

International Journal of Population Data Science

Journal Website: www.ijpds.org



Swansea University
Prifysgol Abertawe

Data resource profile: a nationally representative linked pregnancy cohort in Canada integrating clinical, social, and environmental data

Sabrina Chiodo^{1,2*}, Sonia M. Grandi^{1,2,3}, Jessica Gronsbell⁴, and Laura Rosella^{1,3,5,6}

Submission History

Submitted:	12/09/2025
Accepted:	19/12/2025
Published:	09/03/2026

¹ Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

² Child Health Evaluative Sciences, The Hospital for Sick Children, Toronto, Ontario, Canada

³ ICES, Toronto, Ontario, Canada

⁴ Faculty of Arts & Science, Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada

⁵ Institute for Better Health, Trillium Health Partners, Mississauga, Ontario, Canada

⁶ Laboratory Medicine and Pathobiology, Temerty Faculty of Medicine, University of Toronto, Ontario, Canada

Abstract

Introduction

Perinatal outcomes are shaped by clinical, social, and environmental factors, yet Canada lacks a nationally representative pregnancy cohort capturing these influences at the individual-level. This gap has limited the ability to address multifactorial drivers of maternal and fetal health. To fill this need, we established a linked cohort integrating survey, clinical, and contextual data to support equity-focused, precision public health research in maternal health.

Methods

We linked the Canadian Community Health Survey (CCHS; 2000–2017) to the Discharge Abstract Database (DAD) using Statistics Canada's Social Data Linkage Environment. Eligible participants were female (as defined by the binary CCHS sex variable), aged 15–49 years, with a hospital delivery within two years of their CCHS interview. We excluded multifetal gestations and retained only the first delivery per individual. Area-level and environmental exposures (i.e., neighbourhood inequity, pollution, greenspace, neighbourhood walkability, etc.) were appended via residential postal codes using the Postal Code Conversion File Plus (PCCF+).

Results

The cohort includes 13,360 singleton births. Pre-pregnancy data include sociodemographics, health behaviours, chronic conditions, psychosocial factors, and reproductive history. Contextual measures capture neighbourhood marginalization, air pollution, greenness, and built environment characteristics. In the CCHS, individuals who reported being pregnant at interview and those who did not (but later delivered) had similar characteristics (SMDs < 0.1), except for age and marital status. Data quality is supported by Statistics Canada's survey protocols, CIHI's hospital validation processes, and standardised geocoding.

Conclusion

Approved researchers can recreate this dataset within Statistics Canada's Research Data Centres using reproducible R code, which will become openly available on GitHub. The cohort enables research across descriptive epidemiology, causal inference, predictive modelling, and health equity evaluation, supporting investigations into multilevel determinants of maternal health. Future work should prioritise national mother–child linkages to expand life course research.

Keywords

pregnancy cohort; maternal health; data linkage; social determinants of health; perinatal outcomes

Key features

- **Unique nationally representative resource** – First Canadian population-based pregnancy cohort linking individual-level survey, clinical, social, and environmental data, enabling equity-focused maternal health research at a national scale.
- **Purpose** – Created to address the absence of linked, individual-level data capturing the multifactorial drivers of perinatal health, supporting precision public health approaches and reducing disparities.
- **Population and scope** – Includes 13,360 singleton births (2000–2017) among Canadian Community Health Survey (CCHS) respondents aged 15–49, linked to hospital delivery records across all provinces/territories except Quebec.
- **Data linkages** – Integrates CCHS survey data with the Discharge Abstract Database (DAD) for delivery outcomes, and area-level and environmental exposures (CAN-Marg, CANUE, Can-ALE) via the Postal Code Conversion File Plus (PCCF+).
- **Main data domains** – Using observation windows, captures pre-pregnancy sociodemographics, health behaviours, chronic conditions, psychosocial factors, reproductive history, clinically validated delivery outcomes, and neighbourhood-level marginalization, environmental exposures, and built environment characteristics.
- **Access** – Available to approved researchers through Statistics Canada's Research Data Centres (RDCs); cohort creation R code will become openly available on GitHub (<https://github.com/PopHealthAnalytics>) for reproducibility. Information on the RDC program, including eligibility criteria, application instructions, and proposal templates, can be found at: <https://www.statcan.gc.ca/en/microdata/data-centres/access>

*Corresponding Author:

Email Address: s.chiodo@utoronto.ca (Sabrina Chiodo)

Background

Pregnancy is a critical period in the life course, with lasting implications for both maternal and fetal health [1]. Complications during pregnancy can influence not only immediate outcomes but also long-term risks for chronic disease, mental health challenges, and intergenerational health trajectories [2, 3]. Despite this, Canada currently lacks a nationally representative, population-based pregnancy cohort with linked, individual-level data that captures a spectrum of clinical, social, and environmental influences on maternal and fetal health. This gap in data infrastructure limits our ability to understand and address the multifactorial drivers of pregnancy complications, and hinders the development of equitable, evidence-informed public health strategies.

Canada has several important recruited prospective birth cohort studies, including the Canadian Healthy Infant Longitudinal Development (CHILD) Cohort Study [4] and other regional pregnancy and birth cohorts, for example in Toronto [5], Brant County [6], and Quebec City [7]. These studies are rich in biological, genetic, and longitudinal phenotyping and have generated substantial insights into early-life determinants of child health. However, recruited cohorts are typically limited by sample size, geographic coverage, intensive follow-up costs, and selective participation, which can constrain their utility for national surveillance, health equity research, and population-level policy evaluation. In contrast, population-based electronic cohorts, derived from routinely collected survey and administrative health data, offer broader generalizability, scalability, and the ability to support replication, surveillance, and system-level analyses [8]. These two approaches are therefore complementary rather than competing, and population-based cohorts such as the one described here can serve as platforms for replication, external validation, and bias assessment of findings from recruited cohorts.

Pre-pregnancy population-based cohorts are particularly rare in Canada, and internationally. For example, the Southampton Women's Survey in the United Kingdom remains one of the few deeply phenotyped cohorts that recruited women prior to conception [9]. Again, such cohorts are difficult and costly to replicate at national scale. Population-based electronic pre-pregnancy cohorts offer a pragmatic and scalable alternative for studying fetal and maternal exposures at the population-level, particularly for social and environmental determinants that are not routinely captured in clinical registries.

Social, structural, and environmental inequities are increasingly recognised as critical drivers of perinatal health outcomes. In the United States, the American College of Obstetricians and Gynecologists (ACOG) has issued formal guidance emphasising the role of social determinants in reproductive health [10]. However, similar national efforts to integrate social and structural determinants into perinatal research have been limited in Canada due to siloed data systems and inconsistent availability of sociodemographic information in clinical settings for pregnant individuals [11–13]. While some studies have utilised area-level social determinants [14–16], the lack of individual-level linkage across datasets has hindered a comprehensive study of the risk and protective factors influencing perinatal health outcomes.

Very few studies in Canada have linked national survey data with administrative health records to examine pregnancy-related outcomes, and these efforts remain relatively limited in scope. Existing examples include analyses focused on women with disabilities, which have used Canadian Community Health Survey (CCHS) data linked to the Discharge Abstract Database (DAD) to examine preconception health disparities and pregnancy outcomes [17, 18]. Other linked CCHS–DAD studies have investigated specific questions such as socioeconomic differences in childbirth-related hospital utilisation [19] and caesarean delivery rates among immigrant versus Canadian-born individuals [20]. While these studies demonstrate the feasibility and value of national survey–administrative linkages for perinatal research, they have concentrated on selected exposures or subpopulations and have not fully leveraged the breadth of social determinants available in the survey data. Moreover, none have integrated environmental or contextual measures (e.g., neighbourhood-level marginalization, built environment characteristics, or exposure to pollutants), which are increasingly recognised as critical to perinatal health, particularly in the context of a changing climate.

To address this gap, we aimed to create a novel nationally representative cohort of pregnant individuals by linking the CCHS, a nationally representative survey that includes detailed sociodemographic, behavioural, and health information [21], with the DAD [22], which captures all hospital deliveries in Canada (excluding Quebec). We further enriched the cohort by linking it with environmental and contextual datasets, including the Canadian Marginalization Index (CAN-Marg) [23], the Canadian Urban Environmental Health Research Consortium (CANUE) [24], and the Canadian Active Living Environments (Can-ALE) index [25, 26]. These linkages were facilitated using the Postal Code Conversion File Plus (PCCF+) within Statistics Canada's secure Research Data Centre (RDC) environment [27].

This manuscript provides a detailed description of the linkage strategy, cohort structure, and data sources to enable future researchers to replicate and adapt the cohort for their own investigations. As part of this resource profile, we will also share R code on GitHub to support reproducibility and facilitate cohort creation within Statistics Canada's RDCs. An ancillary objective is to demonstrate the value of integrating individual, social, and environmental data to inform precision public health approaches and promote equitable maternal and perinatal health research across Canada.

This work aligns with recent calls for precision public health approaches, which aim to leverage linked population data to improve targeting of public health interventions [28]. It also fills a critical gap in maternal health research in Canada by enabling a more comprehensive study of risk and protective factors that considers the intersecting influences of income, education, race, healthcare access, neighbourhood conditions, and environmental exposures.

Methods

Data sources

The cohort was created through a multi-step linkage process that integrates population-based survey data with hospital

administrative records and environmental/contextual datasets. The core cohort was developed by linking the CCHS to DAD at the individual-level using the Social Data Linkage Environment (SDLE) at Statistics Canada's Toronto RDC [29]. Additional area-level and environmental exposures were appended based on residential postal codes reported at the time of the CCHS interview.

Primary data sources

Canadian Community Health Survey (CCHS)

The CCHS is a nationally representative, cross-sectional survey conducted by Statistics Canada to collect comprehensive data on health status, healthcare utilisation, and the social determinants of health among Canadians [21]. It is the most widely used population health survey in the country and provides essential information to monitor trends in population health, health behaviours, and healthcare access and utilisation, as well as to support policy development and guide health research at the national, provincial, and sub-provincial levels.

Originally administered biennially, the CCHS adopted an annual format in 2007, sampling approximately 65,000 individuals each year across all ten provinces and three territories. The survey uses a multistage, stratified cluster sampling design based on the Labour Force Survey frame, with recent cycles incorporating targeted sampling from the census to enhance representation of equity-deserving populations. The CCHS excludes individuals living on reserves, full-time members of the Canadian Forces, institutionalised populations, and residents of certain remote regions. Although these groups represent less than 3% of the Canadian population, they are important populations whose health experiences are not captured in the survey.

Data collection is conducted in both official languages (i.e., English and French) via telephone (CATI), in-person (CAPI), or electronic questionnaire (EQ), depending on cycle and region [21]. Typical overall response rates range from approximately 60–75% across cycles, comparable to other large national health surveys. Statistics Canada applies extensive quality control and validation procedures, including interviewer training, real-time logic checks, edit and imputation procedures, and post-collection weighting adjustments to mitigate the impact of nonresponse and coverage error.

Survey content includes annually repeated core modules (e.g., general health, chronic conditions, healthcare access) and rotating thematic modules (e.g., mental health, reproductive health, food insecurity), enabling both comparability over time and responsiveness to emerging public health priorities. It captures a wide range of individual-level variables, including age, sex, race, immigration status, education, income, and health behaviours (e.g., smoking, alcohol use, physical activity, diet).

For the creation of this cohort, the CCHS offers several advantages: with linkage, it captures pre-pregnancy exposures not available in clinical data, enables stratified analyses across sociodemographic groups, and provides rich behavioural and contextual variables within a nationally representative sampling framework.

Discharge Abstract Database (DAD)

The Discharge Abstract Database (DAD), maintained by the Canadian Institute for Health Information (CIHI), is a national administrative dataset that records all hospital separations (discharges, transfers, and in-hospital deaths) from acute care facilities across Canada, with the exception of Quebec [22]. Quebec accounts for approximately 22% of the Canadian population [30], and hospitalisation data for Quebec residents are collected separately by the ministère de la Santé et des Services sociaux and submitted to CIHI for inclusion in the Hospital Morbidity Database (HMD) [22, 31]; consequently, Quebec births are not captured in this cohort. The DAD contains detailed information on obstetrical deliveries, newborns, and stillbirths from acute inpatient hospitals, and in this cohort was used to identify delivery hospitalisations using validated ICD codes. Given that approximately 98% of births in Canada occur in hospitals, the vast majority of CCHS participants who are pregnant—or go on to become pregnant—can be accurately linked to delivery records in the DAD [32]. The small proportion of births not captured by the DAD primarily includes planned home births or deliveries in birthing centres, which are more common among individuals with low-risk pregnancies and those receiving midwifery-led care [33].

Each DAD record contains demographic, administrative, and clinical information abstracted from hospital charts, including patient age and sex, admission and discharge dates, discharge disposition, length of stay, most responsible diagnosis, up to 25 additional diagnoses, and procedures performed during hospitalisation. Clinical data are coded using the International Statistical Classification of Diseases and Related Health Problems and the Canadian Classification of Health Interventions (CCI). Since 2004–2005, all jurisdictions reporting to the DAD have fully adopted ICD-10-CA and CCI coding standards [31]. Prior to this, ICD-9-CA was used. DAD records are subject to validation rules at both the hospital level and at CIHI, including checks for logical consistency, valid code use, and adherence to abstracting standards.

Secondary linked environmental data sources

Canadian Marginalization Index (CAN-Marg)

CAN-Marg is a census-based measure developed by the Centre for Urban Health Solutions at St. Michael's Hospital and now maintained by Public Health Ontario [23, 34]. It was created to capture neighbourhood-level marginalization and facilitate research on health inequities across Canada. Using principal component analysis of census indicators, CAN-Marg generates standardised scores across four key domains—material deprivation, residential instability, dependency, and ethnic concentration. These scores are available at the dissemination area (DA) level, which represents the smallest standard geographic unit used by Statistics Canada for census data (typically comprising 400–700 residents) [35], and have been widely applied in population health studies to investigate the role of social context in health outcomes.

Canadian Urban Environmental Health Research Consortium (CANUE)

CANUE produces standardised, pan-Canadian datasets on environmental exposures that can be linked to health and survey data using residential postal codes [24]. These resources cover multiple environmental domains such as ambient air pollution, access to green space, and climate-related conditions. Depending on the measure, exposures are provided either at the postal code level or at a spatial resolution of 1 km². For this study, exposure assignments were aligned with the timing of each respondent's CCHS interview, using the interview year or the nearest available dataset.

Canadian Active Living Environments Index (Can-ALE)

Can-ALE is an area-based measure that evaluates how supportive neighbourhoods are for physical activity [25]. Developed by the Geo-Social Determinants of Health Research Group at McGill University in collaboration with the Public Health Agency of Canada, Can-ALE incorporates open-source geographic data to generate standardised indicators of walkability and access to local amenities. Data are available for all census dissemination areas (DAs) in Canada for the 2006 and 2016 census years. In this study, Can-ALE values were linked to participants' residential postal codes and matched to the census year closest to the date of their CCHS interview (e.g., a respondent interviewed in 2011 was assigned Can-ALE values from 2006, while a respondent interviewed in 2017 was linked to the 2016 index).

Cohort creation and linkage strategy

We constructed a novel nationally representative pregnancy cohort by linking population-based survey data with hospital discharge records and contextual environmental datasets to capture sociodemographic, behavioural, and environmental exposures prior to delivery. The cohort creation followed four key steps: (1) Cleaning and harmonising the CCHS; (2) Cleaning and restricting the DAD to relevant delivery records; (3) Linking CCHS respondents to delivery hospitalisations within a two-year follow-up window; and (4) Appending area-level and environmental exposures using residential postal codes. Figure 1 visually depicts the cohort creation process, including the data sources, linkage steps, and participant inclusion criteria.

To support transparency and reproducibility, sample R code illustrating each step of the cohort creation process will be made publicly available on GitHub. These scripts are intended as templates and include synthetic example data to demonstrate structure and logic without disclosing confidential information.

Step 1: Cleaning and Harmonisation of CCHS (2000–2017)

We began by restricting the CCHS to individuals who identified as female at birth, aged 15 to 49 years, who completed a survey cycle between 2000 and 2017 and consented to data linkage. This age range aligns with the reproductive lifespan. Given the evolving structure of the CCHS over time (e.g., changes

in sample design, survey content, and variable formats) each cycle was cleaned separately before harmonisation.

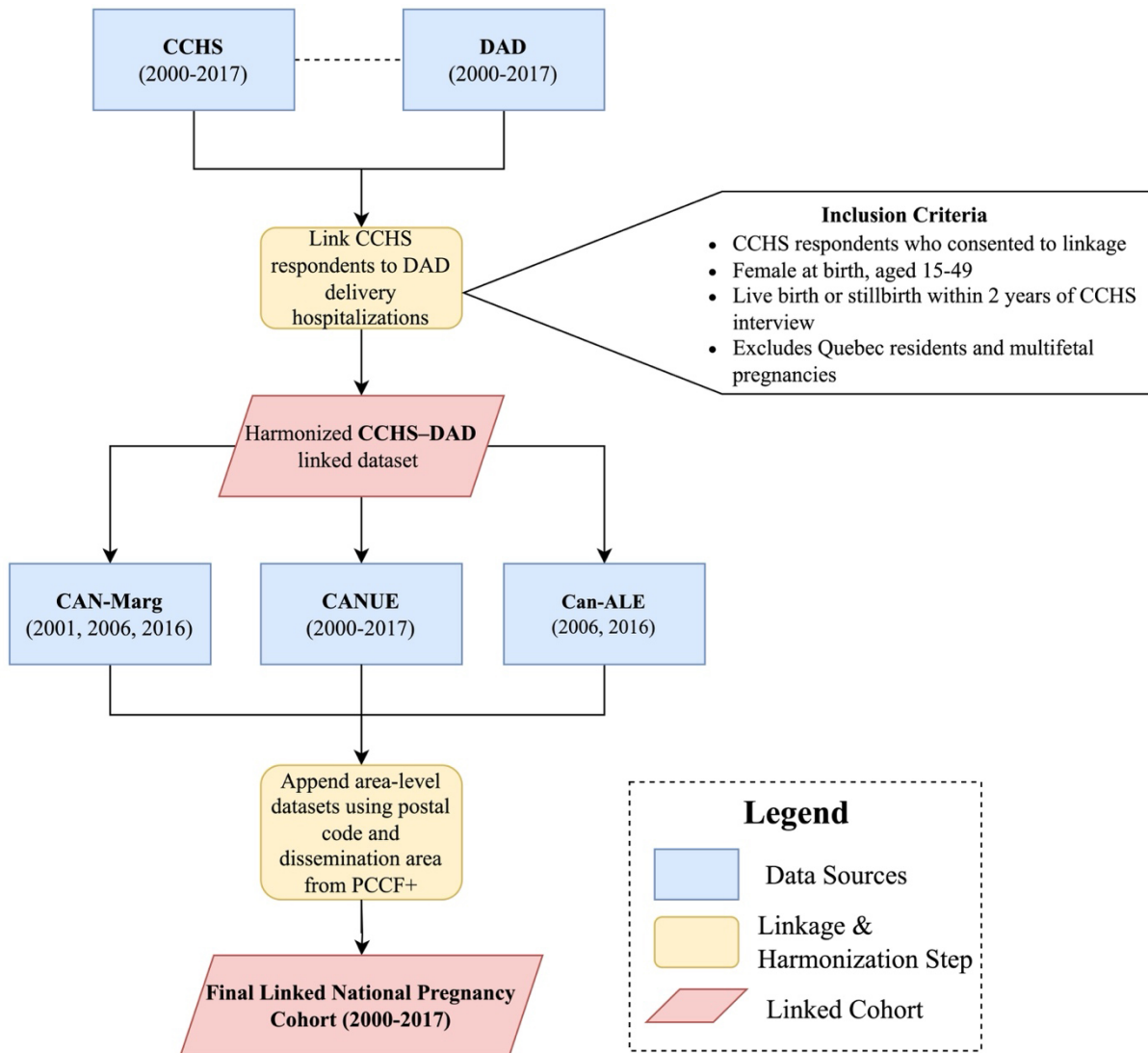
The cleaning process involved recoding and standardising key variables to ensure consistency across cycles. In particular, the 2015 CCHS redesign introduced changes in variable naming conventions, question phrasing, and data collection methodology (e.g., expanded use of electronic questionnaires), necessitating careful cross-cycle harmonisation. Variables were organised into the following conceptual domains:

- **Sociodemographics:** age, education, household income quintiles, immigration status, marital status, household size, employment status, food security, urban/rural residence, and self-identified visible minority status.
- **Chronic Conditions:** self-reported diagnoses of hypertension, diabetes, asthma, arthritis, back problems, migraines, intestinal ulcers, urinary incontinence, bowel disease, mood disorders, anxiety, and multimorbidity.
- **Psychosocial Stress and Perception:** self-rated life stress, sense of community belonging, life satisfaction, perceived mental health, and general health status.
- **Health Behaviours:** smoking status, alcohol use, and physical activity.
- **Healthcare Use and Access:** access to a regular healthcare provider, receipt of a flu shot in the past 12 months, history of Pap smear screening, and prior mental health consultations.
- **Reproductive and Preconception Health:** body mass index (BMI), parity, pregnancy status at the time of interview, folic acid use, and history of miscarriage or preterm delivery.

A complete list of variables included in the cohort, along with their definitions, is provided in Table 1. Because the cohort spans a 17-year period (2000–2017), we undertook systematic cross-cycle harmonisation to ensure temporal consistency of key variables. Where questions differed slightly across cycles, we created harmonised variables based on conceptually equivalent items. Major structural changes to the CCHS occurred in 2007 (transition from biennial to annual cycles) and in 2015 (survey redesign, updated questionnaire structure, and expanded electronic data collection) [21]. All variables were harmonised using standardised coding schemes and consistent derived thresholds to preserve comparability over time. Where direct harmonisation was not possible, variables were collapsed into broader categories to ensure stability across cycles. For example, employment status—measured using different question formats in early cycles—was recoded into consistent categories, and categorical variables such as BMI classification and income were recoded using standardised thresholds.

Although only a small proportion of CCHS respondents reported being pregnant at the time of the interview (approximately 400 per year), we flagged these individuals for internal validation (see Results) and descriptive purposes using cycle-specific pregnancy indicators (e.g., “Are you currently pregnant?”). However, current pregnancy was not

Figure 1: Data linkage and harmonisation process for the nationally representative pregnancy cohort. Flow diagram of Canadian Community Health Survey (CCHS, 2000–2017) respondents linked to Discharge Abstract Database (DAD, 2000–2017) delivery records



Inclusion required consent to linkage, female sex at birth, age 15–49, and a live birth or stillbirth within two years of interview; Quebec residents and multifetal pregnancies were excluded. The harmonised CCHS–DAD dataset was enriched with area-level measures from the Canadian Marginalization Index (CAN-Marg), Canadian Urban Environmental Health Research Consortium (CANUE), and Canadian Active Living Environments (Can-ALE) index using PCCF+ postal code linkage, yielding the final cohort (2000–2017).

a requirement for cohort inclusion, which was instead based on 2-year subsequent linkage to delivery hospitalisations.

After cleaning and harmonisation, all eligible cycles were merged into a single dataset.

Step 2: Cleaning and restriction of DAD delivery records (2000–2017)

To identify relevant delivery hospitalisations, we extracted and cleaned records from the DAD between 2000 and 2017. The DAD includes all acute care hospital separations in Canada (excluding Quebec), and delivery-related records were identified using validated diagnostic and procedural codes from the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CA) and the Tenth Revision,

Canadian Modification (ICD-10-CA). Clinical coding in the DAD transitioned from ICD-9-CA to ICD-10-CA between 2001 and 2004, with full national adoption by 2004–2005 [22]. All delivery and obstetric outcome definitions used in this cohort were developed using validated crosswalks and harmonised code lists spanning both coding systems to ensure consistency of outcome ascertainment over the full study period.

The full list of codes used to identify delivery hospitalisations is provided in Table 2.

We restricted the dataset to delivery records involving either a live birth or stillbirth. To maintain cohort consistency and reduce clinical heterogeneity, we excluded multifetal gestations (e.g., twins or higher-order multiples), identified through ICD-coded plurality indicators. These pregnancies are associated with elevated risk for complications and adverse

Table 1: Variables included in the linked CCHS–DAD pregnancy cohort, with definitions, data sources, and measurement scales

Variable Grouping	Variable	Definition ^{1–3}	Dataset
Sociodemographic & Economic Status	Age	Continuous	CCHS
	Age group	15–19; 20–24; 25–29; 30–34; 35–39; 40–49	CCHS
	Household income	Quintile 1 (lowest 20%) to Quintile 5 (highest 20%)	CCHS
	Household size	Continuous	CCHS
	Education	Less than Secondary; Secondary Graduate; Post-secondary Graduate	CCHS
	Immigration status	Canadian; Recent immigrant (<10 years); Established immigrant (≥10 years)	CCHS
	Food insecurity	Food secure; Food insecure	CCHS
	Visible minority status	White; Visible Minority	CCHS
	Marital status	Married or Common Law; Single or Never Married; Widowed, Separated, or Divorced	CCHS
	Employment status	Employed (worked last week); Employed (absent last week); Unemployed (last week)	CCHS
Health Behaviours	Alcohol consumption	Regular drinker: ≥1x/week; Occasional drinker: <1x/week but >1x/year; Non-drinker: none in past year	CCHS
	Cigarette smoking	Never smoker; Former smoker; Current smoker	CCHS
	Fruit and vegetable intake	<3 servings/day; 3–5 servings/day; >5 servings/day	CCHS
	Physical activity	Inactive; Moderately active; Active	CCHS
Chronic Conditions	Pre-pregnancy diabetes	Yes/No	CCHS
	Hypertension	Yes/No	CCHS
	Asthma	Yes/No	CCHS
	Arthritis	Yes/No	CCHS
	Back problems	Yes/No	CCHS
	Migraines	Yes/No	CCHS
	Intestinal ulcers	Yes/No	CCHS
	Urinary incontinence	Yes/No	CCHS
	Bowel disease	Yes/No	CCHS
Psychosocial Stress & Perception	Multimorbidity	≥2 chronic diseases; <2 chronic diseases	CCHS
	Life stress	Not at all; Not very; A bit; Quite a bit; Extreme stress	CCHS
	Community belonging	Very weak; Somewhat weak; Somewhat strong; Very strong	CCHS
	Life satisfaction	Satisfied; Neither; Dissatisfied	CCHS
	Self-rated health	Poor; Fair; Good; Very good; Excellent	CCHS
	Self-rated mental health	Poor; Fair; Good; Very good; Excellent	CCHS
Reproductive & Preconception Health	BMI	Continuous	CCHS
	BMI group	Underweight (<18.5); Normal weight (18.5–24.9); Overweight (25.0–29.9); Moderately obese (30.0–34.9); Very obese (35.0–39.9); Severely obese (≥40.0)	CCHS
	Parity (last 5 years)	Yes/No	CCHS
	Folic acid use	Yes/No	CCHS
	History of spontaneous abortion	None; 1 previous miscarriage; >1 previous miscarriage.	DAD
	History of preterm delivery	None; 1–3 previous preterm births; ≥4 previous preterm births.	DAD
	History of live birth	None; 1–3 previous live births; ≥4 previous live births.	DAD

Continued

Table 1: Continued

Biomedical & Health Services	Access to regular doctor (past 12 months)	Yes/No	CCHS
	Ever had a pap smear	Yes/No	CCHS
	Flu shot (past 12 months)	Yes/No	CCHS
	Ever had a mental health consult	Yes/No	CCHS
Built Environment⁴	Urbanicity	Urban residence; Rural residence	CCHS
	Intersection density	Number of three-way or greater intersections per km ²	Can-ALE
	Dwelling density	Number of residential units per km ²	Can-ALE
	Transit stops	Density of public transit stops within a given area	Can-ALE
	Active living environment index	Composite measure (intersection density, dwelling density, proximity to destinations, access to public transit), standardised and categorised into 5 classes (C1 to C5)	Can-ALE
Environmental Exposures⁴	PM2.5 concentrations	Annual fine particulate matter ($\mu\text{g}/\text{m}^3$) from satellite and modelled data, calibrated to monitoring stations	CANUE
	NO ₂ concentrations	Annual nitrogen dioxide (ppb) from land-use regression models	CANUE
	Ozone (O3)	Annual average ground-level ozone (ppb) from atmospheric models and monitoring stations	CANUE
	Greenness Index (GRLAN)	Normalized Difference Vegetation Index (−1 to 1), higher values indicate more surrounding green space	CANUE
Area-level Social Determinants⁴	Neighbourhood deprivation	Quintile 1 (least deprived) to Quintile 5 (most deprived)	CAN-Marg
	Ethnic concentration	Quintile 1 (least concentrated) to Quintile 5 (most concentrated)	CAN-Marg
	Residential instability	Quintile 1 (least stable) to Quintile 5 (most stable)	CAN-Marg
	Dependency	Quintile 1 (least dependent) to Quintile 5 (most dependent)	CAN-Marg
	Neighbourhood income	Quintile 1 (lowest) to Quintile 5 (highest)	Census

¹Variable definitions are based on Canadian Community Health Survey (CCHS) content guides, the Discharge Abstract Database (DAD) abstracting manual, and associated documentation for linked contextual datasets (CAN-Marg, CANUE, Can-ALE, Census).

²Measurement scales are presented as collected; derived variables were harmonised across CCHS cycles (2000–2017) prior to linkage.

³*Note:* According to sample size, some variables may need to be collapsed to meet Statistics Canada's vetting requirements for disclosure control. For example, psychosocial variables (e.g., self-rated life stress, community belonging, life satisfaction) can be combined into fewer categories for reporting.

⁴Area-level variables are assigned using respondents' residential postal code at the time of CCHS interview and the Postal Code Conversion File Plus (PCCF+).

outcomes. We also derived supplementary variables where available, such as gestational age at delivery, history of spontaneous abortion, and prior term deliveries.

For the purposes of this manuscript, we have not included any specific pregnancy outcomes, as users of the cohort may define outcomes according to their research objectives. For example, current manuscripts in preparation using this cohort focus on adverse pregnancy outcomes (e.g., gestational diabetes, placental abruption, and preeclampsia) which have been defined using ICD codes and incorporated into the analytic dataset.

Step 3: linking CCHS and DAD records

Using Statistics Canada's SDLE, CCHS respondents were linked to DAD records through a probabilistic record linkage

process conducted by Statistics Canada. This method uses multiple identifiers (e.g., name, date of birth, sex, postal code) to assign an encrypted unique linkage key for each individual and applies standardised quality assurance procedures, including clerical review and validation thresholds, to minimise false matches and missed links [29].

Linkage was restricted to CCHS respondents who provided informed consent for record linkage at the time of survey participation, consistent with Statistics Canada data governance policies. Across survey cycles, over 80% of CCHS respondents consent to linkage. Among consenting respondents, the linkage success rate between the CCHS and the DAD is approximately 85% (excluding Quebec, which does not participate in the DAD) [36, 37].

While individual-level linkage error rates are not released due to confidentiality restrictions, prior methodological

Table 2: Delivery outcome and multifetal pregnancy definitions and corresponding ICD-9-CA¹ and ICD-10-CA² codes used in cohort creation

Condition	ICD Version	ICD Code ³	Definition
Delivery Outcome (Live or Stillbirth)	ICD-9	650	Normal delivery
		V27x	Outcome of delivery
		V30x	Liveborn infant, single birth
	ICD-10	V39x	Liveborn infant, unspecified
		Z37x	Outcome of delivery
		Z38x	Liveborn infants according to place of birth
		O80x	Single spontaneous delivery
Multifetal Pregnancy	ICD-9	O82x	Delivery by elective caesarean section
		V272	Twins, both liveborn
		V273	Twins, one liveborn and one stillborn
		V274	Twins, both stillborn
		V275	Other multiple birth, all liveborn
		V276	Other multiple birth, some liveborn
		V277	Other multiple birth, all stillborn
		651	Multiple gestation
		V31	Twin birth, mate liveborn
		V32	Twin birth, mate stillborn
		V33	Triplets, all liveborn
		V34	Triplets, some liveborn
		V35	Triplets, all stillborn
		V36	Other multiple births, all liveborn
		V37	Other multiple births, some liveborn
	ICD-10	Z37.2	Twins, both liveborn
		Z37.3	Twins, one liveborn and one stillborn
		Z37.4	Twins, both stillborn
		Z37.5	Other multiple birth, all liveborn
		Z37.6	Other multiple birth, some liveborn
		Z37.7	Other multiple birth, all stillborn
		Z38.3	Twin, born in hospital
		Z38.4	Twin, born outside hospital
		Z38.5	Twin, unspecified place of birth
		Z38.6	Other multiple, born in hospital
		Z38.7	Other multiple, born outside hospital
		Z38.8	Other multiple, unspecified as to place of birth

¹ICD-9-CA = International Classification of Diseases, Ninth Revision, Canadian Adaptation.

²ICD-10-CA = International Classification of Diseases, Tenth Revision, Canadian Adaptation.

³The “x” in codes (e.g., V27x, Z37x, O80x) denotes inclusion of all subcategories within that code range.

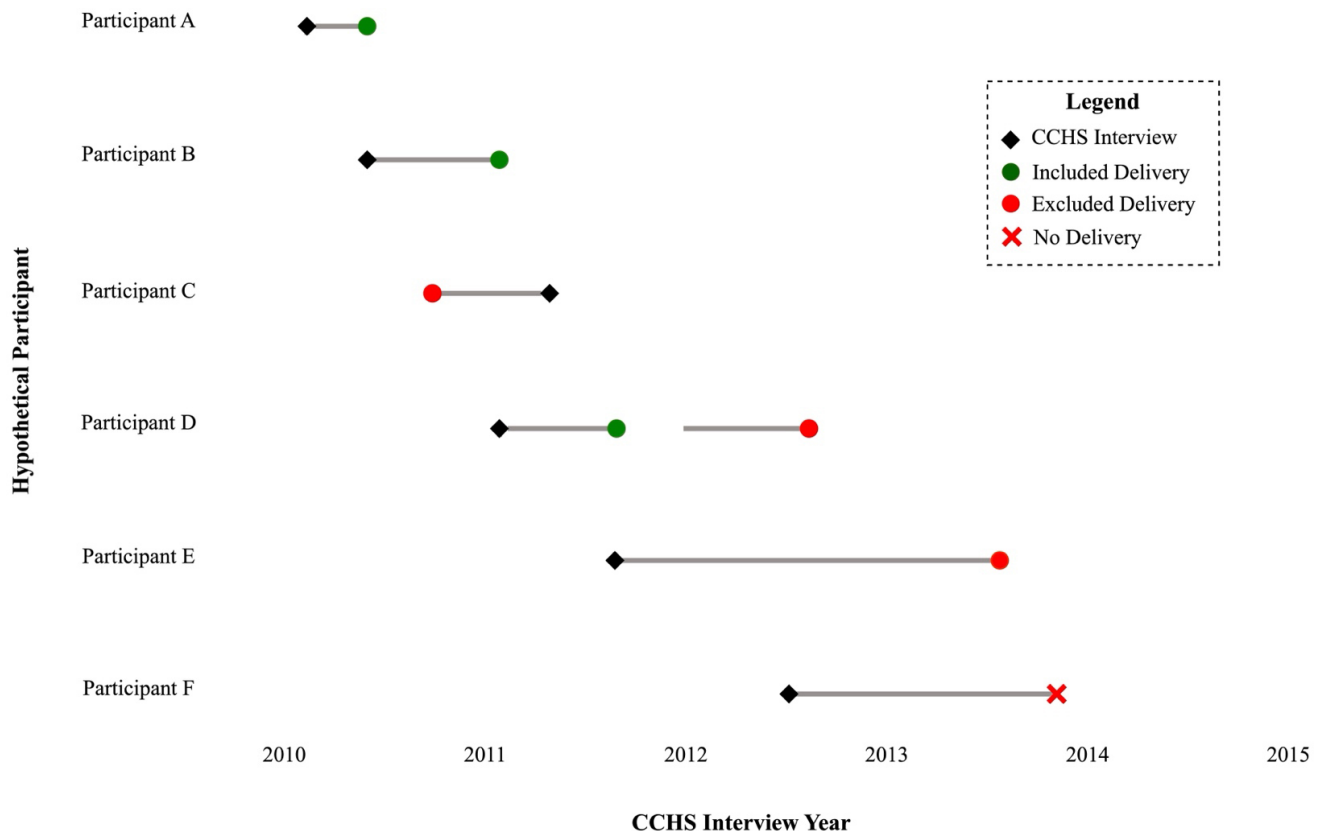
evaluations of CCHS–administrative data linkages have demonstrated high linkage quality and minimal bias at the population-level. For example, a coverage evaluation comparing linked CCHS–hospital records with full hospital discharge data found only a modest under-coverage (~7.7%) of hospitalisations in the linked file, indicating high overall linkage quality at the population-level [38]. Nevertheless, potential sources of linkage bias remain, including differential consent to linkage (not all survey respondents agree to link), and failure to link when personal identifiers are missing or inconsistent. These limitations should be considered when interpreting analyses based on the linked cohort.

For this study, only respondents with a hospital delivery within two years following their CCHS interview date were

included in the analytic cohort. This two-year window was chosen to balance temporal alignment between pre-pregnancy exposures and delivery outcomes, while minimising misclassification. In cases where multiple deliveries occurred within the window, only the first was retained to reflect the pregnancy most proximate to the CCHS interview.

Figure 2 illustrates the inclusion and exclusion logic using hypothetical participant timelines. Participant A is included in the cohort, as they delivered within 9 months of their CCHS interview, suggesting they would likely have reported being pregnant at the time of the survey. Participant B is also included, with a delivery that occurred within the two-year window. Participant C is not included in the cohort as they delivered prior to their CCHS interview. Participant D had

Figure 2: Inclusion and exclusion criteria for cohort linkage using hypothetical participant timelines



Examples illustrate inclusion (e.g., delivery within two years of CCHS interview) and exclusion (e.g., delivery before interview, no linked record, or delivery beyond two years). For individuals with multiple deliveries within two years, only the first was retained.

two deliveries within two years of their CCHS interview; however, only the first delivery is retained, since survey responses—particularly those related to pregnancy—may not correspond to subsequent pregnancies within the two-year follow-up period. Participant E is excluded because their delivery occurred more than two years after the CCHS interview date, falling outside the inclusion window. Finally, Participant F is excluded from the cohort due to the absence of a linked delivery record.

Step 4: linking area-level and environmental exposures

We linked area-level exposures to CCHS participants using their six-digit residential postal code at the time of survey. Postal codes were mapped to standardised geographic units using the Postal Code Conversion File Plus (PCCF+), a probabilistic linkage tool developed by Statistics Canada.

The PCCF+ is an enhanced version of the Postal Code Conversion File (PCCF) [27] and includes a SAS-based program that assigns postal codes to a range of standard geographic identifiers (e.g., dissemination areas, census subdivisions, census metropolitan areas, and health regions). Unlike the PCCF, PCCF+ employs population-weighted random allocation for postal codes that map to more than one geographic area, improving accuracy in urban and multi-jurisdictional settings. It also draws on additional sources, including the Postal Code Population Weight File, Geographic Attribute File, and Health Region boundary files.

Using the CCHS postal code and DA-level generated by the PCCF+, we were able to link CCHS respondents to multiple contextual datasets: CAN-Marg, CANUE, Can-ALE. The PCCF+ itself also generates area-level income variables, which provide geographically standardised proxies for socioeconomic position.

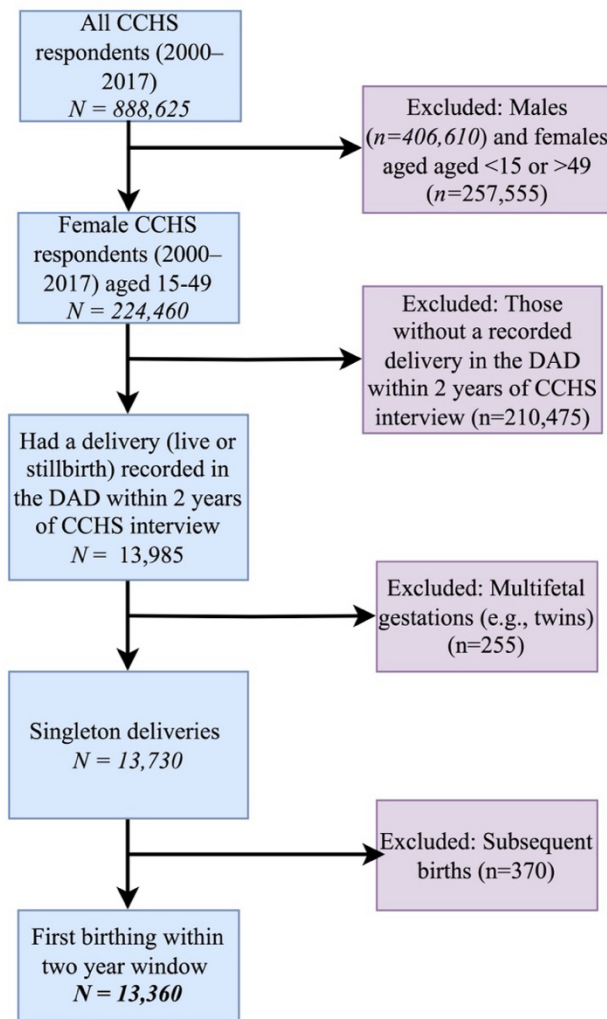
Variables were organised into the following conceptual domains:

- **Built Environment:** Urbanicity, intersection density, dwelling density, transit stops, and active living environment index from Can-ALE.
- **Environmental Exposures:** Fine particulate matter (PM_{2.5}), nitrogen dioxide (NO₂), ozone (O₃), and greenness index (NDVI) from CANUE.
- **Area-Level Social Determinants:** Neighbourhood deprivation, ethnic concentration, residential instability, and dependency from CAN-Marg and neighbourhood income from the census.

Results

There are very few nationally representative cohorts that link survey and administrative data to create a pregnancy cohort, and to our knowledge, this is the first to construct one using a two-year period from CCHS interview to delivery allowing for more births to be included. Between 2000 and 2017, a total of 888,625 individuals participated in the CCHS. After

Figure 3: Cohort derivation flowchart for the linked CCHS–DAD pregnancy dataset (2000–2017)



Flowchart showing the selection of the final analytic cohort from Canadian Community Health Survey (CCHS) respondents (2000–2017) linked to Discharge Abstract Database (DAD) delivery records. Exclusions were applied sequentially for sex and age, absence of a recorded delivery within two years of interview, multifetal gestations, and subsequent births within the two-year window, resulting in 13,360 singleton first births.

restricting to female respondents aged 15 to 49 years, 224,460 individuals were eligible for linkage. Of these, 13,985 were successfully linked to a hospital delivery record (live birth or stillbirth) in the DAD within two years of their CCHS interview date. We excluded 255 multifetal gestations and 370 cases involving multiple deliveries within the follow-up period, retaining only the first delivery per individual. The final analytic cohort includes 13,360 individuals with singleton births. Note that cohort numbers have been rounded to the nearest 5, in accordance with Statistics Canada's vetting and disclosure control rules. The cohort derivation process is illustrated in Figure 3.

As no direct cohort comparisons were available, we compared those who reported being pregnant at the time of their CCHS interview (i.e., flagged via the question "Are you currently pregnant?") with those who were not pregnant at interview but went on to deliver within the two-year follow-up window (Table 3). This comparison was conducted to help validate our approach by assessing whether those pregnant at the time of interview had similar sociodemographic

characteristics to those who became pregnant within the subsequent two years, thereby supporting the plausibility that the two-year window captures a comparable pregnancy population.

When comparing groups, those who reported being pregnant at the time of their CCHS interview ($n=4,605$) were broadly similar in sociodemographic characteristics to those who were not pregnant at interview but delivered within two years ($n=8,755$) (Table 3). Differences included age and marital status: those pregnant at the time of the interview were slightly older on average (29.7 vs. 28.8 years; $SMD=0.163$) and more likely to be married or living common law (88.0% vs. 79.1%; $SMD=0.241$). These minor differences likely reflect individuals who were interviewed mid-pregnancy during their peak reproductive years, when most births occur within partnered relationships. In contrast, those not pregnant at the time of the interview (but who conceived within the two-year follow-up) had more follow-up time and therefore were older and had more time to enter a marriage or common law relationship before becoming pregnant, resulting in a

Table 3: Sociodemographic Characteristics of Cohort Participants by Pregnancy Status¹ at CCHS² Interview

Variable	Pregnancy Status CCHS Interview		Standardised Mean Difference (SMD)
	No	Yes	
n	8755	4605	
Age, years (mean (SD))	28.8 (5.2)	29.7 (5.3)	0.163
Age group, years (%)			0.167
15-19	405 (4.6)	140 (3.0)	
20-24	1380 (15.8)	620 (13.5)	
25-29	2945 (33.6)	1465 (31.8)	
30-34	2820 (32.2)	1555 (33.8)	
35-39	1065 (12.2)	675 (14.7)	
40-49	140 (1.6)	150 (3.3)	
Marital status (%)			0.241
Married or Common Law	6925 (79.1)	4050 (88.0)	
Single or Never Married	1550 (17.7)	490 (10.6)	
Widowed, Separated, or Divorced	280 (3.2)	70 (1.5)	
Highest level of education (%)			0.051
Less than Secondary	400 (4.6)	190 (4.1)	
Secondary Graduate	1370 (15.7)	650 (14.1)	
Post-secondary Graduate	6985 (79.8)	3770 (81.9)	
Immigration status (%)			0.036
Canadian	6555 (74.9)	3420 (74.3)	
Recent Immigrant (<10 years in Canada)	1370 (15.7)	705 (15.3)	
Established Immigrant (≥10 years Canada)	830 (9.5)	485 (10.5)	
Household income quintile³ (%)			0.029
Quintile 1	1705 (19.5)	925 (20.1)	
Quintile 2	2300 (26.3)	1155 (25.1)	
Quintile 3	1975 (22.6)	1065 (23.1)	
Quintile 4	1180 (13.5)	620 (13.5)	
Quintile 5	1595 (18.2)	845 (18.4)	
Visible minority status (%)			0.001
White	6485 (74.1)	3410 (74.1)	
Visible Minority	2270 (25.9)	1200 (26.1)	
Place of residence⁴ (%)			0.021
Urban	7465 (85.3)	3960 (86.0)	
Rural	1290 (14.7)	645 (14.0)	
Employment status in week prior to interview (%)			0.085
Employed (worked last week)	5565 (63.6)	2865 (62.2)	
Employed (absent last week)	915 (10.5)	400 (8.7)	
Unemployed (last week)	2275 (26.0)	1340 (29.1)	
Food Security (%)			0.059
Food Secure	7640 (87.3)	4110 (89.3)	
Food Insecure	1115 (12.7)	500 (10.9)	
Household Size (mean (SD))	3.1 (1.3)	3.1 (1.4)	0.028

¹Pregnancy status was self-reported at the time of CCHS interview ("Are you currently pregnant?").²CCHS = Canadian Community Health Survey.³Household income quintiles are based on province-specific income cut-points adjusted for household size, derived from CCHS.⁴Urban/rural classification is based on CCHS definitions of census metropolitan areas and census agglomerations.

slightly broader distribution of ages and marital statuses in this group.

Differences in other factors, such as education, immigrant status, household income, visible minority status, urban/rural residence, employment, food security, and household size, were minimal (SMDs < 0.10; Table 3).

Taken together, the close balance of sociodemographic characteristics between individuals who were pregnant at interview and those who conceived within the subsequent two years supports the internal representativeness of the linked cohort with respect to key population-level social and economic characteristics captured in the CCHS.

Discussion

This manuscript introduces a novel nationally representative cohort that integrates individual-level survey data, hospital discharge records, and environmental/contextual datasets to support population-based research on adverse pregnancy outcomes in Canada. This linked dataset addresses a longstanding gap in Canadian maternal health research infrastructure by enabling a more holistic and equity-informed examination of pregnancy-related risk factors across clinical, social, and environmental domains.

Strengths

A major strength of this cohort is its ability to capture rich, pre-pregnancy individual-level data through the CCHS, which is typically unavailable in administrative datasets. The CCHS offers detailed information on sociodemographic characteristics, psychosocial perceptions, health behaviours, chronic conditions, and other social determinants, allowing for nuanced analyses of APO risk across diverse population subgroups. When linked to the DAD, which provides comprehensive clinical data on delivery hospitalisations, the dataset supports robust investigations into both individual- and system-level factors associated with maternal and fetal outcomes.

The addition of environmental and contextual data, via linkages to CAN-Marg, CANUE, and Can-ALE, further enhances the cohort's utility. These datasets capture neighbourhood-level marginalization, environmental exposures (e.g., air pollution, greenness), and built environment characteristics known to influence perinatal health. Together, these linkages enable researchers to explore the complex, multilevel determinants of APOs in ways that have previously not been possible at the national level in Canada.

Another key strength is the methodological transparency and reproducibility of the cohort creation process. The cleaning, harmonisation, and linkage strategies were carefully documented, and R code will be provided on GitHub to support replication in other Statistics Canada RDCs. This increases the accessibility of the dataset and encourages widespread uptake by researchers across Canada.

Limitations

Several limitations should be noted. First, although the CCHS is nationally representative, the linked cohort includes only

individuals who consented to data linkage and had a hospital-based delivery within two years of their survey interview. This may introduce selection bias by excluding pregnancies that did not result in a hospital birth—such as early pregnancy losses, home births, or elective terminations—as well as individuals who declined linkage. As a result, the cohort may be skewed toward lower-risk pregnancies. Although approximately 98% of births in Canada occur in hospitals and over 80% of CCHS respondents consent to data linkage, some bias may remain if individuals who decline linkage differ systematically from consenters. Previous evaluations of CCHS linkage suggest that while such differences exist, the linked cohorts remain broadly representative of the target population [38, 39].

Second, the cohort excludes residents of Quebec due to jurisdictional differences in data availability. While this exclusion affects national representativeness, the remaining provinces and territories still provide broad coverage of the Canadian population.

Third, environmental and contextual data may be limited by temporal and spatial resolution; for instance, some CANUE variables were only available for specific years or at coarser geographic scales, which may lead to misclassification.

Fourth, although this cohort includes over 13,000 singleton births and is well-powered for analyses of common adverse pregnancy outcomes and sociodemographic disparities, the sample size may limit its utility for studying rare outcomes, such as specific congenital malformations or uncommon obstetric syndromes. For outcomes with very low incidence, estimates may be unstable, and analyses may require pooling across extended calendar periods, broader outcome groupings, or linkage with external registries where feasible. Users of the cohort should therefore carefully consider statistical power when designing studies focused on rare events.

Fifth, the cohort was restricted to singleton first deliveries within the two-year linkage window. Multifetal gestations were excluded due to their distinct etiologic pathways, clinical management, and outcome distributions, which differ substantially from singleton pregnancies [40]. Likewise, retaining only the first delivery per individual avoids correlated within-person outcomes, reduces time-varying exposure misclassification (as CCHS exposures may not correspond to subsequent pregnancies), and limits structural dependence that can complicate many analytic approaches, including predictive modelling and causal inference methods. While these design choices enhance internal consistency and analytic tractability, they limit the applicability of this resource to multifetal and repeat pregnancies. Importantly, users retain full flexibility to modify these restrictions using the reproducible cohort creation R code should alternative analytic frameworks be required for specific research objectives.

Sixth, because the cohort spans a 17-year period, temporal changes in clinical practice guidelines, obstetric management, and public health policy may influence outcome ascertainment and risk factor distributions over time. For example, screening thresholds for gestational diabetes, hypertensive disorders of pregnancy, and prenatal surveillance practices evolved over the study period [41–44], which may affect observed outcome frequencies independently of underlying risk. While all variables were harmonised to maximise comparability across cycles, residual temporal heterogeneity may remain. Users of

the cohort are encouraged to adjust for calendar time or stratify analyses by era where relevant.

Seventh, while the DAD provides validated clinical outcomes, certain adverse pregnancy outcomes may be underreported or misclassified if they are not documented at the time of hospital discharge. For example, if a woman is diagnosed with preeclampsia in an outpatient setting (such as at a family physician's office), but the diagnosis is not recorded during her hospital stay, it may not be captured in the DAD. While it is rare that this would not be recorded at delivery admission, this example illustrates the potential for misclassification. Additionally, the DAD does not contain information on prenatal care, such as provider type, number of visits, or outpatient interventions, limiting the ability to assess care trajectories across pregnancy.

There are also limitations in the temporal alignment of events. While the CCHS provides valuable pre-pregnancy data, the timing between survey completion and conception varies across individuals. A two-year follow-up window was used to balance temporal proximity with cohort size, and this timeframe is supported by previous literature suggesting that social determinant data remains relatively stable over a two-year period [45, 46]. However, this is an assumption, and the precise alignment of variables with pregnancy onset may still vary. Moreover, pregnancy itself is a known trigger for residential mobility, which may result in changes in neighbourhood-level and environmental exposures between the time of survey completion and the index pregnancy [47, 48]. As a result, geospatial exposure assignments based on postal code at the time of the CCHS interview may misclassify true exposures during pregnancy for some individuals. Researchers using this cohort may wish to explore shorter or longer linkage windows, incorporate time-varying exposure definitions where feasible, or conduct sensitivity analyses depending on the specific exposures, outcomes, and research questions of interest.

Finally, although prior linkage evaluations demonstrate minimal population-level bias, the absence of an external national benchmark containing comparable social and behavioural variables limits our ability to empirically quantify residual selection or linkage bias beyond internal validation.

Current applications & future directions

This cohort is already being utilised in ongoing research to develop predictive models for adverse pregnancy outcome risk that incorporates social and environmental determinants. These projects demonstrate the cohort's capacity to support a wide range of analytic approaches—from descriptive epidemiology and population health surveillance to more advanced inferential analyses and machine learning applications. For example, current studies are exploring how individual- and area-level factors interact to shape the risk of conditions such as preeclampsia, placental abruption, and gestational diabetes, with the aim of developing risk stratification tools that can inform targeted prevention strategies.

Looking ahead, the dataset holds significant potential for researchers, clinicians, and policymakers interested in maternal and child health. It can be used to identify high-risk subgroups based on intersecting sociodemographic, behavioural, and

environmental exposures; monitor disparities in perinatal outcomes across geographic regions or population groups; and evaluate the effectiveness of public health interventions or policy changes over time. In particular, the availability of linked environmental data allows for the investigation of how exposures such as air pollution, green space access, and neighbourhood walkability contribute to perinatal health and health equity. Additionally, researchers can examine regional variations in maternal outcomes and explore how these are shaped by contextual factors such as healthcare access and neighbourhood-level deprivation.

As new sources of environmental and contextual data become available (e.g., through updated census cycles, improved spatial resolution, or longitudinal geocoding) future iterations of the cohort could be expanded and refined to reflect these improvements. Researchers may also consider extending the follow-up period to examine longer-term outcomes, such as postpartum health or the development of chronic conditions in mothers.

An important future extension of this data infrastructure is the creation of parallel general-population comparator cohorts using the CCHS with linkage through Statistics Canada's SDLE. While the present resource is intentionally defined as a pregnancy-based cohort to support analyses of pregnancy-specific risk, prediction, and outcomes, the same linkage framework fully supports the construction of matched or population-based comparator cohorts of individuals without recorded deliveries. Such designs would enable a broader range of research questions, including contrasts in baseline health, social, and environmental exposures between those who do and do not go on to experience pregnancy or specific pregnancy outcomes. We view this as a key opportunity for future research building on the infrastructure described here.

At present, there is no established mother–baby linkage available within Statistics Canada's RDCs, meaning that a child's hospital discharge data or future survey responses (e.g., from subsequent cycles of the CCHS) cannot be directly linked back to their mother's file. This represents a notable limitation for researchers interested in examining intergenerational health outcomes or child development following birth. While such linkage is possible at the provincial level (e.g., through the Better Outcomes Registry & Network [BORN] in Ontario [49] or the British Columbia Perinatal Data Registry [50]) there is currently no national infrastructure to support this at the RDC. Future efforts should advocate for the development of a standardised, national mother–child linkage framework, which would significantly expand the utility of this cohort for life course research and child health surveillance.

Data access

The linked dataset is held securely within Statistics Canada's Research Data Centre (RDC) network and is not available in a public open-access repository due to privacy, confidentiality, and data governance requirements under the Statistics Act. Access is restricted to approved researchers with projects that have undergone peer review and have received approval from Statistics Canada.

Researchers interested in using this cohort must submit a research proposal to the RDC program through the Microdata

Access Portal. Proposals are reviewed for scientific merit, feasibility, and compliance with confidentiality protections. Successful applicants are granted access to the data in a secure RDC facility and must conduct all analyses on-site or through remote desktop connections where available. All outputs are subject to Statistics Canada's vetting and disclosure control procedures before release.

Information on the RDC program, including eligibility criteria, application instructions, and proposal templates, can be found at: <https://www.statcan.gc.ca/en/microdata/data-centres/access>.

To inquire about access, researchers may contact the Statistics Canada RDC program at: Email: statcan.mad-rdc-data-dam-drc-data.statcan@statcan.gc.ca

Reproducible R code used for cohort creation, including cleaning, linkage, and variable harmonisation templates with synthetic example data, will become openly available on GitHub: <https://github.com/PopHealthAnalytics>.

Conclusions

This cohort represents an important step toward integrating multidimensional data to better understand adverse pregnancy outcomes in Canada. By linking nationally representative survey data with hospital discharge records and contextual environmental exposures, it provides a unique platform for precision public health research. Capturing pre-pregnancy sociodemographic and behavioural factors alongside clinically validated outcomes and area-level determinants enables a more comprehensive, equity-informed approach to maternal and perinatal health.

This resource fills a longstanding gap in Canadian health data, allowing examination of how individual, social, and environmental factors shape pregnancy outcomes at a population-level. It establishes a foundation for epidemiologic research, targeted interventions, risk prediction tools, and policies to improve maternal health and reduce disparities. Ultimately, it contributes to building a more responsive, data-driven maternal health system that centers equity, prevention, and population impact.

Acknowledgements

This research was conducted at the University of Toronto, a part of the Canadian Research Data Centre Network (CRDCN). This service is provided through the support of the Canada Foundation for Innovation, the Canadian Institutes of Health Research, the Social Sciences and Humanities Research Council, and Statistics Canada, and through the support of the University of Toronto. All views expressed in this work are our own.

Ethics statement

This study was approved by the University of Toronto Research Ethics Board (Protocol #00045767).

Conflicts of interest

The authors declare no conflicts of interest.

Publication consent

We confirm that we have obtained the necessary permissions to publish analyses using the linked cohort described in this manuscript.

Funding

Laura C. Rosella is supported by a Canada Research Chair (950–230702), and Sabrina Chiodo is supported by a Canadian Institutes of Health Research Doctoral Award and a Studentship Award from the Edwin S.H. Leong Centre for Healthy Children at the University of Toronto and The Hospital for Sick Children.

Data sharing

The data are held within Statistics Canada's Research Data Centre (RDC) network and are not publicly available due to privacy and governance requirements under the Statistics Act. Access is limited to approved researchers with peer-reviewed projects authorised by Statistics Canada. Applications are submitted through the Microdata Access Portal; approved researchers work within secure RDC facilities or, where available, via remote desktop. All analyses occur in the secure environment, and outputs undergo Statistics Canada's vetting before release. Further information is available at <https://www.statcan.gc.ca/en/microdata/data-centres/access>. Environmental data were provided by the Canadian Urban Environmental Health Research Consortium (CANUE) under a study-specific Data Sharing and Use Agreement and cannot be redistributed; access may be requested directly from CANUE.

Contributors statement

Sabrina Chiodo and Laura C. Rosella conceived and designed the study, and contributed to data acquisition, analysis, and interpretation. Sabrina Chiodo led the drafting of the manuscript. Sonia M. Grandi and Jessica Gronsbell provided methodological guidance, contributed to data interpretation, and critically reviewed the manuscript for important intellectual content. All authors read and approved the final version of the manuscript.

References

1. Garr Barry, V., et al., *Adverse Pregnancy Outcomes and Postpartum Care as a Pathway to Future Health*. Clin Obstet Gynecol, 2022. **65**(3): p. 632-647. <https://doi.org/10.1097/grf.0000000000000724>
2. Yee, L.M., E.C. Miller, and P. Greenland, *Mitigating the Long-term Health Risks of Adverse Pregnancy Outcomes*. JAMA, 2022. **327**(5): p. 421-422. <https://doi.org/10.1001/jama.2021.23870>
3. Damm, P., et al., *Gestational diabetes mellitus and long-term consequences for mother and offspring: a view from*

- Denmark. *Diabetologia*, 2016. **59**: p. 1396-1399. <https://doi.org/10.1007/s00125-016-3985-5>
4. Takaro, T.K., et al., *The Canadian Healthy Infant Longitudinal Development (CHILD) birth cohort study: assessment of environmental exposures*. *Journal of exposure science & environmental epidemiology*, 2015. **25**(6): p. 580-592. <https://doi.org/10.1038/jes.2015.7>
5. Anderson, L.N., et al., *The Ontario Birth Study: a prospective pregnancy cohort study integrating perinatal research into clinical care*. *Paediatric and Perinatal Epidemiology*, 2018. **32**(3): p. 290-301. <https://doi.org/10.1111/ppe.12473>
6. Wahi, G., et al., *Aboriginal birth cohort (ABC): a prospective cohort study of early life determinants of adiposity and associated risk factors among Aboriginal people in Canada*. *BMC public health*, 2013. **13**(1): p. 608. <https://doi.org/10.1186/1471-2458-13-608>
7. Morgan, C., et al., *Cord blood vitamin D status and neonatal outcomes in a birth cohort in Quebec, Canada*. *Archives of gynecology and obstetrics*, 2016. **293**(4): p. 731-738. <https://doi.org/10.1007/s00404-015-3899-3>
8. Gregg, E.W., et al., *Use of Real-World Data in Population Science to Improve the Prevention and Care of Diabetes-Related Outcomes*. *Diabetes Care*, 2023. **46**(7): p. 1316-1326. <https://doi.org/10.2337/dc22-1438>
9. Inskip, H.M., et al., *Cohort profile: The Southampton Women's Survey*. *International journal of epidemiology*, 2006. **35**(1): p. 42-48. <https://doi.org/10.1093/ije/dyi202>
10. Committee on Health Care for Underserved Women, *ACOG Committee Opinion No. 729: importance of social determinants of health and cultural awareness in the delivery of reproductive health care*. *Obstetrics and gynecology*, 2018. **131**(1): p. e43-e48. <https://doi.org/10.1097/aog.0000000000002459>
11. Williams-Roberts, H., et al., *Facilitators and barriers of sociodemographic data collection in Canadian health care settings: a multisite case study evaluation*. *International Journal for Equity in Health*, 2018. **17**(1): p. 186. <https://doi.org/10.1186/s12939-018-0903-0>
12. Campbell, E.E., et al., *Socioeconomic status and adverse birth outcomes: a population-based Canadian sample*. *Journal of biosocial science*, 2018. **50**(1): p. 102-113. <https://doi.org/10.1017/s0021932017000062>
13. Lau, R.S., et al., *Siloed mentality, health system suboptimization and the healthcare symphony: a Canadian perspective*. *Health Research Policy and Systems*, 2024. **22**(1): p. 87. <https://doi.org/10.1186/s12961-024-01168-w>
14. Miao, Q., et al., *Agreement assessment of key maternal and newborn data elements between birth registry and Clinical Administrative Hospital Databases in Ontario, Canada*. *Archives of Gynecology and Obstetrics*, 2019. **300**(1): p. 135-143. <https://doi.org/10.1007/s00404-019-05177-x>
15. Heaman, M.I., et al., *Inequities in utilization of prenatal care: a population-based study in the Canadian province of Manitoba*. *BMC pregnancy and childbirth*, 2018. **18**(1): p. 430. <https://doi.org/10.1186/s12884-018-2061-1>
16. Miao, Q., et al., *Association between maternal marginalization and infants born with congenital heart disease in Ontario Canada*. *BMC public health*, 2023. **23**(1): p. 790. <https://doi.org/10.1186/s12889-023-15660-5>
17. Brown, H.K., et al., *Pregnancy Outcomes in Canadian Women With Disabilities: Results From Linked Survey and Health Administrative Data*. *Journal of Obstetrics and Gynaecology Canada*, 2023. **45**(10): p. 102179. <https://doi.org/10.1016/j.jogc.2023.06.010>
18. Forbes, S.M., et al., *Preconception health disparities among reproductive-aged women with and without disabilities in Canada*. *Canadian Journal of Public Health*, 2024. **115**(3): p. 493-501. <https://doi.org/10.17269/s41997-024-00873-x>
19. Mah, S.M., et al., *Childbirth-Related Hospital Burden by Socioeconomic Status in a Universal Health Care Setting*. *Int J Popul Data Sci*, 2018. **3**(1): p. 418. <https://doi.org/10.23889/ijpds.v3i1.418>
20. Hetherington, E., et al., *Cesarean deliveries among immigrant and Canadian-born women in a representative community population in Canada: A retrospective cohort study*. *Journal of Obstetrics and Gynaecology Canada*, 2022. **44**(2): p. 148-156. <https://doi.org/10.1016/j.jogc.2021.07.017>
21. Statistics Canada. *Canadian Community Health Survey - Annual component (CCHS)*. 2021 [cited 2022 October 15]; Available from: https://www23.statcan.gc.ca/imdb/p3Inst.pl?Function=assembleInstr&a=1&=&en&Item_Id=1293153#qb1293208.
22. Canadian Institute for Health Information (CIHI). *Discharge Abstract Database (DAD) metadata*. [cited 2024 January 23]; Available from: <https://www.cihi.ca/en/discharge-abstract-database-dad-metadata>.
23. Matheson, F.I., et al., *Development of the Canadian Marginalization Index: a new tool for the study of inequality*. *Canadian Journal of Public Health/Revue Canadienne De Sante'e Publique*, 2012: p. S12-S16.
24. Doiron, D., et al., *The Canadian Urban Environmental Health Research Consortium (CANUE): a national data linkage initiative*. *International Journal of Population Data Science*, 2018. **3**(4). <https://doi.org/10.23889/ijpds.v3i4.715>

25. Ross, N., et al., *Canadian active living environments database (Can-ALE) user manual & technical document*. Geo-Social Determinants of Health Research Group, Department of Geography, McGill University: Montreal, QC, Canada, 2018.
26. Herrmann, T., et al., *A pan-Canadian measure of active living environments using open data*. Health Rep, 2019. **30**(5): p. 16-25.
27. Statistics Canada. *Postal Code OM Conversion File Plus (PCCF+)* 2025 [cited 2025 June]; Available from: <https://www150.statcan.gc.ca/n1/en/catalogue/82F0086X>.
28. Naumova, E.N., *Precision public health: is it all about the data?* Journal of public health policy, 2022. **43**(4): p. 481-486. <https://doi.org/10.1057/s41271-022-00367-5>
29. Statistics Canada. *Overview of the Social Data Linkage Environment (SDLE)*. 2024 [cited 2025 June]; Available from: <https://www.statcan.gc.ca/en/sdle/overview>.
30. Institut de la statistique du Québec. *How has Québec changed over the past 25 years?* . 2024 [cited 2025 December]. Available from: <https://statistique.quebec.ca/en/communiquer/how-has-quebec-changed-over-past-25-years>.
31. Canadian Institute for Health Information (CIHI), *Data Quality Documentation, Discharge Abstract Database–Multi-Year Information*. 2012, Ottawa, ON.
32. Canadian Institute for Health Information (CIHI), *Hospital births in Canada: A focus on women living in rural and remote areas*. 2013, Ottawa, ON.
33. Darling, E.K., et al., *Outcomes associated with planned place of birth among low-risk pregnancies in Ontario, Canada (2012–2021): A protocol for a population-based propensity score weighted cohort study*. Plos one, 2024. **19**(5): p. e0302489. <https://doi.org/10.1371/journal.pone.0302489>
34. Public Health Ontario. *Ontario Marginalization Index (ON-Marg)*. 2023 [cited 2025 June]. Available from: <https://www.publichealthontario.ca/en/Data-and-Analysis/Health-Equity/Ontario-Marginalization-Index>.
35. Statistics Canada, *Dissemination area: Detailed definition*. 2018 [cited 2025 December]; Available from: <https://www150.statcan.gc.ca/n1/pub/92-195-x/2011001/geo/da-ad/def-eng.htm>.
36. Rosella, L.C., et al., *A study protocol for a predictive model to assess population-based avoidable hospitalization risk: Avoidable Hospitalization Population Risk Prediction Tool (AvHPoRT)*. Diagnostic and Prognostic Research, 2024. **8**(1): p. 2. <https://doi.org/10.1186/s41512-024-00165-5>
37. Statistics Canada. *Canadian Community Health Survey Data (2000 to 2011) Linked to the Discharge Abstract Database (1999/2000-2012/2013)*, 2022 [cited 2025 December]; Available from: <https://www.statcan.gc.ca/en/microdata/data-centres/data/cencchs-dad>.
38. Rotermann, M., *Evaluation of the coverage of linked Canadian Community Health Survey and hospital inpatient records*. Health reports, 2009. **20**(1): p. 45.
39. Raina, P., et al., *Agreement between self-reported and routinely collected health-care utilization data among seniors*. Health services research, 2002. **37**(3): p. 751-774. <https://doi.org/10.1111/1475-6773.00047>
40. Duffy, C.R., *Multifetal gestations and associated perinatal risks*. Neoreviews, 2021. **22**(11): p. e734-e746. <https://doi.org/10.1542/neo.22-11-e734>
41. Boyd, P., et al., *The evolution of prenatal screening and diagnosis and its impact on an unselected population over an 18-year period*. BJOG: An International Journal of Obstetrics & Gynaecology, 2012. **119**(9): p. 1131-1140. <https://doi.org/10.1111/j.1471-0528.2012.03373.x>
42. Cuckle, H. and R. Maymon. *Development of prenatal screening—A historical overview*. in *Seminars in perinatology*. 2016. Elsevier. <https://doi.org/10.1053/j.semperi.2015.11.003>
43. Negrato, C.A. and M.B. Gomes, *Historical facts of screening and diagnosing diabetes in pregnancy*. Diabetology & metabolic syndrome, 2013. **5**(1): p. 22. <https://doi.org/10.1186/1758-5996-5-22>
44. Tanner, M.S., et al., *The evolution of the diagnostic criteria of preeclampsia-eclampsia*. American Journal of Obstetrics and Gynecology, 2022. **226**(2): p. S835-S843. <https://doi.org/10.1016/j.ajog.2021.11.1371>
45. Chagin, K. and A.R. Sehgal, *Stability of Patient Social Determinants of Health With a Focus on Food Insecurity*. Journal of Primary Care & Community Health, 2024. **15**: p. 21501319241273214. <https://doi.org/10.1177/21501319241273214>
46. Fosse, N.E. and S.A. Haas, *Validity and stability of self-reported health among adolescents in a longitudinal, nationally representative survey*. Pediatrics, 2009. **123**(3): p. e496-e501. <https://doi.org/10.1542/peds.2008-1552>
47. Bell, M.L. and K. Belanger, *Review of research on residential mobility during pregnancy: consequences for assessment of prenatal environmental exposures*. Journal of exposure science & environmental epidemiology, 2012. **22**(5): p. 429-438.
48. Saadeh, F.B., et al., *Pregnant and moving: understanding residential mobility during pregnancy and in the first year of life using a prospective birth cohort*. Maternal and

child health journal, 2013. **17**(2): p. 330-343. <https://doi.org/10.1007/s10995-012-0978-y>

49. Murphy, M.S.Q., et al., *Data Resource Profile: Better Outcomes Registry & Network (BORN) Ontario*. Int J Epidemiol, 2021. **50**(5): p. 1416-1417h. <https://doi.org/10.1093/ije/dyab033>
50. Frosst, G., et al., *Validating the British Columbia Perinatal Data Registry: a chart re-abstraction study*. BMC pregnancy and childbirth, 2015. **15**(1): p. 123. <https://doi.org/10.1186/s12884-015-0563-7>

