International Journal of Population Data Science

Journal Website: www.ijpds.org

Addressing Missing Paternal Linkage in Administrative Data – A Candidate Father Approach

Amani F. Hamad[1,2,*], Lin Yan[1,3], Olawale Ayilara[1], Oleguer Plana-Ripoll[4,5], and Lisa M. Lix[1,2,3]

Submission History

Submitted:	14/02/2026
Accepted:	06/09/2026
Published:	xx/xx/20xx

[1] College of Community and Global Health, University of Manitoba, Winnipeg, Canada
[2] The Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, Canada
[3] George & Fay Yee Centre for Healthcare Innovation, University of Manitoba, Winnipeg, Canada
[4] ISGlobal, Barcelona, Spain
[5] National Centre for Register-based Research, Department of Public Health, Aarhus University, Aarhus, Denmark

Abstract

Introduction
Paternal linkages are often missing from administrative health data, posing a challenge for familial health research. We propose and evaluate an approach to identify candidate fathers using residence and demographic information.

Methods
We conducted a population-based cohort study of Manitoba (Canada) adults between 1977 and 2017. We modelled the probability of missing paternal linkage and applied it to remove fathers for offspring with highest predicted probability (10%–50% missing linkages) in pre-1980 births, when missing paternal linkage was lowest. Candidate fathers were identified within the same postal code; up to three were randomly selected using parental age gap weights. We quantified the percentage of candidate sets containing the true father and estimated hazard ratios (HR) for paternal transfer of mental disorders and ischaemic heart disease (IHD) using true versus candidate fathers. Risk estimates were compared with indicator category and complete-case approaches.

Results
The cohort included 140,001 offspring; 25.8% had missing paternal linkages. In the pre-1980 cohort, 27-28% of candidate sets included the true father. Using true fathers, paternal history was associated with mental disorders (HR 1.27, 95% CI 1.23–1.31) and IHD (HR 1.94, 95% CI 1.35–2.79). Using candidate fathers, associations were attenuated, with HRs ranging from 1.17 (95% CI 1.13–1.22) to 1.24 (95% CI 1.20–1.29) for mental disorders and 1.64 (95% CI 0.92–2.89) to 1.91 (95% CI 1.32–2.75) for IHD. Indicator category and complete-case approaches produced estimates close to the true association for mental disorders but deviated for IHD as paternal missingness increased. In the full cohort, estimates were similar for candidate father, indicator category, and complete-case approaches for both outcomes.

Conclusions
Although the candidate father approach had low accuracy in identifying true fathers, it produced attenuated yet comparable risk estimates, particularly at low levels of missing paternal linkage, suggesting utility for familial research.

Keywords
administrative health data, data linkage, paternal linkage, missing data, intergenerational research, familial health

*Corresponding Author:
Email Address: amani.hamad@umanitoba.ca (Amani F. Hamad)

https://doi.org/10.23889/ijpds.v6i1.3426
November 2026 © The Authors. Open Access under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/deed.en)

Introduction

Administrative health data that link individuals to their parents provide a valuable resource for population-based familial health research. Such data have been used to investigate the intergenerational transfer of health conditions and social risks, and the influence of exposures across generations, and to develop risk prediction models [1–14]. Linkages to parents in administrative data rely on personal identifiers, address or household identifiers, birth and marriage registry information, or health claims such as hospital records [15, 16]. Unlike maternal linkages, which can be derived from birth information recorded in the mother's hospital record, paternal linkages cannot be determined from the father's hospital records because paternal identifiers are not systematically captured in birth-related administrative records. Therefore, most existing familial health research focused on maternal and child health and did not consider fathers [7, 17–25]. Fathers play an important role in shaping offspring health through shared genetic risk, environmental exposures, and social factors that may differ from those of mothers. Paternal health is associated with a wide range of offspring health outcomes across the life course, including mental disorders and cardiovascular disease [1, 26, 27]. Consequently, complete and accurate paternal linkage is increasingly important for studies investigating familial disease risk, intergenerational health, and family-based risk prediction using population-based administrative data. In many data sources, paternal linkages are therefore derived using indirect information, such as address or household identifiers [16, 28–30]. This can result in incomplete paternal linkages and varying completeness over time, which presents challenges for using administrative health data in familial research [29, 31]. For example, 10% of births registered in Western Australia between 1980 and 2005 had incomplete paternal information [31]. Similarly, while administrative health data in Manitoba, Canada, contain virtually complete linkages to mothers (over 98%), linkages to fathers, which rely on household information reflecting registered family units, have decreased significantly from 86% in 1974 births to 36% in 2019 births [32].

Methods for addressing missing paternal information should be considered when incorporating fathers into population-based familial health research. Most previous familial research relied on complete-case analysis to address missing parental linkages [2, 10, 11, 14, 26, 33, 34]. This approach excludes offspring who cannot be linked to their parents. As this approach is simple to implement despite complex mechanisms underlying missing paternal linkages, and does not require advanced technical expertise or computational resources, it is commonly used in familial research. However, when paternal linkage is not missing completely at random, it could introduce selection bias and compromise the validity of estimated associations, in addition to reducing sample size and statistical precision [35, 36]. Using an indicator category by assigning an additional category to denote missing information about the variable is an alternative approach that has been previously used to address missing data [1, 37]. This approach is straightforward to apply and has the advantage of retaining cohort size; however, it can also result in biased estimates [36, 38–40].

Given the challenges posed by incomplete paternal linkages in administrative data and the limitations of existing approaches to handling missing paternal information, an approach that could probabilistically impute the missing relationships is needed. We hypothesised that geographic proximity and demographic plausibility, specifically, shared 6-digit postal code and a likely parental age gap, could be used to identify plausible candidate fathers. Because parents generally live together or in proximity at the time of birth and have age differences within a typical range, these characteristics provide reasonable criteria for identifying candidate fathers. Therefore, we aimed to develop and evaluate a novel approach to identify candidate fathers for offspring with missing paternal linkage. To evaluate whether the candidate father approach could preserve father–offspring associations in epidemiologic analyses, we adopted it in two use cases. These use cases tested the association between: (a) a paternal history of mental disorders and offspring mental disorders, and (b) a paternal history of ischaemic heart disease (IHD) and offspring IHD. Mental disorder and IHD were selected as test cases because they have different administrative data accuracy, aetiology and prevalence, and therefore can demonstrate the approach's utility across diverse research scenarios. This novel approach provides a practical method for incorporating otherwise missing fathers by assigning plausible proxy fathers, thereby improving the completeness of paternal information for analysis, reducing bias, and facilitating more robust familial health research.

Methods

Design and Data Source

This was a population-based cohort study using administrative health data from the province of Manitoba, Canada, between April 1, 1974 and March 31, 2020. Due to the universal and publicly funded healthcare system in Manitoba, almost all encounters with the healthcare system, such as hospitalisations and outpatient physician visits, are captured for administrative purposes. The de-identified data are retained for research purposes in the Manitoba Population Research Data Repository (Repository) housed at the Manitoba Centre for Health Policy.

Four datasets were used in this study: the Manitoba Health Insurance Registry (Registry), Hospital Abstracts, Medical Claims, and Statistics Canada Census. The Registry captures all individuals insured for healthcare coverage in Manitoba, and dates back to 1970. It includes information on person-level demographics, composition of family units, residential postal codes, and healthcare coverage periods. Family units are recorded in the Registry using the Family Registration Numbers (FRNs), which are assigned to children and adults residing in the same household. Mothers generally share the same FRN as their children; relationships between fathers and mothers must be reported to *Manitoba Health, Seniors and Long-Term Care* for fathers to share the same FRN as the mother and children. Failure to report the mother-father relationship results in missing paternal linkages in the Registry. FRNs change over time to accommodate family structure changes; for example, children receive their own

FRNs when they turn 18 years of age. For offspring with paternal linkage in the Registry, the father identified through the FRN recorded at birth (birth FRN) was treated as the reference ("true") father in the validation analyses. The birth FRN reflects the legally registered parent–child relationship at birth and is expected to represent the biological father in most births, although biological paternity cannot be directly verified using administrative data. Accordingly, paternal linkage in this study represents the father recorded in the Registry and should be interpreted as capturing both shared genetic and familial environmental influences, rather than biological inheritance alone.

Hospital Abstracts record all hospitalisations in Manitoba since 1970 and capture diagnoses using the International Classification of Diseases (ICD) system. The adapted eighth revision (ICDA-8) was used in Hospital Abstracts between 1970 and 1979; the ninth revision with clinical modifications (ICD-9-CM) between 1980 and 2004; and the tenth revision with Canadian adaptations (ICD-10-CA) since 2005. Medical Claims captures outpatient physician visits with diagnosis and procedure information. Diagnoses were recorded in Medical Claims using ICD-8 between 1970 and 1979, and ICD-9-CM since 1980. Statistics Canada Census is a nationwide survey generally conducted every five years and captures geographic and sociodemographic information. The 2016 Census was used to create area-level income quintiles for the study cohort.

Study Population

The cohort included Manitoba residents (hereafter referred to as offspring) who turned 18 years of age between 1977 and 2017 (corresponding approximately to birth years 1959 to 1999), had residence information (6-digit postal code) recorded within one year of birth, and met the eligibility criteria described below. The index date was defined as April 1^{st} of the year an offspring turned 18 years of age. A few exclusion criteria were applied. First, offspring without three years of healthcare coverage before and after their index date were excluded. Second, offspring were excluded if they were missing maternal linkage or if their mothers or fathers had less than three years of continuous healthcare coverage immediately prior to their index date. Three years of continuous healthcare coverage was required to ascertain prevalent mental disorders and IHD in offspring and parents, to evaluate the performance of the candidate father approach. Offspring with missing maternal linkage were excluded since maternal information was used to identify candidate fathers for offspring missing paternal linkages. Third, offspring with a recorded mental disorder or IHD diagnosis prior to their index date, as defined below, were excluded. Limiting the analysis to incident cases ensures a temporal exposure-outcome relationship, with parental diagnoses preceding offspring diagnoses. Finally, one offspring per mother was randomly selected to reduce within-family correlation. Therefore, maternal siblings were not included in the cohort. Fathers could still contribute to multiple offspring if they had children with different mothers. Offspring were followed from 18 years of age until the earliest of a mental disorder or IHD diagnosis, death, moving out of Manitoba, or the end of the study period (March 31, 2020). Follow-up duration varied by birth year, with more recent birth cohorts contributing shorter follow-up because of administrative censoring at the end of study period. Parental histories of mental disorders and IHD were ascertained prior to offspring's index date to preserve the temporal exposure-outcome relationship.

Study Measures

Mental disorders and IHD were identified for offspring and their parents in Hospital Abstracts and Medical Claims data using established case definitions [41, 42]. ICD codes and their descriptive labels, which were used to construct the case definitions, are listed in Supplementary Appendix 1. Sex, area-level income quintile, region of residence (urban, rural), decade of birth and number of chronic conditions based on the 17 chronic conditions in the Charlson Comorbidity Index (0, 1, 2+) were included as covariates and were ascertained at index date [43]. Income quintile is an area-level socioeconomic measure estimated using 6-digit postal codes of offspring and the average income levels, derived from Canada Census, in dissemination areas (a small area with an average population of 400 to 700 persons) [44]. Postal codes identify the smallest postal delivery unit representing a single address, such as a single building or a group of addresses, for example, one side of a city block. Income quintiles divide the population into five groups (1: lowest, 5: highest), with approximately 20% of the population in each group. Income quintile was assigned a "missing" value when postal codes could not be assigned to one of the five quintiles. Residential postal codes referring to the cities of Winnipeg or Brandon, Manitoba, were defined as urban, while postal codes outside these two urban centres were considered rural. The chronic conditions of Charlson Comorbidity Index were identified in Hospital Abstracts and Medical Claims.

Missing Paternal Linkage

Three strategies were evaluated to address missing paternal linkage: the newly developed candidate father approach, an indicator category approach, and a complete-case approach.

Candidate Father Approach

For each offspring with missing paternal linkage, a candidate father set was created from all eligible males in the Registry. Candidates shared the same 6-digit postal code as the offspring (anytime within one year of birth), had active health registration on offspring's date of birth, were within a specific age range relative to the mother (between the 5th and 95th percentiles of the parental age gap observed among those with linkage to both parents), and had no existing father-child linkage recorder in the Registry. The candidate father approach was designed to identify one or more eligible candidate fathers from the Registry for each offspring with missing paternal linkage. These candidate fathers represent men who met the predefined selection criteria and could correspond to the legally registered father at birth, who is expected to represent the biological father in most births. From the pool of all eligible candidates for a given offspring, up to three were randomly selected with replacement for use in the analysis. We limited candidate sets to three plausible fathers to maintain computational feasibility and restrict candidate

pool size, consistent with guidance that a small number of replacements (e.g., 3–5) is generally sufficient in missing data settings [36]. Selection was weighted by the probability density of the observed parental age gap distribution. Offspring without any candidate fathers were excluded.

Indicator Category Approach

For offspring with missing paternal linkage, a separate "missing" category was created for parental mental disorder and IHD histories. Therefore, paternal history was categorised as present, absent, or missing.

Complete-Case Approach

Offspring with missing paternal linkage were excluded, and analyses were restricted to offspring with observed paternal linkage.

Validation

To evaluate the accuracy of the approaches, missing paternal linkages were simulated in a historical sub-cohort where true paternal linkages were known. Offspring born before 1980 (pre-1980 cohort), with high linkage completeness, were used as a validation population [32]. Missing paternal linkages were induced in this population; the candidate father approach was then applied and evaluated. To ensure the induced missingness patterns were realistic, the probability of missingness was first modelled using the 1980-onward cohort, where missing linkages were more common than before 1980.

The simulation procedure involved the following steps. First, patterns of missing paternal linkage were identified in the recent time period (1980-onward) using a multivariable logistic regression model with missing paternal linkage as the binary outcome. Demographic characteristics (sex, region, income quintile), and maternal and offspring mental disorder and IHD diagnoses were included as predictors. Second, the multivariable logistic regression model in the previous step was applied in the pre-1980 cohort to estimate the probability of missing paternal linkage. Third, offspring in the pre-1980 cohort were ranked by their predicted probability of missing paternal linkage, and paternal linkage of those in the top 20% was artificially assigned to "missing" to approximate the empirically observed missingness level in the full cohort. Sensitivity analyses were conducted by repeating this procedure using exclusion thresholds of 10%, 30%, 40%, and 50%. Finally, for each offspring in the pre-1980 cohort with a simulated missing father, we applied the candidate father approach to identify a candidate set.

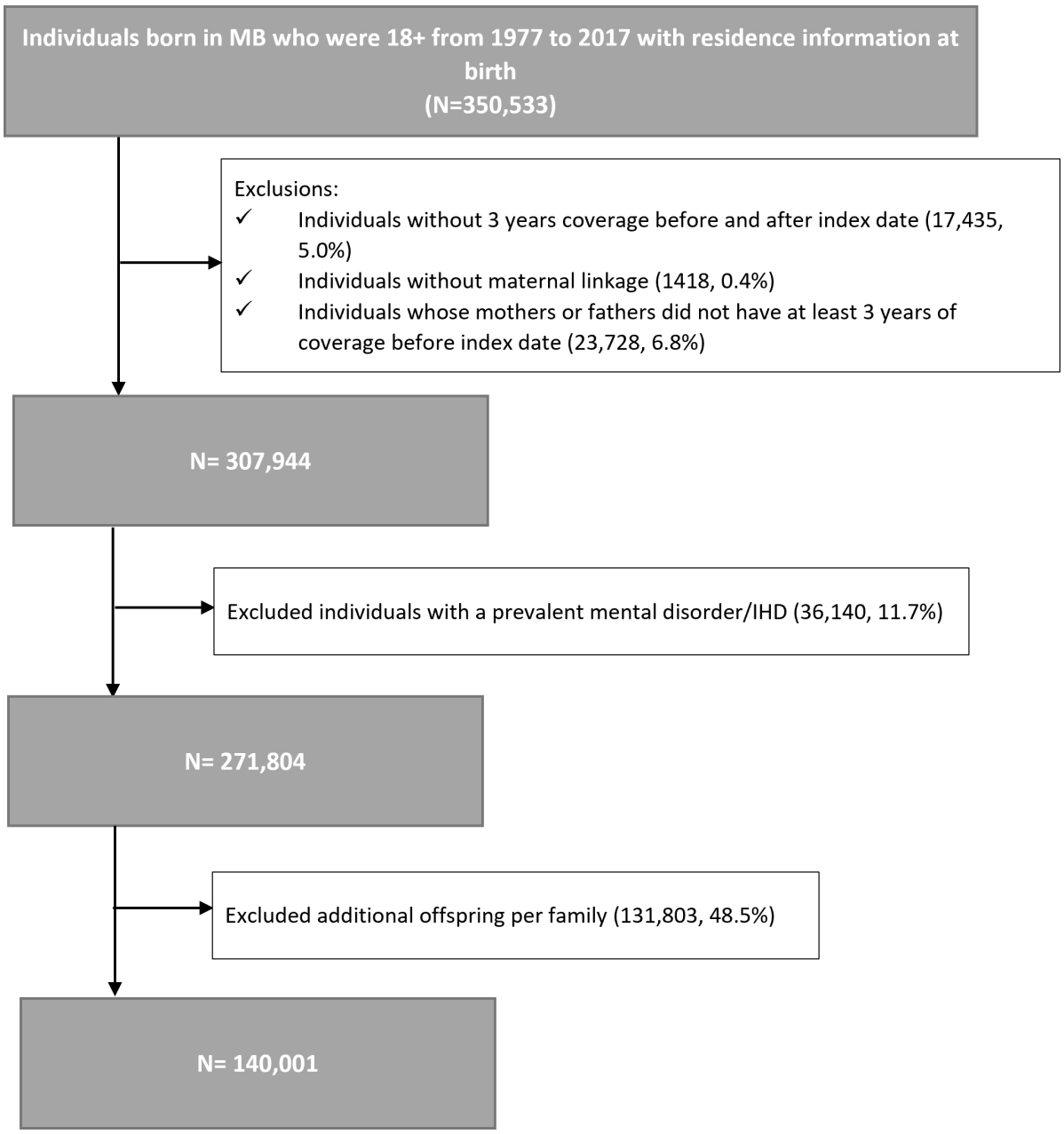
The candidate father approach was evaluated in two settings: a validation cohort and a real-world, full cohort. In the validation cohort, we used a cohort of pre-1980 births, where the true paternal linkages were known. Artificial missing paternal linkage was induced to create simulated missingness levels ranging from 10% to 50%. We quantified the percentage of candidate father sets that contained the true (originally linked) father, both before and after restricting candidate sets to a maximum of three fathers per offspring. The percentage was calculated as the number of candidate father sets that included the true father divided by the total number of candidate father sets. Then, estimates of the paternal transfer of mental disorder and IHD risks were compared across four strategies: a) true fathers, b) candidate fathers, c) an indicator category, and d) a complete-case approach (excluding those with artificial missing linkages). In the real-world setting, the candidate father approach was applied to the full study cohort, where paternal linkages were truly missing for a subset of offspring and the true father was unobserved. Plausible candidate fathers were identified for all offspring with missing paternal linkages. The risk estimates for the paternal transfer of mental disorder and IHD obtained using the candidate father approach were compared with those generated using indicator category and complete-case approaches.

Statistical Analysis

Characteristics of the full study and validation cohorts were reported using frequencies and percentages. Multivariable Cox proportional hazards regression models were used to estimate the association between paternal history and risk of mental disorders and IHD; hazard ratios (HRs) and 95% confidence intervals (CIs) were reported. The models were adjusted for sex, area-level income quintile, region of residence, number of chronic conditions and decade of birth. Within the pre-1980 validation cohort with simulated missing paternal linkages, we generated risk estimates for each outcome and across missingness levels using: 1) true fathers (gold standard), 2) candidate fathers, 3) indicator category, and 4) complete-case approach. For the candidate father approach, the modelling was repeated over three iterations (one for each candidate father); HRs and 95% CIs were pooled across iterations using Rubin's Rules [45]. A few additional analyses were conducted in the validation cohort. First, the candidate father approach was evaluated separately for iterations in which the selected candidate father matched the true father and those in which the selected candidate father did not. Second, to assess agreement between the candidate father and true father classifications, concordance tables summarising paternal disease history (mental disorders and IHD) were generated separately for each candidate father iteration in the primary validation analysis. Finally, to characterise offspring for whom the candidate father approach could not identify an eligible candidate, demographic and clinical characteristics were compared between offspring with at least one eligible candidate father and those with no eligible candidate father in the primary validation analysis (20% simulated missingness).

In the full cohort, we generated risk estimates for each outcome using the candidate father (pooled over three iterations), indicator category and complete-case approaches. In a sensitivity analysis, the primary full -cohort analyses were stratified by the offspring index date (before or after April 1, 2004), corresponding to the transition of hospital diagnosis coding from ICD-9-CM to ICD-10-CA. All analyses were performed using SAS® version 9.4 (SAS Institute, Cary, North Carolina, United States). This study was approved by the Health Research Ethics Board of the University of Manitoba, and data access was approved by the Manitoba Health Information Privacy Committee (now the Provincial Health Research Privacy Committee).

Figure 1: A data flow diagram describing the study cohort creation and exclusion criteria



Abbreviations. IHD: ischaemic heart disease. Note: The index date was defined as April 1st of the year an offspring turned 18 years of age.

Results

Description of Study and Validation Cohorts

Out of 350,533 offspring who met the inclusion criteria, 42,589 were excluded because they or their parents did not meet the healthcare coverage requirement, or they did not have maternal linkages (Figure 1). Additionally, 36,140 offspring were excluded for having a recorded mental disorder or IHD diagnosis prior to their index date, and 131,803 were excluded after selecting one random offspring per mother.

The study cohort consisted of 140,001 offspring; 52.8% were male, and 58.1% were urban residents. A total of 36,184 (25.8%) offspring were missing paternal linkages. Offspring missing paternal linkages were more likely to be born later in the study period or to a younger mother, have lower socioeconomic status, and have a maternal history of mental disorders (Table 1).

The study cohort of 140,001 offspring was split into a pre-1980 validation cohort (37,576 offspring) and a 1980-onward cohort (102,425 offspring). Table 2 compares the characteristics of the offspring whose paternal linkages were artificially assigned to missing in the pre-1980 cohort (simulating a 20% missingness level) with those who truly had missing paternal linkages in the 1980-onward cohort. Offspring with simulated missing fathers in the pre-1980 cohort were more likely to be urban residents, have higher socioeconomic status, have older mothers and have a mental disorder, compared to those with true missing paternal linkage in the 1980-onward cohort.

Table 1: Characteristics of the Full Study Cohort Stratified by Paternal Linkage Status

Characteristic	Paternal linkage		
	NoN = 36,184 (25.8%)	YesN = 103,817 (74.2%)	Total N = 140,001 (100.0%)
Sex			
Female	17,018 (47.0)	49,050 (47.2)	66,068 (47.2)
Male	19,166 (53.0)	54,767 (52.8)	73,933 (52.8)
Birth year			
1950-1969	318 (0.9)	985 (0.9)	1303 (0.9)
1970-1979	3832 (10.6)	32,441 (31.2)	36,273 (25.9)
1980-1989	9702 (26.8)	39,253 (37.8)	48,955 (35.0)
1990-1999	22,332 (61.7)	31,138 (30.0)	53,470 (38.2)
Region of residence			
Rural	14,229 (39.3)	44,378 (42.7)	58,607 (41.9)
Urban	21,955 (60.7)	59,439 (57.3)	81,394 (58.1)
Income quintile			
1 (lowest)	12,607 (34.8)	17,958 (17.3)	30,565 (21.8)
2	9005 (24.9)	25,132 (24.2)	34,137 (24.4)
3	6027 (16.7)	23,468 (22.6)	29,495 (21.1)
4	4879 (13.5)	20,791 (20.0)	25,670 (18.3)
5 (highest)	3025 (8.4)	13,466 (13.0)	16,491 (11.8)
Missing	641 (1.8)	3002 (2.9)	3643 (2.6)
Mental disorder diagnosis	13,665 (37.8)	39,507 (38.1)	53,172 (38.0)
IHD diagnosis	111 (0.3)	433 (0.4)	544 (0.4)
Number of comorbidities*			
None	23,160 (64.0)	67,437 (65.0)	90,597 (64.7)
1	11,451 (31.6)	32,663 (31.5)	44,114 (31.5)
2+	1573 (4.3)	3717 (3.6)	5290 (3.8)
Maternal mental disorder history	25,113 (69.4)	54,756 (52.7)	79,869 (57.0)
Paternal mental disorder history	NA	38,744 (37.3)	38,744 (27.7)
Maternal IHD history	473 (1.3)	1210 (1.2)	1683 (1.2)
Paternal IHD history	NA	5064 (4.9)	5064 (3.6)
Maternal age			
<20	7880 (21.8)	5190 (5.0)	13,070 (9.3)
20-24	11,921 (32.9)	23,234 (22.4)	35,155 (25.1)
25-29	8332 (23.0)	40,359 (38.9)	48,691 (34.8)
30+	8051 (22.3)	35,034 (33.7)	43,085 (30.8)
Paternal age			
<20	NA	1045 (1.0)	1045 (0.7)
20-24	NA	12,992 (12.5)	12,992 (9.3)
25-29	NA	35,295 (34.0)	35,295 (25.2)
30+	NA	54,485 (52.5)	54,485 (38.9)

NA: not available, IHD: ischaemic heart disease.
Note: reported as frequencies (percentages).
*Based on the 17 chronic conditions in the Charlson Comorbidity Index.

For the 6,685 offspring in the pre-1980 cohort with paternal linkage artificially excluded (simulating 20% missingness), the candidate father approach identified at least one plausible candidate for 5,267 (78.8%); the remaining 1,418 (21.2%) had no eligible candidates and were excluded. The number of offspring with candidate fathers identified across all simulated missingness levels (10-50%) is summarised in Supplementary Appendix 2.

The percentage of candidate sets that contained the true father ranged from 26.5% to 28.3% across simulated missingness levels when candidate sets were restricted to a maximum of three candidates. However, when the full, unrestricted pool of eligible candidates was considered (i.e., without limiting sets to a maximum of three candidates), the true father was included in 64.8% – 66.1% of candidate sets. The percentage of candidate father sets that included the true

Table 2: Characteristics of Offspring with Simulated Missing Paternal Linkages (pre-1980 Cohort, 20% Missingness) Versus True Missing Paternal Linkages (1980-Onward Cohort)

Characteristic	Simulated missing linkages (Pre-1980, 20% missingness) N = 6685	True missing linkages (1980-onward) N = 32,034
Male sex	3530 (52.8)	17,020 (53.1)
Urban residence	2957 (44.2)	12,456 (38.9)
Income quintile		
1 (lowest)	1284 (19.2)	10,917 (34.1)
2	1697 (25.4)	7922 (24.7)
3	1538 (23.0)	5417 (16.9)
4	1160 (17.4)	4462 (13.9)
5 (highest)	777 (11.6)	2823 (8.8)
Missing	229 (3.4)	493 (1.5)
Mental disorder diagnosis	3119 (46.7)	11,291 (35.2)
IHD diagnosis	49 (0.7)	43 (0.1)
Number of comorbidities*		
None	4509 (67.4)	20,376 (63.6)
1	1965 (29.4)	10,271 (32.1)
2+	211 (3.2)	1387 (4.3)
Maternal mental disorder history	3119 (46.7)	22583 (70.5)
Maternal IHD history	112 (1.7)	370 (1.2)
Maternal age		
<20	525 (7.9)	6897 (21.5)
20-24	1773 (26.5)	10,410 (32.5)
25-29	2566 (38.4)	7443 (23.2)
30+	1821 (27.2)	7284 (22.7)

Note: reported as frequencies (percentages).
*Based on the 17 chronic conditions in the Charlson Comorbidity Index.

father across missingness levels is presented in Supplementary Appendix 3.

Association of Paternal Histories with Mental Disorder and IHD Risks

Figure 2 and Supplementary Appendix 4 summarise the association of paternal histories with mental disorder and IHD risks in the pre-1980 cohort. For mental disorders, the HR associated with a paternal history was 1.27 (95% CI: 1.23–1.31) when analyses were based on true fathers. Estimates from the candidate father ranged from 1.24 (95% CI 1.20–1.29) at 10% missingness to HR 1.17 (95% CI 1.13–1.22) at 50% missingness. The indicator category and complete-case approaches produced risk estimates similar in magnitude to the true father estimate across missingness levels (Figure 2; Supplementary Appendix 4). For IHD, the risk associated with a parental history was HR 1.94 (95% CI 1.35–2.79) using true fathers. Candidate father approach estimates were broadly similar at lower missingness levels but showed smaller HRs with wider confidence intervals as missingness increased. Indicator category and complete-case approaches showed increasing deviation away from the null with higher missingness, with wider confidence intervals (Figure 2; Supplementary Appendix 4).

The validation analyses were stratified according to whether the selected candidate father matched the true father. For mental disorders, the association was similar to that observed using the true fathers when the selected candidate was the true father (HR 1.26, 95% CI 1.06–1.51), whereas no association was observed when the selected candidate was not the true father (HR 1.01, 95% CI 0.92–1.11). Corresponding risks for IHD could not be reliably estimated because of the small number of events after stratification (Supplementary Appendix 5). Concordance analyses showed that concordant negative classifications were the most frequent for both mental disorders (45.1%–50.2%) and IHD (89.2%–90.1%), while concordant positive and discordant classifications accounted for the remaining observations (Supplementary Appendix 6). The characteristics of offspring with and without an eligible candidate father were generally similar across demographic and clinical characteristics (Supplementary Appendix 7).

In the full cohort, where 25.8% of offspring lacked paternal linkage, the association of paternal history with mental disorder and IHD risk was consistent across the candidate father, indicator category and complete-case approaches, though estimates obtained using the candidate father approach were modestly attenuated toward the null compared to the other two approaches (Table 3). Stratified

Figure 2: Hazard Ratios (HRs) and 95% Confidence Intervals (CIs) for the Association of Paternal History of Mental Disorders or IHD with their Risks in Pffspring using True Ffathers, Candidate Fathers, Indicator Category or Complete-Case Approaches in the Pre-1980 Validation Cohort

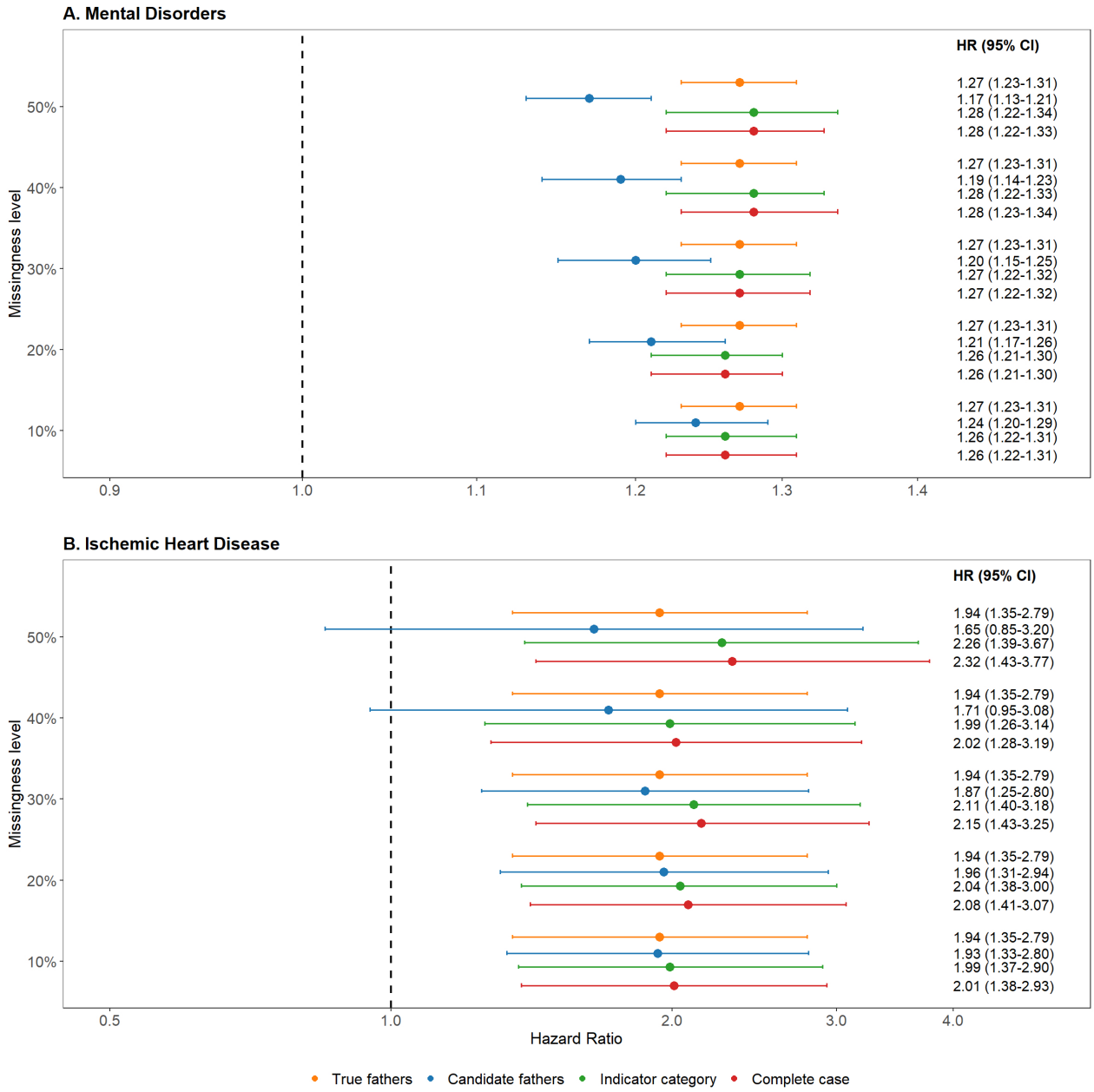


Table 3: Hazard Ratios and 95% Confidence Intervals for the Association of Paternal History of Mental Disorders or IHD with their Risks in Offspring Using Candidate Father, Indicator Category, and Complete-Case Approaches in the Full Cohort

Outcome	HR* (95% CI)		
	Candidate father N = 140,001	Indicator category N = 140,001	Complete-case N = 103,817
Mental disorders	1.21 (1.18-1.24)	1.26 (1.23-1.28)	1.26 (1.23-1.29)
IHD	1.71 (1.23-2.39)	1.83 (1.33-2.51)	1.82 (1.33-2.50)

analyses by ICD coding period produced risk estimates for mental disorders that were similar to the primary analyses. For IHD, the estimates were generally consistent but less precise during the ICD-10-CA period because of the small number of events (Supplementary Appendix 8).

Discussion

This large population-based cohort study developed a novel candidate father approach to address missing paternal linkages in administrative health data and applied it to investigate the

paternal transfer of mental disorders and IHD. Although the candidate father approach had limited accuracy in identifying true fathers, it produced risk estimates that were close to the true estimates, particularly at lower missingness levels. Across outcomes, estimates produced using the candidate father approach showed progressive attenuation toward the null as the missingness level increased. In contrast, indicator category and complete-case approaches produced estimates closely aligned with the true estimate for mental disorders but increasingly deviated away from the null for IHD at higher missingness levels, with wider confidence intervals reflecting increased imprecision. In the full cohort, all three approaches resulted in comparable inferences about the paternal transfer of mental disorders and IHD, despite a modest attenuation of effects using the candidate father approach.

The attenuation in risk estimates using the candidate father approach is consistent with exposure misclassification arising from the use of plausible rather than true fathers. As missingness levels increased, the cohort included a greater proportion of plausible fathers, which may have diluted the measured association between paternal history and offspring outcomes, resulting in smaller effect estimates. This pattern was observed for both mental disorders and IHD outcomes, but the attenuation was stronger for IHD, likely reflecting a lower outcome prevalence and greater imprecision. In contrast, indicator category and complete-case approaches rely on observed linkages and therefore do not introduce exposure misclassification as with the candidate father approach, but they may increase statistical uncertainty and, in the case of complete-case approach, reduce precision due to smaller sample sizes. For mental disorders, these two approaches produced estimates closely aligned with the true estimates, likely due to the high prevalence of the outcome. However, for IHD, a rarer outcome, these approaches showed greater deviation from the null with increased missingness, suggesting their sensitivity to selection bias. In the full cohort with true paternal linkage missingness, all the approaches led to comparable risk estimates, demonstrating the practical utility of the candidate father approach. The consistent attenuation of effects with this approach highlights the trade-off between reducing selection bias by retaining offspring with missing paternal linkage and introducing exposure misclassification. By assigning plausible fathers rather than excluding offspring or classifying paternal history as missing, the candidate father approach preserves cohort size while providing a structured strategy for approximating paternal exposure under incomplete paternal linkage.

Although the unrestricted candidate pool contained the true father for approximately two-thirds of offspring, a substantial proportion of true fathers were not identified because they did not meet the candidate selection criteria. This finding reflects both deliberate design choices to prioritise plausible candidates and maintain computational feasibility, and inherent limitations of the administrative data. For example, some true fathers may not have shared the same postal code around the time of birth or may not have met the age-based eligibility criteria. Future refinements could incorporate additional information on familial relationships, where available, to improve candidate identification. However, the performance of any candidate father approach will ultimately remain constrained by the completeness and accuracy of familial information captured in administrative data. An additional validation analysis showed that the association between paternal and offspring mental disorders was preserved only when the selected candidate father matched the true father, indicating that the observed associations were driven by correct identification of the paternal linkage.

Missing paternal linkages in administrative health data primarily reflect administrative and social processes, rather than true absence of a biological father. In Manitoba, paternal linkage depends on the reporting of marital or common-law relationships for the fathers to share the same FRNs as the mother and children; such reporting is not mandatory. As a result, paternal linkages may be missing in cases of relationship instability, younger maternal age, or failure to report parental relationships, and are influenced by changes in family structure over time [29]. While other scenarios, such as same-sex parenting and assisted reproduction, may contribute to missing paternal information, these are expected to represent a minority of cases and cannot explain the magnitude of missingness observed in this study cohort. Although missing paternal linkage was more common among offspring of younger mothers and those from lower socioeconomic backgrounds, the candidate father approach was designed to improve the handling of missing paternal linkage rather than to characterise specific family structures. Because our study did not directly evaluate performance according to family structure, we cannot determine whether the approach performs differently in more or less stable family contexts. However, it identified at least one plausible candidate father for nearly 80% of offspring with missing paternal linkage, substantially reducing the number of offspring for whom paternal information could not be approximated.

Consistent with previous work, paternal linkage was systematically associated with sociodemographic and familial characteristics, indicating that missingness does not satisfy the assumption of missing completely at random [31]. In the validation analyses with simulated missing paternal linkage, approaches for handling missing paternal linkage performed differently based on the outcome characteristics and missingness level, with indicator category and complete-case approaches showing deviations from the true father estimate for IHD but not for mental disorders. In contrast, in the full cohort with true missing paternal linkage, all approaches resulted in comparable risk estimates for both outcomes, with the candidate father approach producing more conservative effect estimates. These findings highlight the importance of using multiple analytic approaches to assess robustness in familial studies using administrative data with incomplete parental linkage. In addition to analytic approaches for handling missing paternal linkage, there is also potential to improve paternal linkage by integrating administrative health data with complementary data sources. These may include social service or income assistance records that capture familial relationships, prospective cohorts with family information, or surveys that collect self-reported family structure information. Such data sources could enrich administrative datasets by providing additional information on paternal relationships that are not routinely captured. However, these sources are typically available only for selected subpopulations and lack the population-based coverage of administrative

health data, potentially introducing selection bias. As a result, analytic approaches that can accommodate missing paternal information within population-based administrative data remain essential for familial health research.

A major strength of this study is the large, population-based cohort representative of the Manitoba population, enabled by the universal coverage of the healthcare system. Additional strengths include the use of objective administrative data case definitions of mental disorders and IHD in offspring and their parents, as well as the application of a novel validation design using a historical cohort with high paternal linkage completeness to compare multiple approaches against a true father benchmark. However, there are a few limitations to note. First, the approaches were compared in the context of the paternal transfer of mental disorders and IHD; findings may not generalise to other outcomes, exposure definitions or study designs. Second, although a reference standard was available in the validation analyses, no such benchmark existed in the full cohort with true missing paternal linkage, preventing definitive conclusions about which approach produces the least biased risk estimates in applied settings. Additionally, our validation analyses used the father identified through the FRN recorded at birth as the reference standard, which is expected to represent the biological father in most families, as adoptions at birth and other alternative family arrangements are uncommon. However, biological paternity cannot be independently verified using administrative data. The father–offspring associations evaluated in this study should be interpreted in the context of the Registry-defined paternal linkage, which likely reflects a combination of shared genetic susceptibility and familial environmental and social influences rather than biological inheritance alone. The candidate father approach is intended to recover the paternal relationship represented in the administrative data. Finally, the mechanisms and patterns of paternal linkage missingness in the Manitoba Repository may differ from those in other jurisdictions, potentially limiting generalizability of our candidate father approach to settings with different administrative or social contexts. The performance of the approach is likely influenced by the availability and quality of family identifiers, residential address information, and the administrative processes used to establish parent–child linkages. In addition, social and demographic factors, including family structure, partnership reporting practices, migration, and cultural differences in how familial relationships are recorded, may affect both the completeness of paternal linkage and the suitability of the candidate selection criteria. Replication in other jurisdictions with differing administrative practices would help establish the generalizability of these findings.

Conclusion

This newly developed approach used residence and demographic information to identify plausible fathers for offspring with missing paternal linkage. Although this approach identified only a subset of true fathers, it produced estimates of the association between paternal history and offspring mental disorder and IHD risks that were close to those obtained using the true fathers, particularly at lower missingness levels. These findings indicate that the candidate father approach can support population-based familial research using administrative data where incomplete paternal linkage limits the inclusion of paternal information. Across approaches, performance varied by outcome and level of missingness, and no single approach was uniformly superior, highlighting the value of applying multiple strategies to assess the robustness in studies with incomplete parental linkage.

Acknowledgements

We acknowledge the Manitoba Centre for Health Policy for the use of data contained in the Population Health Research Data Repository (HIPC/PHRPC #: 2019/2020–52; MCHP Project #: 2020-005). The results and conclusions are those of authors and no official endorsement by the Manitoba Centre for Health Policy, Manitoba Health, Seniors and Long-Term Care, or other data providers is intended or should be inferred. AFH was supported by a Canadian Institutes of Health Research (CIHR) Fellowship during the study period. AFH is now supported by a Tier 2 Canada Research Chair (CRC-2023-00200). LML is supported by a Tier 1 Canada Research Chair (CRC-2023-00349).

Conflicts of Interest

All authors declare that they have no conflict of interest.

Ethics Approval

The study was approved by the Health Research Ethics Board of the University of Manitoba; data access approval was provided by the Manitoba Health Information Privacy Committee (now the Provincial Health Research Privacy Committee).

Data Availability

Data used in this article were derived from administrative health data as a secondary use. The data were provided under specific data sharing agreements only for approved use at the Manitoba Centre for Health Policy (MCHP). The original source data are not owned by the researchers or MCHP and as such, cannot be provided to a public repository. The original data source and approval for use have been noted in the acknowledgements of the article. Where necessary, source data specific to this article or project may be reviewed at MCHP with the consent of the original data providers, along with the required privacy and ethical review bodies.

AI Disclosure Statement

The authors used ChatGPT (OpenAI) to a limited extent to help simplify the presentation of complex ideas and improve the readability of selected paragraphs. All suggestions were reviewed and revised by the authors, who take full responsibility for the final manuscript content.

References

1. Hamad AF, Monchka BA, Bolton JM, Plana-Ripoll O, Roos LL, Elgendi M, et al. The Intergenerational Transfer of Mental Disorders: A Population-Based Multigenerational Linkage Study: Le transfert intergénérationnel des troubles mentaux : une étude sur les liens multigénérationnels basée sur la population. Can J Psychiatry Rev Can Psychiatr 2024:7067437241255096. https://doi.org/10.1177/07067437241255096

2. Yang S, Leslie WD, Yan L, Walld R, Roos LL, Morin SN, et al. Objectively Verified Parental Hip Fracture Is an Independent Risk Factor for Fracture: a Linkage Analysis of 478,792 Parents and 261,705 Offspring. J Bone Miner Res 2016;31:1753–9. https://doi.org/10.1002/jbmr.2849

3. Bolton JM, Au W, Leslie WD, Martens PJ, Enns MW, Roos LL, et al. Parents Bereaved by Offspring Suicide: A Population-Based Longitudinal Case-Control Study. JAMA Psychiatry 2013;70:158–67. https://doi.org/10.1001/jamapsychiatry.2013.275

4. Hamad AF, Alessi-Severini S, Mahmud S, Brownell M, Kuo IF. Prenatal antibiotic exposure and risk of attention-deficit/hyperactivity disorder: a population-based cohort study. Can Med Assoc J 2020;192:E527–35. https://doi.org/10.1503/cmaj.190883

5. Martens PJ, Shafer LA, Dean HJ, Sellers EAC, Yamamoto J, Ludwig S, et al. Breastfeeding Initiation Associated With Reduced Incidence of Diabetes in Mothers and Offspring. Obstet Gynecol 2016;128:1095–104. https://doi.org/10.1097/AOG.0000000000001689

6. Hamad AF, Yan L, Jafari Jozani M, Hu P, Delaney JA, Lix LM. Developing a prediction model of children asthma risk using population-based family history health records. Pediatr Allergy Immunol 2023;34:e14032. https://doi.org/10.1111/pai.14032

7. Wall-Wieler E, Roos LL, Nickel NC, Chateau D, Brownell M. Mortality Among Mothers Whose Children Were Taken Into Care by Child Protection Services: A Discordant Sibling Analysis. Am J Epidemiol 2018;187:1182–8. https://doi.org/10.1093/aje/kwy062

8. Rom AL, Wu CS, Olsen J, Jawaheer D, Hetland ML, Ottesen B, et al. Parental Rheumatoid Arthritis and Long-term Child Morbidity. Ann Rheum Dis 2016;75:1831–7. https://doi.org/10.1136/annrheumdis-2015-208072

9. Solberg BS, Hegvik T-A, Halmøy A, Skjærven R, Engeland A, Haavik J, et al. Sex differences in parent–offspring recurrence of attention-deficit/hyperactivity disorder. J Child Psychol Psychiatry 2021;62:1010–8. https://doi.org/10.1111/jcpp.13368

10. Dean K, Stevens H, Mortensen PB, Murray RM, Walsh E, Pedersen CB. Full spectrum of psychiatric outcomes among offspring with parental history of mental disorder. Arch Gen Psychiatry 2010;67:822–9. https://doi.org/10.1001/archgenpsychiatry.2010.86

11. Hamad AF, Yan L, Leslie WD, Morin SN, Walld R, Roos LL, et al. Association Between Parental Type 1 and Type 2 Diabetes Diagnosis and Major Osteoporotic Fracture Risk in Adult Offspring: A Population-Based Cohort Study. Can J Diabetes 2022;46:3-9.e3. https://doi.org/10.1016/j.jcjd.2021.04.005

12. Latvala A, Kuja-Halkola R, D'Onofrio BM, Jayaram-Lindström N, Larsson H, Lichtenstein P. Association of parental substance misuse with offspring substance misuse and criminality: a genetically informed register-based study. Psychol Med undefined/ed:1–10. https://doi.org/10.1017/S0033291720002135

13. Bolton JM, Au W, Walld R, Chateau D, Martens PJ, Leslie WD, et al. Parental Bereavement After the Death of an Offspring in a Motor Vehicle Collision: A Population-based Study. Am J Epidemiol 2014;179:177–85. https://doi.org/10.1093/aje/kwt247

14. Hamad AF, Monchka BA, Bolton JM, Roos LL, Elgendi M, Lix LM. Leveraging multigenerational health data to enhance mental disorder risk prediction: a population-based cohort study. BMC Psychiatry 2025;25:862. https://doi.org/10.1186/s12888-025-07323-z

15. Hamm NC, Hamad AF, Wall-Wieler E, Roos LL, Plana-Ripoll O, Lix LM. Multigenerational health research using population-based linked databases: an international review. Int J Popul Data Sci 2021;6:1686. https://doi.org/10.23889/ijpds.v6i1.1686

16. Lut I, Harron K, Hardelid P, O'Brien M, Woodman J. 'What about the dads?' Linking fathers and children in administrative data: A systematic scoping review. Big Data Soc 2022;9:20539517211069299. https://doi.org/10.1177/20539517211069299

17. Atladóttir HÓ, Thorsen P, Østergaard L, Schendel DE, Lemcke S, Abdallah M, et al. Maternal Infection Requiring Hospitalization During Pregnancy and Autism Spectrum Disorders. J Autism Dev Disord 2010;40:1423–30. https://doi.org/10.1007/s10803-010-1006-y

18. Baker R, Kendrick D, Tata LJ, Orton E. Association between maternal depression and anxiety episodes and rates of childhood injuries: a cohort study from England. Inj Prev 2017;23:396–402. https://doi.org/10.1136/injuryprev-2016-042294

19. Burstyn I, Sithole F, Zwaigenbaum L. Autism spectrum disorders, maternal characteristics and obstetric complications among singletons born in Alberta, Canada. Chronic Dis Can 2010;30:125–34. https://doi.org/10.24095/hpcdp.30.4.04

20. Canova C, Zabeo V, Pitter G, Romor P, Baldovin T, Zanotti R, et al. Association of maternal education, early infections, and antibiotic use with celiac disease: a population-based birth cohort study in northeastern Italy. Am J Epidemiol 2014;180:76–85. https://doi.org/10.1093/aje/kwu101

21. Chang Z, Lichtenstein P, D'Onofrio BM, Almqvist C, Kuja-Halkola R, Sjölander A, et al. Maternal age at childbirth and risk for ADHD in offspring: a population-based cohort study. Int J Epidemiol 2014;43:1815–24. https://doi.org/10.1093/ije/dyu204

22. He H, Yu Y, Liew Z, Gissler M, László KD, Valdimarsdóttir UA, et al. Association of Maternal Autoimmune Diseases With Risk of Mental Disorders in Offspring in Denmark. JAMA Netw Open 2022;5:e227503. https://doi.org/10.1001/jamanetworkopen.2022.7503

23. Li M, Fallin MD, Riley A, Landa R, Walker SO, Silverstein M, et al. The Association of Maternal Obesity and Diabetes With Autism and Other Developmental Disabilities. Pediatrics 2016;137:e20152206. https://doi.org/10.1542/peds.2015-2206

24. Ray GT, Croen LA, Habel LA. Mothers of Children Diagnosed with Attention-Deficit/Hyperactivity Disorder: Health Conditions and Medical Care Utilization in Periods Before and After Birth of the Child. Med Care 2009;47:105–14. https://doi.org/10.1097/MLR.0b013e31817e18c0

25. Kong L, Nilsson IAK, Brismar K, Gissler M, Lavebratt C. Associations of Different Types of Maternal Diabetes and Body Mass Index With Offspring Psychiatric Disorders. JAMA Netw Open 2020;3:e1920787–e1920787. https://doi.org/10.1001/jamanetworkopen.2019.20787

26. Hamad AF, Roos LL, Bolton JM, Wall-Wieler E. Familial associations in adolescent substance use disorder: a population-based cohort study. Addiction 2022;117:2720–9. https://doi.org/10.1111/add.15981

27. McAlister FA, Yan L, Roos LL, Lix LM. Parental Atrial Fibrillation and Stroke or Atrial Fibrillation in Young Adults. Stroke 2019;50:2322–8. https://doi.org/10.1161/STROKEAHA.119.025124

28. Dreyer K, Williamson RAP, Hargreaves DS, Rosen R, Deeny SR. Associations between parental mental health and other family factors and healthcare utilisation among children and young people: a retrospective, cross-sectional study of linked healthcare data. BMJ Paediatr Open 2018;2:e000266. https://doi.org/10.1136/bmjpo-2018-000266

29. Hamad AF, Walld R, Lix LM, Urquia ML, Roos LL, Wall-Wieler E. Int J Epidemiol 2022;51:e65–72. https://doi.org/10.1093/ije/dyab195

30. Logan JE, Riley AW, Barker LE. Parental mental and pain-related health and pediatric ambulatory care sensitive emergency department visits and hospitalizations. Health Serv Res 2008;43:656–74. https://doi.org/10.1111/j.1475-6773.2007.00790.x

31. Sims S, O'Donnell M. Identifying families most likely to have missing paternal details in birth registrations using linked data. J Fam Stud 2015;21:163–77. https://doi.org/10.1080/13229400.2015.1020983

32. Hamad AF, Walld R, Lix LM, Urquia ML, Roos LL, Wall-Wieler E. Data Resource Profile: The Manitoba Multigenerational Cohort. Int J Epidemiol 2022;51:e65–72. https://doi.org/10.1093/ije/dyab195

33. Alio AP, Salihu HM, McIntosh C, August EM, Weldeselasse H, Sanchez E, et al. The effect of paternal age on fetal birth outcomes. Am J Mens Health 2012;6:427–35. https://doi.org/10.1177/1557988312440718

34. Durkin MS, Maenner MJ, Newschaffer CJ, Lee L-C, Cunniff CM, Daniels JL, et al. Advanced parental age and the risk of autism spectrum disorder. Am J Epidemiol 2008;168:1268–76. https://doi.org/10.1093/aje/kwn250

35. Acock AC. Working With Missing Values. J Marriage Fam 2005;67:1012–28. https://doi.org/10.1111/j.1741-3737.2005.00191.x

36. Buuren S van. Multiple imputation. In: Flexible Imputation of Missing Data. 2nd ed., Chapman and Hall/CRC; 2018, p. 29–62.

37. Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. Br J Cancer 2004;91:4–8. https://doi.org/10.1038/sj.bjc.6601907

38. Vach W, Blettner M. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. Am J Epidemiol 1991;134:895–907. https://doi.org/10.1093/oxfordjournals.aje.a116164

39. Knol MJ, Janssen KJM, Donders ART, Egberts ACG, Heerdink ER, Grobbee DE, et al. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. J Clin Epidemiol 2010;63:728–36. https://doi.org/10.1016/j.jclinepi.2009.08.028

40. Groenwold RHH, White IR, Donders ART, Carpenter JR, Altman DG, Moons KGM. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. CMAJ 2012;184:1265–9. https://doi.org/10.1503/cmaj.110977

41. Chartier M, Bolton J, Mota N, MacWilliam L, Ekuma O, Nie Y, et al. Mental illness among adult Manitobans. Winnipeg, MB: Manitoba Centre for Health Policy; 2018.

42. Lix LM, Ayles J, Bartholomew S, Cooke CA, Ellison J, Emond V, et al. The Canadian Chronic Disease Surveillance System: A model for collaborative surveillance. Int J Popul Data Sci n.d.;3:433. https://doi.org/10.23889/ijpds.v3i3.433

43. Glasheen W.P., Cordier T., Gumpina R., Haugh G., Davis J., Renda A. Charlson comorbidity index: ICD-9 update and ICD-10 translation. Am Health Drug Benefits 2019;12:188–97.

44. Concept: Income Quintiles - Child Health Income Quintiles | MCHP Concept Dictionary and Glossary for Population-Based Research | Max Rady College of Medicine | University of Manitoba n.d. http://mchp-appserv.cpe.umanitoba.ca/viewConcept.php?conceptID=1161 (accessed October 13, 2022).

45. Rubin DB. Inference and missing data. Biometrika 1976;63:581–92. https://doi.org/10.1093/biomet/63.3.581

Abbreviations

IHD:	Ischaemic Heart Disease
FRNs:	Family Registration Numbers
ICD:	International Classification of Diseases
ICD-8:	The adapted eighth revision of the International Classification of Diseases
ICD-9-CM:	The ninth revision of the International Classification of Diseases with clinical modifications
ICD-10-CA:	The tenth revision of the International Classification of Diseases with Canadian adaptations
HR:	Hazard ratio
CI:	Confidence interval

Supplementary Appendix 1: Case definitions and the International Classification of Diseases diagnosis codes used to ascertain the study outcomes

Outcome	Case definition	ICD codes
Mental disorders	Fulfilling the case definition of either mood and anxiety, psychotic or substance use disorders as described below	
Mood and anxiety disorders	At least one hospitalisation or two physician visits within three years	ICD-8: 296, 300, 307 ICD-9-CM: 296, 311, 309, 300 ICD-10-CA: F30-F34, F38, F40, F41, F41.1, F41.2, F41.3, F41.8, F41.9, F42, F43, F53
Psychotic disorders	At least one hospitalisation or one physician visit	ICD-8: 295, 297, 298, 299 ICD-9-CM: 295, 297, 298 ICD-10-CA: F11.5, F12.5, F13.5, F14.5, F15.5, F16.5, F17.5, F18.5, F19.5, F20-F25, F28, F29
Substance use disorders	At least one hospitalisation or one physician visit	ICD-8: 291, 294, 303, 304 ICD-9-CM: 291, 292, 303-305 ICD-10-CA: F10-19, F55, Z50.2, Z50.3
Ischaemic heart disease	At least one hospitalisation or two physician visits within one year	ICD-8: 410-414 ICD-9-CM: 410-414 ICD-10-CA: I20-I25

ICD: International Classification of Diseases, ICD-8: the eighth revision of ICD, ICD-9-CM: the ninth revision of ICD with clinical modifications, ICD-10-CA: the tenth revision of ICD with Canadian adaptations.

Supplementary Appendix 2: The number and percentage of offspring with and without any candidate fathers in the pre-1980 cohort across missingness levels

Simulated missingness level (%)	10	20	30	40	50
With at least 1 candidate father, N (%)	2638 (78.9)	5267 (78.8)	7773 (77.5)	10,420 (77.9)	13,070 (78.2)
Without any candidate fathers, N (%)	704 (21.1)	1418 (21.2)	2254 (22.5)	2950 (22.1)	3643 (21.8)

Supplementary Appendix 3: The percentage of candidate father sets that included the true father in the pre-1980 cohort, before and after restricting sets to three fathers

Simulated missingness level (%)	10	20	30	40	50
Candidate sets limited to 3 fathers, %	26.5	27.9	28.2	28.3	28.2
Unrestricted candidate sets, %	64.8	65.8	65.7	66.1	66.0

Supplementary Appendix 4: Hazard ratios (HRs) and 95% confidence intervals (CIs) for the association of paternal history of mental disorders or ischaemic heart disease (IHD) with their risks in offspring using true fathers, candidate fathers, indicator category and complete-case approaches in the pre-1980 cohort

Missingness level (%)	HR (95% CI)			
	True fathers N = 33,426	Candidate father approach*	Indicator category (N = 33,426)	Complete-case**
Mental disorders				
10	1.27 (1.23-1.31)	1.24 (1.20-1.29) N = 32,722	1.26 (1.22-1.31)	1.26 (1.22-1.31) N = 30,084
20		1.21 (1.17-1.26) N = 32,008	1.26 (1.21-1.30)	1.26 (1.21-1.30) N = 26,741
30		1.20 (1.15-1.25) N = 31.172	1.27 (1.22-1.32)	1.27 (1.22-1.32) N = 23,399
40		1.19 (1.14-1.23) N = 30,476	1.28 (1.22-1.33)	1.28 (1.23-1.34) N = 20,056
50		1.17 (1.13-1.21) N = 29,783	1.28 (1.22-1.34)	1.28 (1.22-1.33) N = 16,713
IHD				
10	1.94 (1.35-2.79)	1.93 (1.33-2.80) N = 32,722	1.99 (1.37-2.90)	2.01 (1.38-2.93) N = 30,084
20		1.96 (1.31-2.94) N = 32,008	2.04 (1.38-3.00)	2.08 (1.41-3.07) N = 26,741
30		1.87 (1.25-2.80) N = 31.172	2.11 (1.40-3.18)	2.15 (1.43-3.25) N = 23,399
40		1.71 (0.95-3.08) N = 30,476	1.99 (1.26-3.14)	2.02 (1.28-3.19) N = 20,056
50		1.65 (0.85-3.20) N = 29,783	2.26 (1.39-3.67)	2.32 (1.43-3.77) N = 16,713

*Excluded those without any candidate fathers, which varies at each missingness level.
**Excluded those with artificially excluded fathers (simulated missingness).

Supplementary Appendix 5: Hazard ratios (HRs) and 95% confidence intervals (CIs) for the association of paternal history with offspring mental disorders and ischaemic heart disease (IHD) in the pre-1980 validation cohort, stratified by whether the selected candidate father matched the true father (20% simulated missingness)

Outcome	HR (95% CI)*			
	True fathers N = 33,426	All candidate fathers	Candidate father was the true father	Candidate father was not the true father
Mental disorders	1.27 (1.23–1.32)	1.22 (1.18–1.26)	1.26 (1.06–1.51)	1.01 (0.92–1.11)
IHD	1.94 (1.35–2.79)	1.83 (1.25–2.69)	NA**	NA**

*Birth decade was excluded from the models because there was no variation in its distribution among offspring whose selected candidate father matched or did not match the true father.
**Analyses could not be run because of the small number of events.

Supplementary Appendix 6: Concordance of paternal history classification with the true father among offspring with simulated missing paternal linkage (20% simulated missingness) in the pre-1980 validation cohort for mental disorders and ischaemic heart disease (IHD)

Paternal History	Candidate Father Iteration	True Yes/ Candidate Yes	True Yes/ Candidate No	True No/ Candidate Yes	True No/ Candidate No
Mental disorders	1	860 (16.3%)	850 (16.1%)	913 (17.3%)	2644 (50.2%)
	2	603 (12.7%)	928 (19.5%)	1022 (21.5%)	2210 (46.4%)
	3	487 (11.6%)	873 (20.8%)	948 (22.6%)	1895 (45.1%)
IHD	1	82 (1.6%)	227 (4.3%)	214 (4.1%)	4744 (90.1%)
	2	37 (0.8%)	231 (4.9%)	222 (4.7%)	4273 (89.7%)
	3	18 (0.4%)	216 (5.1%)	219 (5.2%)	3750 (89.2%)

Note: reported as frequencies (percentages).

Supplementary Appendix 7: Characteristics of offspring with simulated missing paternal linkage according to whether an eligible candidate father was identified in the pre-1980 validation cohort (20% simulated missingness)

Characteristic	Candidate father identified (5267, 16.5%)	No eligible candidate father (26741, 83.5%)
Male sex	2760 (52.4%)	13886 (51.9%)
Urban residence	2754 (52.3%)	14914 (55.8%)
Income quintile		
1 (lowest)	1016 (19.3%)	4925 (18.4%)
2	1349 (25.6%)	6863 (25.7%)
3	1193 (22.7%)	6150 (23.0%)
4	922 (17.5%)	4890 (18.3%)
5 (highest)	595 (11.3%)	3083 (11.5%)
Missing	192 (3.7%)	830 (3.1%)
Mental disorder diagnosis	2440 (46.3%)	12378 (46.3%)
IHD diagnosis	33 (0.6%)	252 (0.9%)
Number of comorbidities*		
None	3472 (65.9%)	17980 (67.2%)
1	1616 (30.7%)	7807 (29.2%)
2+	179 (3.4%)	954 (3.6%)
Maternal mental disorder history	2505 (47.6%)	12481 (46.7%)
Maternal IHD history	83 (1.6%)	422 (1.6%)
Maternal age		
<20	422 (8.0%)	1922 (7.2%)
20-24	1404 (26.7%)	7338 (27.4%)
25-29	2035 (38.6%)	10013 (37.4%)
30+	1406 (26.7%)	7468 (27.9%)

Abbreviations: IHD: ischaemic heart disease.
Note: reported as frequencies (percentages).
*Based on the 17 chronic conditions in the Charlson Comorbidity Index.

Supplementary Appendix 8: Hazard ratios (HRs) and 95% confidence intervals (CIs) for the association between paternal history of mental disorders and ischaemic heart disease (IHD) and offspring outcomes in the full cohort, stratified by ICD coding period

Outcome/ICD coding period	HR* (95% CI)		
	Candidate father	Indicator category	Complete-case
Mental disorders	1.21 (1.19-1.24)	1.26 (1.23-1.28)	1.37 (1.34-1.40)
Index date <April 1, 2004	1.22 (1.19-1.25)	1.25 (1.21-1.28)	1.39 (1.35-1.44)
Index date >=April 1, 2004	1.20 (1.15-1.25)	1.29 (1.24-1.33)	1.35 (1.31-1.40)
IHD	1.71 (1.23-2.39)	1.83 (1.33-2.51)	1.35 (1.09-1.68)
Index date <April 1, 2004	1.76 (1.25-2.49)	1.82 (1.31-2.53)	1.31 (1.02-1.67)
Index date >=April 1, 2004	1.42 (0.49-4.16)	2.06 (0.74-5.76)	1.52 (0.94-2.47)

*Birth decade was excluded from the models because there was no variation in its distribution among offspring with index date before or after April 1, 2004.