# Using Data About You for Research: Who, How, and Why

A public deliberation with British Columbians

April 7-8 and 21-22, 2018
Vancouver, BC
University of British Columbia

*This booklet was prepared by the BC Data Deliberation research team (listed in alphabetical order):*

*Colene Bentley*
*Michael Burgess*
*Kimberlyn McGrail*
*Kieran O'Doherty*
*Jack Teng*

*April, 2018*
*Vancouver, British Columbia, Canada*

## Table of contents

# 1 | What is the purpose of this booklet?

This booklet was prepared to inform and stimulate discussion for the *Using Data About You for Research: Who, How, and Why* public deliberation. The information in this booklet comes from academic literature, consultations with experts and stakeholders, and media such as newspapers and radio.

The intent of this booklet is to inform you on how **data** about you are collected, and how they are used for research. The regulations that need to be followed for them to be shared for research will also be discussed. The booklet will provide you with a good information base for conversations during the public deliberation. We hope it will encourage ongoing discussion and reflection on this topic.

In this booklet, you will find information describing the nature of data that are used for research and the types of individuals and organizations that collect them. You will also find information on how data from different sources or organizations can be combined to create **linked data** and how using linked data allows **research** to be done. You will find examples of the potential benefits of research using linked data, and you will also find descriptions on the potential risks of using linked data. Finally, you will find descriptions of current data access practices, procedures, and the legislation that must be followed to share data.

> As an important clarification, the data we are referring to in this deliberation are **individual-level data**. Data about individuals can include things like the number of visits to the doctor, number of trips by bus, the number and type of prescriptions for medicine. When data are used for research, **direct identifiers** like names and Personal Health Numbers are removed, that is, the data become **de-identified**. De-identified data can contain quite detailed information, but without an easy or direct way to connect those details to a specific person.

The information in this booklet does not cover all the possible issues surrounding the use of data for research. Of particular note, the information and the issues we are addressing in this deliberation refer to the policies and regulations affecting researchers and public bodies, not private corporations. While we will be discussing data collected by private enterprises (e.g., corporations), the policy suggestions from the deliberation will be targeted at academic researchers who use the data for the benefit of population health and well-being, and public bodies that set the rules around that use.

You will likely have further questions after reading this booklet. We encourage you to bring these questions to the deliberation, as well as any insights and perspectives you may have.

A glossary and list of resources can be found at the end of this booklet. Glossary terms appear in bold lettering throughout the booklet.

## 2 | What is a public deliberation?

A **public deliberation** is a community discussion on issues that affect members of the public such as yourself. It is a democratic process that supports citizens to understand issues and different people's perspectives about those issues. It is also intended to make group recommendations and/or identify disagreement. Policy makers, experts, and stakeholders may provide information or attend the deliberation as observers.

For this deliberation, people from British Columbia have been selected to reflect the diversity of life experiences and perspectives of British Columbians.

## 3 | What is the importance of public deliberations? What is the importance of this deliberation?

Public deliberations are democratic discussions about important societal issues. Instead of experts *telling* the public what they need to know about an issue, they invite the public into *active participation*. Members of the public have an opportunity to identify what is important to them about a societal issue and provide advice, in the form of recommendations, to policy makers. These recommendations are important for issues where there are difficult trade-offs. In this deliberation, the trade-offs involve balancing the ability to conduct research while protecting privacy.

We intend both to educate and seek advice from you and other members of the public during this deliberation about the use of individual-level data for research. Our hope is to inform policy by bringing together people with different backgrounds, opinions, and life experiences. By working together, you will identify what is important to you with regards to the sharing of individual-level data for research. You and other deliberation participants will combine your knowledge, perspectives, and advice to create policy recommendations that reflect the diversity of values, experience and opinions of British Columbians.

A public deliberation is about respecting the diversity of perspectives amongst us and finding ways we can live together. The information you read and hear may inform your opinion, and your opinion may change over time or may not. The intent is to inform and engage citizens as they discuss issues and make recommendations.

# 4 | Introduction to using linked data for research

**Data** about us are collected nearly every moment of our lives. As technology advances and we are increasingly connected to wireless networks, more aspects of our lives are being recorded, tracked, and turned into data. Everything from our daily commute, to the time we wake up, our levels of exercise, and even the type of coffee we enjoy can now be collected as data.

When data are associated with a specific person, we refer to the data as **individual-level data.** Some of this data collection occurs at obvious times and by obvious organizations, such as when the Insurance Corporation of British Columbia (ICBC) records our driver's licence renewal or when the BC Vital Statistics Agency records the birth of a child. Data collection also occurs when we make purchases online, by companies such as Amazon.ca, or when we are admitted to a hospital.

Almost all new data collection is digital, meaning data are stored on computers. Digital data are likely to be stored indefinitely. This means that not only are there more data, but data are more readily accessible, and they can more easily be moved, shared, and used, including for research. When these data are used for research, they are **de-identified** so that they cannot readily be traced back to the individual.

A collection of data from a single source, such as ICBC, is called a **data set**. Data sets from different sources (e.g., ICBC and BC Vital Statistics) can be shared and combined to create **linked data**. Data sets are not automatically linked – not even data sets between government ministries. There are specific legislative requirements that have to be met to link data. These requirements are similar for both the public and private sectors.

In the last few years, creating and using linked data has become easier due to technological improvements and new analytical tools. This development is giving researchers the opportunity to shed light on complex questions that can be achieved only with linked data. For example, linking data about early childhood, education, and employment could help researchers understand the long-term effects of early childhood socio-economic conditions, and where to intervene to improve longer-term outcomes.

But linking data also creates risks. Linked data sets can contain sensitive personal information that could cause harm if used for an inappropriate or unethical purpose or if they were disclosed inappropriately. For example, the linked data set described above could be used to target and stigmatize neighbourhoods or communities with higher rates of poor early childhood outcomes.

There is always the potential for both risks and rewards in research. The rise in data collection and linked data means there are new opportunities and new challenges. In the context of increasing data capture and technological developments that enable data use, it is crucial to discuss how best to manage these risks and rewards for research. Since individual-level data come from British Columbians like you, it is necessary to consult with the public to gain their advice and preferences on how linked data should be used in BC.

There are recent cases illustrating uses of linked data that were not acceptable to the public. In 2014, England's National Health Service (NHS) was planning to implement a database project that would store patients' medical information in a single linked database, called "care.data." This project progressed until the news agency, the Guardian, reported that the patients' information could be sold to insurance or pharmaceutical companies. This public reacted strongly and negatively, and as a consequence the care. data project was completely cancelled in 2016.

> The creation and research uses of linked data will be central topics of this deliberation. You will be discussing who should have access to linked data, how and when these data should be used, and under what conditions sharing and studying these data would be appropriate.

## 5 | What kind of individual-level data are being collected?

At the most basic level, data are any kind of information that can be associated with a place, person, or object. For the purposes of this deliberation, we will be focusing on data about individuals, or individual-level data.

Data can be categorized into the following types:

***Administrative data***: data collected in the course of providing and/or paying for services (e.g. hospital admissions, physician payment information, immigration records)

***Clinical data***: data that includes specific aspects of persons, conditions, and/or care (e.g. blood pressure, lab results, pharmaceutical data)

***Survey data and data from research participation***: data collected directly from and about individuals or groups, which may be for research (e.g., age, ethnicity, education, income, daily activities, opinions)

***Licensure/registry data***: data about specific groups of people for regulatory or monitoring purposes (e.g., Vital Statistics, professional regulatory bodies such as College of Physicians, cancer registries)

***Physical and biological data***: data about specific aspects of human biology (e.g., height, weight, blood pressure, genomics). When these samples are collected in a central location, they are sometimes called "biobanks."

***Surveillance data***: data about individuals in the context of everyday life (e.g., CCTV video capture, Fitbit, web activity, GPS tracking)

## 6 | Who is collecting individual-level data?

In British Columbia, data are collected by both individuals and organizations, such **academics**, **public bodies**, and **private enterprises**. Data collected by public bodies and academics are considered to be **publicly collected data**, while data collected by private enterprises are considered to be **privately collected data**.

### Public bodies

Public bodies comprise a wide range of federal and provincial organizations. Public bodies collect personal information such as driving records, workplace injuries, educational attainment, and income levels. On a provincial level, examples of public bodies include Ministries such as the Ministry of Health, the Ministry of Children and Family Development, and the Ministry of Education. Public bodies also include organizations such as universities, WorkSafeBC, ICBC, and health authorities, such as Fraser Health Authority, Vancouver Coastal Health, and Provincial Health Services Authority (PHSA).

### Academics

Academics are researchers who work at universities, such as the University of British Columbia, and so are part of **public bodies**. Researchers and their students collect a wide-range of data for the purposes of research, including both clinical and survey data. Often, the information they collect can go into great detail, involving in-depth interviews or long-term observation. The data collected by academics frequently involve a smaller number of participants for a very specific topic, such as interviews on experiences growing up in a specific neighborhood, or detailed activity levels before and after an operation.

In some cases, academic data collection can involve a large number of participants and be part of long-range studies. An example of this would be the BC Generations project, where participants agree to share data about their lives for a long period of time so that researchers can understand factors that influence the development of diseases such as cancer.

### Private enterprises

Private enterprises include corporations like Amazon, Loblaws, Facebook, and Google, Apple, and app developers. Private enterprises also include pharmaceutical companies, device manufacturers, and many doctors' offices. To use their services, we often provide private enterprises with our personal information, such as credit card information, phone numbers, addresses, and date of birth. Because of how we interact with them, private enterprises have the capacity to collect detailed information about us. By collecting and keeping information on our purchases and our search histories, our health conditions and our service use, companies can make inferences about our preferences, habits, and even our political inclinations.

Private enterprises can also collect biological samples to provide various kinds of health-related and ancestry information. The company 23andMe© offers gene testing using a saliva sample that you can mail in. They then sequence the genome in the sample and send results on whether it indicates any predispositions to diseases (to the extent that diseases have a known genetic cause).

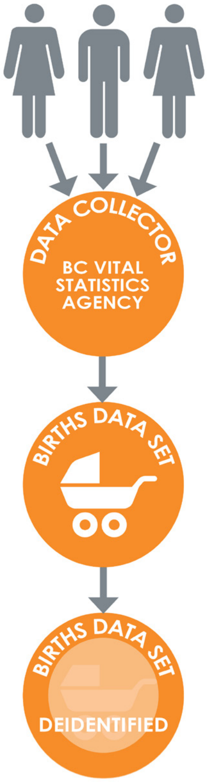
The company AncestryDNA© provides similar services, testing for ancestry using blood samples. Most direct-to-consumer genetic testing companies do not disclose what they do with the genetic information after they have collected it, and in most cases the business model relies on selling those data.

### Examples of individual-level data collection and use

| | **Types of individual level data collected** | **Examples of data use** |
|---|---|---|
| **Public bodies** | Vancouver Coastal Health: Services received from home care agencies, admission and discharge from long-term residential care, admission and discharge from hospitals and all services associated with a hospital stay<br>WorkSafeBC: Number of injury claims; types of accidents; place of employment | Vital Statistics BC could compile a “birth and marriage” data set. This would comprise all the dates of birth and the days of marriage and divorce from everyone in British Columbia. With this dataset, Vital Statistics BC could ask questions like, what is the average age at which that people get married? |
| **Academics** | University researchers: in-depth interviews; detailed description of weekly exercising habits; urine samples after heavy work-outs | Researchers at UBC could conduct a series of interviews with runners from the yearly SunRun, and create a “runners’ experience” data set. They can then ask how many people felt healthier after the run, or how many regretted having done it. |
| **Private enterprises** | Amazon: Types of purchases; frequency of purchases; preferred brands<br>Facebook: Preferred news sources; types of ads clicked on; favorite topics<br>Device manufacturer: All readings and information from insulin pumps routinely worn by people with diabetes | Netflix tracks how often movies and shows are watched, and then examines what kind of individuals tend to watch them. They then suggest movies and shows to watch based on an individual’s viewing patterns. |

## 7 | How are individual-level data used in research?

A single individual's data is often not useful for research. Typically, data about groups of people are what is valuable to researchers. Research is about finding patterns, and trying to understand experiences that can generalize to groups of people.

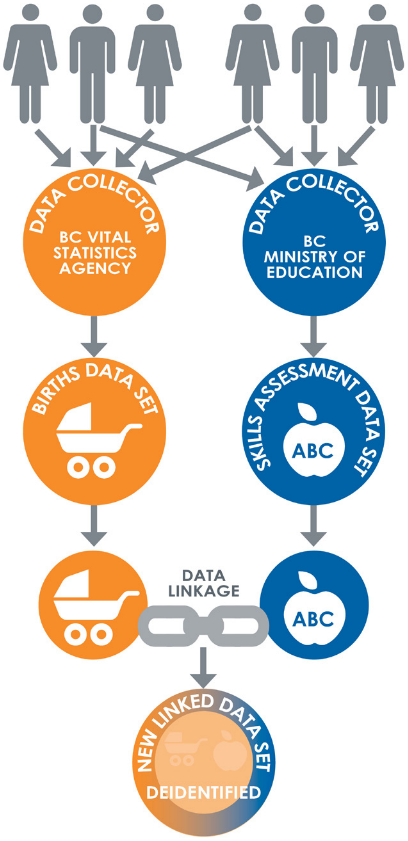


Data are usually **de-identified** before being shared with researchers, to protect the privacy of individuals represented in those data. This process of **de-identification** can include removing names, social insurance numbers and/or other **direct identifiers**. It can also mean altering birthdates, postal codes and other **indirect identifiers**, for example providing only a year of birth rather than a full birthdate. Doing so helps lower the likelihood of any risk or harm to individuals of being identified, and if done with care it does not lower the potential reward of the research.

To study the data, researchers must be able to justify any collection of or request for detailed data, especially data that might be considered identifying. If the researchers are from a university or a health care facility, they must seek review and approval from their **research ethics board** and in some cases a privacy advisor prior to starting their research. A research ethics board reviews the methods and procedures of a research proposal to see if they are ethical and respect the rights and privacy of the participants. Researchers must also seek approval from the scientific peers through **peer review**. This involves other researchers examining the research proposal, and may determine if a proposal is funded. Finally, researchers must also obtain approval for use of existing data from **data stewards**, who have responsibility for making decisions about sharing data. . In some cases, a privacy advisor is required for review as well.

However, while individual data sets have been useful for researchers, they are fundamentally limited by the amount, the type, and the details of the information they contain. For instance, reading-habit data from Indigo may reveal people's preferences on the genres of books they read, but they only do that for books purchased at Indigo. These data will not include purchases from other booksellers (online or otherwise) or books taken out on loan from the library. Indigo data, then, can shed light on individuals' "Indigo reading habits" but not necessarily on their "reading habits" overall. Similarly, researchers are often interested in linking data to get a clearer picture of individuals' experiences. For example, a health researcher studying diabetes would want to link physician visits, hospital visits and prescription medication use over time (and perhaps other data as well) to understand whether current care guidelines are truly producing better outcomes for patients.

## 8 | How are individual-level data combined into linked data?



Since individual data sets are limited, researchers have an interest in combining two or more data sets to form **linked data**. Notably, these data sets would probably not be collected with an intent to link. Because of this, another term used to describe the study of linked data is **secondary uses of data**, meaning a use that is conceived of after the data are already collected.

Linked data are increasingly becoming a focus of research for public bodies, private enterprises, and academics. Research using these data can involve many different methods, from basic descriptions of the data (the number of people, by age group, by sex, and so on) to **regression**, **data mining**, and **machine learning**. These terms refer to the statistical and mathematical techniques that use the power of computers to find patterns that the human eye is unable to see because of the size of data sets. Studying large data sets can sometimes reveal new and unexpected results.

Linking data sets is not an easy process. Sometimes data sets have common information about people, like a Personal Health Number, which makes linking easier. In other cases, there are no common numbers and instead data have to be linked using names and other information. In British Columbia, data linking is typically done for researchers by another body, such as a government Ministry or Population Data BC. Doing so minimizes the exchange of identifiable data and risk of these data being improperly disclosed.

> Notably, the data sets researchers often work with are those from publicly collected sources (i.e., public bodies and academics). An example would be linking an ICBC data set with a WorkSafe BC data set. This is why the use of publicly collected data for research is the main focus of this deliberation. However, researchers can also collaborate with private enterprises and work with privately collected data sets as well. This deliberation will be able to recommend policies that affect what data may be linked and who should have access to those data, but will have less influence on who adopts those policies, particularly outside the public sphere.

### What is Population Data BC?

Population Data BC (PopData) is a multi-university data and education resource. Staff at PopData facilitate interdisciplinary research on the determinants of human health, well-being and development. Services include securely housing and managing data, linking data, de-identifying data, coordinating requests for access, and operating a secure research facility for data analysis.

## 9 | What are some of the benefits of studying linked data?

Linked data are complex and can involve the information from hundreds, if not hundreds of thousands of people. Because of this complexity, they can be very powerful in revealing patterns that are directly relevant to our well-being. These results can help suggest policy solutions to problems affecting British Columbians.

### Example of academic research that could only have been completed using linked data

Dr. Mieke Koehoorn at UBC was interested in looking at impacts of asbestos exposure, which can lead to a type of cancer called mesothelioma. About 95% of cases of this disease are caused by workplace exposure, and are therefore eligible for compensation through WorkSafeBC (if the exposure happened in BC). Mesothelioma develops very slowly, with symptoms generally appearing 30 to 40 years after exposure. Since the disease development happens over such a long time, Dr. Koehoorn wanted to examine whether people affected by mesotheloima were receiving compensation.

To examine this, Dr. Koehoorn worked with BC Cancer and WorkSafeBC to link a cancer occurrence data set and injury files data sets. This meant working with two public bodies to link two data sets and approve the process. The approval process involved not only the public bodies communicating with each other and evaluating Dr. Koehoorn, but also required the research approval from the UBC research ethics board. Once approved, the data sets were linked and de-identified by Population Data BC. See page 18 for a description of the approval process.

Dr. Koehoorn found that less than 50% of mesothelioma patients receive compensation from WorkSafeBC. Instead of the compensation going to the workers, their treatment and health care costs are paid for by the BC Ministry of Health. She also found that women and retired workers are least likely to seek compensation.

These results had direct policy implications. It led to BC Cancer and WorkSafeBC working together to increase the awareness of compensation benefits for people with mesothelioma. They also knew to target women and retired workers.

## 10 | What are the risks of studying linked data?

Bringing a variety of personal and potentially sensitive data together carries some risk. The main risks revolve around the potential for a privacy breach, the possibility of re-identification of individuals, and the potential for unethical research using the data. These risks are summarized and described below:

### Loss of privacy through data breaches or inappropriate sharing of data

With more data being shared and studied, the number of people working with data is increasing. As a result, the chance of mistakes and the possibility of malicious behaviour is also increasing.

Organizations as well as the general public are concerned about the risk of a **data breach**, that is, when data are improperly released (disclosed) or are stolen. Organizations can have their databases hacked, and then have their data stolen and released by a malicious person. Data may be accidentally released by an employee or government employee as well. These data breaches have been happening around the world, and security consultants believe the frequency of these breaches will increase.

One of the more prominent data breaches in British Columbia happened in 2010 and 2012 and involved the BC Ministry of Health. There were three incidents of data breaches that involved the personal information of 38,000 British Columbians. The information had been improperly stored on non-secured USB sticks. Those incidents caused a review of policies and practices surrounding the data sharing among Ministry employees and researchers at the University of Victoria and UBC.

### Loss of privacy by being able to re-identify individuals within the linked data

As previously mentioned, in order for data sets to be linked and shared with researchers, the data are **de-identified** for use in research. However, with enough information, resources, intent, and effort, even data that appear to be de-identified can be used to find specific individuals. This re-identification and loss of privacy may be done by someone with malicious intent. Technology is a factor in this, in the sense of making data analysis faster and easier. The variety and volume of data about individuals is an even more important factor. Linking early childhood, education, health care and workplace injury data can provide a lot of detail about individuals. This, combined with external knowledge (something a researcher knows about someone in those data), can make it possible to identify an individual and possibly learn something new about him or her.

It is important to note that in some cases using identifiable data can be crucial for research to succeed and to produce meaningful results. For example, an epidemiologist tracking the spread of a disease would need to have access to postal codes. They would also need access to birth dates and gender to determine which populations are affected. Hence, there is trade-off between the ability to conduct research, and protecting the privacy of people in the data.

### Unethical uses of research resulting from linked data

In many ways, the dangers of a data breach lie in the data being potentially misused (e.g., identity theft). However, *how* the research from linked data are interpreted and then used is also problematic, as they may be used in unethical manners. In addition, the research itself may be unethical. If the basis of the analysis is flawed, it is possible that researchers may produce inaccurate results. Those results become particularly concerning when they start influencing policy – and even more concerning when those policies affect vulnerable populations such as children, people with disabilities, or First Nations communities.

As an example of the dangers of unethical applications of research, Cathy O'Neil describes in her book *Weapons of Math Destruction* how many credit agencies and lenders are relying on mathematical approaches called algorithms to make their decisions. These algorithms are the result of research using linked data. O'Neil found that African Americans were being disproportionately denied loans based on those algorithms. This was due to the algorithm's dependence on zip codes to identify neighbourhoods with a large number of loan defaults. Since African Americans are more likely to live in low-income neighborhoods with a larger number of loan defaults, they are more likely to be identified as high risk. Hence, they were more often denied loans, opportunities to go to college, or the chance of starting their own businesses.

## 11 | What are the current laws and regulations on data sharing?

There are specific legislative and procedural rules that need to be followed when linking data sets. For instance, each public body in British Columbia collects their data separately and there are often many steps and approvals to be obtained to share and link data sets.

There are two main Acts that apply to data sharing and privacy in British Columbia:

- *The Personal Information Protection Act (PIPA)*: regulates private enterprises and organizations, such as corporations, charities, and co-ops. They also cover doctors' offices, religious organizations, pharmaceutical companies, and unions.
- *The Freedom of Information and Protection of Privacy Act (FIPPA)*: regulates government organizations, such as provincial ministries, and also academics in research institutions, such as universities.

The proper application of PIPA and FIPPA is the responsibility of the Office of the Information and Privacy Commissioner, an independent government organization. Complaints can also be filed through them if an organization is thought to have used data improperly. A recent guidance document from the Office of the Information and Privacy Commissioner stated that the privacy rules in FIPPA apply to data sets with identifiable data, but do not apply to de-identified data sets.

PIPA and FIPPA provide privacy protections for personal information, and place limits on how and when information can be collected or used. In addition, they also specify how and when consent needs to be obtained when collecting data, and also what procedures must be followed to maintain data security. For example, if an online corporation is collecting personal information, they must give individuals an option either to agree or to opt out of providing their personal information. This is often a check box to click. In some cases, obtaining consent is not necessary, such as with administrative data collected by public bodies. For example, WorkSafeBC will make a record of an accident without asking for consent.

PIPA and FIPPA are very crucial pieces of legislation that protect our privacy. However, because of changes in technology and new analytical tools, it is possible that they are no longer meeting our needs adequately. New technological developments that were not thought to be possible when the Acts were written may not be well covered by the legislation. For example, data sets that were considered to be de-identified may be identifiable with new technology. Or, new ways that data sets are being used may raise issues and difficulties that are not addressed by the existing Acts.

**What does the FIPPA say about sharing data for research?**

**There are several parts of the Act that are relevant, but the most specific for research is Section 35:**

Disclosure for research or statistical purposes

35 (1) A public body may disclose personal information in its custody or under its control for a research purpose, including statistical research, only if

(a) the research purpose cannot reasonably be accomplished unless that information is provided in individually identifiable form or the research purpose has been approved by the commissioner,

(a.1) subject to subsection (2), the information is disclosed on condition that it not be used for the purpose of contacting a person to participate in the research,

(b) any data linking is not harmful to the individuals that information is about and the benefits to be derived from the data linking are clearly in the public interest,

(c) the head of the public body concerned has approved conditions relating to the following:

(i) security and confidentiality;

(ii) the removal or destruction of individual identifiers at the earliest reasonable time;

(iii) the prohibition of any subsequent use or disclosure of that information in individually identifiable form without the express authorization of that public body, and

(d) the person to whom that information is disclosed has signed an agreement to comply with the approved conditions, this Act and any of the public body's policies and procedures relating to the confidentiality of personal information.

(2) Subsection (1) (a.1) does not apply in respect of research in relation to health issues if the commissioner approves

(a) the research purpose,

(b) the use of disclosed information for the purpose of contacting a person to participate in the research, and

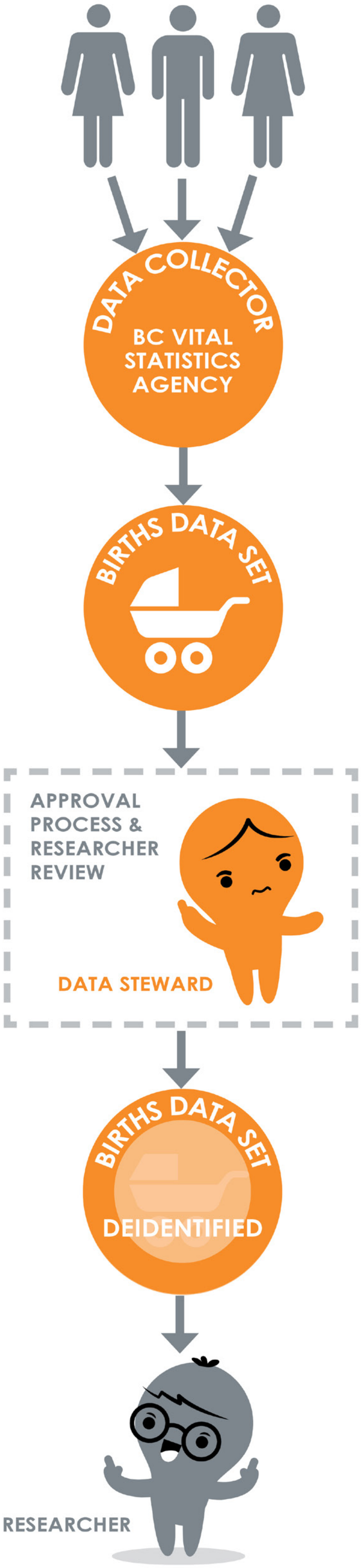
(c) the manner in which contact is to be made, including the information to be made available to persons contacted.

## 12 | Who decides on whether data can be shared? How are data access requests approved?

Each data set is administered by an individual **data steward**. Data sets are not all housed in the same location nor are they automatically linked, even if they are all collected by the provincial government. Each public body (e.g., health authorities, hospitals, provincial agencies, Ministries) houses their data sets separately.

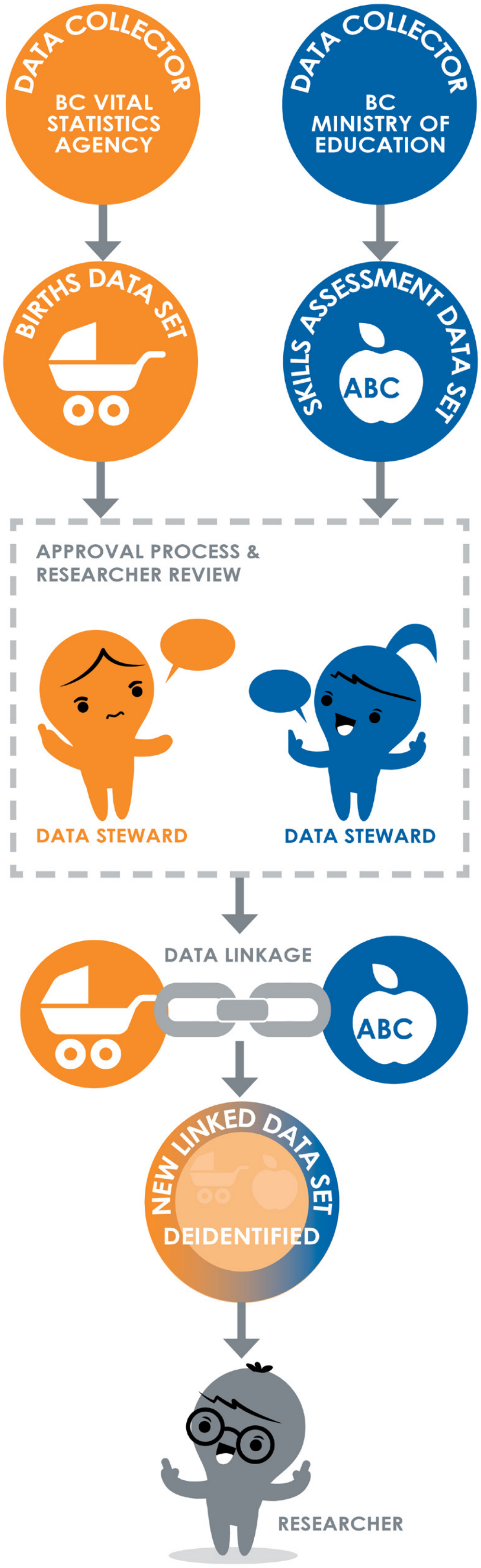
Since data stewards are based in different organizations, the steps they follow vary, even though they are generally covered by the same laws and regulations. Sometimes data stewards work with a committee to review data access requests. Other times, the data stewards make the decisions themselves and within a few days and even a few hours (if the request is very simple). Their data sets can also vary greatly in complexity and size, and they may access different kinds of resources and specialized personnel to support their tasks. These differences can result in different review times for data access requests.

The image opposite describes the general process that data stewards follow to decide if they should grant a researcher access to the data set that they manage. Here, we are showing the approval process involving a data steward. If the researcher is at a university, his or her research proposal must also be approved by the university's research ethics board and through peer review.



When a researcher is requesting access to two or more data sets that require linking, his or her data access request will reflect that. The request will be reviewed by all data stewards involved – for example if the request is for access to three different data sets, three different data stewards will review. The image opposite describes how this happens.

Note that to link data sets, the data stewards of each data set must also agree with each other about the data sharing agreement, which could involve additional negotiations. Currently, the process of reducing the released data to the minimum necessary can take several months to over a year depending on the complexity of the request.



A more detailed approval process can be described in the five stages  below:

### Stage 1

Data stewards receive and review **data access requests** from researchers. This request could require the researcher to submit additional documents or evidence that they have passed ethics and peer review.

### Stage 2

Data stewards review the application, and determine whether researchers may access the data set or which parts of the data set they may safely receive. It's important to know that researchers almost never get access to an entire data set. Data stewards do not necessarily make all of the data they have even potentially available to researchers, and researchers must justify the pieces of data they request.

### Stage 3

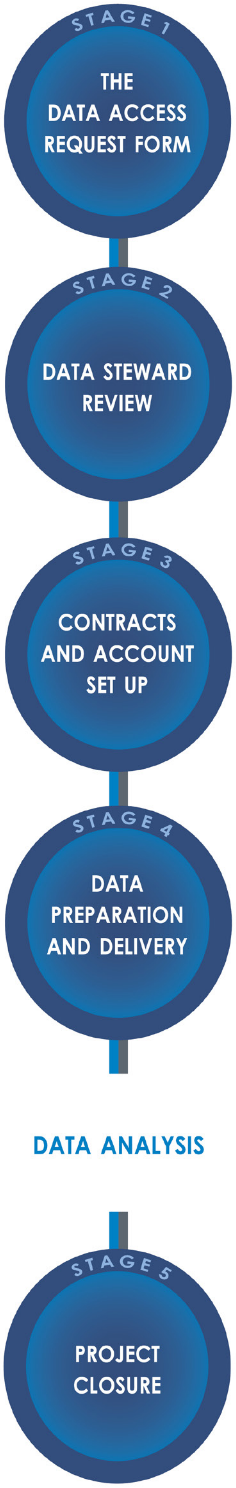
A research agreement is set up with the researchers. This may involve arranging for the researcher to access a **secure research environment.** This means that the researcher would be able to access the data *only* from that controlled environment. They would not be able to take the data somewhere else or copy them. Indeed, they can only use the data for the research purposes they specified and nothing else.

### Stage 4

If more than one data set has been allowed for release, the data sets will be linked and be placed in the secure research environment. The researcher may need to pay an additional fee for the linking to happen since it is a complicated process. Given the expense of the linkage, not all researchers can afford this.

### Stage 5

Once the research is complete, the researcher notifies the data steward and they must begin to close the project. Access to the secure research environment is closed and the data are archived.


STAGE 1
THE DATA ACCESS REQUEST FORM
STAGE 2
DATA STEWARD REVIEW
STAGE 3
CONTRACTS AND ACCOUNT SET UP
STAGE 4
DATA PREPARATION AND DELIVERY
DATA ANALYSIS
STAGE 5
PROJECT CLOSURE

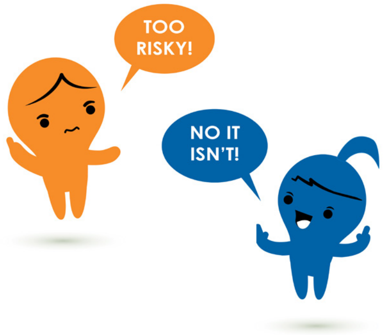
## 13 | What are the current challenges that data stewards encounter?

Deciding on whether data may be shared and linked is determined by data stewards. In the case of academic researchers, there are also research ethics boards and peer review requirements. For every data access request, data stewards must balance privacy risks and the potential benefits and innovations from research. To do so, they work within the legislative guidelines of PIPA and FIPPA.

Legislation can be vague on how to implement the requirements. As a result, data stewards must interpret the legislation to the best of their ability. An excellent example of this is if we re-examine Section 35 of FIPPA on page 15. Although the intent of the legislation is clear, in clause 1b, it states that data linking must be in the "public interest." How is "public interest" defined? How should it be assessed and based on what criteria?

Also in clause 1b is the statement that the data linking should not be "harmful to the individuals that the information is about." What does "harmful" mean and how should it be assessed? What should be done if there is a risk of harm? If the data steward approves the data access request, what kind of measures should be put in place to guard against the risk of harm? What if some people might consider the conclusions of a study to be against their interests, even if the study is accurate (e.g., if a health service is no longer offered because it is not effective)?

And finally, since data stewards are all individuals, it is possible that they may disagree with another steward's interpretation of the legislation. Depending on their organization, they may also have to consult with their institution's privacy advisor. Moreover, the data stewards and the privacy advisors may not all agree. One may think the project is not harmful and could result in beneficial results, while another would not allow the data to be linked, believing that the risks to privacy are too high. One may think the project is asking for too much data, while another is more willing to allow the data to be shared.



These are some of the challenges that data stewards face when reviewing data access requests and that researchers face in requesting access to those data. These challenges can slow down the processing of data access requests. As a consequence, discoveries and innovations from studying linked data sets may be slowed or prevented altogether.

## 14 | Summary

More and more individual-level data are being collected about us every day. This trend will only continue and even accelerate as technology improves. These data can be combined to create linked data, which offer new opportunities for researchers to pursue. Research using linked data can lead to new discoveries that can positively affect individuals and society. However, the same aspects that make linked data powerful also raise issues around maintaining privacy and the ethical use of data and the information that are derived from them.

As one data steward put it, “Taking account of privacy should reduce risk… not eliminate value.” To help figure how to strike this balance, we have brought people from many different backgrounds, and with many different opinions and life experiences. Our aim is to work together with you on this challenging issue.

## 15 | Your role in the deliberation

During the deliberation, you will hear more about linked data, sharing data, and privacy from speakers who have expertise on particular issues and the other deliberants. You and your fellow deliberants will bring your own perspectives to the discussion. You are not expected to be an expert on this topic.

You will be asked to discuss some of the issues related to sharing and researching linked data with the other deliberants. These may include issues such as:

- What criteria should be used to evaluate whether data should be shared and linked?
- How should public interest be evaluated for requests regarding linked data?
- How should the benefits of studying linked data be balanced with the risks to privacy?

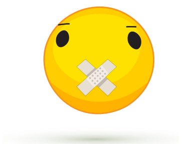
We hope that you will bring your opinions, values, and ideas about data and privacy to the deliberation. You will work together to make recommendations that can be used to more effectively inform policy decisions on data access regulations.

To facilitate discussion, we ask that you follow these ground rules:

**Keep an open mind**

**No eye rolling**

**Listen to others**



**Avoid cross-talk**

**Participate in respectful deliberation**

**Try not to interupt**

**Ask for clarification**

**Try to justify your opinions**

## Glossary of Terms

**Administrative data**: data collected in the course of providing and/or paying for services (e.g. hospital admissions, physician payment information)

**BioBank**: a dedicated institution preserving biological samples (e.g., spit, blood, tissue). These samples are usually used for research, such as at the BC Children's and Women's Hospital.

**Biometric data**: data regarding the physical descriptors of an individual (e.g., fingerprints, retina scans)

**Clinical data**: more detailed information about specific aspects of persons, conditions and/or care (e.g. blood pressure, weight, lab results)

**Data**: any type of information that can be associated to an individual

**Data access request**: forms and other requirements are usually submitted to data stewards to request access to data sets.

**Data breach**: an unauthorized release of data either through hacking or by accident

**Data linking**: the act of combining two data sets into a larger data set

**Data set**: a collection of data that has been gathered using the same criteria

**Data steward**: a designated official responsible with approving or denying data access requests

**De-identified data**: data where the personal identifiers of the individuals have been removed with the intent of minimizing or removing any chance of re-identification

**Genomic data**: DNA-derived data about molecular aspects of human biology

**Individual-level data**: data that are collected from individuals and that can be associated to an individual

**Licensure / registry data**: Information about specified groups for regulatory or monitoring purposes (e.g. Vital Statistics, professional regulatory bodies such as College of Physicians, cancer registries

**Linked data**: a data sets that combines two or more data sets

**Peer review**: subjecting research plans to review and assessment by experts. This often determines whether funding is provided for the research

**Privacy**: the state or condition of having an individual's individual-level data not be shared with unauthorized parties

## Glossary of terms (continued)

**Private enterprises**: any privately owned business or entity, such as a corporation, union, or club.

**Privately collected data**: any data that are collected by a private enterprise

**Public bodies**: provincial or federal organizations, such as Ministries, agencies, universities and hospitals.

**Public deliberation**: a public deliberation is a community discussion on issues that affect members of the public. It is a democratic process in which citizens participate, and policy makers and experts provide information and observe. The results of a public deliberation can inform important policy decisions.

**Publicly collected data**: any data that are collected by a public body or an academic

**Qualitative data**: data that cannot be measured, such as a category (e.g., favourite colour, preferred coffee type) or text (e.g., responses to open ended survey questions)

**Quantitative data**: data that can be measured, such as height and weight

**Research**: the systematic investigation into and study of materials and sources in order to establish facts and reach new conclusions

**Research ethics board**: An organization with universities and other public bodies that determine whether research may be approved based on ethical considerations.

**Secondary use of data**: studying data for another purpose for which it was originally intended

**Survey data**: information collected directly from and about individuals or groups

**Surveillance data**: data that exists about individuals in the context of everyday life, e.g. CCTV video capture, Fitbit, web activity, GPS tracking)