Privacy by Design at Population Data BC: a case study describing the technical, administrative, and physical controls for privacy-sensitive secondary use of personal information for research in the public interest

Caitlin Pencarrick Hertzman,[1] Nancy Meagher,[1] Kimberlyn M McGrail[2]

[1]Population Data BC, School of Population and Public Health, University of British Columbia, Vancouver, Canada
[2]Centre for Health Services and Policy Research, School of Population and Public Health, University of British Columbia, Vancouver, Canada

Correspondence to
Caitlin Pencarrick Hertzman, Population Data BC, School of Population and Public Health, University of British Columbia, 2nd Floor – 2206 East Mall, Vancouver, BC, Canada V6T 1Z3; caitlin.hertzman@popdata.bc.ca

Received 12 April 2012
Accepted 3 August 2012
Published Online First 30 August 2012

ABSTRACT

Population Data BC (PopData) is an innovative leader in facilitating access to linked data for population health research. Researchers from academic institutions across Canada work with PopData to submit data access requests for projects involving linked administrative data, with or without their own researcher-collected data. PopData and its predecessor—the British Columbia Linked Health Database—have facilitated over 350 research projects analyzing a broad spectrum of population health issues. PopData embeds privacy in every aspect of its operations. This case study focuses on how implementing the Privacy by Design model protects privacy while supporting access to individual-level data for research in the public interest. It explores challenges presented by legislation, stewardship, and public perception and demonstrates how PopData achieves both operational efficiencies and due diligence.

INTRODUCTION

Canada's publicly-funded social safety net provides provincial and territorial governments with a wealth of individual-level administrative health and social service data. These data are a rich resource for researchers to study the production and distribution of health and well-being in the population and to provide evidence for public policy development. Secondary research use of data is a cost-effective and efficient alternative to primary collection for individual research projects.[1] Use of secondary data can also reduce bias in findings.[2]

The use of these data for research is in the public interest, improving prospects for health and well-being. At the same time, appropriate safeguards are essential to govern disclosure and use of what is often sensitive information.[3] While the benefits of using linked administrative data are well documented, many jurisdictions experience challenges with access to such data, whether from delays, denials, or data restrictions. These challenges include varying interpretations of complex legislation, fears of privacy breaches, and public perception of the risks inherent in compiling population-wide repositories of sensitive and identifiable data.[4] Data stewards or custodians worry about unauthorized use or disclosure of data by researchers,[5] such as the re-identification of individuals. The public worries about increasing privacy risks related to constant shifts in technology including the development of electronic health records. While many are willing to post personal information online, they retain a specific and enduring concern about the release and misuse of sensitive health information because of the potential for stigma and discrimination.[6]

Population Data BC (PopData) is a resource in British Columbia (BC), Canada that enables research access to linked, longitudinal administrative data. It has been able to do this by building operations around a set of core privacy principles, including Privacy by Design.

BACKGROUND

PopData is a multi-university, nationally active data and education research resource providing data linkage, access, and training to support research on human health, well-being, and development. PopData performs no research itself, and was established in 2009 to broaden the data holdings from the health care services focus of the British Columbia Linked Health Database (BCLHD). The BCLHD, described in 1998,[7] provided the foundation for PopData's linkage and data access models.

PopData facilitates research requests for data, including coordinating applications, producing data extracts for approved projects, and making those extracts available on a secure server. This work is carried out on behalf of public bodies whose data it receives, links, and houses. Data stewards at these public bodies retain control of the data, approving data extracts on a per-project basis. PopData holds data from the British Columbia (BC) Ministry of Health, BC Vital Statistics Agency, BC Cancer Agency, the Human Early Learning Partnership and WorkSafeBC, among others. PopData strategically chose to structure its operations around several features:

- Operating as a trusted third party for data linkage.[8] PopData chose to conduct linkage and forego a research function in order to set itself up as a neutral body that has no other interests in the sensitive data needed for linkage across data providers.[9] PopData receives personally identifiable information for linkage purposes and generally achieves linkage rates above 95%, which limits potential bias in population-based analyses.[10]
- Providing research extract access via virtual private networking (VPN) regardless of location within Canada. Researchers can securely access data directly without the confines of designated offices or terminals such as those utilized by other models including the Institute for

Clinical Evaluative Science[11] (ICES) and the Research Data Centers of Statistics Canada.[12]

- Providing individual-level, de-identified data unlike the practice at organizations that provide data at the aggregate level only.

The vast majority of data holdings in PopData are administrative data that have been collected for the purpose of providing or paying for services. Data sets include physician payments, hospital separations, workplace injuries, and more. Even when direct identifiers are not present, the data include detailed and varied enough information that re-identification may be possible. Privacy professionals in Canada have developed best practice recommendations to mitigate this and other risks, such as the Canadian Government's Tri-Council Policy Statement (TCPS)[13] and the Canadian Institutes of Health Research's Privacy Best Practices.[9] They also endorse the principles outlined in the Canadian Standards Association's Model Code for the Protection of Personal Information[14] and those found in the Privacy by Design framework developed by the Ontario Privacy Commissioner, Ann Cavoukian.[15]

This case study describes how PopData achieved its strategic objectives by responding to the principles described in Privacy by Design.[15] The following description of PopData's approach is further detailed in table 1, which maps the various controls to the relevant foundational principle.

CASE DESCRIPTION AND METHODS: PRIVACY BY DESIGN

Principle One: preventative controls to safeguard privacy

PopData's *physical controls* reflect the first of the seven principles—being proactive and preventative.[15] The physical offices are separated into three zones: purple, red, and yellow. Each zone has different access controls. The yellow zone is locked, but every PopData employee has access. The red zone has a separate lock and alarm and is the zone in which employees with access to raw, identified data work. There is video surveillance of this door, and the walls are fortified with steel to prevent intrusion. The purple zone, which houses the data servers, has fortified walls and a separate alarm. It is only accessible to three named employees.

The zonal approach also applies to PopData's *technical security* structure. Although protected by firewalls and intrusion detection software, yellow zone computers can be used to communicate externally. Red zone computers are dumb terminals used to access the encrypted data stored in the purple zone on a network which is moated within the physical red zone. Programmers use two-factor authentication to log on to these computers. These moating and access controls were devised to meet international standards (such as ISO), but were customized for research when scanning the systems of organizations such as the Western Australian Data Linkage System,[16] ICES and the Manitoba Center for Health Policy.[10]

Secure transfer of data from data stewards (usually annually) is achieved by transferring encrypted data using secure file transfer protocol, avoiding the storage of personal information on removable media and the use of couriers.

Separation of identifiers from content is performed as new data sets and their annual updates are imported. All content variables that may be requested for research are stored separately from the identifiers. This prevents both from being accessed at the same time, including during the data linkage process. Separation is a commonality among data linkage organizations, whether performed at the source or the institution.

Table 1 Privacy by Design principles at Population Data BC

	Implementation at Population Data BC		
Principle	**Physical controls**	**Technical controls**	**Administrative controls**
1. Proactive not reactive; preventative not remedial	Reinforced physical perimeter Roles-based access to physical areas using fobs and alarm codes Motion alarms Video monitoring of entrances/exits	Moated networks Separation of identifier and content data Dummy/dumb computers Two-factor authentication using SecurID Encryption of data at every point, from transfer to storage to backup Secure transfer on SFTP Roles-based access to systems	Criminal record checks Confidentiality agreements for staff and researchers Comprehensive privacy education program for staff including annual training Mandatory privacy training for researchers
2. Privacy as the default setting	Roles-based physical access	Moated networks Separation of identifier and content data Encrypted storage	Confidentiality pledge Privacy training for staff and researchers
3. Privacy embedded into design	Physical controls planned in advance of construction of work and server facilities	Access and network controls developed based on national and international best practices and reviewed in advance of becoming operational	Authentication, access, and other administrative controls outlined in Information Sharing Agreements
4. Full functionality—positive-sum, not zero-sum. Privacy controls accommodate all interests and objectives in a positive sum 'win-win' manner. Privacy is protected without losing functionality		SRE: software, technical support, and systems (including encrypted backup) provided to researchers SRE: no need for data on removable media SRE: logging, auditing, tracking, and other controls provided to data stewards	Individual-level data provided, while meeting all security and privacy requirements of data stewards, government policy, and legislation
5. End-to-end security—full lifecycle protection		Data secured and encrypted from point of transfer through extraction, use, storage, backup, and finally destruction	
6. Visibility and transparency—keep it open	Public notice of video surveillance	Access controls with logging and regular auditing	Availability of documents such as Privacy Impact Assessment, Privacy Policies, and Terms of Use documents; listing of all research projects fulfilled on website
7. Respect for user privacy—keep it user-centric			Annual privacy training of staffBi-annual review of policies in light of current standards and requirements

The table maps controls at Population Data BC against the seven foundational principles found in the document 'Privacy by Design: The 7 Foundation Principles' by Ann Couvakian.[15]
BC, British Columbia; SFTP, secure file transfer protocol; SRE, Secure Research Environment.

Proactive linkage entails linking each data set when it arrives from a data provider, rather than project by project. Once the linkage has been carried out, the results are stored using meaningless but unique IDs. These IDs can then be used to draw together data from multiple data sets as required for approved research projects. Proactive linkage also allows PopData to maintain a population database that helps to improve linkage rates.

Principle Two: privacy by default

Technical controls, such as the separation of content and identifiers and moated network systems, minimize the risk of breaches and their resulting harm. Physical controls limit access to secure locations, preventing accidental exposure to unauthorized users. Administrative controls like pledge of confidentiality agreements and privacy training ensure that staff and researchers are educated in their responsibilities and privacy best practices. This reduces the risk of incidents due to error. These controls operate without requiring ongoing action, thus allowing privacy to be the default setting.

Principle Three: embedding privacy into design

PopData embedded privacy into the design of its secure data facilities from the start. The access systems were also devised with privacy as the main objective. PopData finalized data sharing agreements, policies, and procedures before becoming operational in order to ensure privacy was embedded in the design of the administrative structure as well.

Principle Four: the positive sum approach

Privacy by Design seeks to accommodate all interests and objectives in a positive sum 'win-win' manner.[15] PopData embodies this principle by facilitating the use of individual-level data—which presents more benefits *and* risks than using aggregate-level data[17]—in a manner that can meet data steward and public expectations.

Access to custom-produced de-identified research data extracts is limited to authorized researchers who have signed agreements with data stewards and PopData. For the majority of projects, data are disclosed to and then analyzed by researchers via PopData's Secure Research Environment (SRE), a virtual server system providing storage, processing, encryption, back-ups, and a wide range of analytic software to researchers. Researchers can access the SRE from any location in Canada using two-factor authentication and VPN, an innovative arrangement that ensures privacy protection without requiring researchers to go to specific physical locations.

The SRE provides controls not possible when distributing data directly to researchers on electronic media. For example, this permits access to be suspended if ethics approvals expire or unauthorized use is suspected. The SRE prevents the transfer of individual-level data on and off the system by screening file size, type, and name. SRE staff perform manual monitoring of any files transferred. The system logs all transfers and accesses, including allowable transfers such as downloads of analytic output.

The benefits of the SRE make it a win-win privacy control for both researchers and data stewards, but it is important to note that although access to individual-level data is allowable under governing privacy legislation, there is no absolute solution for inferential disclosure risks. The safeguards of the SRE, the stringent requirements of the research and confidentiality agreements the researcher signs, and the de-identification carried out by PopData all work together to address this privacy risk. This approach is consistent with best practices that recommend that a combination of technologies customized to the risks specific to the data provides the best mitigation strategy.[18 19]

Principle Five: full lifecycle protection

PopData encrypts all data during transfer, storage, use, disclosure, back-up, and destruction. The data's lifecycle is managed according to policies designed to meet and exceed the requirements of data stewards and applicable legislation. This provides end-to-end security, from the moment data are brought in-house until they are destroyed.

Principle Six: transparency through independent verification

Access controls of core systems as well as the SRE allow for auditing of logs, which along with the video surveillance provide data stewards with an independent verification of who is accessing or transferring data and when these activities occur. PopData works with data stewards to implement regular reviews and audits to verify processes and policies in action. This is an important component of visibility and transparency.[15] PopData also makes its Privacy Impact Assessment, Privacy Policy, and Terms of Use documents available to the public.

Principle Seven: respect for the individual's privacy

Respect for privacy is at the heart of PopData's policies and practices. Section 35 of BC's Freedom of Information and Protection of Privacy Act (FIPPA) requires that researchers: (1) justify the need for identifiable data; (2) show that the research is in the public interest; (3) describe how security, destruction, and subsequent use and disclosure are addressed; and (4) sign an agreement regarding these conditions.[20] PopData designed its privacy controls to meet and exceed these requirements. For example, the Data Access Request[8] also investigates requests for sensitive fields or population-based cohorts, if applicable.

Privacy-relevant policies include a research data access framework, privacy policy, data asset management policy, access policies, incident response procedures, and more.[8] PopData requires criminal record checks before hiring new staff, provides employee privacy training, and conducts privacy impact assessments for all new or significantly revised activities. Researchers must have Research Ethics Board approval and complete online privacy training before receiving access to data. These policies and practices are meant to ensure that PopData always treats the confidentiality of the individual's information as a priority.

CONCLUSION

Providing researchers with access to linked administrative data has privacy benefits. As but one example, primary data collection that includes names and other identifiable data have decreased,[21] which minimizes disclosure risk, use of individuals' time, and the cost of research. Research use of these data has enormous potential to provide evidence for policy-making that can improve the population's health and well-being. This use, however, demands careful control, especially as interests broaden and the available data sets expand. PopData employs Privacy by Design to meet the privacy requirements of legislation and data stewards as well as the expectations of the public whose information is being used for research. Nevertheless, significant challenges lie ahead. Legislative changes have the potential, sometimes inadvertently, to make data access and thus research more difficult. Continuing technological developments demand constant refinement of physical and technical infrastructure. And PopData must always be vigilant against human fallibility. PopData will continue to implement Privacy by Design

principles to improve privacy controls while ensuring that research reusing individual-level personal information informs public policy and improves the well-being of the population.

Contributors This case study was authored by CPH and co-authored by KMM and NM.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Roos LL,** Brownell MD, Lix L, *et al*. From health research to social research: Privacy, methods, approaches. *Soc Sci Med* 2008;**66**:117–29.
2. **Gershon AS,** Tu JV. The effect of privacy legislation on observational research. *CMAJ* 2008;**178**:871–3.
3. http://www.aracy.org.au/publicationDocuments/LECTURE_The_Linking_Of_Records_And_Analysis_Of_Population_Data_For_Epidemiological_And_Public_Health_Research_2006.pdf (June 2012).
4. **Wartenberg D,** Thompson WD. Privacy versus public health: the impact of current confidentiality rules. *Am J Public Health* 2010;**100**:407–12.
5. **Kelman CW,** Bass AJ, Holman CDJ. Research use of linked health data—a best practice protocol. *Aust N Z J Public Health* 2002;**26**:251–5.
6. **Ingelfinger JR,** Drazen JM. Registry research and medical privacy. *N Engl J Med* 2004;**350**:1452–3.
7. **Chamberlayne R,** Green B, Barer ML, *et al*. Creating a population-based linked health database: a new resource for health services research. *Can J Public Health* 1998;**89**:270–3.
8. http://www.popdata.bc.ca/aboutus (April 2012).
9. **Canadian Institutes of Health Research.** *CIHR best practices for protecting privacy in health research*. Canadian Institutes of Health Research, 2005.
10. **Jutte DP,** Roos LL, Brownell MD. Administrative Record Linkage as a tool for Public Health Research. *Annu Rev Public Health* 2011;**32**:91–108.
11. http://www.ices.on.ca/webpage.cfm?site_id=1&org_id=119 (June 2012).
12. http://www.statcan.gc.ca/rdc-cdr/process-eng.htm (June 2012).
13. http://www.pre.ethics.gc.ca/eng/policy-politique/initiatives/tcps2-eptc2/Default/ (March 2012).
14. http://www.csa.ca/cm/ca/en/privacy-code/publications/view-privacy-code (June 2012).
15. **Cavoukian A.** *Privacy by design ... take the challenge*. Information and Privacy Commissioner of Ontario, 2009.
16. **Holman CDA,** Bass AJ, Rosman DL, *et al*. A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. *Aust Health Rev* 2008;**32**:766–77.
17. **El Emam K,** Cavoukian A. *A positive-sum paradigm in action in the health sector*. Information and Privacy Commissioner of Ontario, 2010.
18. **Sparks R,** Carter C, Donnelly John B, *et al*. Remote access methods for exploratory data analysis and statistical modelling: privacy-preserving analytics. *CMPB* 2008;**91**:208–22.
19. **O'Keefe CM.** Privacy and the use of health data—reducing disclosure risk. *EJHI* 2008;**3**:1–9.
20. **Freedom of Information and Protection of Privacy Act [RSBC 1996]** Chapter 165, 2012.
21. **Trutwein B,** Holman CDA. Health data linkage conserves privacy in a research-rich environment. *AEP* 2006;**16**:279–80.