

## Proceedings of Statistics Canada Symposium 2024: The Future of Official Statistics

### Development of Linkage-adjusted Weights Accounting for Gender for the 2021 Canadian Census Health and Environment Cohort

by Yubin Sung and Eric Hortop

Release date: September 8, 2025



Statistics  
Canada

Statistique  
Canada

Canada

# Development of Linkage-adjusted Weights Accounting for Gender for the 2021 Canadian Census Health and Environment Cohort

Yubin Sung and Eric Hortop<sup>1</sup>

## Abstract

The latest Canadian Census Health and Environment Cohort (CanCHEC) continues a series of population-based microdata linkages focused on population health research by demographic, social and economic characteristics. The 2021 CanCHEC consists of 95.5% of the 2021 Census long-form sample survey records. The records of survey respondents that could not be linked to the Derived Record Depository and those presumed to be duplicates account for the remaining 4.5%. Linkage-adjusted main and replicate weights allow researchers to estimate and evaluate the variance of summary measures about population health in the presence of missed linked pairs to better understand the experiences of diverse population groups.

Key Words: Record linkage; Linkage weight adjustment; Calibration; Poststratification; Gender.

## 1. Introduction

The 2021 Canadian Census Health and Environment Cohort (CanCHEC) is the latest in a series of population-based microdata linkages focused on population health research by demographic, social and economic characteristics. The 2021 CanCHEC consists of 95.5% of the 2021 Census long-form sample survey records. The records of survey respondents that could not be linked to the Derived Record Depository (DRD) in the Social Data Linkage Environment (SDLE) and those presumed to be duplicates, identified via internal record linkage, account for the remaining 4.5%. The initial version of the 2021 CanCHEC was released on November 27, 2024, and it linked the 2021 Census of long-form sample to the Canadian Vital Statistics – Death database (May 11, 2021, to December 31, 2022) and to the annual postal codes for mailing addresses (1981 to 2022). The 2021 CanCHEC was highly anticipated because the 2021 Census included new content to address emerging trends and issues (sex at birth and gender, Canadian military experience, etc.) (Statistics Canada, 2022).

To allow researchers to estimate and assess the variance of summary measures about the population health in the presence of missed linked pairs between the 2021 Census long-form sample and the DRD, the main and replicate weights produced and associated with persons from their responding households of the long-form sample survey were adjusted and calibrated to produce the linkage-adjusted main and replicate weights. Decisions in the weighting procedures can affect the accuracy of estimates for small population groups such as transgender men, transgender women, and non-binary people. The 2021 Census collected data specifically about gender for the first time, differentiating from sex at birth, enabling an examination of health outcomes among gender-diverse people. The new data points support government and societal efforts to fill known data and knowledge gaps, address inequalities, and promote fair and inclusive decision-making. We compare the population counts estimated from the 2021 CanCHEC data to the population counts enumerated from the 2021 Census long-form sample data, by gender diversity status and select demographic, social and economic characteristics, to evaluate our practical weighting procedures.

## 2. Methodology

### 2.1 Weighting procedures

---

<sup>1</sup>Yubin Sung, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6, [yubin.sung@statcan.gc.ca](mailto:yubin.sung@statcan.gc.ca); Eric Hortop, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6, [eric.hortop@statcan.gc.ca](mailto:eric.hortop@statcan.gc.ca)

One way to study the changing realities and inequalities of diverse groups of people is through detailed disaggregated data and gender diversity is a critical dimension for disaggregation. However, it is important to be aware of statistical biases and errors when analyzing small subpopulations. Linkage biases (and errors) are concerning because biased findings may lead researchers to draw unsupported conclusions about the population. One illustration of this is that our analysis determined that percentages of transgender men and transgender women from the 2021 Census long-form sample data linked to the DRD are 88.4% and 89.7%, respectively. These are lower than those of cisgender men (95.8%), cisgender women (95.8%), and non-binary people (92.4%). To improve estimation from the 2021 CanCHEC, we briefly outline how the weighting procedures for scaling the CanCHEC data up to the population level were modified such that its estimates are representative of the target population of private households.

For the 2021 Census, using the main (final) weights led to generally minimal discrepancies between the population counts enumerated from the census short-form data (i.e., census counts) and those enumerated from the census long-form sample data (i.e., census long-form estimates) (Statistics Canada, 2023). To produce estimates for the population counts from the 2021 CanCHEC data (i.e., CanCHEC estimates), the main weights produced for the long-form estimates were adjusted and calibrated.

The replicate estimator chosen for the 2021 Census long-form sample survey was derived from Fay's balanced half-sample method, which determined the creation of replicates, the calculation of replicate weights and the multiplication factor used to estimate variance (Statistics Canada, 2023). To produce variance estimates for the CanCHEC estimates, the set of 100 replicate weights produced for variance estimates for the census long-form estimates was adjusted and calibrated.

Hortop (2024) described in detail the weighting procedures for adjusting and calibrating the main and Fay's Balanced Repeated Replication (BRR) weights for the 2021 CanCHEC, which were a modification of the weighting procedures for the 2016 CanCHEC developed by Hortop and Aubin (2020). The weighting procedures for creating the 2021 CanCHEC's main and replicate weights consisted of two steps:

1. Response homogeneous groups were first constructed using the linkage propensity score from a logistic model fitted for the long-form sample to adjust the main and replicate weights.
2. A cell calibration procedure was used to further adjust the main and replicate weights such that the weighted counts for a multivariate tabulation of the 2021 CanCHEC match the weighted counts from the 2021 Census long-form sample survey. The tabulation variables included the two-category gender variable ("men+" includes men (and/or boys), as well as some non-binary persons; and "women+" includes women (and/or girls), as well as some non-binary persons) among other characteristics.

In the first step, we assumed a linear relationship between the log odds of the event that the 2021 Census long-form sample survey records were included in the 2021 CanCHEC (i.e., linked to the DRD and duplicates were deleted) and the following independent variables:

- Age,
- Province or Territory of residence on Census day,
- Gender (two-category or five-category),
- School attendance,
- Document type classification (2A-L long-form questionnaire or 2A-R long-form questionnaire),
- Year of immigration,
- Indigenous identity,
- Mobility status – Place of residence one year ago (2020),
- Indigenous people; racialized groups; and non-racialized, non-Indigenous population,
- Labour force status.

After fitting the models, predicted linkage propensity scores were calculated for each census long-form sample survey record. Two sets of response homogeneous groups were then constructed using the predicted linkage probabilities from two logistic models fitted for the long-form sample to adjust the main and replicate weights. The first logistic model (model G-2) was used to predict linkage probabilities based on a given data set of independent variables that

included the two-category gender variable (men+ and women+). For the second logistic model (model G-5), the five-category gender variable ([cisgender men](#), [cisgender women](#), [transgender men](#), [transgender women](#), and [non-binary people](#)) replaced the two-category gender variable as one of the independent variables. The response homogeneous groups were constructed for ranges of predicted probabilities such that:

- Each group contained at least 1 000 total records,
- The weight adjustment factor must be less than 2,
- The difference between the end points of the range was greater than the following value:

$$\begin{cases} 0.05 & \text{if } (1 - p_\ell)/2 > 0.05, \\ 0.01 & \text{if } (1 - p_\ell)/2 < 0.01, \\ (1 - p_\ell)/2 & \text{otherwise,} \end{cases}$$

where  $p_\ell$  is the lowest observed linkage probability in the range. This process yielded 15 response homogeneous groups for both sets.

In the second step, the calibration cells were the Cartesian product of the calibration variables. The same set of calibration variables was used for the cell calibration procedure and they were:

- Age group (12 categories),
- Gender (two-category),
- Province or Territory of residence on Census day (11 categories with the Territories collapsed),
- Indigenous identity indicator,
- Immigration indicator.

This simple modification in the data modelling procedure makes our weighting procedure practical, as a standard procedure in poststratification by collapsing cells based on the sample size thresholds is sensitive to increasing or decreasing the variance of an estimate and may simultaneously increase its bias (Kim et al., 2007).

In this manuscript, we compare the two sets of CanCHEC estimates, produced based on the main and replicate weights adjusted using the two models (G-2 and G-5) and calibrated using the cell calibration procedure, to the census long-form estimates. Confidence intervals were constructed as the variance-based quality indicator to accompany the census long-form estimates, as well as the two sets of CanCHEC estimates. The confidence interval method for counts used is described in Section 2.2.

## 2.2 Modified Wilson confidence interval for counts

For the 2021 Census long-form estimates of counts, Statistics Canada (2023) recommends the modified Wilson confidence interval method because of its generally superior coverage (Neusy et al., 2021) and its practicality for implementation. For counts by gender diversity status and a secondary variable of interest, the Wald and Student confidence intervals would perform poorly for transgender men and women, and non-binary people due to small sample size. For the 2021 CanCHEC, the modified Wilson confidence interval method is used for estimated counts. The lower bound (LB) and the upper bound (UB) of a 95% modified Wilson confidence interval for a count  $Y$  are given by:

$$LB = \hat{Y} + t^2 \frac{1}{2} \frac{\widehat{Var}(\hat{Y})}{\hat{Y}} - \sqrt{t^2 \widehat{Var}(\hat{Y}) + \left[ t^2 \frac{1}{2} \frac{\widehat{Var}(\hat{Y})}{\hat{Y}} \right]^2},$$

$$UB = \hat{Y} + t^2 \frac{1}{2} \frac{\widehat{Var}(\hat{Y})}{\hat{Y}} + \sqrt{t^2 \widehat{Var}(\hat{Y}) + \left[ t^2 \frac{1}{2} \frac{\widehat{Var}(\hat{Y})}{\hat{Y}} \right]^2},$$

where

- $\hat{Y} = \sum_{k \in s} w_k$  is an estimate of  $Y$ ,
- $t$  is the 97.5<sup>th</sup> percentile of the Student's t-distribution with  $R$  degrees of freedom,
- $\widehat{Var}(\hat{Y})$  is the estimated variance of  $\hat{Y}$ . The equation of the variance estimator of an estimator  $\hat{Y}$  of a count  $Y$  from a set of  $R$  replicates used is shown below:

$$\widehat{Var}(\hat{Y}) = \frac{1}{R/2} \sum_{r=1}^R (\hat{Y}^{(r)} - \hat{Y})^2,$$

where

- $\hat{Y} = \sum_{k \in s} w_k$ ,
- $\hat{Y}^{(r)} = \sum_{k \in s} w_k^{(r)}$ ,
- $R = 100$  replicates.

The main weight of the CanCHEC is represented by  $w_k$ ,  $w_k^{(r)}$  is the CanCHEC's weight of replicate  $r = 1, \dots, R$ ,  $k$  is a cohort member with characteristics of interest (e.g., Black and non-binary intersection), and  $s$  is the CanCHEC (i.e., a subset of the Census long-form sample).

Note that the confidentiality rules for the 2021 CanCHEC were developed based on the confidentiality standards and guidelines for the dissemination of 2021 Census of Population data and those for the dissemination of administrative data. Various confidentiality rules were applied to the estimates presented in this manuscript to prevent the publication or disclosure of any information deemed confidential.

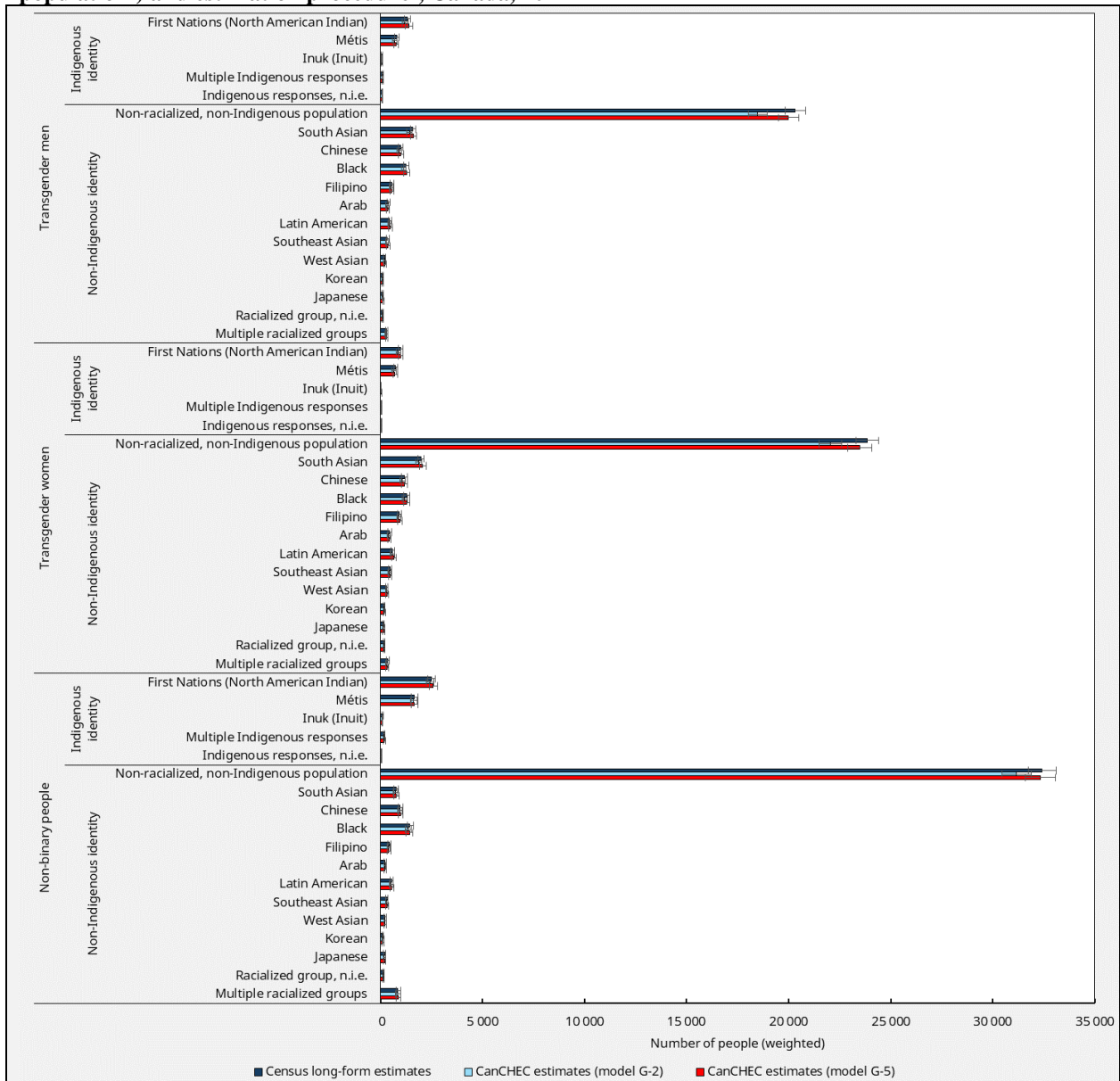
### 3. Results

The estimated counts of transgender men, transgender women, and non-binary people aged 15 and older in Canada, in private households, by population group (Indigenous identity population and non-Indigenous identity population) and estimation procedure, are shown in Figure 3-1. The 95% modified Wilson confidence intervals for the counts are also shown with whiskers over the bars. As the discrepancies between the estimated counts of cisgender men and cisgender women aged 15 and older in Canada, in private households, by population group and estimation procedure, are negligible, they are not shown.

For each intersection of the variables shown in Figure 3-1, widths of the confidence intervals around the three estimates are similar. Moreover, based on the length of the confidence intervals that overlap between the census long-form estimate and each of the CanCHEC estimates (models G-2 and G-5), we found that accounting for gender diversity (i.e., five-category gender variable) when adjusting main and replicate weights in the modelling procedure enabled generally better agreement between the census long-form estimates and the CanCHEC estimates (model G-5). The discrepancies between the census long-form estimates and the CanCHEC estimates (model G-2) are particularly notable among non-racialized, non-Indigenous transgender men, transgender women, and non-binary people, where the confidence intervals for all except for non-racialized, non-Indigenous non-binary people are non-overlapping.

Our analysis also compared the estimated counts of cisgender men and women, transgender men and women, and non-binary people aged 15 and older in Canada, in private households, by each of the following census long-form variables and estimation procedure: immigrant status (three categories), highest certificate, diploma or degree (six categories), and labour force status (three categories, derived). Again, accounting for gender diversity when adjusting main and replicate weights in the modelling procedure yielded generally better agreement between the census long-form estimates and the CanCHEC estimates (model G-5).

**Figure 3-1**  
**Estimated counts of transgender men, transgender women, and non-binary people aged 15 and older, in private households, by population group (Indigenous identity population<sup>1</sup> and non-Indigenous identity population<sup>2</sup>) and estimation procedure<sup>3</sup>, Canada, 2021**



1. "Indigenous identity" refers to whether the person identified with the Indigenous peoples of Canada. This includes those who identify as First Nations (North American Indian), Métis and/or Inuk (Inuit), and/or those who report being Registered or Treaty Indians (that is, registered under the *Indian Act* of Canada), and/or those who have membership in a First Nation or Indian band. Aboriginal peoples of Canada (referred to here as Indigenous peoples) are defined in the *Constitution Act, 1982, Section 35 (2)* as including the Indian, Inuit and Métis peoples of Canada. The abbreviation "n.i.e." refers to "not included elsewhere."

2. In this figure, the concept of "racialized group" within "non-Indigenous population" is based and derived directly from the concept of "visible minority" in the 2021 Census of Population. The *Employment Equity Act* defines visible minorities as "persons, other than Aboriginal peoples, who are non-Caucasian in race or non-white in colour." It consists mainly of the following groups: South Asian, Chinese, Black, Filipino, Arab, Latin American, Southeast Asian, West Asian, Korean and Japanese. The abbreviation "n.i.e." refers to "not included elsewhere."

3. The two sets of CanCHEC estimates, produced based on the main and replicate weights adjusted using the two models (G-2 and G-5) and calibrated using the cell calibration procedure, are compared to the census long-form estimates. A 95% modified Wilson confidence interval for the count was used as the variance-based quality indicator to accompany each estimated count for the 2021 Census long-form sample and the 2021 Canadian Census Health and Environment Cohort.

**Sources:** Statistics Canada, Census of Population, 2021; Statistics Canada, Canadian Census Health and Environment Cohort, 2021.

## 4. Discussion

This manuscript emphasizes the necessity of regularly assessing the quality of record linkage and reviewing the subsequent weighting procedures. We found that decisions in the weighting procedures can impact the accuracy of estimates for small population groups, including gender-diverse people. We compared the 2021 CanCHEC estimates to the 2021 Census long-form estimates using the data only collected through the 2021 Census long-form questionnaire (forms 2A-L and 2A-R). It was observed that accounting for gender diversity (i.e., five-category gender variable) when adjusting main and replicate weights in the modelling procedure enabled generally better agreement between the census long-form estimates and the CanCHEC estimates (model G-5).

A known limitation in our propensity modelling and calibration procedure is that the main and replicate weights only consider the linkage of the census long-form sample data to the DRD while many population health analyses would need to account for additional linkages to health administrative data.

## References

- Hortop E. (2024), “Adjusting and calibrating Fay’s Balanced Repeated Replication (BRR) weights for the 2021 Canadian Census Health and Environment Cohort (CanCHEC)”, internal report, Ottawa, Ontario: Statistics Canada.
- Hortop E., and Aubin P. (2020), “Adjusting and calibrating Fay’s Balanced Repeated Replication (BRR) weights for the 2016 Canadian Census Health and Environment Cohort (CanCHEC)”, internal report, Ottawa, Ontario: Statistics Canada.
- Kim, J. J., Li J., and Valliant R. (2007), “Cell collapsing in poststratification”, *Survey Methodology*, 32(2), pp. 139-150, Catalogue no. 12-001-X. Ottawa, Ontario.
- Neusy E., Savard, S.-A., Hidioglou, M., and Martin, V. (2021), “Modified Wilson Intervals for Estimated Counts with Application to Census 2021 Long Form Estimation”, presentation to the Advisory Committee on Statistical Methods, May 2021, internal report, Ottawa, Ontario: Statistics Canada.
- Statistics Canada. (2022), “Guide to the Census of Population, 2021”, Catalogue no. 98-304-X2021001. Ottawa, Ontario. Version updated November 30, 2022.
- Statistics Canada. (2023), “Sampling and Weighting Technical Report, Census of Population, 2021”, Catalogue no. 98-306-X2016001. Ottawa, Ontario.