Stanford
University Privacy Office

DATA DE-IDENTIFICATION OVERVIEW AND GUIDANCE

Background

Stanford routinely de-identifies data before disclosure to third parties, in order to comply with laws and protect the privacy of individuals. This document briefly describes what de-identification is and the risks of re-identification, and provides basic guidance to data stewards at Stanford. Subsequent papers from the University Privacy Office (UPO) will describe other aspects and challenges of de-identification at Stanford.

The identifiability of all data is on a spectrum, from directly identifiable on one end to completely anonymous on the other – with many gradations of identifiably in between. For example, a dataset with names, contact information and photos would be on one side of the spectrum, with data that can be directly associated with specific, unique individuals. On the other side of the spectrum, a dataset with only the number of registered Democrat, Republican and Independent voters in each state would be anonymous – with data that cannot be associated with any specific individual. An example of a dataset in the middle of the spectrum would be a listing of zip codes, dates of birth, occupation and other demographic information of a population – which when combined with other data could potentially be used to indirectly identify specific individuals.

In many circumstances, Stanford desires to share data, but also wants to avoid disclosing the identity of the individuals on which the data is based. For example, Stanford may want to share percentages of faculty with a particular demographic characteristic (e.g., sexual orientation) in order to increase diversity, without disclosing the particular details of any single individual. Or a research team may want to study how ethnicity may correlate to likelihood of having a particular disease – but desires not to reveal the ethnicity of any particular person. In these cases, Stanford may de-identify the data before disclosure.

De-identification is the removal or alteration of information in a dataset in order to make it more difficult or impossible to identify specific individuals. In other words, de-identification moves a dataset along the identifiability spectrum, which protects privacy by reducing the ability to associate data with specific individuals.

There are numerous ways to de-identify data, which are beyond the scope of this paper. But it's helpful to note that de-identification is not a single, standardized process. Rather, de-identification is a myriad of methods and algorithms that can be applied to data, each with different results and levels of effectiveness.

In some cases, the more that data is aggressively de-identified, the more difficult it is to draw individual conclusions from the data – and thus the less useful a researcher may find the data to be. In other words, researchers sometimes encounter a tension between de-identifying data to protect privacy, versus using identifiable data to maintain its utility.

Various laws and regulations worldwide have different definitions of de-identification. For example, the US Health Insurance Portability and Accountability Act (HIPAA) covers health data processed by Stanford – but does not restrict disclosure of defined "de-identified" information, where there is "no reasonable basis to believe that the information can be used to identify an individual." HIPAA contemplates several specific methods that entities like Stanford can use to de-identify data in compliance with the statute. *[See Attachment 1 below for more information.]*

Beyond HIPAA, other statutes in the US and worldwide have very different definitions of de-identification. For example, where Stanford processes personal data in the EU, we have to comply with the EU General Data Protection Regulation (GDPR) – which uses the terms "anonymous" and "pseudonymous." In general, information that has been rendered irreversibly anonymous in such a way that an individual is no longer identifiable is not subject to the requirements of GDPR. In contrast, if information has been "pseudonymized" in such a way that it can still be used to identify or re-identify a person individually, then it remains "personal data" and continues to fall within the scope of GDPR. *[See Attachment 1 below for more information.]*

Re-identification is one key risk that Stanford considers when disclosing de-identified data. Specifically, unless data is rendered irreversibly anonymous on the far end of the spectrum (which experts agree is a very difficult standard to meet in light of evolving technology and expanding availability of external datasets), there is a risk an attacker or other third party could re-associate data back to the specific individuals to which the data relates.

Guiding Principles and Position

Consistent with Stanford's Minimum Privacy Standards (minpriv.stanford.edu), research and operational teams can reduce risks of disclosure of de-identified data in several ways:

- De-identify datasets to the extent possible and practical under the circumstances (e.g., by removing all unnecessary personal data, or by using aggregation, tokenization, or other anonymization techniques).

- Limit the collection and use of personal data to the minimum that is directly relevant and necessary to accomplish the specified purpose. For example, do not process or publish data unless necessary for scientific analysis or research validation purposes.

- Whenever possible and practical, disclose de-identified data only to reputable third parties that contractually commit (1) to use the data only for legitimate research purposes and the advancement of general knowledge, (2) not to re-identify, disclose or use the data for any purposes other than as set forth in the agreement, and (3) not to combine the dataset with any other external data, without Stanford's prior consent.

Ultimately, the designated Stanford owner of a dataset (typically the Principal Investigator, in the context of research) is responsible for ensuring that relevant datasets are de-identified to the maximum extent possible and practical under the circumstances, and consistent with applicable privacy and data protection laws.

The above guidance is intended to apply in addition to all applicable law and Stanford policies and standards.

Stanford faculty, staff and students can submit any questions or comments to UPO at https://privacyrequest.stanford.edu/.

* * *

Attachment 1

HIPAA

Stanford research teams in the School of Medicine often use the HIPAA "Safe Harbor" method to de-identify health data – which requires the removal of ***all*** of the following eighteen identifiers of individuals (and also for any of their relatives, employers and household members, if applicable):

1. Names
2. All geographic subdivisions smaller than a state, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census:
    a) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and
    b) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
4. Telephone numbers
5. Fax numbers
6. Email addresses
7. Social security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers and serial numbers, including license plate numbers
13. Device identifiers and serial numbers
14. Web Universal Resource Locators (URLs)
15. Internet Protocol (IP) address numbers
16. Biometric identifiers, including finger and voice prints
17. Full face photographic images and any comparable images
18. Any other unique identifying number (includes accession numbers), characteristic, or code (includes study codes)

If any of the above identifiers is included, the data is not de-identified for HIPAA purposes – and continues to be considered PHI and "High Risk" data under Stanford's risk classification system (https://uit.stanford.edu/guide/riskclassifications).

Notes:

- If the ***only*** HIPAA identifiers in a dataset are study codes ***and*** the data recipient does not have access to the key which links the study codes to the study participant PHI, we consider the dataset to be de-identified. However, the study codes cannot be derived from information linked to an individual. (For example, a study code cannot be derived from the study participant's initials and year of birth.)
- In order for data to be deemed de-identified, the research team (and everyone else at Stanford) must not have actual knowledge that the data could be used alone or in combination with other data to identify an individual who is a subject of the data.

GDPR

Under GDPR Recital 26, anonymous information is "information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable." The recital further states that "To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments."

In other words, when considering if a particular dataset includes only "anonymous" data, we should consider all means reasonably likely to be used – including what means and other available datasets might be used to re-identify an individual. Entities don't need to be able to prove definitively that it's impossible for any individual to be identified. But rather, entities have to show that it's unlikely that an individual will be identified given the circumstances and the state of technology.

* * *