Technical Documentation: Annual Homelessness Cohort Data Set

Table of Contents

PREPARED BY:
Data Science Partnerships & Population Research
BC Stats
BC Data Service
Ministry of Citizens' Services

DATE: 2026-01-23

BRITISH COLUMBIA | BCStats

Acknowledgements

We gratefully acknowledge the First Nations traditional territories across the province on which we live and work, the Métis Chartered Communities, and Inuit living in B.C.

The methods discussed in this document were originally developed by the Preventing and Reducing Homelessness project, run by the Ministry of Housing, Province of British Columbia.

The following data sets were used in this study: BC Housing's Shelter and Homeless Outreach Private Market Rent Supplements, Social Development and Poverty Reduction's BC Employment and Assistance, and Health's Central Demographics Files. You can find further information regarding these data sets by visiting the BC Data Catalogue.

All inferences, opinions, and conclusions in these materials are those of the authors. They do not reflect the opinions or policies of the provider(s) of the data upon which they are based.

Purpose

This document summarizes the analytical method used to create the "Annual Homelessness Cohort Data Set", a row-level homeless cohort research data set using the analytical methods first developed and implemented by the Preventing and Reducing Homelessness Integrated Data Project (Province of B.C., 2021). The annual cohort methodology remains focused on implementing a cross-agency analytic definition of homelessness based on administrative data available through the Data Innovation Program. This document serves to inform researchers about the underlying data sets and how they were used to create the homelessness cohort data set. If you are a researcher and want to work with this data set contact the Data Innovation Program here: https://dpdd.atlassian.net/servicedesk/customer/portal/2.

Overview

The population who experienced homelessness in B.C. is estimated using Ministry of Social Development and Poverty Reduction's (SDPR) *Employee and Assistance Program* usage data and integrating it with BC Housing's (BCH) *Emergency Shelter Program* usage data. More information about how the homeless population was identified can be found in the Homelessness Definition section. Several attribute variables are defined to describe the demographic and geographic characteristics of this population. geographic location.

The project is enabled by the Data Innovation Program, a data integration and analytics program for the B.C. government. While every B.C. ministry collects and manages its own data, the Data Innovation Program can securely link and de-identify data from multiple ministries, giving government analysts a broader understanding of complex issues.

Data sets are linked and provisioned into a secure analytics environment by Population Data BC, an internationally recognized academic organization that has facilitated population-based research for over 20 years.

Data

All administrative data was provisioned though the Data Innovation Program (DIP). Geography summaries were derived using open licensed data from Statistics Canada. Below is a list of all the data sets used in the cohort data set creation.

B.C. Employment and Assistance Data (SDPR)

Involvement Data

- filenames:
    - bcea_involvements_replacement_for_2023.csv
    - bcea_involvements_update_for_2024.csv
- columns used:
    - ym – year and month
    - fileid – unique case file identifier, used to link to nfa data
    - deprltncd – dependent relationship code
    - birthdt – birth date year and month
    - gender – values include: M, F, X, U
    - studyid – unique identifier, used to link to other datasets

No Fixed Address (NFA) Data

- filenames:
    - bcea_nfa_replacement_for_2023.csv
    - bcea_nfa_update_for_2024.csv
- columns used:
    - ym – year and month
    - fileid – unique case file identifier, used to link to involvement data
    - nfa – flag for no fixed address
    - postcd – first three characters of postal code (often of the ministry office associated with the case)
    - csdname – census subdivision name associated with the case
- rows used: only rows where nfa = 1

BRITISH COLUMBIA | BCStats

Shelter and Homeless Outreach Data (BCH)

Client Data

- filename: Clients.csv
- columns used:
    - clientId – client identifier in BCH database
    - Gender – values include: Female, Male, Non-binary, Other, Transgender, Transgender - Female, Transgender - Male, Two-spirit, Unknown
    - DOB – birth date year and month
    - studyid – unique identifier, used to link to other datasets

Shelter Stays Data

- filename: ShelterStays.csv
- columns used:
    - clientId – client identifier in BCH database
    - ShelterStayStartDate – book-in date to a shelter
    - ShelterStayEndDate – book-out date from a shelter
    - ShelterStayPostalCode – first three characters of shelter postal code
    - ShelterCensusSubDivision – census subdivision name associated with the shelter
    - studyid – unique identifier, used to link to other datasets

Client Roster (Ministry of Health)

- filename: client-roster-sn_dtl_dm_clnt_pi_2025_20250328.csv
- columns used:
    - KEY – values in this field were replaced with studyid by Population Data BC
    - CLNT_GENDER_CD – values include: F, M, I, U
    - MRG_CLNT_BRTH_DATE – birth date year and month

2021 Geographic Attribute File (Statistics Canada)

- filename: 2021_92-151_X.csv
- columns used:
    - CSDuid, CSDname, CSDtype – id, name, and type for census subdivisions
    - CDuid, CDname, CDtype – id, name, and type for census divisions
    - PRuid – id for provinces

Data Processing

Homelessness Definition

SDPR data

Based on the SDPR data, an individual was considered to have experienced homelessness if they had at least 3 months of consecutive income assistance with no fixed address.

To find this population, the nfa data was linked with the involvement data using fileid and ym as a primary key. The merged nfa-involvement data was subset for only rows with no fixed address (denoted by nfa = 1). Each row represents a monthly income assistance payment. For a given month, if an individual received income assistance and had no fixed address for that month and the prior two months, they were included in the population for that month.

BCH data

Based on the BCH data, an individual was considered to have experienced homelessness if they spent any time in a shelter.

To find the population who experienced homelessness each month, overlapping intervals of shelter visits were merged to create one continuous interval. Then, all the intervals were partitioned by month. An individual was considered to have experienced homelessness each month that intersected with a shelter stay interval.

Merging BCH and SDPR data

The two populations were merged using studyid, year and month as the primary key, keeping track of whether an individual was present in the BCH data, SDPR data or both.

Attributes

Data Source

The means of entry into the population was recorded to determine whether an individual entered the study population via service usage definitions from the SDPR data, BCH data or both. At the monthly resolution, this measure directly results from whether a person entered into the population by their usage of income assistance, shelters or both.

Homeless Category (Chronicity)

A chronic homelessness sub-population, distinct from the non-chronic homelessness population, was also defined. This category was defined using an individual's past 12 months of service usage for a given month. For some months, this includes reaching into the previous year's shelter and income assistance data. We implemented the following criteria to define homeless categories:

1. Calculate the longest period of consecutive monthly income assistance for those individuals that reported no fixed address over the past 12 months
2. Calculate the cumulative number of nights spent in a shelter over the past 12 months
3. Calculate the cumulative number of shelter visits separated by at least 30 days from a previous visit (unique visits) over the past 12 months
4. Apply the following definition for a given month:
    - Non-Chronic Homelessness: three to five months of consecutive income assistance reporting no fixed address OR fewer than 180 days in a shelter OR one or two unique visits to a shelter
    - Chronic Homelessness: six or more months of consecutive income assistance reporting no fixed address OR 180 days or more in a shelter OR three or more unique visits to a shelter.
5. In instances where the above criteria resulted in differing non-chronic homelessness and chronic homelessness outcomes between the services, a studyid was associated with chronic homelessness.

To calculate the number of unique shelter visits, it was necessary to merge overlapping shelter stay intervals to create one continuous interval. A small portion of the BCH shelter data contained individuals with overlapping time intervals at different locations. For instances where an individual was present in two census subdivisions, the intervals were not merged. This is a very small number of shelter visits and therefore represents an acceptable loss of data accuracy.

Demography data

Age data

Age data was derived from three ranked sources in the following order:

1. Health data: most authoritative source for accurate birth data
2. SDPR data: largely also derived from registry data but provided more values due to slightly more complete data
3. BCH data: least accurate age data as it was entirely self-reported

Age was calculated as of December 31 of a given year, which is the end of the accrual window, and then rounded down to the nearest integer. Studyids with two birthdays were removed from the demographic data. To address concerns of statistical disclosure, age was aggregated to categories consistent with what is used by Statistics Canada (0-24, 25-34, 35-44, 45-54, 55+), and age was removed for records with unknown gender.

Gender data

Gender data was derived from three ranked sources in the following order:

1. BCH data: gender was collected using open ended response and was therefore the priority gender indicator
2. SDPR data: collected as binary gender, but was the most current data and therefore was the next priority
3. Health data: also collected as binary gender, and was the least current data source and therefore, while still very reliable, was the lowest priority

Values of Male/Female were recoded to use gender classifications (e.g., Man, Woman, Non-Binary) following the Guidelines to the Gender and Sex Data Standard (Province of B.C., 2023). This was to correct for the historical conflation of the terms sex and gender in data collection (i.e., by using the binary sex classifications Male and Female to collect gender data). To address concerns of statistical disclosure, the category of Non-Binary was ultimately removed from the dataset. All instances of this gender category were in the BCH data and therefore replaced by either SDPR or demographic data.

Geographic Data

Both BCH and SDPR data were provisioned with census subdivision, however because of slight differences in the names of geographic locations between BCH, SDPR and the Statistics Canada data, a crosswalk table was created and linked to the Statistics Canada Geographic Attribute File to resolve this issue. Geographic location of service provision was aggregated to the census division level, adopting a conservative approach to statistical disclosure and acknowledging that geographic locations were limited to issuing offices and shelter locations. Further, the following census divisions were combined due to small cell sizes:

- Mt. Waddington and Central Coast
- Northern Rockies and Peace River
- Kitimat/Stikine and Stikine

Final Data Sets

The above work resulted in the following two final data sets available through the Data Innovation Program and four summary tables available on the BC Data Catalogue.

Monthly Data

- filename: homeless_cohort_by_month.csv
- columns available:
    - studyid
    - year
    - month
    - data_source – the service used to define homelessness for the individual; nfa, nfa_shelter, shelter

  - hl_category – chronic homelessness (CH) or non-chronic homelessness (NCH) defined by level of service usage
  - gender – Man, Woman
  - age_group – age group of the individual at the end of the calendar year: 0-24, 25-34, 35-44, 45-54, 55+
  - core_version – core-snapshot version of the source data

Geographic Data

- filename: homeless_cohort_geo.csv
- columns available:
  - studyid
  - year
  - month
  - location_defined_by – the service used to define homlessness for the individual at the provided location; nfa, nfa_shelter, shelter
  - CDname – census division where the individual visited a shelter or received income assistance with no fixed address
  - core_version – core-snapshot version of the source data

Summary Tables

Four summary tables will be produced annually and exported. They will be made available in the BC Data Catalogue. The tables include counts by homelessness chronicity (chronic/non-chronic) by year, counts by age group and gender by year, count by census division by year, and counts by data source (nfa, shelter, nfa and shelter) by month and year.

For the annual estimates of homelessness chronicity, an individual is considered to have experienced chronic homelessness if they were categorized as chronic in any month within the year, otherwise they are categorized as non-chronic.

For the annual estimates of homelessness by census division, the following hierarchical rules were used to assign a census division to individuals that received services (income assistance with no fixed address or visited a shelter) in more than one census division:

1. The census division most frequented - highest number of months - was assigned
2. If the most frequented census division was “NA”, then the most frequented non-NA census division was assigned
3. If more than one census division has the highest number of months, then the census division closest to the end of the cohort year was assigned (i.e., the most recent location)

4. If more than once census division meets all of the above criteria, then one of those census divisions was assigned at random (this option was used for less than 2% of the individuals)

Software

This cohort creation analysis is implemented in the R programming language (R Core Team 2021). The code used to generate this analysis was reviewed by two data scientists. Key tools used to complete this work include the Apache Arrow project (Richardson et al. 2021), the tidyverse (Wickham et al. 2019), dipr (Albers and Hazlitt 2020) and the targets R package (Landau 2021) for project organization. All code is stored under the git version control system and available inside the DIP secure environment in a GitLab repository:

- Homelessness cohort data set generation: https://projectsc.popdata.bc.ca/shares/standalone-hl-cohort

This code is based on previous work from the Preventing and Reducing Homelessness Integrated Data Project also available inside the DIP secure environment in GitLab repositories:

- Parquet and restaging: https://projectsc.popdata.bc.ca/shares/hl-data-to-parquet
- Application of definition and creation of output group: https://projectsc.popdata.bc.ca/shares/hl-cohort

The above URLs are provided for researchers inside the DIP secure environment and are not accessible outside of the DIP secure environment. If you have questions, please contact BC Stats here: https://dpdd.atlassian.net/servicedesk/customer/portal/12.

References

Albers, Sam, and Stephanie Hazlitt. 2020. *Dipr: Provide Functions to Efficiently Import SRE Data*.

Landau, William Michael. 2021. "The Targets R Package: A Dynamic Make-Like Function-Oriented Pipeline Toolkit for Reproducibility and High-Performance Computing." *Journal of Open Source Software* 6 (57): 2959. https://doi.org/10.21105/joss.02959.

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.

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.

BC Housing. [creator]. Shelter and Homeless Outreach Private Market Rent Supplements. E06. Data Innovation Program, Province of British Columbia [publisher]. Data Extract. Approver Year (2024).

Ministry of Social Development and Poverty Reduction. [creator]. BC Employment and Assistance (BCEA). E07. Data Innovation Program, Province of British Columbia [publisher]. Data Extract. Approver Year (2024).

Ministry of Health. [creator]. Central Demographics Files. E06. Data Innovation Program, Province of British Columbia [publisher]. Data Extract. Approver Year (2024).

Statistics Canada. 2022. *2021 Geographic Attribute File*. Catalogue no. 92-151-X.

Province of British Columbia. 2023. *Guidelines to the Gender and Sex Data Standard.* https://www2.gov.bc.ca/gov/content/data/gender-sex-data-standard.

Province of British Columbia. 2021. *Preventing and Reducing Homelessness Integrated Data Project.* http://gov.bc.ca/homelessness-cohort.