Tanzania Health Microdata Statistical Disclosure Control Process

External Report

1. 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 and in many cases legal objections. Furthermore, when confidentiality in not guaranteed, current and potential future respondents are less likely to be willing to respond in future surveys.

Statistical Disclosure Control (SDC) refers to (i) a set of methods to measure the risk in a dataset that confidential information may be disclosed when releasing a dataset and (ii) to the set of methods to treat the data in order to prevent the release of confidential information when releasing the dataset. SDC is used in many statistical offices to anonymize data before release. Templ et al. (2014) give a concise introduction to SDC. We refer to Hundepool et al. (2012) and the references therein for a complete overview of the SDC process and a detailed description of the methods.

In this report we describe the methods applied to mitigate this risk with the emphasis on the implications for the user. Anonymization leads to changing the dataset.

2. SDI Tanzania health data set

The SDI Tanzania health dataset contains information on a sample of 403 facilities. The statistical objects in this dataset are health centers and health workers. The main concern for re-identification and confidentiality are the health workers. However, since the data is hierarchical, i.e. health workers belong to health facilities, the re-identification of a health facility might lead to the re-identification of a health worker too.

The datasets consists of four modules that contain information on:

- Module 1: Facility information
- Module 2: Staff roster
- Module 3: Patient case simulations (to measure provider knowledge)
- Module 4: Facility expenditure, resources, and governance (not released)

3. Risk before anonymization

Risk in the SDC 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 correctly 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 (SUF), public use file (PUF), or other ways of release and the sensitivity of the data. This file will be released as PUF and hence needs a higher level of protection. Also, the potential harm caused by disclosure should be taken into account when determining the acceptable risk level. The SDI health dataset is not a typical example of a health microdata set, since it does not contain information on the health status of the respondents.

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. In case of the SDI health survey, the period between the survey and the data release introduces already uncertainty into several variables, such as number of hospital beds. It should be noted that a complete elimination of disclosure risk is not possible.

For categorical variables, risk measures are based on the frequency count of **keys**, the combination of values for categorical key variables. Key variables are the categorical variables that are known to the intruder or available in external datasets and can be used for re-identification. Disregarding the sample weight, the higher the frequency of a key, the lower the risk. This principle leads to the idea for the risk measure **k-anonymity**; it is a count of the number of records with keys that are not shared by at least k records. K-anonymity can also be used as a threshold when anonymizing a dataset. If 2-anonymity is violated we speak of a **sample unique**. A sample unique has a unique combination of values for the selected categorical key variables in the dataset and is at high risk of re-identification (depending on the sample weight).

Sample uniques can be further classified as **special uniques**: these are records that do not need the complete set of selected key variable to reach uniqueness in the dataset, but instead a subset of the key variables suffices to create uniqueness of the record. The **SUDA** (special uniques detection algorithm) score is based on this definition of special uniques. Special uniques are more likely to be re-identified by a possible intruder.

Based on the disclosure scenarios and the selected key variables, we can evaluate the disclosure risk for all scenarios. The risk measures are based on the prepared data.

4. Overview of anonymization methods

In this section we give an overview of commonly used anonymization methods. In case the disclosure risk is too high in the data set, there are several methods to reduce this risk before the release. The most common approach is to reduce the detail in the data by **recoding**, which is a method to reduce the number of categories in the data by combining several categories. An example is to combine several regions or industry types in more aggregate categories. Often this can be realized by using a higher level geographical or more aggregate industry type without losing valuable information for data users. A variation to recoding is **top coding**, where high values of a certain variable, which are often outliers, are replaced by one common value. An example is to replace the age values over 60 with 60. Continuous variables can be transformed into bands, e.g. income bands.

In case recoding does not reduce the risk sufficiently, individual values can be removed by using **local suppression.** Local suppression algorithms seek to suppress values that cause uniqueness of a record. Suppressing or removing these values guarantees that these records cannot anymore be identified based in these values. Here also (rare) combinations of values are considered.

There are several other methods to introduce uncertainty in the microdata, which prevent an intruder knowing whether an identity disclosure is correct or not. These methods perturb the data and are called **perturbative** methods. One popular method for categorical variables is **PRAM**, which changes the values in a categorical variable at random. **Noise addition**, which is based on adding small distortions to continuous variables, is often used for creating uncertainty around values of continuous variables. Perturbative methods are generally only reverted to if non-perturbative methods do not provide sufficient protection. The reason is that perturbative methods might lead to a high level of information loss.

Removing variables

The released microdata set has only a selected number of variables contained in the initial survey. Not all variables could be released in this PUF without breaching confidentiality rules. Direct identifiers (e.g. facility and health worker names, enumerator names), other variables that include identifying information (e.g. destination of last ambulance trip) and geographical information were removed. Table 1 gives an overview of the removed variables.

Table 1 Overview of removed variables by module

Variable	Description
Module 1	
m1saq3	First Administrative Level
m1saq4	Second Administrative Level
m1saq4a	Third Administrative Level (Ward)
sub_reg_1	Province
sub_reg_2	County

m1saq5	Health Facility Name
m1saq6	Health Facility Survey Code
m1saq8	GPS Position S
m1saq9	GPS Position E
m1saq11	Name of enumerator during first visit
m1saq15	Name of enumerator during second visit
m1saq18c	Supervisor's name
m1sbq1	Name of respondent
m1sbq2a	Permission to collect contact (Phone No.)
m1sbq2	Respondents contact (Phone No.)
m1scq5e	Source of power, other
Module 2	
m2saq4	Personnel ID (Visit I)
m2saq5	Name of Personnel (Visit I)
m2sbq1	Personnel ID (Visit II)
m2sbq2	Name of Personnel (Visit II)
m2sbq3	Personnel ID in Visit I
Module 3	
m3saq1	Name of Health Facility
m3saq2	Health Facility number
m3saq3	Health Facility code
m3saq5	Clinician identifier
m3saq7a	Enumerator playing Observer
m3saq7b	Enumerator playing Patient
m3saq10	The reason of refusal for participation
Module 4	
No release	

Randomization

To further anonymize the data, following variables are randomized: ID of facilities within the state and staff_ids within the fac_ids. Randomization allows to use these variables for the evaluation of fixed effects.

Reducing detail in variable by recoding and top-coding

In several variables, the detail is reduced by recoding values or top-coding. The following table gives an overview of the variables that have been changed.

Table 2 Overview of recoded variables

Variable name	**Description**	**Approach used**	**Before**	**After**
Health facility level				
m1saq7 & m1saq3	Rural/urban & Region	Recode rural, urban and capital to strata	1; 2,3 & region_ids	Strata variable: 1; 2; 3
m1saq12-13, m1saq16-17,	Start- and end-time of survey	AM/PM	Exact time HH:MM	1; 2

m1sbq4, m1sbq5, m1sbq6, m1sbq7	Ownership	Recode to public/ private /other	1; 2-5; 9	1, 2, 9
m1sbq10	Travel time	Recoded as <=2, 3 and 4+ hours	Exact time HH:MM	<=2; 3, 4+
m1sbq11	Days per week open	Recode 0-6 to 5 (as median);	0,1,2,3,4,5,6; 7	5; 7
m1sbq12	Consultation time	Recoded as <=9, 10-12, 13-23, 24 hours	1-24	9; 12; 23; 24
m1sbq38	Travel time	Recoded and topcoded	Exact time HH:MM	1; 2; 3+
m1scq1	Main power source	Recode to no electricity, electricity	1,9; 2,3,4,5	0; 1
m1scq6	Main water source	Recode to unimproved, improved	1,6,8,10,11,12,13, 14, 15; 2,3,4,5,7,9	0; 1
m1scq10	Walking distance	Recoded and topcoded	Exact time HH:MM	1; 2; 3+
m1scq12-16	Number of toilets	Topcoded at 6		0,1,2,4,5,6+
Health worker level				
m2saq6, m2sbq4	Position	Recoded to management and staff	1, 2, 3; 4, 5, 9	1; 2
m2saq11, m2sbq10	Age	<30; 30-39; 40-49; 50-59; 60+	14-82	<30; 30-39; 40-49; 50-59; 60+
m2saq2, m2saq3	Number of staff	topcoded at 50 and 30		1-49, 50+ 1-29, 30+

To guarantee confidentiality, the several other variables are recoded or removed:

- Dates m1saq10, m1saq14 and m1saq18b are recoded relative to date of the first survey.
- Variable m1sbq13 (outpatient attendance) is recoded as ratio over the number of staff who regularly consults and top-coded to 500.
- Variable m1sbq15 (number of inpatient) and m1sbq16 (number of inpatient bed days) are recoded as ratio over the number of staff who regularly consult and top-coded to 200, respectively 300.
- Variables m1sbq17 (number of beds), m1sbq18 (number of beds for hospitalization) are recoded as ratio over the number of staff who regularly consult and top-coded to 20.
- Variables m1sbq19 (number of beds for observation), m1sbq20 (number of beds for deliveries) are recoded as ratio over the number of staff who regularly consult and top-coded to 5.
- Variable m1sbq33 (number of deliveries) is recoded as ratio over the number of staff who regularly consults and top-coded to 10.
- Variable m1sbq34 (number of caesars) is recoded as ratio over the number of staff who regularly consults and top-coded to 50.
- Variable m1sbq35 (number of deliveries ref. to other facilities) is recoded as ratio over the number of staff who regularly consults.

- Variable m1sbq36 (women passed away during delivery) is recoded as ratio over the number of staff who regularly consults and top-coded to 20.
- Variables m1scq5a-e (second source of electricity) were removed.

Local suppression

We further anonymize the data by suppressing certain values. This process is done once for the variables at the health facility level and once for the variables at the health worker level. Values of certain variables for particular health facilities and health workers were deleted. The following table gives an overview of the number of suppressed values per variable.

Table 3 Overview of number of supressions

Variable name	Description	Number of suppressions
Health facility level		
m1sbq8	Type of health facility	7 (1.7%)
m1sbq11	Days per week open	6 (1.5%)
m1scq1, m1scq4, m1scq5b-e	Main power source	4 (1.0%)
m1scq6- m1scq9	Main water source	6 (1.5%)
m1scq34a-g, m1scq35, m1scq36	Ambulance	13 (3.2%)
Health worker level		
fac_id, staff_id	Health facility ID	122 (2.3%)*

*The number of 122 suppressions corresponds to the amount of suppressions in Module 2. IDs of those facilities were also suppressed in Module 3, where 19 (3.3%) fac_ids and staff_ids were removed.

5. Note of information loss

SDC methods lead to a loss of information or utility in the data. To check the validity of the data for research purposes, several indicators were computed from the original data and from treated data. The indicators values computed from the released dataset do not significantly differ from the values in the original dataset and hence the dataset is valid for research purposes.

6. Bibliography

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., et al. (2012). *Statistical Disclosure Control.* Chichester, UK: John Wiley & Sons Ltd.

Templ, M., Meindl, B., Kowarik, A., & Chen, S. (2014, August 1). *Introduction to Statistical Disclosure Control (SDC).* Retrieved July 14, 2015 from http://www.ihsn.org/home/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf