# Technical Documentation
# BC Demographic Survey: DIP Data Linkage Rates

## Table of Contents



**PROJECT:** Evaluating the BC Demographic Survey Data in the DIP


**PREPARED BY:**
Amelia Lowery, Julie Hawkins & Lindsay Fredrick
Data Science Partnerships Program
BC Stats BC Data Systems
Ministry of Citizens' Services


**DATE:** 2024-04-19

## Purpose

This document summarizes the analytical methods, caveats, and technical software used to generate the linkage summaries provided in the dashboard BC Demographic Survey: DIP Data Linkage Rates and made available on the **Error! Hyperlink reference not valid.**

## Project Overview

In 2023, BC Stats conducted the  BC Demographic Survey. More than 200,000 people responded to the voluntary survey, providing information about many aspects of their identity (such as race, ethnicity, ancestry, gender and many others).

The  Data Innovation Program (DIP) securely links and de-identifies data from multiple ministries, organizations or agencies in a secure platform. Many DIP datasets also contain demographic related information. However, the number of datasets with demographic information is limited, and are only partially complete pictures. The BC Demographic Survey aims to improve our understanding of how people with varying backgrounds interact with public services by broadening the scope of available demographic information to each data (and therefore service) provider.

The analysis presented in this dashboard used the secure platform to access available datasets from the DIP, and linked these datasets, where possible, to the data from the BC Demographic Survey. Overall linkage rates, as well as more specific demographic linkage rates were investigated. The methods, tools, and caveats associated with the dashboard are explored more fully below.

## Methodology

## Overview

To compare the BC Demographic Survey to each individual DIP dataset, a list of unique StudyIDs (which represent unique individuals as determined by PopData) was created for each individual dataset. The BC Demographic Survey list was compared to the individual DIP dataset, to determine:

- The number of individuals in each DIP dataset.
- The number of individuals within a DIP dataset that have a survey record.

## Survey Demographics

After comparing to individual DIP datasets at a broad level, a deeper dive exploring DIP demographic data was completed as well. Six key demographic components from the survey were included for this dataset. These include:

- **Gender**
  - This was cleaned in the demographic survey to include: Man/Boy, Woman/Girl, Non-Binary Person, but also catered for those who preferred to remain unknown (with I don't know/I am unsure, or Prefer not to answer).
- **Racial Identity**
  - This was cleaned in the demographic survey to include: the 14 non-Indigenous options made available on the survey, the two unknown options, as well as the option for someone to have reported multiple races.
  - Note that the methodology used here does not align with the racialized population groups reported by Statistics Canada. See the Education/Health technical reports for more details **Error! Hyperlink reference not valid.**
- **(Distinctions Based) Indigenous Identity**
  - This was cleaned in the demographic survey to include: First Nations, Inuk (Inuit), Métis, Not Indigenous, the two unknown options, as well as the option for someone to have reported multiple Indigenous identities.
  - See the Health technical report for more details **Error! Hyperlink reference not valid.**
- **Indigenous Identity**
  - Because current DIP datasets are limited in their capability to provide Indigenous demographics, we also rolled up the Distinctions based demography to a singular Indigenous/Non Indigenous option.
  - This was done to facilitate easier comparison with what is currently available in various DIP datasets, but it is still recommended that a distinctions based approach to using Indigenous data is followed wherever possible.
- **Disabilities**

  - This was cleaned in the demographic survey to include: has a condition that is always, often, or sometimes a disability, no condition that is a disability, as well as the two unknown options.
- **Date of Birth (DOB) Status**
  - This was cleaned both in the demographic survey and in DIP datasets to identify solely if there was a valid record or not, due to the inherent number of options that would be available otherwise.
  - Individual demographic comparisons are not made available for this category, only summary level comparisons.

While the BC Demographic Survey contains many more demographic details, we believe that this dashboard will provide sufficient information in order for researchers to make informed choices in producing high quality research questions. It is important to note that the survey variable names used in this dashboard were determined by the BC Stats analysts of this project specifically and may differ from published results related to the BC Demographic Survey elsewhere.

## Linked Variables Summary

Each dataset from DIP was evaluated separately for both overall linkage rates as well as more specific demographic linkage rates. This second process involved reviewing the demographics available from the BC Demographic Survey, then reviewing the specific dataset for any comparable demographics using associated metadata. If a given demographic within the DIP dataset was determined to be reasonably comparable, the comparison was done at a summary-level first. To make this comparison:

- Data from DIP datasets were manipulated to reduce every StudyID to a single demographic variable per demographic category.
  - For example, someone with multiple DIP records that all indicate ‘female’ as their given gender would be reduced to a single record that indicates ‘female’. If more than one type of demographic category was listed for a given ID, it would be recorded as ‘multiple reported’.
- Every StudyID in the DIP dataset was then compared to those in the BC Demographic Survey, and sorted into one of four possible summary categories:
  - **Survey Only**: this person only has demographic information for this category available from the survey. This may be because the value

recorded in the DIP dataset was **NULL**, or because the DIP dataset does not contain this particular demographic information.

- **DIP Only**: this person only has demographic information available from the DIP dataset. This may be because they did not respond to the survey, or they skipped the relevant associated questions on the survey.
- **DIP and Survey**: this person has demographic information from both sources. This does not guarantee that the information is matching in both sources, simply that it exists. The Linked Individual Demographics tab contains more information on the alignment of variables.
- **Neither Source**: this person has a DIP dataset record, but it does not contain any viable demographic information, nor does the survey.

## Linked Individual Demographics

The analysis was then also done at the StudyID (person) level, where a cross-tabulation of what demographic category a person is associated with in the DIP dataset is compared to the demographic category the person self-identified with in the BC Demographic Survey. The result could be:

- Additional information: the DIP dataset provided no information, but the survey does.
  - Example: a health dataset indicates an unknown date of birth, while this is provided by the survey.
- Contradicting information: the DIP dataset indicates a different value than what is provided by the survey.
  - Example: an education dataset indicates the gender of a student as female, while the survey indicates the gender of the student is non-binary.
- Aligned information: the DIP dataset and the survey provide generally agreeable demographics.
  - Example: a child in care is recorded in the MCFD dataset has non-Indigenous, which agrees with the value provided by the survey.

## Caveats

**This dashboard and data source serves as a guide only. Researchers are ultimately responsible for determining the feasibility and reliability of the**

**variables themselves prior to using the BC Demographic Survey variable information for their project.**

Here we list some important distinctions to keep in mind when browsing the dashboard and related datasets:

- Not every DIP dataset has a StudyID
- Some DIP datasets contain multiple StudyIDs
- Because DIP datasets were only manipulated to reduce every StudyID to a single demographic variable, some 'Known Dip' percentages may appear artificially high, as the variable may include 'Unknown' or 'Prefer not to Answer' responses. Determining the exact content of each demographic variable in each dataset was out of scope for the creation of this dashboard, and should be carefully examined by any DIP researcher.
- BC Stats analysts used their discretion in determining comparable demographics between the DIP dataset and the associated BC Demographic Survey variable. Additionally, some demographics within a given DIP dataset may have been excluded from analysis without intending to do so.
- Many DIP datasets include several years worth of data, and not just point-in-time information, therefore, even within a given dataset there could be contradicting information available for a single record. The result is seeing "multiple reported" within the 'Value in DIP'.
- All counts provided here are unweighted. DIP researchers are responsible for providing their own weights where necessary.

## Software

This analysis is implemented in the R and python programming languages Van Rossum and Drake Jr (1995). The code used to generate this analysis was reviewed by three data scientists. Key tools used to complete this work include the Apache Arrow project (Richardson et al. 2021), the tidyverse (Wickham et al. 2019) and the interal SAE package dippy (Fredrick 2023).

### GitLab

All code is stored under the git version control system and shared inside the secure environment in these GitLab repos:

- Creation of Demographic Survey Variables: https://projectsc.popdata.bc.ca/shares/arda-demographic-survey

- Creation of Linkage Statistics: https://projectsc.popdata.bc.ca/shares/linkage-litmus-test

## References

Fredrick, Lindsay. 2023. *Dippy: Provide Functions to Efficiently Import SRE Data*.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Jonathan Keane, Romain François, Jeroen Ooms, and Apache Arrow. 2021. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Van Rossum, Guido, and Fred L Drake Jr. 1995. *Python Tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.