore, please describe how GIS/GPS data VALUE/USABILITY can be retained. | *List all GPS and/or GIS data.* |  | *Describe process for de-identification. For example: introduce random errors into geographic data (GPS, GIS, etc.).* *Displace urban points 0-2 km, rural points 0-5 km, and additional 1% of rural points 0-10 km[[2]](#footnote-3).* |  |
| 7. | **Attribute Disclosures:** What variables have OUTLIERS that create INDIRECT identifiers are in the raw data? | *List the identifying items/variables* |  | *Describe top/bottom coding: set upper & lower bounds to remove outliers for continuous. Specify: are values set to the median, or other?* *For large categories/datasets, the OMB suggests top coding at least the highest .5%; for smaller categories/datasets, top code the highest 3-5%. The same principles apply to bottom coding.[[3]](#footnote-4)* |  |
| *Describe any variables that require collapse and describe construction of new variable* |  |
| *Describe any global re-coding to group observations into categories (e.g., age 0-5, 5-10, 65+, etc.). Ensure that the categories are neither too broad nor too narrow.* |  |
| 8. | **Attribute Disclosures:** What variable combinations produce UNIQUE observations that create INDIRECT IDENTIFIERS (for example: individuals with high incomes, ages, or unique combinations, such as 17-year old widowers or contextually unusual racial/ethnic backgrounds) | *List the identifying items/variables:* |  | *For each identified rare data, describe the local suppression techniques employed to remove unique and rare data. Specify: are values set to missing, the median, or other?* |  |

1. As stated in NIST 2016, de-identification practitioners should assume that de-identified US government datasets will be subjected to sustained, world-wide re-identification attempts, and they should gauge their de-identification requirements accordingly. Although a specific dataset may not be seen as sensitive, de-identifying that dataset may be an important step in de-identifying another dataset that is sensitive. Alternatively, the adversary may merely wish to embarrass the US government agency or its partners. Thus, adversaries may have a strong incentive to re-identify datasets that are seemingly innocuous. [↑](#footnote-ref-2)
2. ICF International, Demographic & Health Surveys [↑](#footnote-ref-3)
3. Office of Management and Budget, Checklist on Disclosure Potential of Proposed Data Releases ([current link](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&ved=0CCsQFjAA&url=http%3A%2F%2Fwww.fcsm.gov%2Fcommittees%2Fcdac%2Fcdac.html&ei=UN9vUpvxDZWt4APZyYD4Bg&usg=AFQjCNFwhGwSnNTfTDllVwmYgpJ2rdKEsg&bvm=bv.55123115,d.dmg)) [↑](#footnote-ref-4)