T-CAIREM De-Identification Policy

Introduction

To protect the privacy of research subjects within the Health Data Nexus ("the HDN"), all datasets are expected to be de-identified prior to submission. De-identification is the general process of removing personal identifiers from a dataset to prevent the identification of an individual therein. The Personal Health Information Protection Act (PHIPA, 2004) defines "de-identify" as "to remove any information that identifies the individual or for which it is reasonably foreseeable in the circumstances that it could be utilized, either alone or with other information, to identify the individual". Guidance from the Information and Privacy Commissioner (IPC) of Ontario highlights the utility of de-identification, as a dataset which does not contain personal information cannot violate the privacy of individuals. The goal of this document is to summarize the policy of T-CAIREM for accepting de-identified datasets. De-identification guidance is adapted from guidelines provided by the IPC. Reading this document in full is highly recommended for institutions aiming to submit de-identified data to the HDN[1].

Upon submission of a dataset to the HDN, a detailed description of the de-identification approach used in the creation of the dataset must be provided by the Data Holder. This includes the anticipated Personal Health Information (PHI) entities within the dataset, which identifiers were removed, the method of their removal, specific de-identification software packages used (if any), and considerations for re-identification attacks, including which publicly available datasets may be used for linkage attacks. Further, evidence should be provided regarding the efficacy of the de-identification approach. This evidence should include information about any quasi-identifiers considered, and how they were handled. If perfect de-identification cannot be guaranteed, provisions within the Data Use Agreement must outline a process to be followed in the event that the dataset is discovered to contain PHI.

Data release

Datasets are made available under a restricted access model within the HDN. Only Authorized Data Users (individuals who have verified their affiliation with a research or education institution, signed both the Code of Conduct and dataset-specific Data Use Agreement, and completed training in research with human participants) may have access. The restricted access nature of HDN datasets provides additional protection

[1] https://www.ipc.on.ca/wp-content/uploads/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf

against re-identification risk for datasets and should be considered when designing the de-identification process.

De-identification Process

All data provided to the HDN must be *de-identified*, that is, it must not contain individually identifiable information. PHIPA defines individually identifiable information as information that identifies an individual or for which it is reasonably foreseeable in the circumstances that it could be utilized, either alone or with other information, to identify an individual.
There are two broad classes of PHI: **direct identifiers** and **indirect identifiers** (**quasi-identifiers**). Direct identifiers immediately reference the individual: these include name(s), medical record number, social insurance number(s), email addresses, and so on. Indirect identifiers do not immediately reference the individual but could reasonably be used to ascertain their identity. Indirect identifiers include dates of procedures, profession, demographic information (age, gender, ethnicity), and so on.

Data which could reasonably lead to participant identification **must** be removed prior to contributing a dataset to the HDN. This includes identifiers relating to an individual patient, their families, providers, and all other individuals involved in the patient's care.

T-CAIREM requires the following identifiers to be removed:

- Names
    - Includes names of providers, family members, and other individuals affiliated with the individual's care
- Dates
    - Includes birth date, admission/discharge date, death date, date of service, etc.
- Location
    - Geographic subdivisions including city, province, country, street address, county, postal code, referring to home, work, institution, or etc.
    - Institution names (e.g. Hospital for Sick Children, SickKids)
    - Nationalities where that nationality implies a location (e.g. Columbian)
    - Does **not** include generic locations within an institution such as emergency department, medical ward, etc.
- Contact information
    - Telephone numbers, fax numbers, email addresses, IP addresses, web URLs
- Individual specific identifiers

    - Driver’s license number, social insurance number, passport number, medical record number, health plan identifiers, insurance identifiers, account numbers, device identifiers, serial numbers
- Biometric identifiers
    - Voice prints, finger prints
- Full-face photographs and any comparable images
- Any other uniquely identifying characteristics
    - Examples include specific profession (e.g. executive of a specific company), or membership in a small group (e.g. significantly advanced age)

In addition to the removal of identifiers, T-CAIREM also requires **an assessment of quasi-identifiers** in the dataset, and justification regarding their processing. For example, if quasi-identifiers may exist but are deemed a theoretical risk which cannot be practically protected against, the de-identification approach must acknowledge this and describe the non-technical measures used to account for these entities (e.g. clauses within the Data Use Agreement).

Methods of removal

A number of alternative methods for de-identification of datasets are acceptable within the T-CAIREM platform. These methods are:

- Deletion of PHI
- Encryption of PHI
- Replacement of PHI with surrogate information

The approach to de-identification of a single dataset may adopt one or all three of these methods, depending on the specific context and entities being considered.

Deletion of PHI involves the removal of all PHI entities from the given dataset. For example, a column of medical record numbers could be deleted from a tabular dataset of patient demographics.

Encryption of PHI involves replacing PHI entities with an encrypted value which cannot be deciphered by users, but which could be used by the originator of the dataset to reidentify individuals. For example, medical record numbers could be replaced with random integer identifiers, a form of format preserving encryption which allows data originators to reidentify de-identified records. PHI encryption may be useful as a means of consistently assigning an anonymized identifier to the same individual, and particularly so in cases where the dataset may be expanded in the future. Care must be taken to ensure that the encryption process is robust against nefarious actors, and it should be well described in the de-identification report.

Finally, PHI may be replaced by surrogate values which do not convey meaning regarding the individual’s care but help to obscure potential errors in the de-identification approach (“hiding in plain sight”). For example, an individual’s last name could be replaced by a random name (e.g. “Jones” to “Smith”).

The choice of de-identification method is to be made by the Data Holder and depends on several factors including the type of data submitted, risks associated with the underlying entity, and the difficulty of the de-identification task. It is up to the Data Holder to justify their decisions regarding the approaches they have undertaken.

Linked datasets and methods of managing quasi-identifiers

The dataset may include quasi-identifiers which are unlikely to lead to any direct identification, but may, along with other potential linked information, lead to identification. As part of the evaluation of quasi-identifiers, research should be carried out into publicly available data that may possibly be linked to the provided dataset, such as publicly available hospital data or genetic data which may be linked to other research datasets. While linking to external datasets at the individual level is forbidden by the terms of the Data Use Agreement, it is worthwhile understanding how the data could be used if the agreement were breached.

Two additional strategies may be employed to manage quasi-identifiers:

- Aggregation
- Suppression

Aggregation involves grouping into categories data which may be uniquely identifiable but may also provide useful information. For example, age or year of birth may be used to identify those of significantly advanced age (over 90), but instead arranging the dataset in groups (90+, 85-89, …) may alleviate some of these concerns while still providing useful information.

On the other hand, suppression involves the targeted removal of individual entries in the dataset rather than entire entries—for example, if language spoken is recorded in the dataset and one individual speaks a unique or uncommon language. However, since suppression entails the loss of data, it should be approached cautiously.

To ensure that quasi-identifiers are managed appropriately and do not lead to re-identification, one common method is to first group together all relevant data fields (age, gender, ethnicity, etc.) and treat as identical any two individuals who have the same data in all categories. A threshold of uniqueness must then be satisfied in any public dataset which links to the research data (e.g. census data, cellphone records, etc). A typical threshold is to ensure that there are at least **20** individuals in each category, or an **average of 20** individuals across all classes and a minimum of at least 2 individuals per class.

De-identification "cheat sheet"

This "cheat sheet" lists the information provided above, along with choices for how to approach different data fields.

- Name
  - (1) Replace with a unique identifier for each individual
  - (2) [Optional] A dataset *may* be retained by the health data custodian which links unique patient IDs back to patient names. If so, this may not be provided to T-CAIREM.
- Contact Information
  - Delete
- Health Number and Other Specific Identifiers
  - Delete
- Biometric ID or Identifying Photographs
  - Delete
- Locations
  - (1) Code locations which may be relevant with unique identifiers, such as hospital name/ward name/etc
  - (2) [else] Delete
- Identifying dates (birth, death, procedure)
  - (1) Scramble dates so that they are adjusted by a random amount for each patient but internally consistent for a single patient
  - (2) [else] Delete
- Age
  - (1) Aggregate (e.g. 5-year ranges, including 15 and below, 90 and over)
  - (2) Suppress individual outlier values
  - (3) [else] Delete
- Demographic info (gender, ethnicity, language, etc)
  - Combination of:
  - (1) Aggregate into larger categories

- (2) Suppress individual values if they are uniquely identifiable
- (3) Delete columns if they are unhelpful, particularly sensitive (e.g. religion, criminal history), and/or sufficient aggregation cannot be maintained