Anonymization Protocol

1. Objectives

In order to release the SDI Health dataset as a Public Use File, it is necessary to ensure the privacy of its participants. To this end, and with the intention of avoiding the re-identification of the health facilities and healthcare providers, the following measures are taken:

1. Deletion of variables or labels that contain confidential information or could lead to re-identification.
2. Deletion of value labels that contain confidential information or lead to re-identification.
3. Recoding of variables that could lead to re-identification of the observations.
4. Trimming and/or censoring of specific unique values and outliers that might allow re-identification.

2. Results

The resulting datasets keep the usefulness of the data intact while greatly protecting the privacy of the respondents and reducing the identification risk. Most harmonized variables (>80 %) are unaffected by the anonymization process. In very few instances, some variables are relabeled to avoid the publication of the exact questions/answers of confidential assessment instruments.

All identifiers are being cleaned (deletion of value labels). Some variables are being recoded into categorical variables to ensure confidentiality.

3. Disclosure risk and confidentiality protection

Microdata often contain confidential or sensitive information, which makes release of these datasets in their original form impossible. Release of the data could reveal this confidential information and lead to a breach of privacy of the respondents. This has ethical issues and, in many cases, legal implications. Furthermore, when confidentiality in not guaranteed, current and potential future respondents are less likely to be willing to respond to future surveys.

The aim of this process is to create a Public Use File (PUF), which is a dataset that is freely accessible to the greater public, while taking into consideration the privacy issues. This PUF must also minimize as much as possible unnecessary disturbances to the original to preserve the usability and quality as much as possible when possible.

Risk in the Statistical Disclosure Control context is the probability or likelihood that disclosure by an (hypothetical) intruder of a record occurs. Disclosure can be **identity disclosure**, when the identity of an individual or entity in the dataset is revealed, or **attribute disclosure**, when the intruder gains new (confidential) information from the dataset. Identity disclosure can imply attribute disclosure. The risk is dependent on several factors, amongst others the frequency of **keys** (i.e. combinations of values of key variables), sample size and sampling weights as well as

the availability of external information to intruders to use for re-identification. The disclosure scenarios for a particular dataset describe these parameters and the way an intruder can use a dataset to gain new information.

The acceptable level of risk depends on the release type, e.g. scientific use file, public use file, or other ways of release and the sensitivity of the data. This dataset was prepared to be released as **PUF and hence needs a higher level of protection**. Also, the potential harm caused by disclosure should be considered when determining the acceptable risk level.

In similar microdata releases with, for instance, business survey data, the geographical level is highly reduced, large companies are suppressed and the level of detail in the data is reduced to protect the records. Generally, the period between the survey and data release is also specified, e.g. 1 year. It should be noted that a complete elimination of disclosure risk is not possible.

4. SDI Health dataset

The SDI Health dataset consist of a series of country-year surveys, each containing information on health facilities and healthcare providers. The main concern for re-identification and confidentiality are the healthcare providers. However, since the data are hierarchical, i.e. healthcare providers belong to health facilities, the re-identification of a health facility might lead to the re-identification of healthcare providers too.

5. Actions taken

The Anonymization process is done with the aid of the statistical software Stata. All anonymization steps are reproducible with the Stata script for each of the detailed datasets. The process starts from the harmonized dataset: before the start of the anonymization process, any final data quality corrections are made.

Each anonymization script covers the following steps:

1. Identification of ready-to-release variables.
2. Identification and removal/anonymization of variables due to the sensitivity of its data. Identification and removal/anonymization of variables due to the distribution of the data that could lead to high risks of re-identification.
3. Identification and recoding of variables into categories to deal with outliers (top recoding), sample unique and continuous variables whose values represent high risk of reidentification (special unique).

5.1 Ready to release variables.

There is a subset of variables that do not imply considerable risks of disclosure. These variables were identified and revised. We proceeded to select them considering:

- General information: Country, survey year, urban, etc. This information is preserved and

there is no risk associated with the release.

- Randomized id/keys: region, district, province, etc. This information is shared considering it doesn't contain descriptive information. In general terms, it preserves the variability of the data but doesn't provide an associated value label.
- Specific information: a subset of infrastructure variables, a subset of assessment variables, other facility and provider characteristics, etc. These variables gathered through SDI survey represent a high risk of breach of confidentiality. We analyze each of these and their distributions to check if unique values allow for reidentification.

5.2 Identification of variables to delete

The SDI surveys gather data that includes sensitive information: names, financial information and specific descriptions, among others, that should not be available to the public. To avoid its disclosure, some of them are directly removed from the database and others transformed to a "Confidential" or ".c" value, allowing only the knowledge of their existence (with release to be considered upon request).

There are other variables that do not represent a risk of disclosure but include important information that could damage future waves of surveys. These variables contain detailed descriptions of the assessments and their correct answers. All of them are removed, keeping only general information on the type of question, its result/score and the label associated to interpret the latter.

We also identify variables with distributions that represent high risks of reidentification and for which recoding and other perturbative methods cannot account for the mentioned risks. These variables contain sample unique (a unique combination of values for the selected categorical key variables in the dataset and is at high risk of re-identification) and are represented mainly by descriptions/details of "Other" categories. It is not possible to share the information contained in any of the respective variables due to the specificity of the answers.

5.3 Recoding of variables

Some variables represent a risk of reidentification because of their composition and distribution. In order to share the contents, it is necessary to recode them in categories. The recoded variables cover both continuous (e.g. age, number of employees, etc.) and specific categorical variables in which certain categories are too scarce (e.g. position in the establishment, etc.).

The recoding consists in trimming tails of distribution or top recoding, transforming continuous values into ranges (e.g. using decades instead of years), and/or broadening categories to group possible answers (e.g. "Owner/Director/Facility-in-charge" into one value). Lastly, rather than grouping the whole distribution of values into categories, few variables' values are censored or anonymized when their specificity/uniqueness (e.g. excessively large values, rare values, etc.) might allow re-identification. As part of the recoding process, we ensure that all identifiers have no value labels that could contain specific information that leads to the reidentification of facilities or healthcare providers.