Counter Trafficking Data Collaborative (CTDC)

The Global Victim-Perpetrator Synthetic Data Codebook

Version 1 – December 2022

1 DESCRIPTION OF DATA

These data consist of information on victims of trafficking in persons (TIP) (the object of the crime) and persons facilitated the trafficking process (the subject of the crime – hereafter, perpetrators). There are 16 variables in the dataset – 7 variables on victims and 9 variables on victims' account of perpetrators. The data includes information on the socio-demographic profiles of victims and perpetrators (such as gender, age), the region of citizenship and exploitation of victims and perpetrators, the type of exploitation, the relationships between victims and perpetrators (such as family, intimate partner, friend, stranger), and the role of perpetrators (such as recruiter, broker, habourer, abuser, exploiter, etc.).

The Global Victim-Perpetrator Synthetic Dataset is available on the Counter Trafficking Data Collaborative (CTDC) website. This dataset describes the relationship between victims and perpetrators. It includes IOM case data from over 17,000 victims and survivors of trafficking identified across 123 countries and territories, and their accounts of over 37,000 perpetrators who facilitated the trafficking process from 2005 to 2022. While administrative data of perpetrators are sometimes available from the court and law enforcement systems at an aggregate level, the data are rarely linked with victims of trafficking.

1.1 DATA DE-IDENTIFICATION

Since July 2019, IOM has been working with Microsoft Research on this de-identification solution through the Tech Against Trafficking (TAT) accelerator program. Prior to the collaboration with Microsoft, the CTDC team used k-anonymization, which is another de-identification method, to protect the privacy and safety of victims and survivors. The limitation of the k-anonymization approach is that by redacting case records with rare combinations of quasi-identifiers, the overall size of the dataset can be reduced dramatically in ways that greatly distort data statistics. In the case of the victim-perpetrator original data, 40% of case records (i.e., victim's account of over 14,000 perpetrators) would need to be suppressed to protect the safety and privacy of survivors.

With this latest approach, the dataset undergoes two stages of de-identification. In the first stage, all names and identifying details are removed from the data. In the second stage, the dataset is processed through a privacy-preserving pipeline enabled by a long-term collaboration with Microsoft Research on how to support safe data sharing of victim data – informing evidence-based policy while protecting the privacy and safety of victims. The main advantage of the Microsoft algorithm is that it overcomes the challenge of reduced sample size by synthesizing a new dataset in which records do not correspond to actual individuals, but which preserves the structure and statistics of the original data.

This is the second synthetic dataset derived from victim of trafficking case records, and the first to provide the guarantee of differential privacy. Differential privacy was developed at Microsoft Research in 2006, and today represents the gold standard in privacy protection. The idea is that if you get a similar answer to any data query whether or not any individual data subject is in the dataset used to answer the query, then you cannot infer the presence of that individual in the dataset. This is true no matter your background knowledge, including knowing the answers to earlier queries, or how many subjects have been added to the dataset since those queries. In short, it addresses the theoretical privacy gap left by k-anonymity whenever there is an expectation of multiple overlapping data releases over time.

More information on the approach is available through the open-source software and a documentation on differential privacy available via GitHub. Please refer to **Appendix I: Microsoft Research's Approach to Generating Synthetic Data with Differential Privacy**. Please refer to the Global Synthetic Data Codebook for more information on synthetic data generated with k-anonymity.

1.2 THE SOURCE DATA: IOM'S CASE MANAGEMENT SYSTEM

IOM's case management data is the only source of the Global Victim-Perpetrator Synthetic data. IOM's case data are collected by caseworkers at the screening and registration stage from potential beneficiaries. It requires informed consent at the screening and registration stage to register cases. The data are collected at the individual and household level. In this synthetic data, the unit of observation is at the individual (trafficking victim) level. The information, including descriptions of perpetrators, is self-reported by victims. Due to the hidden nature of this crime and that data are only available where IOM is operational, this dataset cannot be considered as a random sample of trafficking victims in the world.

IOM's data collection processes have evolved over time. IOM has recorded case data on victims of trafficking through assistance programmes since 2002. Records from 2002 to 2011 are called "legacy records". In 2012, IOM rolled out a web-based case management system called the "Migrant Management Operational System Application" (MiMOSA); IOM's missions gradually moved to MiMOSA to record identified and assisted victims of trafficking. By 2014, all missions switched to entering case data using the MiMOSA webform. MiMOSA has many data fields – from socio-demographic to route data. By 2019, MiMOSA has been upgraded and some questions in the webform have been updated. IOM's data have been cross-referenced across different databases in-house and over time. More information on IOM's direct assistance activities and data is available here.

By 2019, caseworkers record victim's account of perpetrators systematically. The data collection method and institutional history has informed the way year of registration is grouped in this synthetic data. In the MiMOSA web form (post 2019), trafficking victim can report the characteristics of up to six perpetrators. The victim may perceive that the perpetrator plays multiple roles (such as recruiter, broker, habourer, abuser, exploiter). Although multiple trafficking victims could refer to the same perpetrator, the data are based on each victim's personal account; we cannot identify if the same perpetrator is involved with multiple victims. Please refer to the **Appendix II: Perpetrator Module** to see the web form.

2 LIST OF VARIABLES

1 yearRegister
2 gender
3 majorityStatus
4 isForcedLabour
5 isSexualExploit
6 UN_COO_Region
7 UN_COE_Region
8 IP_RecruiterBroker
9 IP_TransactionProcess
10 IP_ControlAbuseKidnap
11 IP_Exploiter
12 IP_Relation
13 IP_Gender
14 IP_ageBroad
15 IP_citizen_UNRegion
16 IP_PayMoney

Notes: The following variables are available in the CTDC Global Synthetic and K-anonymized datasets: *gender, majorityStatus, isForcedLabour, isSexualExploit*. Th following variables are only in the CTDC K-anonymized data: *UN_COO_Region* and *UN_COE_Region*. These datasets report *yearOfRegistration* (as a continuous variable) while the Global Victim-Perpetrator Synthetic data report year ranges.

3 DETAILED DESCRIPTION OF VARIABLES

Before undertaking exploratory analysis, please consider that the number of observations is highly corrected with IOM's operational efforts. The sample and responses in this dataset cannot be considered as a random sample of all victims of TIP in the world.

Variable 1

Variable label: yearRegister
Type: string
Values and categories:

- NULL *[missing data]*
- 2005 – 2013
- 2014 – 2018
- 2019
- 2020
- 2021 – 2022

Definition: The year in which the victim was registered and assisted by IOM. The year bands are grouped in relation to IOM's database development timeline. Cases registered from 2005 and 2013 includes information from the "legacy records". Cases registered from 2014 to 2018 are data from the old MiMOSA webform. Cases registered beyond 2019 uses the latest webform. This is when the perpetrator module is broadly implemented across IOM's country locations. Cases registered from 2021 to 2022 are grouped together as we have not consolidated all country-level data into the MiMOSA system. In general, bigger group size (in this case, by grouping observations from 2021 to 2022 together) can improve the synthetic data quality.

Variable 2

Variable label: gender
Type: string
Values and categories:

- NULL *[missing data]*
- Male [*the victim identifies with a gender role most usually attributed to a man by relevant culture and society*]
- Female [*the victim identifies with a gender role most usually attributed to a woman by relevant culture and society*]

Definition: It describes the characteristics of women and men that are socially constructed. IOM collects data on another variable called "sex" that refers to the biologically determined sex. Even though IOM collects data on other gender identities that considers the psychological, behavioral, social, and cultural aspects of being transgender feminine, transgender masculine, non-conforming, or not specified/unknown, the number of observations is too small to be included in this dataset.

Variable 3

Variable label: majorityStatus
Type: string
Values and categories:

- NULL *[missing data]*
- Adult [*the victim is 18 years old or older*]
- Minor [*the victim's age is less than 18 years*]

Definition: It designates the victim's majority status at the time of registration and assistance by IOM. Following the United Nations Convention on the Rights of the Child, children (i.e., minor) is defined as persons up to the age of 18.

Variable 4

Variable label: isForcedLabour
Type: binary numeric
Values and categories:

- NULL *[missing data]*
- 1

Definition: It indicates that the victim was trafficked for the purpose of forced labour. Trafficking for the purpose of forced labour refers to the victim i) participating in any types of work, services, and employment, ii) being under the threat of penalty, and iii) working involuntarily. Sexual services are excluded from this definition.

Variable 5

Variable label: isSexualExploit
Type: binary numeric
Values and categories:

- NULL [missing data]
- 1

Definition: It indicates that the victim was trafficked for the purpose of sexual exploitation. This includes forced commercial sexual exploitation such as prostitution.

Variable 6

Variable label: UN_COO_Region
Type: string
Values and categories:

- NULL *[missing data]*
- Africa *[the victim is a citizen of a country in Africa]*
- Americas/Oceania *[the victim is a citizen of a country in the Americas or Oceania]*
- Asia *[the victim is a citizen of a country in Asia]*
- Europe *[the victim is a citizen of a country in Europe]*

Definition: It indicates the UN continental region of a country that the victim is a citizen.

Variable 7

Variable label: UN_COE_Region
Type: string
Values and categories:

- NULL *[missing data]*
- Africa *[the victim's country of exploitation is in Africa]*
- Americas/Oceania *[the victim's country of exploitation is in the Americas or Oceania]*
- Asia *[the victim's country of exploitation is located in Asia]*
- Europe *[the victim's country of exploitation is located in Europe]*

Definition: It indicates the UN continental region of a country that the victim is exploited in.

Variable 8

Variable label: IP_RecruitBroker
Type: binary numeric
Values and categories:

- NULL *[missing data]*
- 1

Definition: It indicates that the victim considers the perpetrator as a recruiter or a broker.

Variable 9

Variable label: IP_TransactionProcess
Type: binary numeric
Values and categories:

- NULL *[missing data]*
- 1

Definition: It indicates that the victim considers the perpetrator as someone who facilitates in the transaction process, including transporting, transferring, harbouring, smuggling, selling, buying, or receiving the victim.

Variable 10

Variable label: IP_ControlAbuseKidnap
Type: binary numeric
Values and categories:

- NULL *[missing data]*
- 1

Definition: It indicates that the victim considers the perpetrator as a controller, abuser, or kidnapper.

Variable 11

Variable label: IP_Exploiter
Type: binary numeric
Values and categories:

- NULL *[missing data]*
- 1

Definition: It indicates that the victim considers the perpetrator as an exploiter.

Variable 12

Variable label: IP_Relation
Type: string
Values and categories:

- NULL *[missing data]*
- FamilyIntimatePartner *[the perpetrator is family member or intimate partner of the victim]*
- FriendAcquaintance *[the perpetrator is a friend or acquaintance of the victim]*
- StrangerUnknownOther *[the perpetrator is a stranger to the victim, or that the perpetrator has other kinds of relationship, or an unknown relationship with the victim]*
- FamilyIntimatePartner;FriendAcquaintance *[at least one perpetrator is a family member/intimate partner; at least one perpetrator is a friend/acquaintance]*
- FamilyIntimatePartner;StrangerUnknownOther *[at least one perpetrator is a family member/intimate partner; at least one perpetrator is a stranger]*
- FriendAcquaintance;StrangerUnknownOther *[at least one perpetrator is a friend/acquaintance; at least one perpetrator is a stranger]*

Definition: It indicates the relationship between the victim and the perpetrator. **Recall:** Each victim can report up to six perpetrators. Each row in the dataset represents a victim. Thus, different perpetrators may relate to the victim differently.

Variable 13

Variable label: IP_Gender
Type: string
Values and categories:

- NULL *[missing data]*
- Male [*the perpetrator identifies with a gender role most usually attributed to a man by relevant culture and society]*
- Female [*the perpetrator identifies with a gender most usually attributed to a woman by relevant culture and society*]
- Female;Male [*at least one perpetrator was identified as male; at least one perpetrator was identified as female*]

Definition: It describes the perpetrator's characteristics of women and men that are socially constructed. IOM collects data on another variable called "sex" that refers to the biologically determined sex. Even though IOM collects data on other gender identities that considers the psychological, behavioral, social, and cultural aspects of being transgender feminine, transgender masculine, non-conforming, or not specified/unknown, the number of observations is too small to be included in this dataset. **Recall:** Each victim can report up to six perpetrators. Each row in the dataset represents a victim. Thus, it is possible that a victim's trafficking process involved male and female perpetrators.

Variable 14

Variable label: IP_ageBroad
Type: string
Values and categories:

- NULL *[missing data]*
- 0 –29
- 30 –38
- 39 – 47
- 48 +
- 0 –29;30 –38
- 0 –29;39 –47
- 30 –38;39 –47
- 30 –38;48 +
- 39 –47;48 +

Definition: It categorizes the age of perpetrators (reported by the victim) into 4 groups. Ages are grouped following the same rule as other CTDC datasets published since 2017. **Recall:** Each victim can report up to six perpetrators. Each row in the dataset represents a victim. Thus, a victim may report more than one perpetrator from different age ranges.

Variable 15

Variable label: IP_citizen_UNRegion
Type: string
Values and categories:

- NULL *[missing data]*
- Africa *[the perpetrator is a citizen of a country in Africa]*
- Americas/Oceania *[the perpetrator is a citizen of a country in the Americas or Oceania]*
- Asia *[perpetrator is a citizen of a country in Asia]*
- Europe *[perpetrator is a citizen of a country in Europe]*
- Oceania *[perpetrator is a citizen of a country in Oceania]*

Definition: It indicates the citizenship of the perpetrator (from the perspective of victims) by UN continental region.

Variable 16

Variable label: IP_PayMoney
Type: string
Values and categories:

- NULL *[missing data]*
- Yes
- No
- No;Yes [*the victim paid at least one perpetrator in one instance, while the victim did not pay at least one perpetrator in another instance*]

Definition: It indicates if the victim paid money to the perpetrator during the trafficking process.
Recall: Each victim can report up to six perpetrators. Each row in the dataset represents a victim. Thus, it is possible that a victim paid a perpetrator in one instance, but the same victim did not pay another perpetrator in another instance.

Appendix I: Microsoft Research's Approach to Generating Synthetic Data with Differential Privacy

One aspect of victim safety is ensuring the privacy of data subjects. Such privacy means that traffickers are prevented from identifying known victims in published datasets, making those victims safe from reprisals. Another aspect of victim safety is ensuring the accuracy of data statistics. Such accuracy means that downstream activities of data-driven decision making and policy making are based on the best available data, leading to the most appropriate actions.

The challenge for safe data sharing is that the methods used to preserve the privacy of data subjects typically distort data statistics, and if they are distorted in the wrong ways then this could lead to misguided actions that compromise the safety of the broader victim population. For example, if a privacy method greatly over- or under-reported a given case pattern – or fabricated it entirely – this could mislead decision makers into misallocating scarce resources in ways that fail to tackle the actual problems observed.

Microsoft Research's approaches are based on the idea that rather than redacting sensitive data to create privacy, one can instead generate synthetic data that is private by design, yet accurately captures the structure and statistics of the underlying sensitive dataset.

In the September 2021 release of the Global Synthetic Dataset, IOM used a new algorithm from Microsoft Research that generated synthetic data with k-anonymity for all combinations of attributes, not just the subset of attributes labelled in advance as quasi-identifiers. This addressed both the privacy and accuracy limitations of the k-anonymization method and earlier release – all combinations of attributes included in the synthetic and aggregate datasets appeared at least k times in the original sensitive dataset and were reported precisely in the aggregate dataset (rounded down to the closest k).

While synthetic data with "full" k-anonymity represented a major improvement over standard k-anonymization, the nature of the privacy guarantee only extends to a single release. Across a series of multiple releases (where each version builds on the data in the previous release), k-anonymity provides no theoretical guarantees about what an attacker could potentially learn by comparing the differences between reported counts and the records they know (or assume) have been added to the dataset. Although a large enough k and a sufficient record increment is likely to guard against all practical attacks, there is another privacy paradigm that is explicitly designed to combat such "differencing" attacks: differential privacy.

Differential privacy was developed at Microsoft Research in 2006, and today represents the gold standard in privacy protection. The idea is that if one gets a similar answer to any data query, whether or not any individual data subject is in the dataset used to answer the query, then one cannot infer the presence of that individual in the dataset. This is true no matter one's background knowledge, including the answers to earlier queries, or how many subjects have been added to the dataset since those queries. In short, differential privacy addresses the theoretical privacy gap left by k-anonymity whenever there is an expectation of multiple overlapping data releases over time.

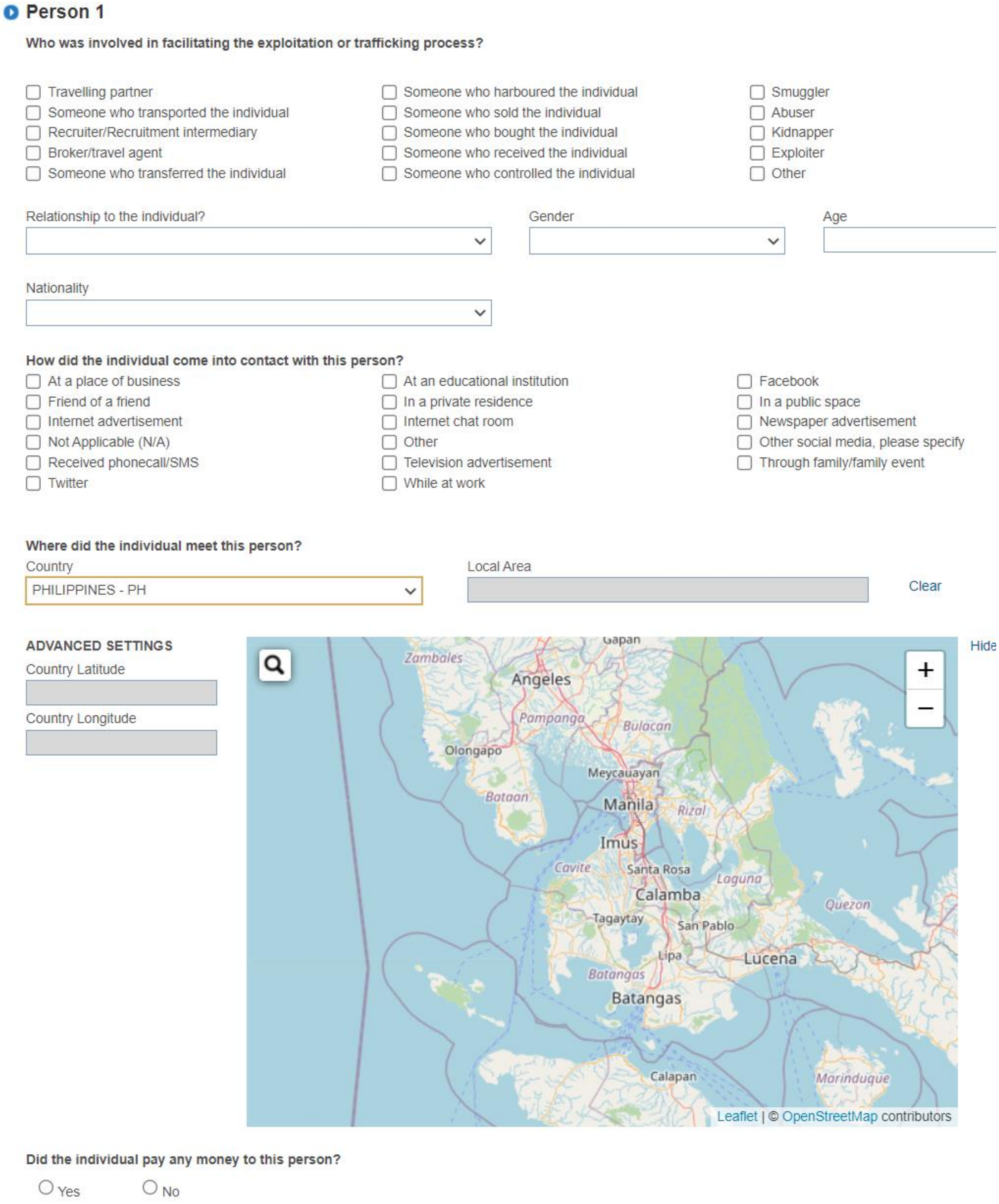
A central concept in differential privacy is the idea of quantifiable privacy loss – the extent to which the answers to arbitrary data queries are allowed to vary, probabilistically, based on the presence or absence of individual data subjects. The parameter that captures this concept is called epsilon. It serves as a budget for the allowable privacy loss across all queries. Each time the sensitive data is queried, calibrated noise is added to the answer in ways that control the possible privacy loss, and part of this budget is consumed. Once the budget is exhausted, no more queries can be answered.

To create synthetic data with differential privacy, the new method from Microsoft Research first uses the privacy budget to query the counts of cases matching all short combinations of case attributes. The results of these queries are released as aggregate data with differential privacy. Synthetic records are then constructed by sampling these combinations based on their noisy counts until all attributes in the sensitive dataset (based on the noisy counts in the aggregate dataset) have been accounted for. The resulting synthetic dataset retains the same degree of differential privacy as the aggregate data used an input, and the worst-case privacy loss across a series of releases is simply the sum of the individual privacy budgets used to generate them. The release of the synthetic case records and aggregate data enables the evaluation of synthetic data accuracy as well as the retrieval of accurate counts (e.g., for official reporting).

For more details on the overall approach, including proof of differential privacy, see https://github.com/microsoft/synthetic-data-showcase.

Appendix II: Perpetrator Module

Figure 1 IOM's MiMOSA webform to collect information on victim's account of perpetrator(s)



Person 1

Who was involved in facilitating the exploitation or trafficking process?

□ Travelling partner
□ Someone who transported the individual
□ Recruiter/Recruitment intermediary
□ Broker/travel agent
□ Someone who transferred the individual
□ Someone who harboured the individual
□ Someone who sold the individual
□ Someone who bought the individual
□ Someone who received the individual
□ Someone who controlled the individual
□ Smuggler
□ Abuser
□ Kidnapper
□ Exploiter
□ Other

Relationship to the individual?

Gender

Age

Nationality

How did the individual come into contact with this person?

□ At a place of business
□ Friend of a friend
□ Internet advertisement
□ Not Applicable (N/A)
□ Received phonecall/SMS
□ Twitter
□ At an educational institution
□ In a private residence
□ Internet chat room
□ Other
□ Television advertisement
□ While at work
□ Facebook
□ In a public space
□ Newspaper advertisement
□ Other social media, please specify
□ Through family/family event

Where did the individual meet this person?

Country: PHILIPPINES - PH

Local Area

Clear

ADVANCED SETTINGS

Country Latitude

Country Longitude

Hide

Did the individual pay any money to this person?

○ Yes ○ No

Notes: This is how the case management webform looks from the case worker's perspective. If an individual has been identified as a victim of trafficking, they would be asked to provide more information for IOM to better assist them (e.g., refer them to protection services locally). After the first stage of de-identification (i.e., removing names and identifying details from the data), some information is recoded into broader categories. It includes the following perpetrator variables: role, age, nationality/citizenship, and relationship with the victim.