Common Limitations of External Data Sources

Table of Contents

Introduction **2**

Data Gaps and Limitations **2**

Disaggregated data is not available *2*

Indigenous-specific data is not available *3*

Legislative barriers exist to accessing Indigenous data *4*

Privacy legislative barriers may become more onerous in the future *4*

The data environment is fractured *5*

Available First Nations or Inuit surveys vary in coverage, timeliness and trending *5*

Data may not exist in the general population as comparators *5*

Data collection is not aligned with an Indigenous need *6*

Indigenous survey data may be of variable accuracy *6*

Data governance protocols limit access to data *7*

Data is not timely *7*

Lack of standardization of community data collection systems *7*

INTRODUCTION

External data sources are organizations that hold information about your citizens. This document provides information on some common limitations with using external data sources in your analyses.

DATA GAPS AND LIMITATIONS

Disaggregated data is not available

In a closed system, such as a country, or to some extent a province, data required for its operation is either possessed by government or accessible by it. Regulations and legislation exist to facilitate the retention and protection of data for the ultimate benefit of government. Data is sequestered in government databases, and generally only high-level analysis is available to external parties. This analysis commonly will not differentiate between different groups in the population.

In 2020, the BC Human Rights Commission looked at the need and power of disaggregating data as a way to support governments and other interested parties. Disaggregated data is data that provides sub-categories of information, for example by ethnic group, gender, occupation or educational status. Disaggregation is often the means to obtain information on Indigenous populations. Disaggregated data is usually anonymized or de-identified—stripped of personal information like name or date of birth—before being used in statistical analysis. Unlike aggregated data, which groups categories of information together, disaggregated data can reveal inequalities and relationships between categories.[1]

Data which has been disaggregated to allow the identification of Indigenous information is a strong tool for promoting systemic change by highlighting inequities with the dominant population and being a catalyst for policy and program change in government, whether it be at the provincial level or within the SGIG jurisdictions. As the BC Human Rights Commissioner points out, disaggregated data is also necessary to support human rights legal frameworks.[2] In the most recent federal Speech from the Throne, the Government of Canada has also committed to, "building a whole-of-federal-government approach around better collection of disaggregated data".[3]

Data analysis using disaggregated data may be coloured by the beliefs and perceptions of the government, and may be used not only for the benefit of the population covered, but for other politically inspired purposes. In particular, data and measurement under the control of non-Indigenous governments have historically been used as an instrument of colonialism and system oppression against Indigenous peoples. The historical practice of government, intentionally or inadvertently, has been to use data to reinforce the perception of Indigenous people as ill, inferior, and vulnerable – unable to care for themselves as individuals and govern themselves as a people. There is an absence of data that acknowledge and celebrate the

[1] BC Human Rights Commission. 2020. Disaggregated demographic data collection in British Columbia: The grandmother perspective.
[2] ibid.
[3] Government of Canada. (2020). A Stronger and More Resilient Canada: Speech from the Throne to open the Second Session of the Forty-Third Parliament of Canada. https://www.canada.ca/en/privy-council/campaigns/speech-throne/2020/stronger-resilient-canada.html

strengths of Indigenous communities and culture, and more often there is an inability of SGIGs to access and reframe data in order to change its narrative. There may also be no accountability for the ultimate benefit and use of data identified as Indigenous – for example, where no government or organization will take ownership for what the data shows and work to ameliorate health or other deficits highlighted by this data. Even in situations where positive action is a desired outcome of a deficit-oriented analysis, governments do not typically consider community harm from stigmatization, reinforcement of stereotypes, and negative profiling.

SGIGs and other Indigenous communities are increasingly interested in establishing ongoing availability of meaningful disaggregated data that will facilitate good governance of services for their citizens and result in increased investments, and at the same time comply with the needs of funders and applicable legislation (e.g., public health acts). This discourse goes beyond the production of disaggregated data to consider the needs and rights of SGIGs for row level data to ensure analysis is conducted from an appropriate paradigm, and create flexibilities to undertake new investigations or data linkages that allow for deeper insight into population health.

Indigenous-specific data is not available

One of the most significant limitations to the access of data describing Indigenous populations is that Indigenous identification in databases is not possible. This inability to identify can be for diverse reasons:

- A mechanism is not available which can identify data which is Indigenous. There are, however, some available strategies to make Indigenous peoples visible in some data. In BC and Manitoba, data linkages with demographic databases allow the identification of First Nations in government health administrative databases. In Yukon, many communities have majority Indigenous populations, and commonly, data has been reported for the population as a whole without a breakout for Indigenous data. First Nations-specific data from DAD and physician databases is available in the NWT as an Indigenous self-identifier is embedded in the health care cards used to access health services. In Yukon, current privacy legislation has been cited as preventing the release of similar First Nations specific data. An additional complication for SGIGs is that their citizenship registries may not align in totality with other Indigenous identifiers (e.g., status number).
- Indigeneity has not been recorded as a data field. In addition, a non-demographic field, such as a personal health number (PHN), is not available to reliably link databases which could result in identification of Indigenous data. Although this is not usually an issue with health databases, social services may not utilize a common variable such as a PHN, preventing data linkages from occurring.
- Data resides in privately administered health services, such as psychologist offices, and there are no systems in place to draw data from individual offices into a common database which could then be linked.
- Data has not been collected, or it has been collected on a subset of the population. To illustrate, Statistics Canada surveys have generally not included First Nations communities in their health surveys, focusing instead on off-reserve Indigenous populations.

- There appears to be little appetite in either provincial or federal governments for the design of data collection strategies which are based on Indigenous needs for data, or which incorporate uniquely Indigenous themes. For example, labour surveys generally look only at work for pay, and do not attempt to quantify the many non-monetary forms of work and volunteerism in Indigenous communities which provide essential services, and often food, to their citizens.

Legislative barriers exist to accessing Indigenous data

The second most significant limitation to the access of data by SGIGs is that Indigenous-specific data exist but row-level data cannot be accessed. Historically, government has restricted access to data by Indigenous organizations – in one particular case, the lack of agreement between Statistics Canada and First Nations and Inuit organizations on access to row level data led to the creation of the initial First Nations and Inuit Regional Health Survey in 1997. SGIGs require not just disaggregation of data into Indigenous-specific subsets, but the flexibility afforded by analyzing data at the row (per person) level for the full expression of their governance responsibilities. Privacy legislation and statistics acts protect personal information and establish powerful legislative barriers preventing Indigenous governments from accessing row level data on their populations, even on an anonymized basis. For example, in BC there are two sets of privacy legislation – one for public institutions and one for private organizations – neither of which have mechanisms for row level data-sharing with Indigenous governments or their representative organizations. The only available remedies are not system-wide, but rather tend to be work-arounds to address specific scenarios, such as seconding staff to government departments in order to access row level, anonymized data.

Where row level data is inaccessible, SGIGs may be able to request that the data steward conduct a custom analysis.

Privacy legislative barriers may become more onerous in the future

Governments are now messaging that anonymization of data does not give sufficient protection against identification of individuals, and there is a high risk of re-identification such as through data linkages which would be considered a violation of privacy. Data re-identification or de-anonymization is the practice of matching anonymous data with publicly available information, or auxiliary data, in order to discover the individual to which the data belong. Organizations with privacy policies, health care providers, and financial institutions may release the data they collect after the data has gone through the de-identification process. Re-identification can occur in combination with other pieces of available data and basic computer science techniques. As an example, Researchers at MIT and the Université catholique de Louvain, in Belgium, analyzed data on 1.5 million cellphone users in a small European country over a span of 15 months and found that just four points of reference, with fairly low spatial and temporal resolution, was enough to uniquely identify 95 percent of them. In other words, to extract the complete location information for a single person from an "anonymized" data set of more than a million people, all that was needed was to place the person within a couple of hundred yards of a cellphone transmitter, sometime over the course of an hour, four times in one year.[4]

[4] Hardesty, Larry. "How hard is it to 'de-anonymize' cellphone data?". MIT news. Retrieved 14 January 2015., referenced in: https://en.wikipedia.org/wiki/Data_re-identification

The data environment is fractured

Indigenous data may be in different government departments or organizations, and if combined with a lack of standardization of data collection or of Indigenous identification among databases, there may not be complete congruence between them which limits data's interpretation and usefulness. The data may represent different Indigenous sub-populations, even within the same organization (e.g., with the First Nations Health Authority, Health Benefits administrative data, chronic disease prevalence data, and RHS survey data have different population definitions, even though all represent the First Nations population).

Standardization may not be appropriate in all situations. Although it seems straightforward that data should be collected the same way with similar definitions and methodologies to enable a valuable comparability function, a monothematic approach may not be congruent with community ownership of data and unique local needs.

Available First Nations or Inuit surveys vary in coverage, timeliness and trending

Generally, First Nations and Inuit-led survey cycles at regional or national levels are infrequent, and do not always cover the same population each cycle. The most predictable is the Regional Health Survey (RHS) which has a quintennial cycle, and the new Inuit Health Survey promises to have the same cycle length. The lack of frequent refreshes results in a reliance on data which will be optimally only two years old, but can extend to six to seven years before a new survey has been administered, analyzed and has progressed through all the steps related to data governance and distribution.

Surveys may incorporate changes to questions reflecting evolving priorities in the years between cycles, which can impact the ability to look for trends over a time period. In addition, these surveys do not provide 100% coverage of all communities in a given region; if an SGIG was not selected to participate as part of the sampling approach, or determined not to participate, then the overall results are of limited relevance and there is not SGIG-specific data available for the SGIG's use. Where possible, an SGIG may choose to participate in these surveys so that data are available for their communities.

Data may not exist in the general population as comparators

Some large databases (hospital and physician services) represent information drawn from the overall population in that jurisdiction. For these databases, it is possible to compare First Nations data against non-First Nations data (i.e., other residents).

For First Nations and Inuit-led data collection efforts, such as the RHS, FNREEES (First Nations Regional Early Childhood, Education and Employment Survey), and FLNED (First Nations Labour and Employment Development Survey), as well as the upcoming Inuit Health Survey, comparisons to the non-Indigenous population cannot be directly drawn. The RHS does have some questions which are analogous to Statistics Canada survey products such as the Canadian Community Health Survey, which allows partial comparability. In most cases, the data is 'stand alone,' and comparative analyses are focused internally on time series trending through the three survey cycles.

Data collection is not aligned with an Indigenous need

Data residing in non-Indigenous organizations may not have been purpose-driven when it was collected, and more often than not, may not reflect Indigenous reality or priorities. The data also represents services largely not modifiable by Indigenous peoples as they are not designed or administered by them.

Available data may have been formatted and presented for other populations or organizations' use. For example, age group stratifications in health care are structured based on the general population which have a different aging profile, progression of disease, and life expectancy than do Indigenous peoples. Additionally, geographic measures in non-Indigenous organizations' data infrequently aligns with historic and territorial boundaries. For example, SGIG data may exist in two or more health authorities, or more than one SGIG may be in the same geographic division and the data are inseparable.

Data may not exist that can advance social equity of Indigenous people. Essentially, government data lends itself to deficit-focused analysis, where Indigenous peoples represent the deficiency. As the BC Human Rights Commission report notes, there is no positive obligation in privacy legislation to gather data in support of furthering equity objectives. Similarly, Statistics Acts provide for the collection and analysis of data on the social and economic conditions of people in BC, but do not reference the advancement of public interest.[5] Reframing to positivity and strengths would require major systemic shifts not only in access to data or the creation of new data, but also to systems and services to reflect the needs of Indigenous communities.

Indigenous survey data may be of variable accuracy

A common way of collecting Indigenous data is through individuals self-identifying by answering pre-set questions, either at the point of service or in surveys. In BC, a self-identification data standard ensures comparability and consistency in how self-identification data is collected. Answering is voluntary. Without an understanding of the reliability of self-identification data collection, it is unknown to what extent the data is representative of the Indigenous population. Self-identification has benefits in that it helps in identifying the non-status population; even though, there is a risk of inconsistent counting, undercounting or biased counting. Due to historical processes in which the state used Indigenous identity to discriminate against and control the lives of Indigenous peoples, there may be fear and mistrust on the part of some Indigenous clients to be identified as Indigenous. Further, as extensively examined in the Review of Indigenous-specific discrimination in the BC health system, the continued existence of systemic and interpersonal racism makes many Indigenous peoples reluctant to self-identify, due to fear of discriminatory treatment.[6]

Databases may be limited by the shifting definitions and policies adopted by the federal government through time. For example, if identification is based on the Indian Register or on eligible beneficiaries, a segment of the SGIG's citizenship or overall population (e.g., which may be non-status) may be missed.

Survey data generally only measure a proportion of the population, and there will always be a degree of uncertainty in survey results which require assumptions of randomness of respondents when analyzing data. Surveys support respondents to share their own perspectives. However, surveys – particularly those that

[5] BC Human Rights Commission. 2020.

[6] Review of Indigenous-specific Discrimination in the BCBC Health System. 2020. https://engage.gov.bc.ca/app/uploads/sites/613/2020/11/In-Plain-Sight-Full-Report.pdf

address sensitive matters – may also introduce biases through the structuring of questions, method of data collection, and data interpretation.

Data governance protocols limit access to data

SGIGs are interested in sourcing data specific to their communities. This will introduce issues of privacy and confidentiality, given that the numbers will be very small for some of the data points of interest. For example, the FNHA has implemented a cell size suppression threshold of 10, meaning that any data points below 11 are not released. This may drastically limit the available information to SGIGs and typically requires seeking the data at a sub-regional level or aggregating SGIG data collectively, both of which may help avoid the cell size suppression issue.

Previous analyses of BC First Nations data have illustrated that generally a population size of 2,000 is necessary in order to obtain a rudimentary analysis of data, with 5,000 population providing an ability to be more comprehensive in reporting. With lower populations, the common stratifications of age group, health condition, geography, gender and other variables of interest result in low cell sizes that are suppressed due to data governance policies, and result in an inability to report on results. As well, small cell sizes can lead to extreme rate fluctuations from year to year, which reduce their accuracy and usefulness.

Data is not timely

The necessary dependence of SGIGs on external data sources, at least for the short term, means that data is generally not timely. Time is required by organizations to prepare and validate the data for distribution, to conduct a data linkage and associated quality checks, and to undertake all data governance processes in the event a data request has been received. These activities can mean that data is at minimum two to three years old by the time it is received by SGIGs.

Lack of standardization of community data collection systems

Optimally, data stewarded within communities will be in electronic form not paper-based files, and linkages will be possible with data held in provincial or territorial databases, particularly where services originated in one jurisdiction, then responsibility was transferred to another (e.g., home care in an Indigenous community after discharge from hospital). For paper storage of data, time will be needed to perform chart reviews or otherwise manage the transference of data to a digital format. This also requires investment of financial and human resources.