J-PAL GUIDE TO DE-IDENTIFYING DATA

Sarah Kopper, Anja Sautmann, and James Turitto

J-PAL Global

January 2020

Abstract: Researchers who plan to publish data on human subjects should take careful steps to protect the confidentiality of study participants through data de-identification—a process that reduces the risk of re-identifying individuals within a given dataset. This guide provides further details on the de-identification process, including various procedures for de-identifying a dataset, a list of common identifiers that need to be reviewed, and sample code that can be used to de-identify data intended for publication. It is intended to be used alongside the accompanying Guide to Publishing Research Data.

Please contact research-resources@povertyactionlab.org with comments, questions, or feedback.

Acknowledgements: We thank Sonia Barbosa, Shawn Cole, Mary-Alice Doyle, Laura Feeney, and William Parienté for helpful comments and Antonn Park for copy-editing assistance. Any errors are our own.

TABLE OF CONTENTS

Disclaimer: This document is intended for informational purposes only. Any information related to the law contained herein is intended to convey a general understanding and not to provide specific legal advice. Use of this information does not create an attorney-client relationship between you and MIT. Any information provided in this document should not be used as a substitute for competent legal advice from a licensed professional attorney applied to your circumstances.

KEY POINTS

- It is important to think of **de-identification as a process that reduces the risk of identifying individuals**, rather than completely eliminating the potential for re-identification.
- To protect human subjects, de-identification should occur **as early as possible** in the research process. This means de-identifying data after data collection steps that require finding respondents, such as back-checks, are complete.
- Data should always be de-identified **before being published**. This is a requirement for projects that are carried out or are funded by J-PAL and is also typically required by Institutional Review Board (IRB) protocols. It is also a legal requirement in many countries.
- Data that is stripped of sensitive information limits the analysis that can be conducted, so there is a tension between de-identification and data usability. As every case is different, the de-identification process requires thought and good judgment.
    - For example, hospital admission data may be important in analysis but could be used to identify individuals. To preserve both data usability and privacy of individuals, a variable for the number of days between hospital admission and treatment could be created and published instead.
    - In cases where re-identification risks privacy violations or potential harm for the subjects, indirect identifiers that will not be used in analysis should be removed or redacted.
- The possibility of **re-identification can almost never be fully ruled out**. Before publishing survey data, the researcher (possibly in collaboration with an IRB) should assess the risk associated with re-identification—both the probability that re-identification may occur and the potential consequences for the subject if identified. Alternatives to data publication include restricted repositories and formal approval processes for re-use.

This guide includes further detail, practical guidelines, and sample code that can be used to de-identify data intended for publication.

RATIONALE: WHY DE-IDENTIFICATION?

Researchers who plan to publish their data are generally required to ensure the privacy of study participants. This is in accordance with ethics standards and IRB protocols and is often a legal requirement: many countries recognize privacy as a universal human right and have guidelines that restrict the publication of personal data. For example, in the United States, restrictions such as the 1996 US Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule and the Family and Educational Records Privacy Act (FERPA) apply to research using individual-level health and education records, respectively. But these requirements do not apply if data has been sufficiently de-identified such that there is "no reasonable basis to believe that the information can be used to identify an individual" (US DHHS 2012). As of May 2018, the EU's General Data Protection Regulation (GDPR) set the stricter guidelines of complete anonymity, stating that it does not apply to data that "does not relate to an identified or identifiable natural person or to data rendered anonymous in such a way that the data subject is not or no longer identifiable" (EU GDPR Recital 26).

There are a number of **potential harmful outcomes** that could result from study participants being identified. Some examples include **identity theft, political or legal repercussions**, **embarrassment or social stigma** (e.g., with STD infections), **loss of benefits** (e.g., medical history can affect access to insurance), and **personal or family repercussions** (e.g., with sexual history). That said, there is a trade-off between data de-identification and usability. Perfectly anonymous data can often not be used for meaningful analysis, and researchers must weigh the balance of privacy against usability. Publishing data that can be used for secondary analysis can promote new research and lower entry costs, especially for young researchers.

Even with de-identified data, there is always a risk that subjects can be identified. Risk of re-identification is especially high for characteristics that are "outliers" in some way—such as an individual who has lost eyesight—or with data that is detailed enough, so that almost every observation becomes unique (e.g., the combination of hair color, eye color, exact height, age, etc. may identify a person, even in a large group). Given the high bar and trade-offs in perfectly anonymized data, it is important to think of **de-identification as a process that reduces the risk of identifying individuals**, rather than completely eliminating the potential for re-identification.

Before publishing research data, the researcher (possibly in collaboration with an IRB) should assess the risk associated with re-identification, based on the identifiability and the potential harm for the subject. Identifiability refers to the likelihood an individual study participant in the dataset could be identified by looking at the data. Harm refers to the consequences to the human subject if the data were disclosed. The potential for harm is context-dependent. For example, sharing data about participants' favorite type of beer would likely cause negligible harm in the United States. In a country where alcohol is outlawed, however, the potential harm from sharing this seemingly innocuous data would be considerably higher.

The relationship between identifiability and harm creates the framework by which to assess where and how to publish the data in question. Some data can only be made available through more secure data archives, described further in J-PAL's Data Publication guidelines, because the risk to the study participants is too great to make the data publicly available. Alternatives to data publication include restricted repositories and formal approval processes for re-use.

ABOUT PERSONALLY IDENTIFIABLE INFORMATION (PII)

Personally identifiable information refers to information that contains identifiers specific to individuals—this includes **direct identifiers** (such as name, social security number, birth ID number, government ID number, etc.) and **indirect identifiers** (such as birth date/month/year, municipality/city, gender, etc.). Individual indirect identifiers may not be unique to the person but in combination create a unique profile. The US's HIPAA guidelines label 18 variables as direct identifiers. The HIPAA list is not exhaustive, especially when considering data collected outside the US; researchers should consider the type of data they are collecting, how identifiable certain variables might be, and the legal framework that applies to their data.

Table 3 at the end of this guide lists a set of common direct and indirect identifiers and provides a recommended method for de-identification.

ONGOING DATA PROTECTION

It is best to plan from the beginning how much you will de-identify at each stage. At a minimum, this includes the steps shown in table 1 below:

Table 1

PROJECT LIFE CYCLE STAGE	DATA PROTECTION STEP
While preparing your IRB proposal and informed consent forms	• Describe how data will be de-identified. • Informed consent forms should include a statement describing the extent, if any, to which confidentiality of records identifying the subject will be maintained.
During data collection	• Direct identifiers should be removed or redacted as soon as back-checks and other data quality checks are complete that require re-interviewing study participants. • Data containing PII should be encrypted. • If the study participants will be re-interviewed in future survey waves, be sure to retain the (secured) identified data so that they can be found again. The IRB may also require that you keep records of who was interviewed in the case of audits or adverse events.
Before data cleaning	• Make decisions about indirect identifiers, bearing in mind how they may be combined to identify individuals. • Do not forget to check for outliers that may also identify individuals. • Document all de-identification steps, as changing the data can affect inference and regression results. Typically, these steps should be reversible, except in cases of extremely sensitive data where PII may have to be destroyed (e.g., if subjects incriminate themselves by answering survey questions).
Before data publication	• Do a final check to ensure respondents cannot reasonably be re-identified from the data files intended for publication. • If you are publishing your data solely for replication purposes, this final check may include dropping (for the published version) all variables not needed for the replication. • Even if you are publishing your data so that it can be used for other analyses, dropping variables that will not plausibly be used for analysis, or for which the risk of re-identification is high, will help ensure privacy. • As a general best practice, you should re-run your data analysis on the data that will be published.

DE-IDENTIFICATION FOR DATA PUBLICATION

Step 1: Identify all variables containing PII

Perform a manual check by browsing for any variables that may contain PII, including string variables, numeric variables, and strings encoded as numeric variables. While automated scanning tools such as the J-PAL PII scanner or the International Household Survey Network's (IHSN) sdcMicro[1] can quickly identify variables or labels that commonly contain PII, a manual check is important to find variables and labels that would not be caught by an automated tool, such as those with uncommon names.

[1] http://www.ihsn.org/software/disclosure-control-toolbox

Step 2: Encode, redact, or remove direct identifiers

A list of direct and indirect identifiers can be found in table 3 at the end of this guide. Direct identifiers must be kept hidden from non-authorized users. There are three main options for doing so: encoding values with random IDs or codes, redacting labels or values, or removing variables altogether. Encoding values has the benefit of preserving the structure of the data and is recommended practice, while redaction and removal render the variable unusable in analysis. Table 2 illustrates these three techniques:

Table 2

ORIGINAL DATA		VALUES REDACTED		VALUES REPLACED WITH CODES		VARIABLES REMOVED: CITY AND STATE WOULD NOT APPEAR
City	**State**	**City**	**State**	**City**	**State**	
Somerville	MA	XXXX	XXXX	11	1	
Cambridge	MA	XXXX	XXXX	12	1	
Boston	MA	XXXX	XXXX	13	1	
Concord	NH	XXXX	XXXX	21	2	
Nashua	NH	XXXX	XXXX	22	2	

ENCODE DIRECT IDENTIFIERS

Encoding identifiers with anonymous ID numbers preserves connections between data points and information about the unit of observation and is thus recommended practice for dealing with identifiers. For example, districts in (randomly numbered) province 1 could be labeled with "1" as a prefix (e.g., district 11, district 12, district 13, etc.). Villages in district 11 could then be labeled with "11" as a prefix (e.g., village 1101, village 1102, etc.). ID numbers must be assigned at random and not linked to a sort order (e.g., by alphabet) or any pre-existing ID variable from another database; otherwise, the encoded variable can be decoded.

Also note that ID number formats should be chosen thoughtfully, with a fixed number length, to avoid confusion that would arise from, for example, district 11 in province 1 being indistinguishable from district 1 in province 11 (e.g., 111 and 111 versus 0111 and 1101). In order to use leading zeros, the variable should be formatted as a string. If the coded variable does not contain relevant information about connections between data points or the unit of observation (such as individuals' names), then encoding has the same effect as removing the variable.

REMOVING OR PARTITIONING

Removing or **partitioning** identifiers consists of separating identified and de-identified data, where identifiers are replaced with randomized IDs in the de-identified data. As this process does not preserve data structure, it should only be done on direct identifiers that serve no purpose in understanding the data, such as individuals' names.

Sample Stata code for partitioning direct identifiers is shown below. The code generates randomized IDs and splits the data into three parts: (1) original data, (2) de-identified data, (3) data with the randomized ID that links the original data with the de-identified data. After this de-identification process, the original dataset with PII should be stored in a secure, encrypted repository/folder, while only the de-identified data is published or shared with users who are not on the project's IRB or have not signed a data use agreement.

```
* Step 1: Determine variables that define an identified observation
global Obsvarlist "District Address”
/* Step 2: Create a cross-walk file "IDLink.dta" with direct identifiers and a new
random ID number: */
preserve
        * keep only the ID variables:
        keep $ObsVarlist
        * keep only one instance of each observation:
        bysort $ObsVarlist: keep if _n == 1
        * generate new IDs:
        egen ID_obs = rank(runiform()), unique // ensures the ID is randomly
       sorted, rather than created based on sort order (e.g., in ascending order)
        label var ID_obs "Unique observation ID"
        save IDlink_obs.dta, replace // this file needs to be protected just like
       the original data
restore

* Step 3: merge the new IDlink file with the raw data file:
merge m:1 $ObsVarlist using IDlink_obs.dta, nogen assert(3)
* note: m:1 is needed because we kept only one instance of each observation

/* Step 4: Drop all the direct identifiers from the raw data and save de-
identified data */
drop $ObsVarlist
order ID_obs // order the variables in your dataset so IDs are first and label
them
save DataAnon.dta // save de-identified file. This file does not need to be
encrypted.
```

REDACTING VALUES

This involves replacing variables that contain PII with string characters that are obviously generic (e.g., "ANY TOWN") or, for example, "XXXX" or "CONFIDENTIAL." This is the approach typically taken for the World Bank's Living Standards Measurement Surveys (LSMS) for variables such as respondent name. Values that are redacted serve no purpose in analysis, as typically every single observation will take on the same value (e.g., "XXXX"). As such, variables that may be important to understand the data, such as the unit of observation, should not be redacted and should instead be encoded, as described above.

Step 3: Make decisions about indirect identifiers

When de-identifying datasets, it is important to keep in mind that **combinations of variables** can be used to identify a study participant. For instance, in the United States, birthdate, a zip code, and an individual's gender identity cannot independently identify an individual; however, if you have access to all three variables, then there is an 87% chance that an individual could be identified (Sweeney 2000). In a study using de-identified Netflix data, researchers were able to re-identify users with 68% success knowing only two movies the user had reviewed, combined with the rating and date of rating (+/– 3 days) (Ohm 2010). In cases where re-identification risks privacy violations or potential harm for the subjects, indirect identifiers that are not used in the final analysis should be removed or redacted.

The tension between data privacy and data usability described above is acute for indirect identifiers, especially geographic variables. While replacing identifiers with codes, as described above, is an effective way of de-identifying the data, this approach removes information that you may want to use, such as geocoded data or employer name. You may want to preserve this information so that you can link it to external sources, such as rainfall data or company revenues. For these types of identifiers, it is then important to think of different ways to de-identify data to reduce harm that can be inflicted upon the subjects in the data while also preserving the usability of the data. J-PAL recommends two approaches: **aggregation** and **geographic masking**.

AGGREGATION

With aggregation, variables that contain PII are summarized and replaced by aggregated or descriptive statistics. Examples are to group birth dates (e.g., by keeping only the birth month, quarter, or year), geographic locations (e.g., aggregate GPS coordinates to the village or county level), or employers (e.g., code industry or firm size). A related approach is top or bottom coding, which is particularly useful for outliers. For example, individuals with an annual income exceeding $250,000 could be grouped into an "over $250,000" income category. While aggregation lowers the probability of re-identification, the aggregated variable is generally less useful in analysis.

In choosing when to aggregate or top code, knowledge of the local context is useful. You should ask how unusual outlying observations are within their context and if it is general knowledge who was surveyed within a specific neighborhood or village. For instance, if there is one person in your dataset with especially high or low income relative to other survey participants, they may appear to be an outlier, but if they live in a village where there are many (non-surveyed) people with this income level, and others in their village do not know who was surveyed and who was not, then they may not be identifiable.

GEOGRAPHIC MASKING (JITTERING)

Preserving information is especially important for geographic data, where researchers may wish to match survey data with, for example, rainfall or temperature data from third party sources. The main method used to de-identify spatial data is **geographic masking** (also known as **jittering** or **displacement**), where points are offset in a systematically random nature to decrease the probability of re-identification. For example, in the USAID Demographic and Health Survey's (DHS) household surveys, data from the same enumeration area is aggregated to a single point coordinate. For urban clusters, the aggregated coordinates are then displaced up to 2 km in any direction. For rural clusters, the aggregated coordinates are displaced up to 5 km, with an additional, randomly selected 1% of clusters displaced up to 10 km, again in any direction (Burgert et al. 2013). This "random direction, random distance" procedure is also followed by the World Bank's LSMS team using a custom-built Python tool in ArcGIS—code for this displacement process can be found in Appendix B of DHS Spatial Analysis Reports 7. Random displacement can be generated using the `runiform` command in Stata or the `runif` function in R but is not easily combined with geospatial data.

Again, knowledge of the local context is important, as you may want greater displacement of coordinates if you are using data from a sparsely populated area. For example, it could be that coordinates that are jittered by up to 5 km point to an unpopulated area with only one nearby village, in which case the village would be identified and coordinates would need to be offset by a larger amount. More details on the DHS approach to geographic displacement can be found in Burgert et al. (2013); both the DHS report and Zandbergen (2014) provide information on other jittering methods.

Geographic masking (jittering) has the advantage of allowing researchers to match locations to other geocoded data, such as satellite imagery, but the risk of re-identification with certain identifiers (for example, households' GPS coordinates) is high. Greater displacement/perturbation reduces the usefulness of the geographic information while potentially creating the illusion of precision *and* still retaining a re-identification risk. **Aggregation by geographic units** such as "village" or "zip code" is often the preferred method.

As such, J-PAL recommends that all coordinates at the household level or below be aggregated to at least the level of the next-lowest geographic unit (e.g., village). Before aggregation, researchers can create variables that may be important in future analyses, such as distance to the nearest road, school, or health clinic, or match survey data with georeferenced climate data to create variables such as average rainfall, rainfall variability, and temperature.

For village level or above, J-PAL recommends a combination of aggregation or masking, depending on the data. Following HIPAA guidelines, a town or city of **at least 20,000 inhabitants** does not need to be jittered or aggregated (just as its name would not need to be encoded or masked). The World Bank's LSMS team jitters (but does not aggregate) coordinates at the enumeration area level (roughly village level), following the DHS procedure described above. As with other identifiers, it is important to consider whether combinations of indirect identifiers could be used to identify individuals, even if the geographic unit consists of over 20,000 inhabitants.

ACCESSING IDENTIFIED DATA

Some types of analyses are not possible with de-identified data. One option for research teams is to make their personally identified data available to other researchers who sign a data use agreement (DUA) and obtain IRB approval (an alternative is to add these researchers as key personnel to the project's IRB). The DUA must include provisions to address the permitted uses and disclosures of the personally identified data, identify who may use or receive the data, and prohibit data users from identifying or contacting individuals. Note that allowing additional researchers to access your data is only in your control if you (the research team) own the data.

The Federal Demonstration Partnership (FDP), a US-based initiative among 10 federal agencies and 154 institutions (including MIT and other leading universities), has created a template DUA, which can be accessed following the FDP link below. While most institutions prefer using their own template, many member institutions have agreed to use this template as a fallback option.

Alternatively, services such as the Inter-University Consortium for Political and Social Research (ICPSR) allow researchers to use identified data while maintaining confidentiality, without having to go through the original researcher. ICPSR has the capacity to host restricted use datasets in cases where de-identifying the data is either not feasible or would significantly impact data usability. Researchers can request controlled use of restricted use data through an application process, which includes agreeing to follow strict legal and electronic requirements to preserve confidentiality.

RE-IDENTIFICATION AND RESPONSIBILITIES

Researchers and data users have a responsibility not to use data to try to identify human subjects. Doing so is not only unethical but can have legal repercussions and financial penalties. For example, following Section 1106(a) of the Social Security Act, the US's Centers for Medicare and Medicaid Services specifies in its standard DUA that unauthorized disclosures of information can be penalized by up to $10,000 in fines or up to five years' imprisonment or both (US DHHS Form CMS-R-0235). Services such as ICPSR and others that provide publicly available data require that users protect the privacy of research participants and report breaches of participant confidentiality (i.e., report to the data owner or repository if sensitive content is found in de-identified data). Phillips et al. (2017) review the debate surrounding legal penalties for re-identification in biomedicine—a debate that is relevant to the same issues in the social sciences.

While we can never fully ensure that an individual remains completely anonymous when we collect identifying data about them, de-identification lowers—though does not eliminate—the risk of re-identification and allows the secondary use of data for further research studies and other uses. For additional details on different de-identification methods, see Altman (n.d.) and Green (2018).

Table 3

IDENTIFIER TYPE	DIRECT IDENTIFIER	STRONG INDIRECT IDENTIFIER	INDIRECT IDENTIFIER	HIPAA IDENTIFIER	J-PAL RECOMMENDED DE-IDENTIFICATION METHOD
Personal ID number	x			x	Data partitioning
Full name	x			x	Data partitioning
Date of Birth		x		x	Aggregation
Year of birth		x		x[1]	Aggregation or top/bottom coding if few observations
Age			x	x[1]	
Gender			x		
Marital status			x		
Household composition			x		
Occupation		(x)	x		Aggregation if few observations
Industry of employment			x		
Employment status			x		
Education			x		Aggregation or top/bottom coding if few observations
Ethnicity			x		Aggregation if few observations

IDENTIFIER TYPE	DIRECT IDENTIFIER	STRONG INDIRECT IDENTIFIER	INDIRECT IDENTIFIER	HIPAA IDENTIFIER	J-PAL RECOMMENDED DE-IDENTIFICATION METHOD
Nationality			x		Aggregation if few observations
Workplace/Employer		(x)	x		Aggregation
Phone number	x			x	Data partitioning
Email address	x	(x)		x	Data partitioning
Audio file or video file displaying person(s)	x			x	Data partitioning
Photograph of person(s) (if full face or comparable)	x			x	Data partitioning
Bank account number		x		x	Data partitioning
IP address		x		x	Data partitioning
Vehicle registration number	x			x	Data partitioning
Web page address		(x)	x	x	Data partitioning
Student ID number		x		x	Data partitioning
Insurance number		x			Data partitioning
Postal code			x	x	Aggregation
Major region			x		
Geographic area with less than 20,000 inhabitants			x	x if <20,000	Replace with ID (encode)
Household location (GPS coordinates)	x			x	Aggregation, jittering
Village/town GPS coordinates			x	x if <20,000	Jittering if less than 20,000 inhabitants

Notes: [1] If individual over age 89. In some cases, the identifier may be considered a strong indirect identifier (e.g., an uncommon occupation), as denoted by (x). This table draws heavily on guidelines from the Finnish Social Science Data Archive (2009).

RESOURCES

- Altman, Micah. n.d. “Data Security and Privacy: Key Concepts.” Lecture for J-PAL102x Micromasters Course. Last accessed August 18, 2017. https://drive.google.com/file/d/0B6NSujurHRIVc0J0MkJTdHhBTzQ/view
- Burgert, Clara R., Josh Colston, Thea Roy, and Blake Zachary. 2013. "Geographic Displacement Procedure and Georeferenced Data Release Policy for the Demographic and Health Surveys." DHS Spatial Analysis Reports No. 7. Calverton, Maryland: ICF International. https://dhsprogram.com/pubs/pdf/SAR7/SAR7.pdf
- EU GDPR. 2016. “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.” Last accessed December 3, 2019. https://gdpr-info.eu/
- FDP. n.d. “Data Stewardship.” Last accessed December 3, 2019. http://thefdp.org/default/committees/research-compliance/data-stewardship/
- Finnish Social Science Data Archive. 2009. "Data Management Guidelines—Anonymisation and Personal Data. Data Archive." Last accessed August 17, 2017. https://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html
- Green, Joe. 2018. “Data De-Identification Stata Programs and Demonstration. BITSS Research Transparency and Reproducibility Training (RT2), Los Angeles.” Last accessed December 3, 2019. https://osf.io/tx3af/
- Ohm, Paul. 2010. “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization.” UCLA Law Review 57. https://www.uclalawreview.org/pdf/57-6-3.pdf
- Phillips, Mark, Dove, Edward S., and Bartha M. Knoppers. 2017. “Criminal Prohibition of Wrongful Re-Identification: Legal Solution or Minefield for Big Data?” Journal of Bioethical Inquiry 14(4): 527–539. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5715031/
- Sweeney, Latanya. 2000. “Simple Demographics Often Identify People Uniquely.” Carnegie Mellon University, Data Privacy Working Paper 3. Last accessed December 10, 2019. http://ggs685.pbworks.com/w/file/fetch/94376315/Latanya.pdf
- US Department of Education. FERPA.
- US Department of Health and Human Services. 2012. "Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.” Last accessed December 10, 2019. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html
- US Department of Health and Human Services. “Health Information Privacy.” Last accessed December 10, 2019. https://www.hhs.gov/hipaa/index.html
- US Department of Health and Human Services Centers for Medicare & Medicaid Services (US DHHS CMS) Form CMS-R-0235.
- Zandbgergen, Paul A., 2014. "Ensuring Confidentiality of Geocoded Health Data: Assessing Geographic Masking Strategies for Individual-Level Data." Advances in Medicine 2014. doi:10.1155/2014/567049.