The Network for Public Health Law

Ideas. Experience. Practical answers.




DE-IDENTIFICATION
Fact Sheet

Statistical or Scientific De-Identification Fact Sheet

"Statistical or scientific de-identification" is an important tool to assist public health in negotiating its dual and sometimes conflicting missions – maintaining the privacy of the information it collects and sharing the information broadly with the community in a legal and privacy protective manner. As opposed to prescriptive methods, which delineate the removal of specific direct and indirect identifiers from the data set, this approach involves removing direct identifiers, like name and Social Security number, and balancing the utility of the inclusion of indirect identifiers, such as dates and geographies, with the risk of re-identification; this approach yields multiple solutions and provides flexibility. Statistical or scientific de-identification allows the expert, in consultation with the data steward, to determine which method(s) to apply to the data set to de-identify the indirect identifiers.

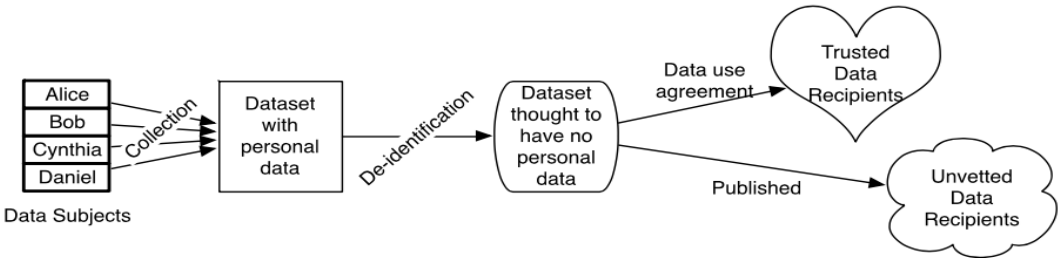
De-identification provides public health with many benefits[1]:

- If data are de-identified at the point of collection, the risk of a privacy breach while data are retained, is significantly decreased.
- When data are de-identified prior to sharing, technical and policy controls may be minimized.
- De-identification affords public health with the ability to share data widely with communities and others.
- This fact sheet is intended to be used by privacy officers, public health practitioners, data managers and their attorneys to provide awareness of these methods. See the Resources document, which is part of this toolkit, for technical resources.

This fact sheet provides an overview of statistical and scientific de-identification methods of structured data, such as lab values and patient demographics, where the data are entered utilizing pre-defined fields from within the record. This fact sheet is not intended for de-identification of unstructured data, such as narrative reports or multimedia. Additionally, detail regarding methods for creation of synthetic data or data enclaves are beyond the scope of this fact sheet.[2]

Understanding De-identification

De-identification is "'the general term for any process of removing the association between a set of identifying data and the data subject.'"[3] De-identification is a process applied to data by either trained individuals or by computers, or both, and includes technical and policy solutions.[4]



Because of the importance of both individual privacy and data utility, de-identification is first a matter of governance and management.

1. Public health should define its purpose in de-identifying the data[5]:
   - Identify goals.
   - Determine types of data to be de-identified.
   - Decide potential uses of de-identified data.
2. Next, an expert at de-identifying health information will assess whether the health information can be re-identified. Experts, with statistical, mathematical or scientific backgrounds may be found within the public health agency or come from another organization.[6]
3. Then, the expert will identify which statistical or scientific methods should be applied to the data to reduce the risk to an acceptable level. With input from the public health agency, the expert will execute on the methods.[7]
4. Next, the expert will test the resulting health information to evaluate the risk of individual re-identification against the agreed upon risk level. An example of such a risk level, is that the risk must be no more than very small. It may take several iterations for the expert to achieve the appropriate risk level.[8]
5. Finally, the expert should document the method and results.

The results of the risk assessment should inform the method utilized to share the data:

1. Publication on the internet.
2. Disclosure to trusted individuals under a Data Use Agreement.
3. Creation of synthetic data, which lack identifying information but reflect statistical properties of the original data set.
4. Presentation of de-identified data in a physical or virtual enclave that allows one to query the data and receive results, but not to export them.[9]

With regard to de-identifying data to be published on the internet or disclosed to trusted individuals, there are a variety of methods that may be applied to transform the data set containing personal data, to a data set thought to have no personal data.

First, remove or obscure direct identifiers. Direct identifiers include "'data that can be used to identify a person without additional information or with cross-linking through other information that is in the public domain.'"[10] Examples include names, Social Security numbers and email addresses; the National Institute of Standards and Technology recommends treating medical record numbers and phone numbers as direct identifiers, as their use for identification is ubiquitous.[11]

Direct identifiers may be:

- Removed.
- Replaced with generic data or symbols.
- Replaced with random values. Where the same identity appears twice, it receives different values.
- Replaced with pseudonyms, allowing for matching.[12]

Second, review and address the inclusion of indirect identifiers. Indirect identifiers are “identifiers that by themselves do not identify a specific individual but can be aggregated and ‘linked’ with other information to identify data subjects.”[13] Examples include birthday, ZIP code and sex. While direct identifiers’ removal from the data set does not generally impact utility, removal or transformation of indirect identifiers has significant impact on utility, as well as privacy, and the balance must be performed with great care. A resulting data set with a low re-identification risk, but with low utility, suggests that the expert should identify additional options.[14]

There are several methods available to de-identify indirect identifiers. Several methods are illustrated utilizing the model table below.[15] In a table, the columns are known as features and the rows represent records:

Table 2. An example of protected health information.

Age (Years)	Gender	ZIP Code	Diagnosis
15	Male	00000	Diabetes
21	Female	00001	Influenza
36	Male	10000	Broken Arm
91	Female	10001	Acid Reflux

Suppression

This is a common technique used to de-identify data. This method removes data if it is determined to be too risky. Suppression may occur at any level; a feature, a specific value or an entire record may be removed. Table 3 illustrates the application of suppression by the black shaded cells.[16]

Table 3. A version of Table 2 with suppressed patient values.

Age (Years)	Gender	ZIP Code	Diagnosis
[redacted]	Male	00000	Diabetes
21	Female	00001	Influenza
36	Male	[redacted]	Broken Arm
[redacted]	Female	[redacted]	Acid Reflux

Generalization

This technique is also commonly utilized in public health. This method transforms specific data into a more abstract representation. For example, a patient's exact age may be transformed into an age range and a five-digit ZIP Code may be generalized into a three-digit ZIP Code. Table 4 illustrates the application of generalization by the grey shaded cells.[17]

Table 4. A version of Table 2 with generalized patient values.

Age (Years)	Gender	ZIP Code	Diagnosis
Under 21	Male	0000*	Diabetes
Between 21 and 34	Female	0000*	Influenza
Between 35 and 44	Male	1000*	Broken Arm
45 and over	Female	1000*	Acid Reflux

Perturbation

This technique replaces data values with specific, but different data values. Certain statistical properties relative to the data set's values are maintained. Table 5 reflects application of perturbation in the gray shaded cells.[18]

Table 5. A version of Table 2 with randomized patient values.

Age (Years)	Gender	ZIP Code	Diagnosis
16	Male	00002	Diabetes
20	Female	00000	Influenza
34	Male	10000	Broken Arm
93	Female	10003	Acid Reflux

After application of statistical or scientific methods are applied to the data set and the resulting risk of re-identification remains higher than acceptable, policy controls such as a Data Use Agreement may also be applied to mitigate the privacy risk to an appropriate level.

SUPPORTERS

Robert Wood Johnson Foundation

The Network for Public Health Law is a national initiative of the Robert Wood Johnson Foundation.

This document was developed by Sallie Milam, JD, CIPP/US/G, Deputy Director, Network for Public Health Law – Mid-States Region Office, and reviewed by Denise Chrysler, JD, Director, Network for Public Health Law – Mid-States Region Office. The

Network for Public Health Law provides information and technical assistance on issues related to public health. The legal information and assistance provided in this document does not constitute legal advice or legal representation. For legal advice, please consult specific legal counsel.

[1] NIST, De-Identification of Personal Information. NISTIR 8053. Simson L. Garfinkel p. 4. (October 2015). http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf

[2] See, DRAFT NIST Special Publication 800-188 (2nd Draft), De-Identifying Government Datasets, p. 48 (2016), available at http://csrc.nist.gov/pu blications/drafts/800-188/sp800_188_draft2.pdf.

[3] NISTIR 8053 at 2.

[4] Id. at 9; figure below at 9.

[5] NIST 800-188 at 14; see the Memphis Community Health Record Project for a discussion of use case development.

[6] Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, p. 10. (November 26, 2012). https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf.

[7] NIST 800-18. at 15-20. On February 27, 2017, Harvard University's Berkman Klein Center published a framework to manage privacy risk involving publication of de-identified data on the internet. Open Data Privacy: A risk-benefit, process-oriented approach to sharing and protecting municipal data (February 2017), available at https://cyber.harvard.edu/publications/2017/02/opendataprivacyplaybook.

[8] HHS Guidance at 12.

[9] NIST 800-188 at ix.

[10] NISTIR 8053 at 15 (citation omitted).

[11] Id.

[12] Id. at 15-16. For additional information regarding pseudonymization, see pp. 16-17.

[13] Id. at 19 (citation omitted).

[14] See, HHS Guidance at 18.

[15] Id. at 18.

[16] Id. at 19.

[17] Id.

[18] Id. at 19-20.

February, 2019