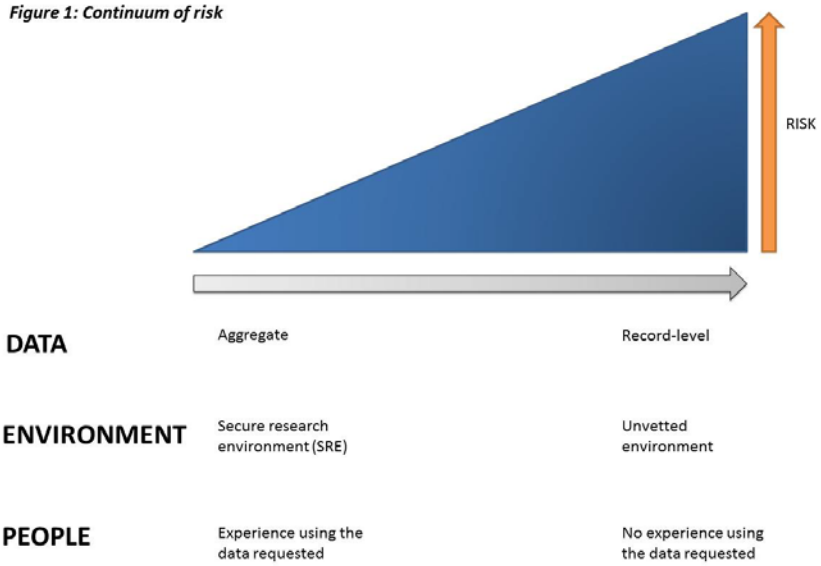
Research Privacy Tip Sheet: Common Terms and Tips to Reduce the Risk of Exposing Identifiable Personal Information[1]

This Tip Sheet references guidance from key Canadian documents that must be followed when conducting research within PHSA supplemented with advice from a range of data experts from both within and outside of PHSA.

- Intended audience: Researchers and research staff who are involved in the conduct of research involving human participants or data on human participants.
- How this guidance can be used: Researchers and research staff who seek assistance in meeting privacy requirements can use this guidance to understand common privacy terms and the steps they can take to reduce the risk of exposing private information.

It is important to keep in mind that the steps you take when sharing, using, or storing data must be at a standard that is reasonable given the circumstances. This is subjective, but necessarily so as there may be a higher reasonable standard for some situations as opposed to others. You can never eliminate all possible risks under all circumstances when sharing, using, or storing data. Figure 1 below provides an illustration for how to think about the continuum of risk associated with sharing, using, or storing data.

Figure 1: Continuum of risk



[1] We gratefully acknowledge all the data experts who contributed to this Tip Sheet.

Definitions

- **Aggregate information:** Summed or categorized data (e.g., total number of visits to the Emergency Department in the last year). The data have been compiled from record-level data to a summary level that ensures the identities of individuals or groups cannot be determined by a reasonably foreseeable method (Source: PHSA Performance Measurement & Reporting (PMR) glossary). Aggregate data is typically used for the purposes of reporting or statistical analysis. It is important to note that aggregate data is not synonymous with de-identified data (Source: Information from the Ministry of Health (MoH)).
    - TCPS2[2] offers an example of the types of risks aggregate data can pose to participants and should be considered by researchers even though these data may be difficult to link to individuals. For example, aggregate data provided to authorities about research on illicit drug use in a penitentiary may pose risks of reprisal to the prisoners, even though they are not identified individually. An additional example is provided when conducting research with indigenous communities. Whatever the nature of the research, it shall be designed to include safeguards for participant privacy and measures to protect the confidentiality of any data collected. Small Aboriginal communities are characterized by dense networks of relationships. As a result, coding individual data is often not sufficient to mask identities, even when data are aggregated. Finally, TCPS2 asks Research Ethics Boards (REBs) to be responsive to new safeguarding mechanisms when reviewing aggregate data from genome-wide association studies to reduce the risks of re-identification.

- **Anonymized information:** The data product result after anonymization. The information is irrevocably stripped of direct identifiers, and information for which there is a reasonable expectation that it could be used, either alone or with other information, to identify the individual (i.e. indirect identifiers). A code is not kept to allow future re-linkage, and risk of re-identification of individuals from remaining indirect identifiers is low or very low (Source: TCPS2 (2014) supplemented with information from the MoH).

- **Anonymous information:** The information never had identifiers associated with it (e.g., anonymous surveys) and risk of identification of individuals is low or very low (Source: TCPS2 (2014)).

- **Coded information:** Direct identifiers are removed from the information and replaced with a code. Researchers with access to the code may re-identify specific participants, if necessary (e.g., the principal investigator retains a list that links the participants' code names with their actual name so data can be re-linked if necessary) (Source: TCPS2 (2014) with additions/modifications from reviewers).

[2] The Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS or the Policy) is a joint policy of Canada's three federal research agencies – the Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Social Sciences and Humanities Research Council of Canada (SSHRC), or "the Agencies." To be eligible to receive and administer research funds from the Agencies, institutions must agree to comply with a number of Agency policies set out as schedules to an Agreement between the Agencies and institutions. This Policy is referenced in Schedule 2 to that Agreement. Institutions must therefore ensure that research conducted under their auspices adhere to this Policy. Researchers are expected, as a condition of funding, to adhere to the TCPS. Institutions should support their efforts to do so (Source: TCPS2 (2014) page 3).

- **Crosswalk file:** A file that contains identifiable information that can be linked back to the coded data in a separate location that is password protected and encrypted (e.g., a separate file in a secure server or a separate project in REDCap) (Source: Reviewers).

- **Data linkage:** Identifiable data based on human participants (i.e., administrative data) may be linked to other research, administrative, or public databases. Such data linkage can be a powerful research tool and a valuable resource for monitoring the health of populations, understanding factors influencing disease, and evaluating health services and interventions. However, data linkage also raises separate privacy issues and obligations under, for example, the *Freedom of Information and Protection of Privacy Act (FIPPA)* that need to be complied with.

  Data linkage must be carefully reviewed by REBs as discussed in Article 5.7 of TCPS2. This Article states that researchers who propose to engage in data linkage shall obtain REB approval prior to carrying out the data linkage, unless the research relies exclusively on publicly available information as discussed in Article 2.2. (Although it is also pointed out in TCPS2 that the REB should consider if data linkage of two or more datasets of anonymous information may present risks of identification (see Article 2.4 or Article 9.22). The application for approval shall describe the data that will be linked, what variables will be used for linkage, how the data will be linked, who will conduct the linkage, and the likelihood that identifiable information will be created through the data linkage. Where data linkage involves or is likely to produce identifiable information, researchers shall satisfy the REB that:

  (a) the data linkage is essential to the research; and
  (b) appropriate security measures will be implemented to safeguard information. (Source: TCPS2 (2014), FIPPA, and reviewers).

- **Data Management Plan:** A data management plan or DMP is a formal document that outlines the lifecycle of data including how and by whom the data are to be handled both during a research project, and after the project is completed. The goal of a data management plan is to consider the many aspects of data management, metadata generation, data transfer, data storage, data preservation, and analysis before the project begins; this ensures that data are well-managed in the present, and prepared for preservation in the future (Source: MoH with additions from reviewers).

- **Data Steward:** Refers to a public body, such as PHSA, that has ultimate responsibility for a given data source. In practice, an individual is typically named as having the authority to approve or reject research requests involving that data, typically called “the / a Data Steward.” (Source: PMR glossary).

- **De-Identified information:** Data modified to remove the association between a set of identifying data and the data subject (the human participant or patient). Records are modified so that the identity of the individual or group cannot be determined by a reasonably foreseeable method. De-identification methods can include removing personal identifiers (e.g., names and personal health numbers); providing age groups instead of date of birth; and providing the fiscal period of admission instead of admission date. While the data may undergo de-identification methods, it may include preserving identifying information which could only be re-linked by a trusted party in certain situations.

It is possible for de-identified data to be either (i) data that may be subject to re-identification or (ii) anonymized data – where the likelihood of re-identification is very small. Once personal information is de-identified to the extent that the de-identified data is considered to no longer contain personal information, it is considered anonymized data (or coded data if a code is maintained) and the privacy protection provisions no longer apply (see the *de-identification as a tool* section below for a more complete discussion of this process as a tool to reduce the potential for re-identification) (Source: Reviewers and the PMR glossary).

- **Direct identifiers:** Data elements that identify an individual without additional information (e.g., name, address, telephone numbers). Direct identifiers are considered personal information (Source: MoH).

- **Directly identifying information:** Information that identifies a specific individual through direct identifiers (e.g., name, social insurance number, personal health number (PHN), medical record number (MRN)) (Source: TCPS2 (2014) with additions from reviewers).

- **Indirect identifiers:** Data that can identify an individual indirectly. Data that can help connect pieces of information to single out an individual (e.g., date of birth, place of residence, or unique personal characteristic). As an example, Canadians have a high risk of re-identification using only their date of birth and postal code. While indirect identifiers individually may not be personal information, they are considered personal information if they can be combined together to single out an individual (Source: MoH and additions from reviewers).

- **Indirectly identifying information:** The information can reasonably be expected to identify an individual through a combination of indirect identifiers (e.g., date of birth, place of residence or unique personal characteristic) (Source: TCPS2 (2014)).

- **Personal information:** According to FIPPA of BC, the definition of personal information is quite broad. Schedule 1 of this Act defines it as follows: *"recorded information about an identifiable individual <u>other than contact information</u>."* However, it is important to note that the definition excludes contact information, which is defined as *"information to enable an individual at a place of business to be contacted and includes the name, position name or title, business telephone number, business address, business email or business fax number of the individual"* So in other words, the information associated with a person's place of business is not considered personal information (Source: FIPPA BC).

- **Record-level data:** Each individual case or patient encounter represents one record (Source: PMR glossary).

- **Re-identification:** The reversal of de-identification or the process of re-identifying individuals that have been de-identified (as defined above) by matching the de-identified data in the dataset to an individual (Source: MoH).

- **Secure Data Environment:** A physical data environment designed to meet a set of defined security and privacy criteria (Source: MoH).

Justifying data requests

When requesting data from a registry, Data Stewards will generally not release direct identifiers such as name, address, Personal Health Number (PHN), phone number, etc. without a very strong justification. They will also want indirect identifiers like full date of birth, and full postal code to be justified since these elements can be combined to identify participants. It is also important to justify the level of data requested. Can the request be fulfilled using aggregate data instead of record level data, for example? Similarly, REBs will also require significant justification for the use of these direct or indirect identifiers for the same reasons. Please check with your Data Steward to determine which identifiers are necessary, justified, and are available for your study before submitting your ethics application to the REB for review.

> As stated in TCPS2, technological developments have increased the ability to access, store and analyze large volumes of data. These activities may heighten risks of re-identification, such as when researchers link datasets, or where a dataset contains information about a population in a small geographical area, or about individuals with unique characteristics (e.g., uncommon field of occupational specialization, diagnosis of a very rare disease). Various factors can affect the risks of re-identification, and researchers and REBs should be vigilant in their efforts to recognize and reduce these risks. The REB should consider if the data linkage of two or more datasets of anonymous information may present risks of identification.

De-identification as a tool

The process of de-identification is a tool used to reduce the sensitivity of data, and to reduce the potential for re-identification. However, de-identification of data is not a straightforward process. There are limits to the ability to de-identify some data and it may be impossible to reasonably remove the possibility of re-identification in some cases. When de-identifying data, it is good practice to remove all direct identifiers, and to either remove or pool indirect identifiers, subject to the needs of the research. For example, 6-digit postal code can be rolled up to Health Authority or Health Service Delivery Area, age at visit or month and year of birth can be used instead of the individual's full birth date (day/month/year), or one could use the mechanism of date shifting where the dates in the database are shifted so that they do not correspond with the actual dates. The participant code used for the data should be unique and not include any identifying information about the participant.

You should then keep the crosswalk file that contains identifiable information that can be linked back to the coded data in a separate location that is password protected and encrypted (e.g., a separate file in a secure server or a separate project in REDCap). The person selected to be responsible for the crosswalk must be capable of ensuring the security of the crosswalk and that access to the crosswalk is only granted in accordance with the Data Management Plan. This person should also be listed in the Data Management Plan. Please check with PHSA IMITS or your responsible research IT office for consultation on data security tools and methods. Your Data Stewards can also help you develop a sensible de-identification process that is appropriate for the data you plan to use in your study. You can also reach out to the PHSA Research Privacy Advisor for support and assistance.

How one measures the risk of re-identification is a key factor in this process that is often contentious. It may be useful to employ scales or frameworks to help determine what constitutes a low, medium, or high risk situation given a particular context. In high or very high risk situations, it would be appropriate to use more stringent methods of de-identification.

In some cases, you may want to draw on tools like the one developed by Dr. Khaled El Emam below to help determine the quantitative % for the risk of re-identifiability. However, the substantive difference regarding how to treat data that has a 15% versus a 22% risk of re-identification is difficult to determine. In addition, it would not be reasonable to insist that a researcher use these types of tools for most study-related databases. So like other processes discussed here, this is a tool to help minimize risk in particular situations and not a solution to eliminate all possible re-identification risks in all cases. It is best practice to request the level of detail in data, especially direct and indirect identifiers, that is required to address your research questions and nothing more.

- The Canadian company - Privacy Analytics. Dr. Khaled El Emam is the founder of Privacy Analytics Inc. Khaled is also a senior scientist at the Children's Hospital of Eastern Ontario (CHEO) Research Institute and Director of the multi-disciplinary Electronic Health Information Laboratory (EHIL) team, conducting academic research on de-identification and re-identification risk.
  Website: https://privacy-analytics.com/

Data management plans

A Data Management Plan (DMP) is a document you create that stipulates how you will organize, transfer, store, preserve or destroy, and share your research data at each stage in your project. Many ethics issues can emerge throughout the lifecycle of these data. For example, it is important to anticipate how sensitive data that must be retained for many years can be stored safely and securely long after your research study is complete.

A DMP is a living document that can be modified to accommodate changes in the course of your research. A DMP is a high level plan: no private data is exposed. For step-by-step guidance in creating a Data Management Plan, we recommend the following online tools:

- DMP Assistant (Portage initiative) to create data management plans for Canadian funders. http://researchdata.library.ubc.ca/plan/
- DMP Tool to create data management plans for US funders, such as NIH or NSF. http://researchdata.library.ubc.ca/plan/

It is important to provide research service providers with all the information they need to evaluate your study. For studies that plan to share and link data, it is important to include a data management plan that details the entire lifecycle of these data including the preservation or destruction phase. Failure to include such a plan with your ethics application, for example, makes it very difficult for the REB and other reviewers to evaluate your study and may result in a deferral.

Secure research environments and security measures

Many privacy risks can also be managed by using a secure research environment (SRE) to analyze data such as the one provided by Population Data BC (PopData). That said, it would not be reasonable to insist that researchers use SRE's for all research-related data and for all studies. In most cases, it is sufficient to use a range of security measures like the ones listed below.

- Collect only the data needed for the research study.

- De-identify data as soon as possible after collection and/or separate identifiable data from coded data. If this is not possible, a higher level of data security must be observed to safeguard the original data.
- Keep the key to re-identify coded data separate and in a secure location. Only share it with others when truly necessary.
- Store and access research data only on computers connected to our secure networks: PHSA, BC Cancer Agency-CRC, BC Cancer Agency-GSC or BC Children's Hospital Research Institute.
- Use secure data encryption if identifiable information will be stored on a networked computer, stored or transmitted via the web, or stored on a portable device such as a laptop or USB flash drive.
- If there is a need to create research databases or applications that use personal identifiable information (and this activity has been approved by a research ethics board), the appropriate privacy office should be consulted.
- When disposing or transferring ownership of computers, CDs, USB keys, and any other form of electronic storage, make sure sensitive data is irretrievably deleted by trained IT professionals following industry standards.

You can also help to manage risks involved in data sharing by using a Secure File Transfer service such as the one below and then ensuring that files that contain de-identified data are password protected and encrypted and held in secure computers that are also password protected. One should only move data when they need to be moved. Privacy risks can be managed by not linking data to any other information without proper permission and by not emailing and downloading it unnecessarily or storing it in multiple locations without proper justifications.

- The IMITS Secure File Transfer service allows anyone with VCH, PHC or PHSA email credentials to share documents and other files with acceptable external recipients. You can use it in place of Dropbox or other cloud-based file sharing services.
  Website: http://imitsinfocentre.healthbc.org/resources/secure-file-transfer

** For database and application development to support your research study, please consult IMITS or with your responsible research IT office for secure and sustainable research systems.