TEXAS A&M UNIVERSITY
SAN ANTONIO

Institutional Review Board

GUIDELINE #3: DATA DE-IDENTIFICATION FOR SOCIAL BEHAVIORAL SCIENTISTS

I. PURPOSE

This guideline is to ensure that human subjects' research conducted complies with federal, state and local laws, regulations, directives and instructions.

II. STATEMENT

All human subjects research, irrespective of the source of funding, conducted by A&M- SA faculty, staff and students must be submitted and reviewed in accordance with the Federal research regulations, Texas A&M System Guidelines, A&M-SA IRB policies and local consideration.

III. SCOPE

This guideline applies to all research conducted where the A&M- SA IRB serves as the Reviewing IRB.

IV. SUMMARY

De-identification is a tool that organizations can use to remove personal information from data that they collect, use, archive, and share with other organizations.

De-identification is not a single technique, but a collection of approaches, algorithms, and tools that can be applied to different kinds of data with differing levels of effectiveness. In general, privacy protection improves as more aggressive de-identification techniques are employed, but less utility remains in the resulting dataset.

De-identification is especially important for government agencies, businesses, and other organizations that seek to make data available to outsiders. For example, significant medical research resulting in societal benefit is made possible by the sharing of de-identified patient information under the framework established by the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, the primary US regulation providing for privacy of medical records.

As long as any utility remains in the data derived from personal information, there also exists the possibility, however remote, that some information might be linked back to the original individuals on whom the data are based. When de-identified data can be re-identified the privacy protection provided by de-identification is lost. The decision of how or if to de-identify data should thus be made in conjunction with decisions of how the de-identified data will be used, shared or released, since the risk of re-identification can be difficult to estimate.
Risks to individuals can remain in de-identified data. These risks include allowing inferences about individuals in the data without re-identification and impacts on groups represented in the data.

The HIPAA Privacy Rule states that once data has been de-identified, covered entities can use or disclose it without any limitation. The information is no longer considered Protected Health Information (PHI), and does not fall under the same regulations and restrictions as PHI.

Some common direct identifiers of the individual or of relatives, employers, or household members of the individual that a data set cannot include if wanting to be categorized as de-identified are:

Direct Identifiers:

- Names
- 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 the ZIP code if, according to the current publicly available data from the Bureau of the Census:
  (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and
  (2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000
- All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, 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
  - Telephone numbers
  - Vehicle identifiers and serial numbers, including license plate numbers
  - Fax numbers
  - Device identifiers and serial numbers
  - Email addresses
  - Web Universal Resource Locators (URLs)
  - Social security numbers

- Internet Protocol (IP) addresses
- Medical record numbers
- Biometric identifiers, including finger and voice prints
- Health plan beneficiary numbers
- Full-face photographs and any comparable images
- Account numbers
- *Any other unique identifying number, characteristic, or code
- Certificate/license numbers
- The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

* unique identifying number, characteristic, or code

Identifying Number:

There are many potential identifying numbers. For example, the preamble to the Privacy Rule at 65 FR 82462, 82712 (Dec. 28, 2000) noted that "Clinical trial record numbers are included in the general category of 'any other unique identifying number, characteristic, or code.'

Identifying Code:

A code corresponds to a value that is derived from a non-secure encoding mechanism. For instance, a code derived from a secure hash function without a secret key (e.g., "salt") would be considered an identifying element. This is because the resulting value would be susceptible to compromise by the recipient of such data. As another example, an increasing quantity of electronic medical record and electronic prescribing systems assign and embed barcodes into patient records and their medications. These barcodes are often designed to be unique for each patient, or event in a patient's record, and thus can be easily applied for tracking purposes. See the discussion of re-identification.

Identifying Characteristic:

A characteristic may be anything that distinguishes an individual and allows for identification. For example, a unique identifying characteristic could be the occupation of a patient, if it was listed in a record as "current President of State University."

Indirect Identifiers

In addition to the direct identifiers listed above, indirect identifiers may still remain. Indirect identifiers are data elements that may make it possible to identify an individual deductively. Examples of indirect identifiers include:

- Gender
- Race
- Ethnicity
- Religion
- Age

- Marital Status
- Household Composition
- # of Children
- Place of Birth
- Education
- Major
- Income
- Job Title
- Place of Work
- Medical Condition
- Dates (of graduation, arrest, marriage)
- Uncommon Characteristics
- Direct Identifiers of Household Members

It is important to note that the more indirect identifiers investigators collect, the higher the risk of re-identification. In addition, the investigator must keep in mind that there may be information publicly available on the subjects that are unconnected from the data collected for the specific research in question and cumulatively, this information may be used to re-identify the subjects. Family Educational Rights and Privacy Act (FERPA) Data: **Data De-identification—Key Concepts and Strategies**

Privacy of individual student records is protected under FERPA. To avoid unauthorized disclosure of Personally Identifiable Information (PII) from education records, students' data must be adequately protected at all times. For example, when schools, districts, or states publish reports on student achievement or share students' data with external researchers, these organizations should apply disclosure avoidance strategies, to prevent unauthorized release of information about individual students. To ensure successful data protection, it is essential that techniques are appropriate for the intended purpose and that their application follows the best practices.

A vital step in deciding which method to apply involves evaluating available disclosure limitation techniques against the desired level of data protection. To aid educational agencies and institutions with making these decisions and to help ensure consistency of the terminology used by the educational community, Privacy Technical Assistance Center (PTAC) is providing an alphabetized glossary of some frequently used terms. The list includes techniques commonly used to protect privacy of individual student records and types of redacted data files that can be produced by applying these techniques. The accompanying figure provides a summary overview of the main types of data typically managed and disseminated by educational organizations, by level of sensitivity and associated need for protection.

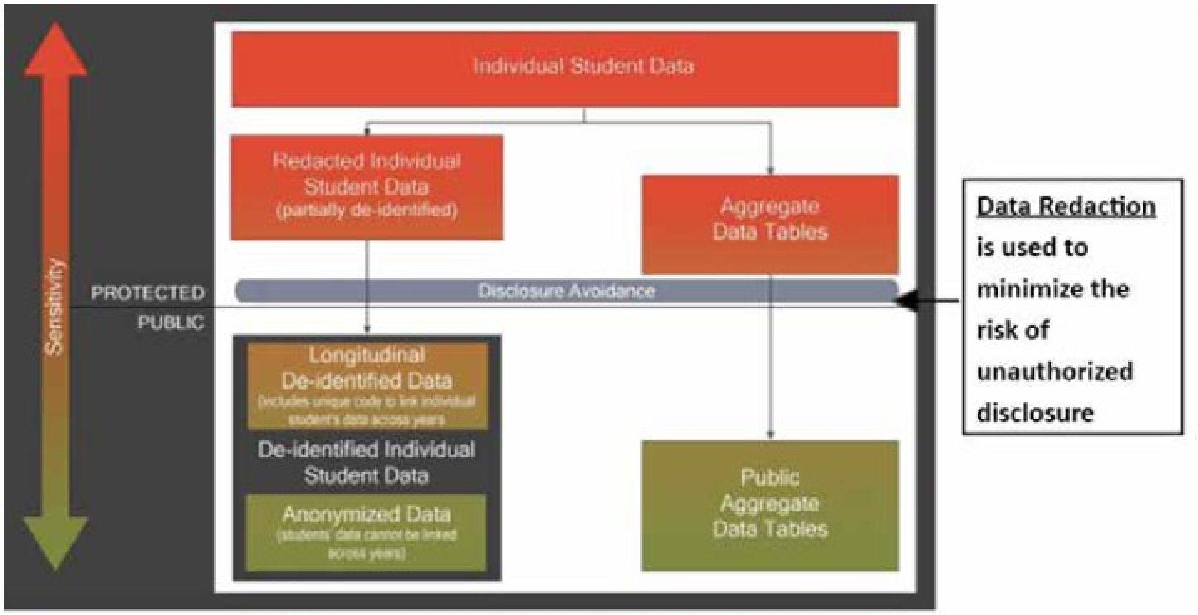
Sensitivity
Individual Student Data
Redacted Individual Student Data (partially de-identified)
Aggregate Data Tables
PROTECTED
PUBLIC
Disclosure Avoidance
Longitudinal De-identified Data (includes unique code to link individual student's data across years
De-identified Individual Student Data
Anonymized Data (students' data cannot be linked across years)
Public Aggregate Data Tables
Data Redaction is used to minimize the risk of unauthorized disclosure

Figure: Types of data by sensitivity and need for protection from unauthorized or inadvertent disclosure.

V. DEFINITIONS

Anonymization [of data] refers to the process of data de-identification which produces de-identified data, where individual records cannot be linked back to an original student record system or to other individual records from the same source, because they do not include a record code needed to link the records. As such, anonymized data are not useful for monitoring the progress and performance of individual students; however, they can be used for other research or training purposes. An anonymized data file could be produced from the de-identified file that contains record codes by removing the codes and reviewing the resulting file to ensure that additional disclosure limitation methods do not need to be applied. The documentation for the anonymized data file should identify any disclosure limitation techniques that were applied and their implications for the analysis.

Blurring is a disclosure limitation method which is used to reduce the precision of the disclosed data to minimize the certainty of individual identification. There are many possible ways to implement blurring, such as by converting continuous data elements into categorical data elements (e.g., creating categories that subsume unique cases), aggregating data across small groups of respondents, and reporting rounded values and ranges instead of exact counts to reduce the certainty of identification. Another approach involves replacing an individual’s actual reported value with the average group value; it may be performed on more than one variable with different groupings for each variable.

De-identification [of data] refers to the process of removing or obscuring any personally identifiable information from student records in a way that minimizes the risk of unintended disclosure of the identity of individuals and information about them. Specific steps and methods used to de-identify information (see disclosure limitation method for details) may vary depending on the circumstances, but should be appropriate to protect the confidentiality of the individuals. While it may not be possible to remove the disclosure risk completely, de-identification is considered successful when there is no reasonable basis to believe that the remaining information in the records can be used to identify an individual.

De-identified data *may be* shared without the consent required by FERPA (34 CFR §99.30) with any party for any purpose, including parents, general public, and researchers (34 CFR §99.31(b)(1)). These data are typically released in the form of aggregated data (such as tables showing numbers of enrolled students by race, age, and sex) or microdata (such as individual-level student assessment results by grade and school). Individual-level data may be released with or without an attached record code (record code cannot be based on the student's personal information), which allows education researchers to track performance of individual students, without students' identity being revealed to them (34 CFR §99.31(b)(2)). The researchers can use the code only to match individual records across previously de-identified data files from the same source (e.g., to compare student assessment results from the same school district over several years); the researchers cannot use the code to access the original data source without consent. (Note that coded individual-level data can only be released for the purposes of education research and are subject to certain conditions—see record code for more information.) De-identified data which do not include a record code and cannot be linked to the original data source are referred to herein as anonymized.

It is important to note that PII may include not only direct identifiers, such as names, student IDs or social security numbers, but also any other sensitive and non-sensitive information that, alone or combined with other information that is linked or linkable to a specific individual, would allow identification. Therefore, simple removal of direct identifiers from the data to be released DOES NOT constitute adequate de-identification. Properly performed de-identification involves removing or obscuring all identifiable information until all data that can lead to individual identification have been expunged or masked.

Further, when making a determination as to whether the data have been sufficiently de-identified, it is necessary to take into consideration cumulative re-identification risk from all previous data releases and other reasonably available information, including publicly-available directory information and de-identified data releases from education records as well as other sources. In particular, care should be taken to monitor new releases of de-identified individual-level student data that are released with an attached record code.

Disclosure means to permit access to or the release, transfer, or other communication of PII by any means (34 CFR §99.3). Disclosure can be authorized, such as when a parent or an eligible student gives written consent to share education records with an authorized party (e.g., a researcher). Disclosure can also be unauthorized or inadvertent (accidental). An unauthorized disclosure can happen due to a data breach or a loss (see PTAC's Data Security: Top Threats to Data Protection brief at https://studentprivacy.ed.gov/ for more information and security tips). An accidental disclosure can occur when data released in public aggregate reports are unintentionally presented in a manner that allows individual students to be identified.

It is important to note that the release of education records that have been de-identified is not considered a "disclosure" under FERPA, since by definition de-identified data do not contain PII that can lead to identification of individual students. This statement holds true regardless of whether de-identified data have been released with an attached record code or without it; however, releases of coded de-identified data are subject to certain conditions (see record code for more information).

Disclosure avoidance refers to the efforts made to de-identify the data in order to reduce the risk of disclosure of PII. A choice of the appropriate de-identification strategy (also referred to as disclosure limitation method) depends on the nature of the data release, the level of protection offered by a specific method, and the usefulness of the resulting data product. The two major types of data release are aggregated data (such as tables showing numbers of enrolled students by race, age, and sex) and microdata (such as individual-level student assessment results by grade and school). Several acceptable de-identification methods exist for each type of data (see *disclosure limitation method* for more details).

De-identification strategy. See *Disclosure limitation method.*

Disclosure limitation method (also known as disclosure avoidance method) is a general term referring to a statistical technique used to manipulate the data prior to release to minimize the risk of inadvertent or unauthorized disclosure of PII. Entities releasing data should apply a consistent de-identification strategy to all of their data releases of a similar type (e.g., tabular and individual-level data) and similar sensitivity level. It is advised that organizations document their data reporting rules in the documents describing their data reporting policies and privacy protection practices, such as a Data Governance Manual. (See PTAC's Data Governance and Stewardship brief at https://studentprivacy.ed.gov/resources/issue-brief-data-governance-and-stewardship for more information on best practices in data governance.)

The major methods used by the U.S. Department of Education for disclosure avoidance for tabular data include defining a minimum cell size (meaning no results will be released for any cell of a table with a number smaller than "X" or else cells are aggregated until no cells based on one or two cases remain) and controlled rounding (meaning that cells with a number smaller than "X" require that numbers in the affected rows and columns be rounded so that the totals remain unchanged). Whenever possible, data about individual students (e.g., proficiency scores) are combined with data from a sufficient number of other students to disguise the attributes of a single student. When this is not possible, data about small numbers of students are suppressed.

For releases of student-level data, the primary consideration is given to evaluating whether the proposed release contains any individuals with unique characteristics whose identity can be deduced by the combination of variables in the file. If such a condition exists, one of a number of methods is employed. These include data blurring, such as "top-coding" a variable (e.g., test scores above a certain level are recoded to a defined maximum), and applying various data perturbation techniques. For additional guidance on specific steps and acceptable methods for de-identifying student data, see the list of Resources at the end of the paper.

Masking is a disclosure limitation method that is used to “mask” the original values in a data set to achieve data privacy protection. This general approach uses various techniques, such as data perturbation, to replace sensitive information with realistic but inauthentic data or modifies original data values based on pre-determined masking rules (e.g., by applying a transformation algorithm). The purpose of this technique is to retain the structure and functional usability of the data, while concealing information that could lead to the identification, either directly or indirectly, of an individual student. Masked data are used to protect individual privacy in public reports and can serve as a useful alternative for occasions when the real data are not required, such as user training or software demonstration. Specific masking rules may vary depending on the sensitivity level of the data and organizational data disclosure policies.

Perturbation is a disclosure limitation method which involves making small changes to the data to prevent identification of individuals from unique or rare population groups. Data perturbation is a data masking technique in that it is used to “mask” the original values in a data set to avoid disclosure. Examples of this statistical technique include swapping data among individual cells to introduce uncertainty, so that the data user does not know whether the real data values correspond to certain records, and introducing “noise,” or errors (e.g., by randomly misclassifying values of a categorical variable).

Record code refers to the unique descriptor that can be used to match individual-level records across de-identified data files from the same source (e.g., for the purposes of comparing performance of individual students over time). FERPA (34 CFR §99.31(b)(2)) allows an educational agency or institution, or a party that has received education records or information from education records, such as a State educational authority, to release de-identified student-level data (microdata) from education records for the purpose of educational research by attaching a code to each record that may allow the researcher to match information received from the same source under the specified conditions. These conditions require that the coded de-identified microdata are used only for educational research purposes, that the party receiving the data is not allowed any access to the information about how the descriptor is generated and assigned, and that the code cannot be used to identify the student or to match the information from education records with data from any other source. Furthermore, a record descriptor may not be based on a student’s social security number or other personal information.

Redaction is a general term describing the process of expunging sensitive data from the records prior to disclosure in a way that meets established disclosure requirements applicable to the specific data disclosure occurrence (e.g., removing or obscuring PII from published reports to meet federal, state, and local privacy laws as well as organizational data disclosure policies). (See *disclosure limitation method* for more information about specific techniques that can be used for data redaction.)

Suppression is a disclosure limitation method which involves removing data (e.g., from a cell or a row in a table) to prevent the identification of individuals in small groups or those with unique characteristics. This method may often result in very little data being produced for small populations, and it usually requires additional suppression of non-sensitive data to ensure adequate protection of PII (e.g., complementary suppression of one or more non-sensitive cells in a table so that the values of the suppressed cells may not be calculated by subtracting the reported values from the row and column totals). Correct application of this technique generally ensures low risk of disclosure; however, it can be difficult to perform properly because of the necessary calculations (especially for large multi-dimensional tables). Further, if additional data are available elsewhere (e.g., total student counts are reported), the suppressed data may be re-calculated.

VI. APPLICABLE REGULATIONS AND GUIDELINES

HIPPA Guidelines:

https://studentprivacy.ed.gov/sites/default/files/resource_document/file/data_deidentification_terms.pdf

References:

https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf

Guideline: # 3	**Version: V.1**
Title: Data de-identification for Social Behavioral Scientists	
Authorized: Dr. Vijay Golla, Vice Provost for Research and Graduate Studies	
Date Approved by IO: March 29, 2023	