International Journal of Population Data Science

Journal Website: www.ijpds.org

Describing a complex primary health care population to support future decision support initiatives

Jacqueline K. Kueper[1,2,*], Jennifer Rayner[3,4], Merrick Zwarenstein[1,3], and Daniel J. Lizotte[1,2]

Submission History

Submitted:	22/04/2022
Accepted:	22/08/2022
Published:	24/100/2022

[1]Department of Epidemiology and Biostatistics, Schulich School of Medicine & Dentistry, Western University, London, Ontario, Canada
[2]Department of Computer Science, Faculty of Science, Western University, London, Ontario, Canada
[3]Department of Family Medicine, Schulich School of Medicine & Dentistry, Western University, London, Ontario, Canada
[4]Alliance for Healthier Communities, Toronto, Ontario, Canada

Abstract

Introduction
Developing decision support tools using data from a health care organization, to support care within that organization, is a promising paradigm to improve care delivery and population health. Descriptive epidemiology may be a valuable supplement to stakeholder input towards selection of potential initiatives and to inform methodological decisions throughout tool development. We additionally propose that to properly characterize complex populations in large-scale descriptive studies, both simple statistical and machine learning techniques can be useful.

Objective
To describe sociodemographic, clinical, and health care use characteristics of primary care clients served by the Alliance for Healthier Communities, which provides team-based primary health care through Community Health Centres (CHCs) across Ontario, Canada.

Methods
We used electronic health record data from adult ongoing primary care clients served by CHCs in 2009-2019. We performed traditional table-based summaries for each characteristic; and applied three unsupervised learning techniques to explore patterns of common condition co-occurrence, care provider teams, and care frequency.

Results
There were 221,047 eligible clients. Sociodemographics: We described 13 characteristics, stratified by CHC type and client multimorbidity status. Clinical characteristics: Eleven-year prevalence of 24 investigated conditions ranged from 1% (Hepatitis C) to 63% (chronic musculoskeletal problem) with non-uniform risk across the care history; multimorbidity was common (81%) with variable co-occurrence patterns. Health care use characteristics: Most care was provided by physician and nursing providers, with heterogeneous combinations of other provider types. A subset of clients had many issues addressed within single-visits and there was within- and between-client variability in care frequency. In addition to substantive findings, we discuss methodological considerations for future decision support initiatives.

Conclusions
We demonstrated the use of methods from statistics and machine learning, applied with an epidemiological lens, to provide an overview of a complex primary care population and lay a foundation for stakeholder engagement and decision support tool development.

Keywords
learning health system; primary health care; epidemiology; artificial intelligence; unsupervised machine learning

*Corresponding Author:
Email Address: jkueper@uwo.ca (Jacqueline K. Kueper)

https://doi.org/10.23889/ijpds.v7i1.1756
October 24, 2022 © The Authors. Open Access under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/deed.en)

Introduction

Increasing amounts of "everyday data", such as that created as a by-product of healthcare delivery, combined with advancements in computing resources and analytical techniques, provide unprecedented opportunities to increase knowledge and support health [1–7]. We are focused on the use of electronic health record (EHR) data from a particular healthcare organization to develop decision support tools whereby the findings or end-product are intended to benefit the organization or the population it serves. This paradigm can support entire populations, including subgroups who have historically been excluded from medical research and clinical guideline development, such as those with complex health needs or barriers to participation [8–11]. Using care-derived data to develop tools that will support the same population(s) that gave rise to the data can happen at any scale: a single clinic, entire organization encompassing several clinics, or other health system notion that represents a defined population [1–4].

Primary care, first contact care provided in a community setting over the life course, is inherently complex [12, 13]. In Canada primary care is covered by provincial healthcare plans, with several different reimbursement models. The Alliance for Healthier Communities employs salaried family physicians who work with other provider types and offer team-based primary health care through 72 Community Health Centres (CHCs) across Ontario. Each CHC is community-governed and motivated by geographic or social need; in general, their focus is on serving clients who face barriers to care and challenges, such as poverty and mental illness, that increase their risk for poor health [14–16]. Compared with other primary care models in Ontario, the population served by CHCs tends to include a larger proportion of people in low income neighbourhoods, who are new to Canada, and have serious mental illness and/or chronic conditions, while exhibiting lower than expected emergency department rates [14, 17]. Population health is a central element of the Alliance care model, which officially adopted a learning health system model in October 2020, demonstrating a formal commitment to using their data to inform and improve care [18, 19].

Multiple different types of decision support initiatives may be pursued. A first step towards any initiative is identifying needs of clients and providers, which is often driven by internal stakeholders [4]. Descriptive epidemiology is instrumental in outlining health states and needs of populations [20], and may be beneficial to add into these early stages of problem selection and project development both to identify new areas to explore and to support existing ideas. For example, describing how clients are represented in EHR data at a population level may complement clinical experience to identify potential bias or misrepresentation that analyses need to account for to obtain meaningful results [21–23]. In addition to proposed benefits for organizational initiatives, descriptive studies are needed to contribute towards closing the gap in understanding about the basic functions of primary care in general [24].

Historically, more methodological research and attention has been given to analytic epidemiology than descriptive epidemiology, despite descriptive studies being common and valuable in research and public health contexts [25]. To properly understand complex EHR data, we propose using both simple statistical techniques traditionally used in descriptive epidemiology and more complex techniques from machine learning, applied with an epidemiological lens. Simple techniques alone may provide an oversimplified or incorrect view of certain characteristics, which could lead to ineffective or harmful decisions later-on. So, in pursuing our primary purpose of better understanding care provided by the Alliance, we explore the suitability of a variety of techniques for epidemiology of a separate primary care system with its own EHR.

We present the first large-scale descriptive and exploratory study of data for a large longitudinal cohort receiving primary care services (in our case, 220,000+ clients served by the Alliance) using a combination of traditional statistical and machine learning methodology. Our *objective* was to summarize sociodemographic, clinical, and health care use characteristics of this population. We used unsupervised learning techniques to identify patterns of multimorbidity, care provider teams, and care access frequency. Findings provide a foundation for future decision support and machine learning initiatives at the Alliance, including those related to their existing interest in using EHR data to segment populations and tailor care. In addition to substantive findings, this work more generally demonstrates the application of an epidemiological lens and use of a variety of methods from statistics and machine learning to effectively describe a complex population and contribute to early stages of an organization's journey to harness value from their EHR data.

Methods

Study population and data source

The Alliance Business Intelligence and Reporting Tools team prepared a de-identified extract of the centralized, structured EHR database from all CHCs for our research use; clients were given unique identifiers to allow tracking of care over time. All CHCs use the same standardized consent for clients, which is on an opt-out basis for research reporting aggregate results such as ours. All providers are expected to record their client interactions into the EHR. This is the legal chart record and employees would have this in their job description and work expectations. All CHCs shifted to a common EHR and data standards in 2000. Not everyone immediately started using the EHR fully, but all standardized data was required at that time. Issues addressed during care are recorded using Electronic Nomenclature and Classification Of Disorders and Encounters for Family Medicine (ENCODE-FM) [26] and International Classification of Disease (ICD)-10 vocabularies [27].

Primary care EHRs represent an open cohort; Supplementary Appendix 1 (Supplementary Figure 1) shows the cohort size along calendar- and observation-based time definitions. Clients eligible for inclusion were over 18 years old in 2009, indicated a CHC as their primary care provider, and had at least one encounter at a CHC in 2009 to 2019. Any additional eligibility for specific analyses is described as needed below.

We followed RECORD reporting guidelines (Supplementary Appendix 2) [28]. Stakeholder engagement throughout the research included the Alliance Director of Research and Evaluation (JR) being involved in all stages of planning and conduct of this research, as well as some preliminary one on one conversations with care providers at CHCs. Two authors (JR and MZ) were also involved in a parallel qualitative study investigating aspects of the Alliance's learning health system journey [29, 30].

General analysis plan

Sociodemographic, clinical, and health care use characteristics are defined in Supplementary Appendix 3 (Supplementary Table 1). Methods specific to each category are described below; we performed "table-based summaries" for all, whereby categorical variables were summarized by counts and percentages, and continuous variables by the range, median, mean, and standard deviation. Where specified, findings were stratified by client multimorbidity status (defined below) or CHC "Urban At-Risk" status, referring to CHCs located in major urban geographical areas that serve priority populations defined by homelessness and/or mental health and substance use challenges [17]. Given their unique care mandate, these CHCs are considered a distinct "peer group" within the larger organization; another CHC peer group includes those with Rural Geography. CHCs without special designation still focus on clients with barriers to care but may serve those in rural or urban settings and do not solely serve clients with the aforementioned complexities [17].

Sociodemographic characteristics

We conducted table-based summaries for select fields from the structured EHR client characteristic table and for certain ENCODE-FM-derived variables, treated as present if ever recorded in a client's EHR. Missingness of the former occurred at the 1) CHC or provider level, whereby a client was not asked about the characteristic and 2) client level, whereby a client was asked and preferred to not respond. Results are presented overall and stratified by Urban At-Risk CHC and multimorbidity status.

Clinical characteristics

We investigated 20 chronic conditions that define multimorbidity in primary care research [31–33] and an additional four conditions of interest identified by Alliance stakeholders (Hepatitis C, smoking or tobacco use, substance use, lonely or isolated). For each condition, clients were assumed to receive related care upon the first record of a relevant code. We explored conditions in single, composite, and pairwise manners.

Prevalence and incidence

To provide different perspectives on clinical complexity, we calculated two measures of prevalence and one measure of incidence for each of the 24 conditions. We also calculated prevalence of multimorbidity. Our primary multimorbidity definition, including for stratification, was presence of at least three of the 20 chronic conditions [31–33]. We also looked at multimorbidity of at least two conditions, as this is another commonly used definition [32].

1) *Eleven-year period prevalence*, based on calendar time, to assess the burden of conditions over the entire observation period (2009–2019). For each condition, we divided the number of clients who ever received a condition indication by an estimate of the average population size (technical details in Supplementary Appendix 3). Sensitivity analyses included the largest possible denominator: total number of eligible clients, and the smallest reasonable denominator: starting with the middle calendar year (2014), additional clients with at least one visit in adjacent years were added until no prevalence estimate was over 100%. Results are shown overall and Urban At-Risk CHC-stratified.

2) *Observation-based period prevalence*, based on length of client observation, to assess the burden of conditions dependent on the number of years clients received care at a CHC. To calculate this, we separated clients into 11 sub-cohorts based on the number of years (consecutive 365.25 day intervals, rounded up) between their first and last recorded events. For each sub-cohort and condition, we divided the number of clients who ever received a condition indication by the number of clients in the sub-cohort. Results are presented as bar graphs.

3) *Cumulative incidence*, to assess the rate of condition indications by days of observation. We plotted cumulative incidence curves using the R package *survival* [34] . To prioritize capture of incident condition-related care, we excluded clients with conditions recorded in 2009 from this analysis.

Condition co-occurrence patterns

To assess co-occurrence for each pair of conditions while adjusting for all of the other conditions, we estimated an *Ising model* (unsupervised machine learning technique) using R package *MRFcov* [35, 36] for all conditions except Hepatitis C (Alliance-suggested condition that overlaps with one of the 20 chronic conditions). We converted coefficients, representing the strength of association between each condition pair adjusted for all other conditions, to odds ratios and interpreted size using Chen et al. (2010) guidelines [37]. We also viewed the top frequency-based co-occurrences.

Health care use characteristics

We performed table-based summaries of provider and care access characteristics overall and stratified by Urban At-Risk CHC, Rural Geography CHC, and client multimorbidity status.

Providers involved

To identify common care provider teams that clients were exposed to across their care histories, we used *non-negative matrix factorization (NMF)* [38] (unsupervised machine learning technique) to identify frequently-occurring: 1) *"Ever-seen" teams* whereby dummy variables were used to indicate

whether each provider type was ever involved in care, and 2) *Relative "amount-seen" teams* based on volume of care whereby the number of events associated with each provider type was normalized within clients. For each version, we ran analyses allowing 2,3,5,10, and 15 topics (provider teams) with the Python package *sklearn.decomposition.NMF* and the Kullback-Leibler divergence distance metric [39]. We interpreted resulting topics by visual inspection. Provider types were maintained as recorded in the EHR except "Other," "Unknown," and "Undefined" were combined. We also summarized the top frequency-based provider types involved in care and referrals. Eligible clients required at least one provider type indication in their EHR.

Care access patterns

We measured *complexity of care* as the number of events (distinct issues addressed or types of care received) per visit (calendar day of access) to a CHC, and *care frequency* as the number of calendar days at least one event was recorded per year (365.25 day intervals) and per quarter-year (90.30 day intervals). To investigate frequency of care in terms of magnitude and shape (changes in magnitude across care histories), we performed *time series clustering* (unsupervised machine learning technique) with the K-Medoids algorithm and dynamic time warping distance metric [40] for 1) *short-term clients* with 2–3 observation years and 2) *long-term clients* with 8–10 observation years. For each time interval and cohort, we used R package *dtwclust* [41] to identify 2,3,4, and 5 clusters. Performance was assessed using the silhouette score and visual inspection.

Results

There were 221,047 eligible clients (Supplementary Appendix 3). Note that we may only be using a subset of the CHC care history for the 64,504 (29.18%) who had at least one care indication in 2009, 141,627 (64.07%) in 2019, and 40,704 (18.4%) who received care in both of these "cohort end" years.

Sociodemographic characteristics

Sociodemographic characteristics are described in Table 1, with remaining sub-strata in Supplementary Appendix 3 Supplementary Table 2. The Urban At-Risk CHCs tended to provide care to clients who were more commonly English-speaking, and had lower levels of education, household income, immigration, stable housing, and/or food security. Clients with multimorbidity tended to be older and more commonly female, reside in rural locations, and had lower levels of education, immigration, stable residence, and/or food security.

Clinical characteristics

Prevalence and incidence

Eleven-year period prevalence estimates ranged from 1.48% (Hepatitis C) to 80.97% (multimorbidity of two conditions) overall, with generally higher estimates in Urban At-Risk CHC strata (Table 2). The low sensitivity estimate for the denominator was based on 2012–2015 (n = 148 595).

Observation-based period prevalence estimates tended to increase with length of observation; however, *cumulative incidence* plots for the 156,543 (70.82%) clients without care recorded in 2009 showed the rate of condition indications notably decreased after the first year of observation. Sample plots are in Figure 1; all are in Supplementary Appendix 1 (Supplementary Figures 2, 3).

Condition co-occurrence patterns

Among the 103,172 (46.7%) clients with multimorbidity of at least three chronic conditions, there were 25,162 unique combinations ranging in frequency from 1 (<0.1%) to 845 (0.4%) clients. Figure 2 presents the *Ising model* results. Pairwise associations between conditions on the log-odds scale ranged from -0.82 (Osteoporosis—Obesity) to 2.93 (Kidney disease or failure—Chronic urinary problem). There was 1 large, 5 medium, 40 small, and 207 very small associations based on odds ratio magnitude. The five largest positive associations were 1) Kidney Disease or Failure—Chronic Urinary Problem, 2) Smoking or Tobacco Use—Substance Use, 3) Cardiovascular Disease—Heart Failure, 4) Hypertension—Hyperlipidemia, and 5) Hypertension—Kidney Disease or Failure. In contrast, the top 5 co-occurring conditions based on raw frequency were 1) Hyperlipidemia—Chronic Musculoskeletal, 2) Hypertension—Chronic Musculoskeletal, 3) Hyperlipidemia—Hypertension, 4) Chronic Urinary Problem—Chronic Musculoskeletal, 5) Asthma or COPD or Chronic Bronchitis—Chronic Musculoskeletal. These directly correspond to the conditions that had the highest marginal frequencies.

Health care use characteristics

Table-based summaries of health care use characteristics are in Supplementary Appendix 3 (Supplementary Table 3). In general, Urban At-Risk CHC and multimorbidity strata had higher health care use while rural geography CHCs were closer to the overall population.

Providers involved

There were 19,394 unique combinations of the 68 distinct provider types seen across the 220,806 (99.9%) clients with at least one provider type recorded. In terms of referrals, 102,088 (46.2%) clients had at least one internal and 143,922 (65.1%) had at least one external referral recorded. Note internal referrals may not have captured "hallway referrals", whereby a nearby provider provides a quick consult that is not formally recorded.

Figure 3 shows results of the *NMF analysis*, listing the highest-weighted provider types in each topic down to a weight of 3. For the *ever-seen* provider team analysis, physician and nursing provider types emerged most prominently overall. In general, as the number of topics increased, additional provider types emerged and then split apart to dominate separate topics. Exceptions were the high-weighted pairings

Table 1: Sociodemographic characteristics

Characteristic	Values	All clients n (%)	Urban at risk CHC[a] n (%)	Multimorbidity n (%)
Number of clients		221 047	35 998	103 172
Age in 2015	25-34	55 505 (25.11)	7976 (22.16)	9346 (9.06)
	35-44	45 646 (20.65)	7540 (20.95)	15 542 (15.06)
	45-54	44 653 (20.2)	8186 (22.74)	23 982 (23.24)
	55-64	37 848 (17.12)	6790 (18.86)	25 578 (24.79)
	65-74	23 162 (10.48)	3644 (10.12)	17 780 (17.23)
	75+	14 233 (6.44)	1862 (5.17)	10 944 (10.61)
Geography	Rural	49 275 (22.29)	6131 (17.03)	26 818 (25.99)
	Urban	167 728 (75.88)	28 538 (79.28)	75 011 (72.70)
	Missing	4044 (1.83)	1329 (3.69)	1343 (1.30)
Sex	Female	127 070 (57.49)	18 699 (51.94)	59 946 (58.10)
	Male	93 294 (42.21)	17 151 (47.64)	43 124 (41.80)
	Other	331 (0.15)	43 (0.12)	19 (0.02)
	Missing	352 (0.16)	105 (0.29)	83 (0.08)
Gender	Female	41 352 (18.71)	5509 (15.30)	21 831 (21.16)
	Gender diverse	340 (0.15)	112 (0.31)	144 (0.14)
	Male	29 366 (13.28)	4585 (12.74)	14 733 (14.28)
	Prefer not to answer	1001 (0.45)	51 (0.14)	376 (0.36)
	Missing	148 988 (67.4)	25 741 (71.51)	66 088 (64.06)
Sexual Orientation	Bisexual	1578 (0.71)	285 (0.79)	690 (0.67)
	Gay	708 (0.32)	192 (0.53)	306 (0.30)
	Heterosexual	57 065 (25.82)	8447 (23.47)	29 105 (28.21)
	Lesbian	485 (0.22)	70 (0.19)	244 (0.24)
	Queer	323 (0.15)	34 (0.09)	91 (0.09)
	Two-Spirit	128 (0.06)	80 (0.22)	61 (0.06)
	Other	246 (0.11)	34 (0.09)	143 (0.14)
	Do not know	924 (0.42)	201 (0.56)	485 (0.47)
	Prefer not to answer	7561 (3.42)	877 (2.44)	4078 (3.95)
	Missing	152 029 (68.78)	25 778 (71.61)	67 969 (65.88)
Highest Level of Education	Post-secondary or equivalent	84 888 (38.4)	12 056 (33.49)	35 763 (34.66)
	Secondary or equivalent	61 831 (27.97)	11 783 (32.73)	32 617 (31.61)
	Less than high school	18 941 (8.57)	3266 (9.07)	10 618 (10.29)
	Other	8507 (3.85)	719 (2.00)	4078 (3.95)
	Do not know	4860 (2.20)	1318 (3.66)	2350 (2.28)
	Prefer not to answer	2950 (1.33)	422 (1.17)	1585 (1.54)
	Missing	39 070 (17.67)	6434 (17.87)	16 161 (15.66)
Primary Language	English	167 163 (75.62)	31 658 (87.94)	79 599 (77.15)
	French	22 547 (10.20)	944 (2.62)	11 091 (10.75)
	Other	26 847 (12.15)	2948 (8.19)	10 710 (10.38)
	Missing	4490 (2.03)	448 (1.24)	1772 (1.72)
Race and Ethnicity	Black	8861 (4.01)	725 (2.01)	3757 (3.64)
	East/Southeast Asian	3739 (1.69)	484 (1.34)	1545 (1.50)
	Indigenous	2944 (1.33)	1577 (4.38)	1641 (1.59)
	Latino	4350 (1.97)	206 (0.57)	1708 (1.66)
	Middle Eastern	2046 (0.93)	344 (0.96)	838 (0.81)
	Other	567 (0.26)	148 (0.41)	306 (0.30)
	South Asian	3597 (1.63)	323 (0.90)	1852 (1.80)
	White	38 464 (17.4)	4531 (12.59)	21 504 (20.84)
	Do not know	838 (0.38)	151 (0.42)	487 (0.47)
	Prefer not to answer	2649 (1.20)	261 (0.73)	1513 (1.47)
	Missing	152 992 (69.21)	27 248 (75.69)	68 021 (65.93)

Continued

Table 1: Continued

Characteristic	Values	All clients n (%)	Urban at risk CHC[a] n (%)	Multimorbidity n (%)
Years Since Arrival in Canada	0to5yr	13 654 (6.18)	1191 (3.31)	3047 (2.95)
	6+	51 815 (23.44)	4940 (13.72)	22 722 (22.02)
	None recorded	155 578 (70.38)	29 867 (82.97)	77 403 (75.02)
Household Income	$0 to $14,999	40 519 (18.33)	8729 (24.25)	17 757 (17.21)
	$15,000 to $24,999	21 102 (9.55)	3555 (9.88)	11 081 (10.74)
	$25,000 to $39,999	20 877 (9.44)	2988 (8.30)	10 736 (10.41)
	$40,000 to $59,999	17 245 (7.80)	2421 (6.73)	8671 (8.40)
	$60,000 or more	28 494 (12.89)	3862 (10.73)	12 868 (12.47)
	Do not know	15 408 (6.97)	2658 (7.38)	6264 (6.07)
	Prefer not to answer	27 621 (12.50)	4130 (11.47)	14 890 (14.43)
	Missing	49 781 (22.52)	7655 (21.27)	20 905 (20.26)
Household Composition	Couple with children	53 398 (24.16)	6759 (18.78)	20 713 (20.08)
	Couple without child	39 664 (17.94)	5945 (16.51)	22 950 (22.24)
	Extended family	7632 (3.45)	1123 (3.12)	3581 (3.47)
	Grandparents with grandchild(ren)	1746 (0.79)	247 (0.69)	1183 (1.15)
	Siblings	1622 (0.73)	250 (0.69)	669 (0.65)
	Single parent	14 445 (6.53)	2527 (7.02)	6348 (6.15)
	Sole member	32 782 (14.83)	7445 (20.68)	18 597 (18.03)
	Unrelated housemates	8622 (3.90)	1567 (4.35)	2849 (2.76)
	Other	8913 (4.03)	1476 (4.10)	4202 (4.07)
	Do not know	2475 (1.12)	643 (1.79)	1279 (1.24)
	Prefer not to answer	3727 (1.69)	491 (1.36)	1927 (1.87)
	Missing	46 021 (20.82)	7525 (20.90)	18 874 (18.29)
Stable Residence	True	199 349 (90.18)	28 227 (78.41)	90 479 (87.70)
Food Insecurity	True	10 985 (4.97)	2947 (8.19)	7323 (7.10)

[a]CHC = Community Health Centre.

of nurse and physician and of registered practical nurse and nurse practitioner. Overall, 18 of the 68 possible provider types emerged prominently in at least one topic; only one (respirologist) did not also appear in the amount-seen analysis.

The *amount-seen* provider team analysis had greater weight distributions between provider types within topics. For example, the first of the three-topic analysis had an approximate 1:1:1:6 ratio of care provided by nurse practitioner:nurse:registered practical nurse:physician. In both versions, about half of clients had a non-zero weight for only one of the first two topics; in the amount-seen analysis more clients maintained a non-zero weight on only one topic as the number of topics increased, e.g., 16.6% versus 2.5% at five topics. In general, results suggest most clients received the majority of care from physician, nurse practitioner, or nurse provider types, usually in combination with other provider types at a lower volume of care and with heterogeneous co-occurrence. An example of patterns that emerged for other provider types include differences in timing and weight of dietician/nutritionist and social worker providers between the two analyses. Interpreted alongside the most common provider and referral types (Supplementary Appendix 3 (Supplementary Table 4)), findings suggest referrals to dietitian/nutritionist were more common than to social worker, but frequent or longer-term care was more commonly provided by social workers.

Care access patterns

Complexity of care from a CHC-perspective was primarily low with 80.4% of client-visits associated with a single-issue and under 1.0% with over five issues addressed (higher intensity); however, from a client-perspective, 24,204 (11.0%) experienced at least one visit with over five issues while 38 533 (17.4%) experienced a maximum of one issue per visit across their care history. The mean *care access frequency* was 6 days per year (standard deviation = 7.4). While 29,191 (13.2%) clients experienced at least one year with over 25 days, 7,455 (3.4%) averaged over 25 days per year across their entire care history. There were 8,700 (3.94%) clients with at least one frequent care period (year with over 25 days care accessed) and complex care episode (visit with over 5 issues addressed).

For the *time series clustering* analyses, the short-term cohort included 37,920 clients and 93,625 client-years of observation; the long-term cohort included 42,855 clients and 387,035 client-years of observation. The silhouette score was always highest for two clusters (Supplementary Appendix 3 (Supplementary Table 5)). Visual inspection of plots (Figure 4) showed high variability within and between clients.

Table 2: Eleven-year period prevalence

Condition	All clients n (%)	Urban at risk CHC[a] n (%)
Denominator[b]	165 125	27 256
Hypertension	68 177 (41.29)	12 304 (45.14)
Depression or anxiety	23 828 (14.43)	5533 (20.30)
Chronic musculoskeletal	104 304 (63.17)	18 842 (69.13)
Arthritis	37 201 (22.53)	6906 (25.34)
Osteoporosis	11 462 (6.94)	1950 (7.15)
Asthma or COPD[c] or chronic bronchitis	43 837 (26.55)	9190 (33.72)
Cardiovascular disease	23 311 (14.12)	4673 (17.14)
Heart failure	7994 (4.84)	1564 (5.74)
Stroke or TIA[d]	2967 (1.80)	585 (2.15)
Stomach problem	36 175 (21.91)	7620 (27.96)
Colon problem	24 949 (15.11)	4974 (18.25)
Chronic hepatitis	13 288 (8.05)	2954 (10.84)
Diabetes	35 704 (21.62)	6912 (25.36)
Thyroid disorder	24 793 (15.01)	4217 (15.47)
Any cancer	14 024 (8.49)	2636 (9.67)
Kidney disease or failure	8290 (5.02)	1555 (5.71)
Chronic urinary problem	59 677 (36.14)	11 131 (40.84)
Dementia or Alzheimer's disease	4776 (2.89)	898 (3.29)
Hyperlipidemia	67 175 (40.68)	11 659 (42.78)
Obesity	38 408 (23.26)	6455 (23.68)
Hepatitis C	2436 (1.48)	1173 (4.30)
Smoking or tobacco use	37 355 (22.62)	9597 (35.21)
Substance use	20 853 (12.63)	7508 (27.55)
Lonely or isolated	17 947 (10.87)	5149 (18.89)
Multimorbidity 2+	133 704 (80.97)	24 129 (88.53)
Multimorbidity 3+	103 172 (62.48)	19 237 (70.58)

[a]CHC = Community Health Centre.
[b]Denominator is the approximated average population size across all years (2009–2019).
[c]COPD = Chronic Obstructive Pulmonary Disease.
[d]TIA = Transient Ischemic Attack.

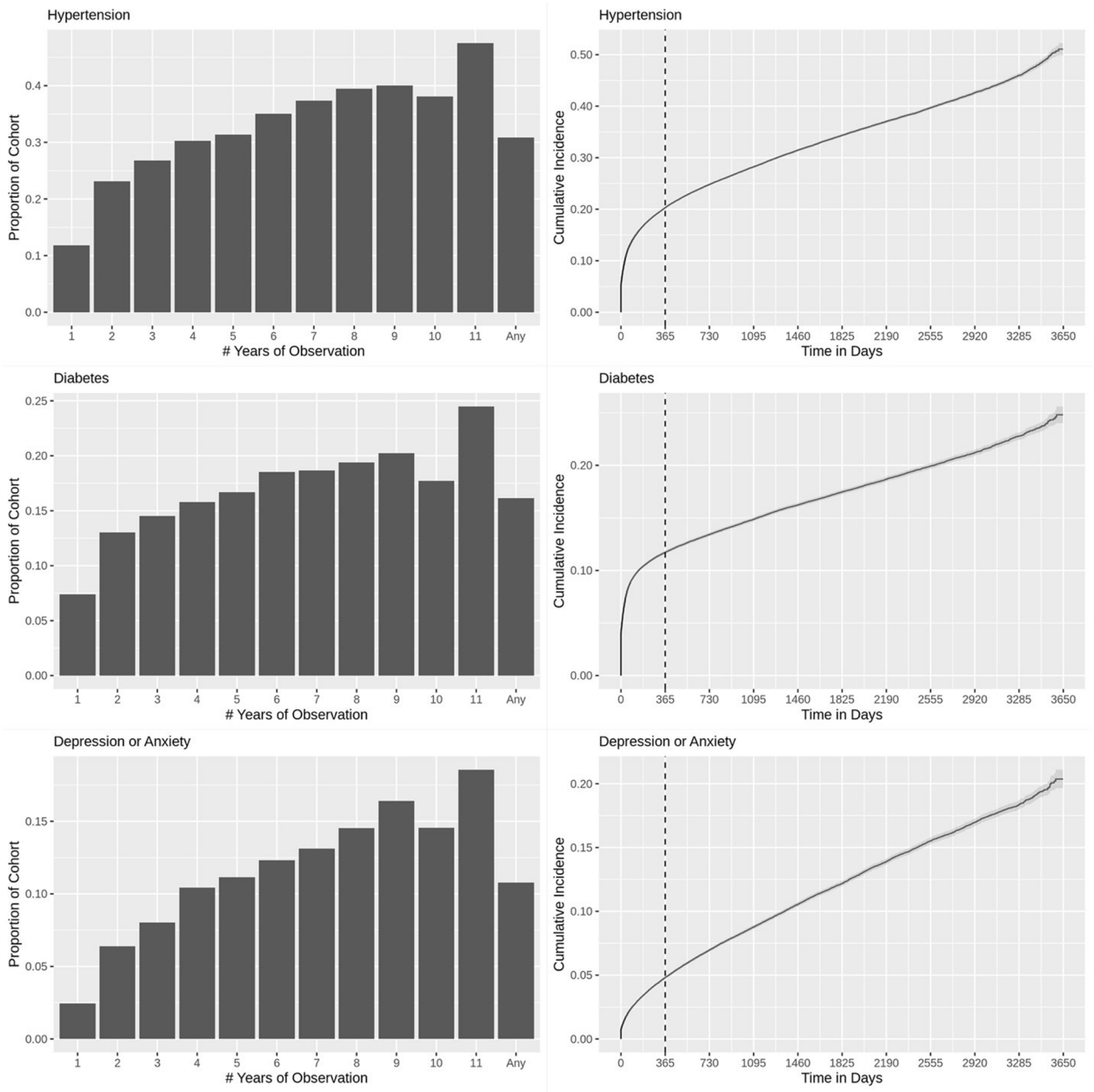
Discussion

We used statistical and machine learning techniques to summarize sociodemographic, clinical, and health care use characteristics captured in the EHRs of ongoing primary care clients served by the Alliance. Substantive findings can motivate new topics for future decision support initiatives, or help to refine existing ideas and selection of performance measures for long-term evaluation of implemented interventions. Methods-related findings may inform the approaches used in these endeavours. While our discussion focuses on decision support initiatives, as with any epidemiological study, substantive results may be immediately useful to the population of interest, e.g., to inform clinic-level case management and onboarding of new clients.

Sociodemographic characteristics

The CHC EHRs contain rich sociodemographic information, both the presence and absence of which is informative. Social determinants of health that may increase risk of poor health including lower household income, education, residence instability, and food insecurity were more prevalent in Urban At-Risk CHC and multimorbidity strata (Table 1 and Supplementary Table 2). There appears to be evidence for the healthy immigrant effect [42], assessed by viewing the proportion of people in each category of the years since arrival in Canada variable across the multimorbidity strata: a lower proportion of people with 0-5 years in Canada had multimorbidity as compared to those with 6 or more years in Canada or no arrival information recorded (missing or born in Canada). Completeness rates varied by characteristic and may be due to client, provider, or CHC level decisions. For example, of the 72,059 (32.60%) clients asked about gender only 1,001 (1.39%) preferred to not answer. In contrast, more clients, 171,266 (77.48%), were asked about household income but there was a higher tendency to not answer, 27,621 (16.13%). These findings align with a framework to assess selection bias in EHR data that suggests multiple mechanisms are usually responsible for missingness so the focus should be on "what data are observed [instead of missing] and why?" [43]. While provider-level decisions may be due to inferring certain characteristics or prioritizing information needed for them to direct care, completeness rates are important for decision support tool performance, which can improve with social determinants of health information

Figure 1: Example observation-based period prevalence and cumulative incidence plots



Left column: Observation-based period prevalence. Right column: Cumulative incidence by days of observation.

[44, 45]. In addition to completeness, the value of these data would be further improved if time-stamped and if changes in mutable characteristics like household income could be easily traced over time. Only the sociodemographic variables constructed with ENCODE-FM codes were associated with time of recording in the data extract we used (Supplementary Table 1).

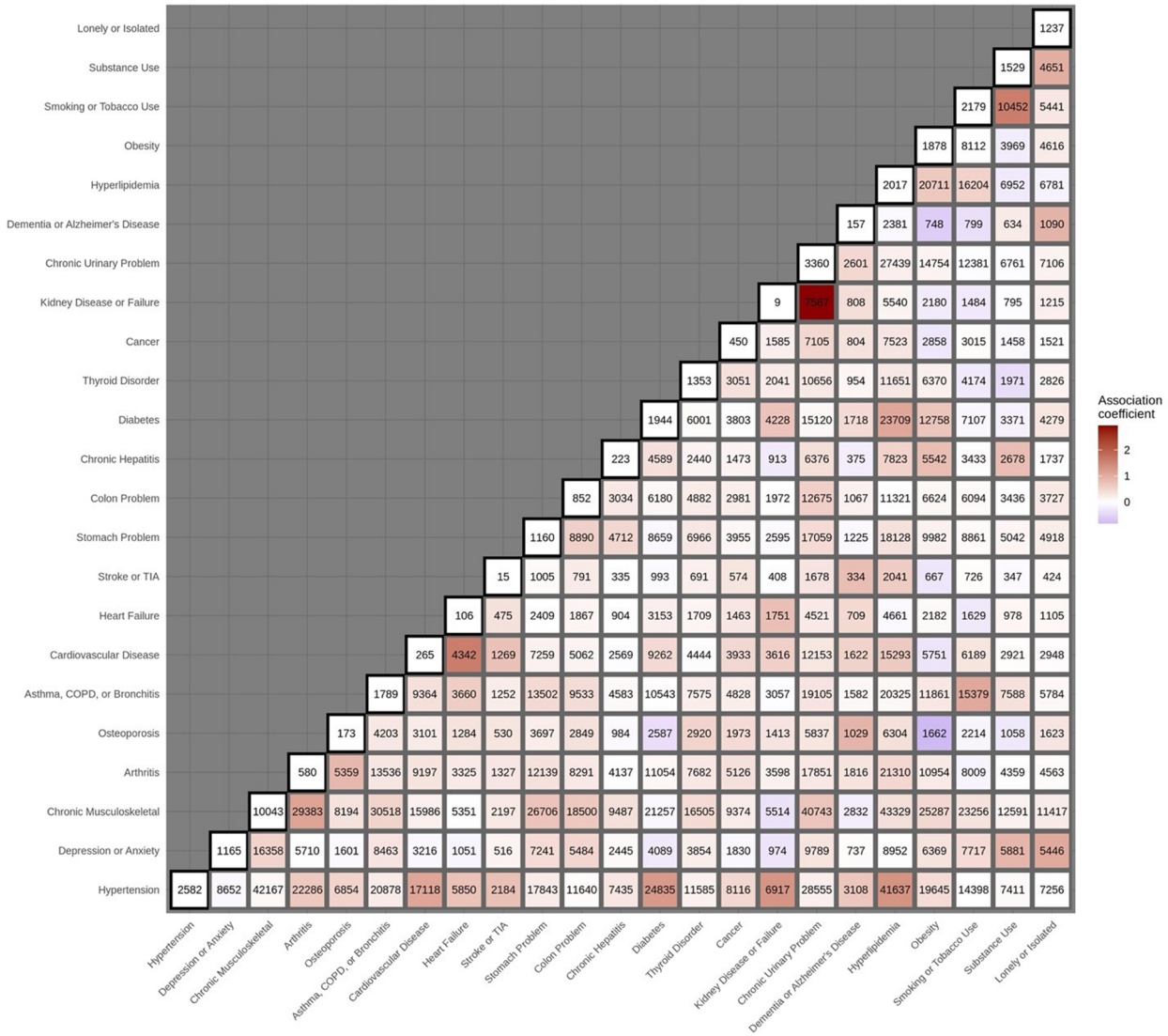
When assessing data quality and completeness, which is emphasized by learning health system and machine learning for EHR guidelines [2, 4, 22, 46, 47], the implications of pursuing decision support initiatives at different levels should also be considered. For example, a subset of CHCs capture self-reported measures of health, which are valuable research outcomes [48]. While these measures are not suitable for analyses with data from all CHCs (CHC population-level initiatives), they should be considered for initiatives specific to the collecting CHCs. Even for sociodemographic variables now collected by all CHCs, there are variable completeness rates between CHCs partly because a subset of CHCs started collecting the information earlier than others. CHCs with higher completeness rates have improved opportunity to increase understanding about typically less well represented groups such as those identifying as gender diverse.

Clinical characteristics

Prevalence and incidence

In operationalizing morbidity measures, the denominator must be defined with the intended end-goal in mind. The *eleven-year period prevalence* (Table 2) estimates relate to a CHC-based perspective and are useful for long-term system-level planning or developing tools to address amount-based condition priorities, while the *observation-based period prevalence*

Figure 2: Condition co-occurrence patterns



Heatmap representing the results of the Ising model. *Interpretation:* Shading is relative to the edge weights or strength of condition co-occurrence adjusting for the other conditions; red indicates a positive association and purple indicates a negative association. The numbers indicate raw counts in the data; diagonal counts represent clients who only have that single condition. *Legend*: TIA = Transient Ischemic Attack; COPD = Chronic Obstructive Pulmonary Disease.

(Figure 1 and Supplementary Figure 2) estimates are more aligned with a client-based perspective and absolute measure of risk. Another consideration is that just as ICD-10 or ENCODE-FM codes do not guarantee true condition presence, the absence of care does not verify absence of conditions [49]. For example, clients may not seek primary care when they are healthy, hospitalized, or experiencing barriers to care.

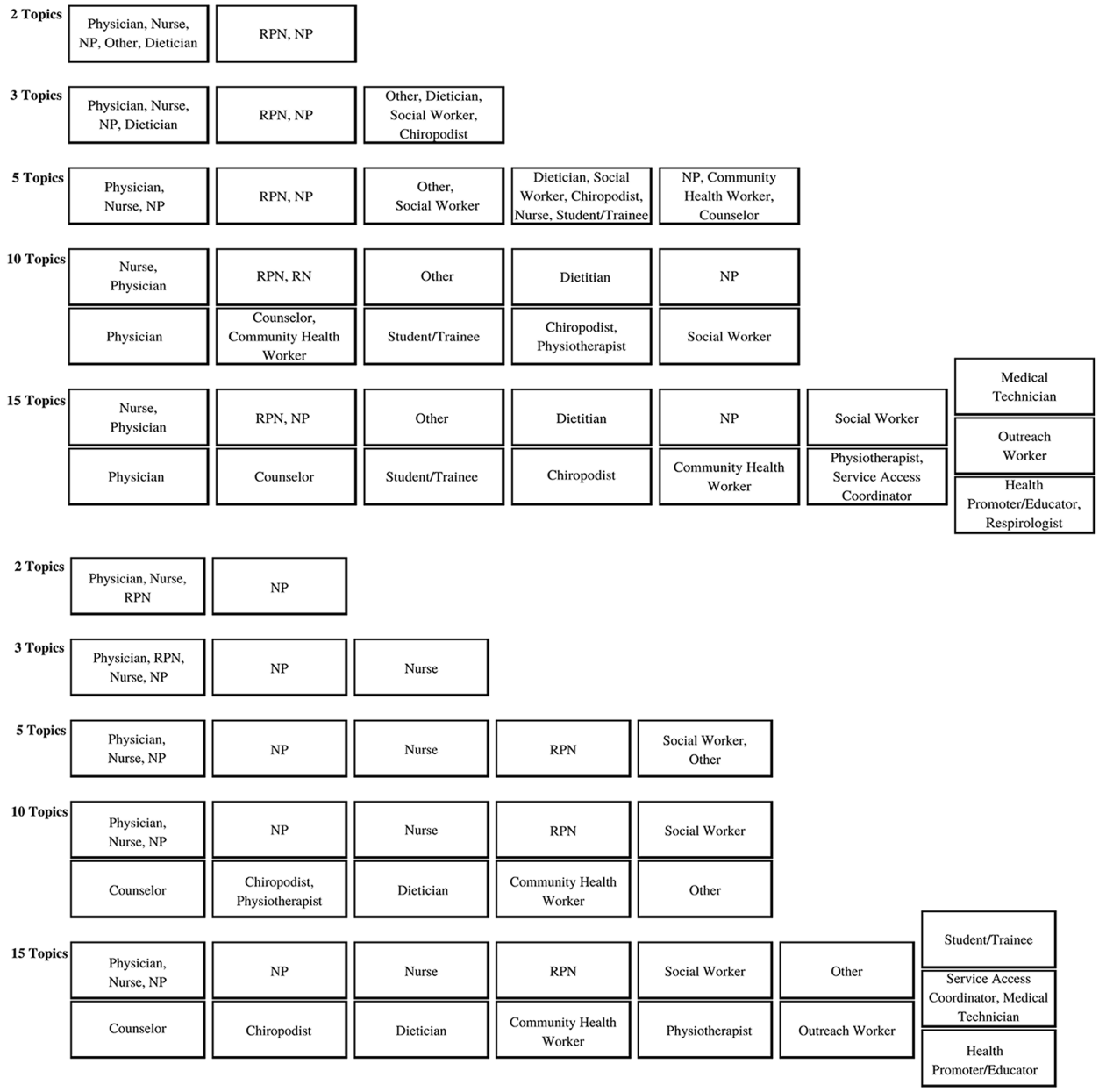
The *cumulative incidence* plots (Figure 1 and Supplementary Figure 3) demonstrate that "risk" of condition codes is highest in the first year of observation. Clinically this makes sense, as new clients may have a build-up of unmet care needs. Nonetheless, there are important takeaways for initiatives that require cohort construction. For example, predictive models developed for decision support need to account for the almost qualitative change in risk related to being a new client: likely a diagnostic model is more useful in the first year of care and a prognostic model thereafter. Although this care pattern is somewhat unique to primary care settings, methods developed for related problems may be useful. For example, accounting for variable lengths of stay in intensive care unit EHRs [50], or handling cold-starts and sparse data for recommender systems [51].

Condition co-occurrence patterns

There was a high prevalence of multimorbidity, but with thousands of different multimorbidity "compositions" it is hard to see how to make use of the category of multimorbidity.

The standard statistical table-based summaries (e.g., prevalence estimates in Table 2) provide information on the amount of each condition present in the population; the Ising model (Figure 3) provides information on the tendency for any two conditions to co-occur, adjusting for the presence or absence of the other conditions. The conditions with the strongest positive associations were not the same

Figure 3: Common care provider teams



Boxes represent the topics resulting from the non-negative matrix factorization analysis for (top) Ever-seen provider team analysis and (bottom) Relative amount seen provider team analysis. *Interpretation:* Provider types are listed in order starting with the highest weighted provider; for any given topic, provider types with a weight less than three are not show. Provider types appearing in the same topic tend to co-occur in client records. *Legend*: NP = Nurse Practitioner, RPN = Registered Practical Nurse.

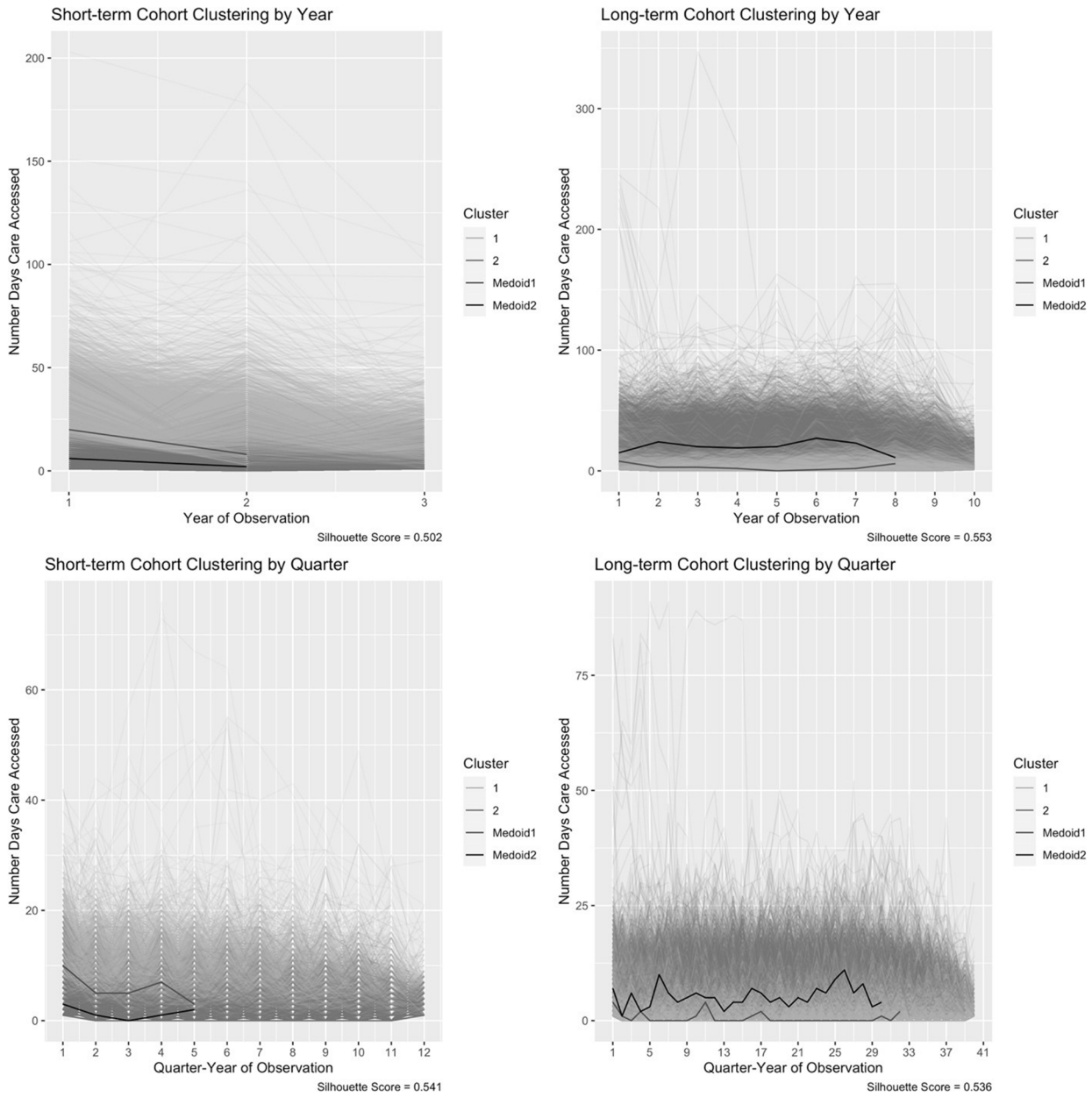
conditions with the highest frequency of occurrence or co-occurrence, so decision support tools targeted at a subset of conditions may consider whether to prioritize conditions accounting for the largest burden of disease in terms of amount or in terms of tendency for co-occurrence. In both cases, multimorbidity presents a long tail problem, with few combinations that are very prominent. Primary care decision support tools will face the challenge of making recommendations on many different and possibly co-occurring conditions. The majority of existing decision support tools and clinical guidelines focus on a single condition at a time; new techniques for providing evidence-based guidelines or recommendations for these vast numbers of combinations are needed [52–55].

Health care use characteristics

Providers involved

Care for ongoing primary care clients was typically led by physicians or by nurse practitioners, with several other provider types possibly also involved in care. These findings align with the general strategy at CHCs to match clients with primary care providers partly on service need and partly on preference. The relative-amount seen NMF analysis (Figure 3, bottom) suggests two prominent care models: nurse practitioners always

Figure 4: Care frequency clusters



Results from the four time series clustering analyses for each cohort and data-representation combination. *Interpretation:* Medoids (cluster representative/center) are shown as highlighted lines over the raw time series data, shaded by cluster number, for the number of clusters that resulted in the highest silhouette score (SS).

dominate their own topic (associated clients receive all or almost all care from nurse practitioners) whereas physicians always appear with nurse and nurse practitioner provider types, at lower weights (associated clients receive the most care from physicians, but usually with some amount of nurse and nurse practitioner care). An implication is that decision support tools targeted at common conditions or situations in primary care, such as chronic disease management, may have different provider type end-users; all types of end users should be included in the development process [56]. Other primary health care provider types emerged in their own topics separate from physician and nurse practitioner dominated topics, suggesting that both of the above prominent care models refer to or work with various combinations of other provider types, most commonly social workers, dietitians/nutritionists, chiropodists, community health workers, counselors, and physiotherapists.

The NMF analyses more easily identified prominent patterns of commonly seen provider types and teams than manually sifting through extensive count-based tables created using simple statistical techniques. Another use for NMF is dimensionality reduction or data pre-processing, whereby data are summarized to reduce the number of variables that need to be included in an analysis [38]. For example, NMF-derived topics could be used as inputs to a predictive model instead of separate variables to represent each provider type or specific, manually selected combinations.

Care access patterns

Complexity of care from a CHC system-level perspective was primarily low intensity (few issues addressed per visit, recorded using ENCODE-FM codes; Supplementary Table 3), although this may be partly due to data quality such as if only one issue was recorded in the EHR when multiple were actually addressed in an appointment. Of note, the CHC model of care and compensation permits coding of multiple issues per visit. The subset of clients who experienced higher care complexity (many issues recorded per visit) did not tend to also have high frequency of care. Sporadic visit patterns may be due to unstable living arrangements or demanding life responsibilities; when there is uncertainty about when a client will return, providers may pack together multiple types of care. The marginal distribution of *care frequency* was right-skewed without a distinct break; most clients experienced lower care frequency, but higher frequencies were also observed. We used time series clustering (Figure 4 and Supplementary Table 5) to try and go beyond statistical techniques or simple cut-off points to find more refined groupings of clients in terms of care frequency patterns. In contrast to expectations, we did not identify consistent, distinct client groupings through the time-series clustering, e.g., to indicate a subpopulation of "frequent visitors". This may be due to restrictions in our clustering approach and the types of similarity that dynamic time warping captures; however, it may also be that, while cognitively desirable, distinct client clusters in terms of care frequency do not exist. Future analyses could further investigate by trying a different similarity metric or including covariates to account for baseline variability.

Learning health systems

The paradigm that our study is intended to support or inform—using care-derived data in decision support tool development for the same population that gave rise to those data—fits within a learning health system framework, which is defined by Menear et al. (2019) as "dynamic health ecosystems where scientific, social, technological, policy, legal and ethical dimensions are synergistically aligned to enable cycles of continuous learning and improvement to be routinised and embedded across the system, thus enhancing value through an optimised balance of impacts on patient and provider experience, population health and health system costs" [1]. Large-scale descriptive studies like ours may be useful in early stages of learning health system development to inform types of initiatives beyond decision support tool development, such as quality improvement or research [2–4, 7].

The Alliance for Healthier Communities is one of the first documented primary care learning health systems in North America [7, 30]. Our next steps will include presenting our results to stakeholders through the Alliance's public Lunch 'n' Learn webinar series as well as engaging in targeted conversations with stakeholders who have self-identified interest in future research engagement. By presenting our results of how the Alliance ongoing primary care population is represented in EHR data, we hope to learn what does (not) align with expectations, and to identify next steps for decision support initiatives. Altogether this will set the stage for co-development of tools that can move forward towards pilot testing and implementation in CHCs.

Strengths and limitations

Strengths included the strong interdisciplinary approach (authors from different disciplines and methodology that connects epidemiology, computer science, and primary care) used to assess complex, longitudinal EHR data. We used chronic condition definitions recommended for primary care research [31–33], although the algorithms have not been validated for CHCs specifically. Our broad cohort definition supported a high-level overview of the population, but may not be appropriate for specific research questions. Inherent in any study using EHR data are limitations related to data quality [43, 57, 58], so it is important to remember our results showcase how clients and care are represented in the data, which may not fully reflect experiences. For example, by including clients with variable care histories the sociodemographic characteristic profiles may be skewed from the average client population profile; our prevalence estimates use a single-code assumption of presence which prioritizes sensitivity over specificity; and we did not adjust analyses for changes in policy or coding practices over time that may have impacted care access or data entry over time.

Conclusions

We demonstrated the use of simple statistics and machine learning techniques, applied with an epidemiological lens, to describe EHR data and inform future primary care decision support tool development initiatives. Substantive findings lay a foundation for future Alliance initiatives and may be informative for other organizations serving complex primary care populations.

Key suggestions for future initiatives include the need to carefully deliberate the level of analysis, or who a given tool should be targeted at (e.g., all or a subset of CHCs, one or many clinical presentations, all or some providers), and the associated implications for how clients will be represented in the data. Representation will depend on analytical-, system-, provider-, and client-level factors. Decision support initiatives need to consider heterogeneity in conditions and care access patterns, including non-uniform risk of condition indications across observation history.

Acknowledgements

This work was supported by the Canadian Institutes of Health Research Canadian Graduate Scholarship-Doctoral to JKK with supervisor DJL.

Statement on conflicts of interest

None declared.

Ethics statement

This study was approved by Western University Review Ethics Board project ID 111353.

Supplementary appendices

Appendix 1 includes extended results presented through figures.

Appendix 2 includes the RECORD reporting guideline checklist.

Appendix 3 includes extended results presented through tables and technical details.

References

1. Menear M, Blanchette M-A, Demers-Payette O, et al. A framework for value-creating learning health systems. *Health Res Policy Syst* 2019; 17: 79. https://doi.org/10.1186/s12961-019-0477-3

2. Foley T, Horwitz L, Zahran R. *Realising the potential of learning health systems*. Newcastle University: The Learning Healthcare Project.

3. Delaney BC, Peterson KA, Speedie S, et al. Envisioning a learning health care system: The Electronic Primary Care Research Network, a case study. *Ann Fam Med* 2012; 10: 54–59. https://doi.org/10.1370/afm.1313

4. Lindsell CJ, Gatto CL, Dear ML, et al. Learning from what we do, and doing what we learn: A learning health care system in action. *Acad Med* 2021; 96: 1291–1299. https://doi.org/10.1097/ACM.0000000000004021

5. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017; 2: 230–243. https://doi.org/10.1136/svn-2017-000101

6. Friedman CP, Allee NJ, Delaney BC, et al. The science of Learning Health Systems: Foundations for a new journal. *Learn Health Syst* 2017; 1: e10020. https://doi.org/10.1002/lrh2.10020

7. Nash DM, Bhimani Z, Rayner J, et al. Learning health systems in primary care: a systematic scoping review. *BMC Fam Pract* 2021; 22: 126. https://doi.org/10.1186/s12875-021-01483-z

8. Robinson JM, Trochim WMK. An examination of community members', researchers' and health professionals' perceptions of barriers to minority participation in medical research: an application of concept mapping. *Ethn Health* 2007; 12: 521–539. https://doi.org/10.1080/13557850701616987

9. George S, Duran N, Norris K. A systematic review of barriers and facilitators to minority research participation among African Americans, Latinos, Asian Americans, and Pacific Islanders. *Am J Public Health* 2014; 104: e16–e31. https://doi.org/10.2105/AJPH.2013.301706

10. Odierna DH, Schmidt LA. The effects of failing to include hard-to-reach respondents in longitudinal surveys. *Am J Public Health* 2009; 99: 1515–1521. https://doi.org/10.2105/AJPH.2007.111138

11. Bonevski B, Randell M, Paul C, et al. Reaching the hard-to-reach: a systematic review of strategies for improving health and medical research with socially disadvantaged groups. *BMC Med Res Methodol* 2014; 14: 42. https://doi.org/10.1186/1471-2288-14-42

12. Starfield B. *Primary care. Balancing health needs, services, and technology*. New York, NY: Oxford University Press, Inc., 1998.

13. CIHR. *CIHR Primary Healthcare Summit 2010 Final Report Summary*. Toronto, Ontario: Canadian Institutes of Health Research, 2010.

14. Glazier RH, Zagorski B, Rayner J. *Comparison of primary care models in Ontario by demographics, case mix and emergency department use, 2008/09 to 2009/10*. ICES Investigative Report, Toronto, Ont.: Institute for Clinical Evaluative Sciences, https://www.deslibris.ca/ID/232144 (2012, accessed 6 May 2020).

15. Booth RG, Richard L, Li L, et al. Characteristics of health care related to mental health and substance use disorders among Community Health Centre clients in Ontario: a population-based cohort study. *CMAJ Open* 2020; 8: E391–E399. https://doi.org/10.9778/cmajo.20190089

16. Albrecht D. Community health centres in Canada. *Leadersh Health Serv* 1998; 11: 5–10. https://doi.org/10.1108/13660759810202596

17. Glazier RH, Rayner J, Kopp A. *Examining community health centres according to geography and priority populations served, 2011/12 to 2012/13: an ICES chartbook*. Toronto, Ontario: Institute for Clinical Evaluative Sciences in Ontario, http://www.deslibris.ca/ID/248807 (2015, accessed 20 January 2020).

18. Alliance for Healthier Communities. Moving Forward as a Learning Health System. *Alliance for Healthier Communities*, 18 November 2020, https://myemail.constantcontact.com/EPIC-News–Issue-1.html?soid=1108953382524&aid=uzy8bphr91U (18 November 2020, accessed 23 November 2020).

19. Alliance for Healthier Communities. *Towards a Learning Health System: Better Care Tomorrow When We Learn from Today*. 2020: Alliance for Healthier Communities. https://www.allianceon.org/sites/default/files/documents/Learning%20Health%20System%20report%202020-10-20%20-%20FINAL_JR.pdf

20. Cameron D, Jones IG. John Snow, the Broad Street Pump and Modern Epidemiology. *Int J Epidemiol* 1983; 12: 393–396. https://doi.org/10.1093/ije/12.4.393

21. Thuraisingam S, Chondros P, Dowsey MM, et al. Assessing the suitability of general practice electronic health records for clinical prediction model development: a data quality assessment. *BMC Med Inform Decis Mak* 2021; 21: 297. https://doi.org/10.1186/s12911-021-01669-6

22. Verma AA, Murray J, Greiner R, et al. Implementing machine learning in medicine. *CMAJ* 2021; 193: E1351–E1357. https://doi.org/10.1503/cmaj.202434

23. Lee S, Xu Y, D'Souza AG, et al. Unlocking the potential of electronic health records for health research. *Int J Popul Data Sci* 2020; 5: 02. https://doi.org/10.23889/ijpds.v5i1.1123

24. Westfall JM, Wittenberg HR, Liaw W. Time to invest in primary care research—commentary on findings from an independent congressionally mandated study. *J Gen Intern Med* 2021; 36: 2117–2120. https://doi.org/10.1007/s11606-020-06560-0

25. Fox MP, Murray EJ, Lesko CR, et al. On the need to revitalize descriptive epidemiology. *Am J Epidemiol* 2022; kwac056. https://doi.org/10.1093/aje/kwac056

26. ENCODE-FM. Electronic Nomenclature and Classification Of Disorders and Encounters for Family Medicine. *ENCODE-FM*, http://aix1.uottawa.ca/∼fammed/fmcencod.htm (2020, accessed 6 April 2020).

27. World Health Organization. ICD-10 Version:2019. *World Health Organization*, https://icd.who.int/browse10/2019/en (2020, accessed 6 April 2020).

28. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLOS Med* 2015; 12: e1001885. https://doi.org/10.1371/journal.pmed.1001885

29. Nash DM, Brown JB, Thorpe C, et al. The Alliance for Healthier Communities as a learning health system for primary care: A qualitative analysis in Ontario, Canada. *J Eval Clin Pract* 2022; jep.13692. https://doi.org/10.1111/jep.13692

30. Nash DM, Rayner J, Bhatti S, et al. The Alliance for Healthier Communities' journey to a learning health system in primary care. *Learn Health Syst* 2021; n/a: e10321. https://doi.org/10.1002/lrh2.10321

31. Fortin M, Almirall J, Nicholson K. Development of a research tool to document self-reported chronic conditions in primary care. *J Comorbidity* 2017; 7: 117–123. https://doi.org/10.15256/joc.2017.7.122

32. Lee ES, Lee PSS, Xie Y, et al. The prevalence of multimorbidity in primary care: a comparison of two definitions of multimorbidity with two different lists of chronic conditions in Singapore. *BMC Public Health* 2021; 21: 1409. https://doi.org/10.1186/s12889-021-11464-7

33. Lee YAJ, Xie Y, Lee PSS, et al. Comparing the prevalence of multimorbidity using different operational definitions in primary care in Singapore based on a cross-sectional study using retrospective, large administrative data. *BMJ Open* 2020; 10: e039440. https://doi.org/10.1136/bmjopen-2020-039440

34. Therneau T. A Package for Survival Analysis in R, https://CRAN.R-project.org/package=survival (2021).

35. van Borkulo CD, Borsboom D, Epskamp S, et al. A new method for constructing networks from binary data. *Sci Rep* 2014; 4: 5918. https://doi.org/10.1038/srep05918

36. Clark NJ, Wells K, Lindberg O. Unravelling changing interspecific interactions across environmental gradients using Markov random fields. *Ecology* 2018; 99: 1277–1283. https://doi.org/10.1002/ecy.2221

37. Chen H, Cohen P, Chen S. How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Commun Stat - Simul Comput* 2010; 39: 860–864. https://doi.org/10.1080/03610911003650383

38. Wang Y-X, Zhang Y-J. Nonnegative Matrix Factorization: A comprehensive review. *IEEE Trans Knowl Data Eng* 2013; 25: 1336–1353. https://doi.org/10.1109/TKDE.2012.51

39. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res* 2011; 12: 2825–2830.

40. Aghabozorgi S, Seyed Shirkhorshidi A, Ying Wah T. Time-series clustering – A decade review. *Inf Syst* 2015; 53: 16–38. https://doi.org/10.1016/j.is.2015.04.007

41. Montero P, Vilar JA. TSclust: An R package for time series clustering. *J Stat Softw*; 62. Epub ahead of print 2014. https://doi.org/10.18637/jss.v062.i01

42. McDonald JT, Kennedy S. Insights into the 'healthy immigrant effect': health status and health service use of immigrants to Canada. *Soc Sci Med* 2004; 59: 1613–1627. https://doi.org/10.1016/j.socscimed.2004.02.004

43. Haneuse S, Daniels M. A general framework for considering selection bias in EHR-based studies: What data are observed and why? *eGEMs*; 4. Epub ahead of print 31 August 2016. https://doi.org/10.13063/2327-9214.1203

44. Chen M, Tan X, Padman R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: A systematic review. *J Am Med Inform Assoc* 2020; 27: 1764–1773. https://doi.org/10.1093/jamia/ocaa143

45. Zhao Y, Wood EP, Mirin N, et al. Social determinants in machine learning cardiovascular disease prediction models: A systematic review. *Am J Prev Med*; 0. Epub ahead of print 27 July 2021. https://doi.org/10.1016/j.amepre.2021.04.016

46. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019; 25: 1337–1340. https://doi.org/10.1038/s41591-019-0548-6

47. Arbet J, Brokamp C, Meinzen-Derr J, et al. Lessons and tips for designing a machine learning study using EHR data. *J Clin Transl Sci* 2020; 5: 1–10. https://doi.org/10.1017/cts.2020.513

48. CIHI. Patient-reported outcome measures (PROMs). *Canadian Institute for Health Information*, https://www.cihi.ca/en/patient-reported-outcome-measures-proms (2022, accessed 7 February 2022).

49. Bagley SC, Altman RB. Computing disease incidence, prevalence and comorbidity from electronic medical records. *J Biomed Inform* 2016; 63: 108–111. https://doi.org/10.1016/j.jbi.2016.08.005

50. Zhang L, Chen X, Chen T, et al. DynEHR: Dynamic adaptation of models with data heterogeneity in electronic health records. In: *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. 2021, pp. 1–4. https://doi.org/10.1109/BHI50953.2021.9508558

51. Alyari F, Jafari Navimipour N. Recommender systems: A systematic review of the state of the art literature and suggestions for future research. *Kybernetes* 2018; 47: 985–1017. https://doi.org/10.1108/K-06-2017-0196

52. Moons KGM, Altman DG, Vergouwe Y, et al. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ* 2009; 338: b606. https://doi.org/10.1136/bmj.b606

53. O'Caoimh R, Cornally N, Weathers E, et al. Risk prediction in the community: A systematic review of case-finding instruments that predict adverse healthcare outcomes in community-dwelling older adults. *Maturitas* 2015; 82: 3–21. https://doi.org/10.1016/j.maturitas.2015.03.009

54. Goldstein BA, Navar AM, Pencina MJ, et al. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017; 24: 198–208. https://doi.org/10.1093/jamia/ocw042

55. Guthrie B, Boyd CM. Clinical guidelines in the context of aging and multimorbidity. *Public Policy Aging Rep* 2018; 28: 143–149. https://doi.org/10.1093/ppar/pry038

56. Kellogg KC, Sendak M, Balu S. AI on the front lines. *MIT Sloan Manag Rev*, https://sloanreview.mit.edu/article/ai-on-the-front-lines/ (2022, accessed 17 May 2022).

57. Terry AL, Stewart M, Cejic S, et al. A basic model for assessing primary health care electronic medical record data quality. *BMC Med Inform Decis Mak* 2019; 19: 30. https://doi.org/10.1186/s12911-019-0740-0

58. Gianfrancesco MA, Goldstein ND. A narrative review on the validity of electronic health record-based research in epidemiology. *BMC Med Res Methodol* 2021; 21: 234. https://doi.org/10.1186/s12874-021-01416-5

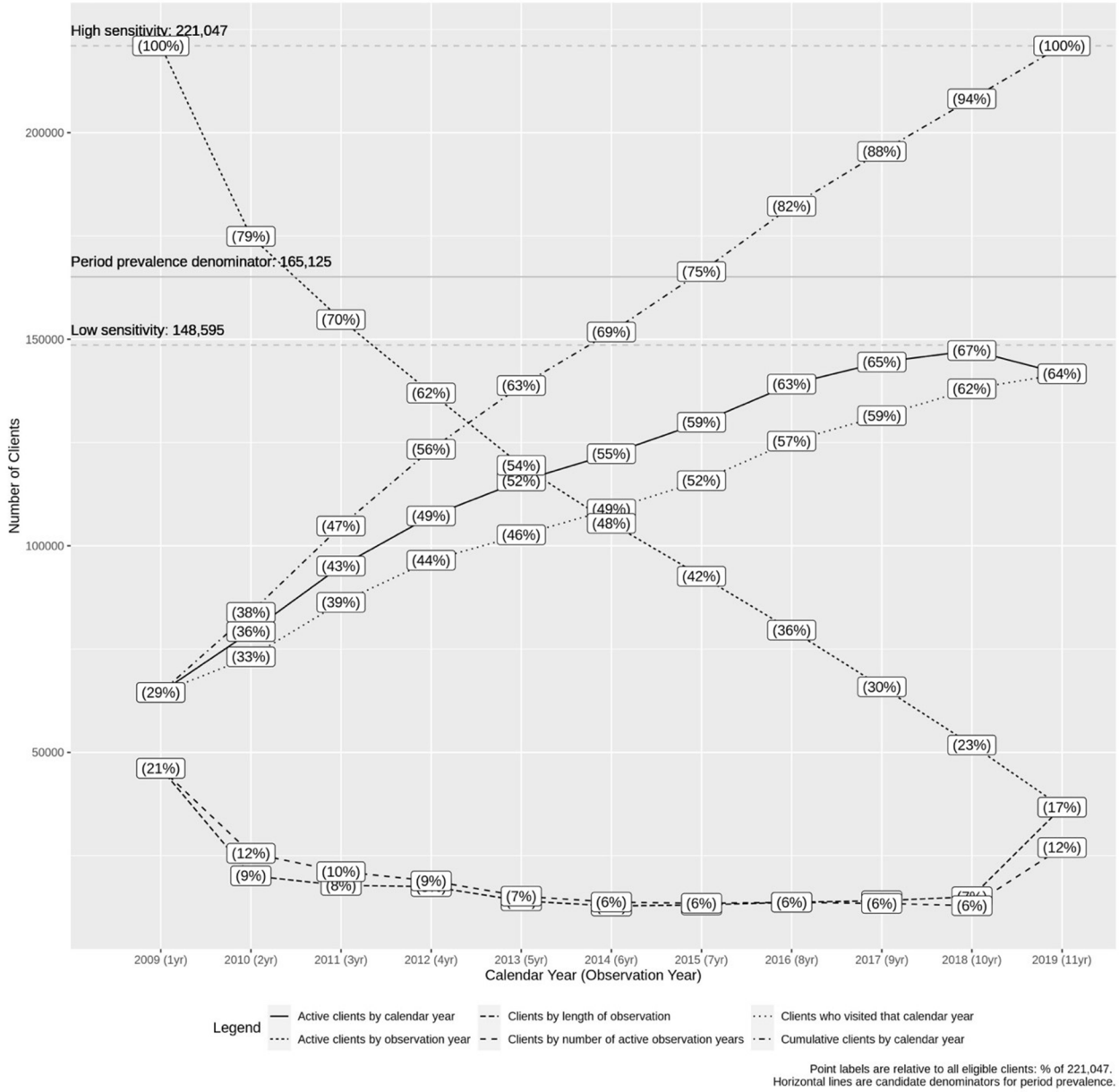
Abbreviations

CHC:	Community Health Centre
EHR:	Electronic Health Record
ENCODE-FM:	Electronic Nomenclature and Classification Of Disorders and Encounters for Family Medicine
ICD-10:	International Classification of Disease - Version 10
NMF:	Non-negative Matrix Factorization

Supplementary material

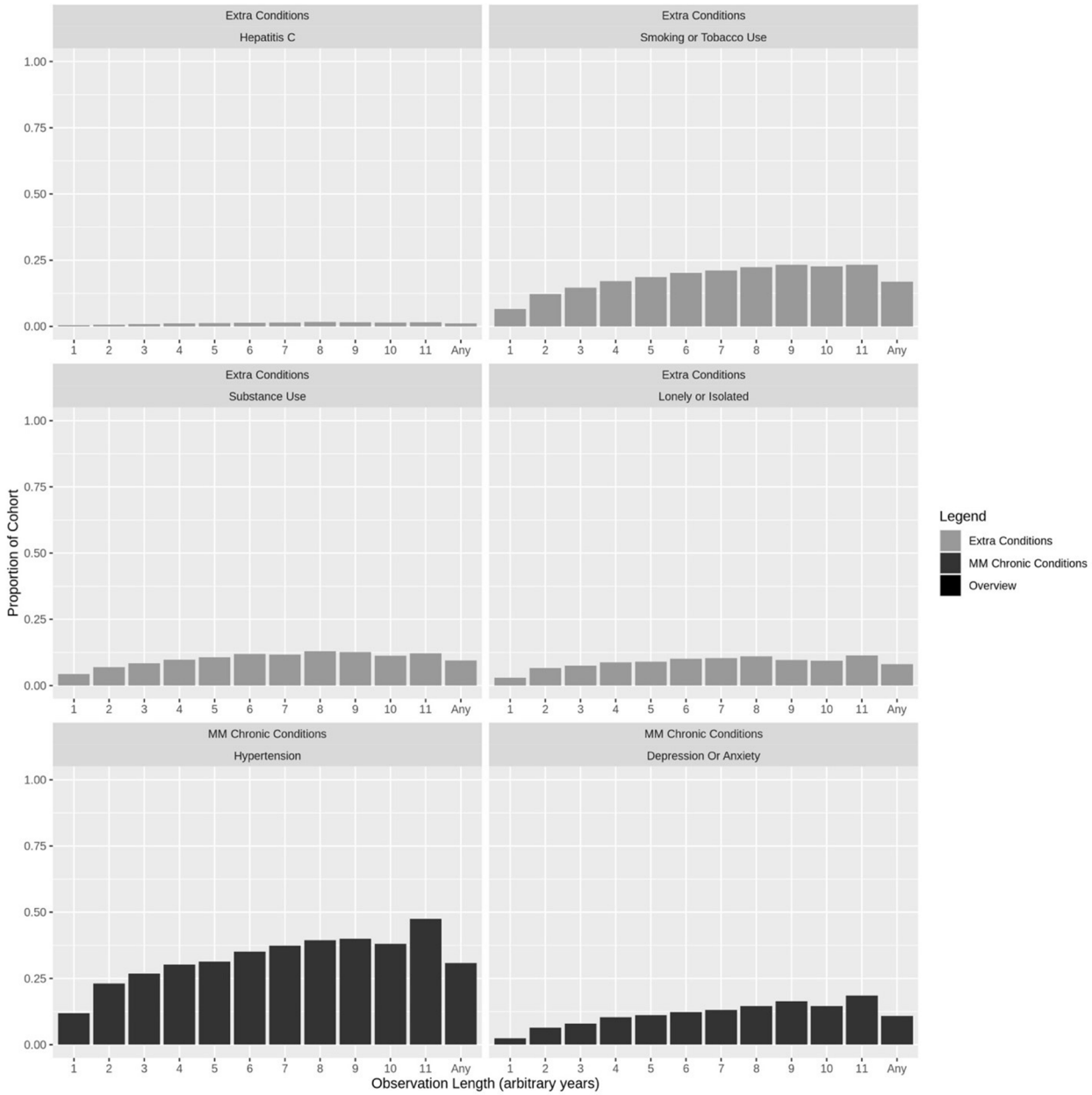
Appendix 1

Supplementary Figure 1: Cohort size by calendar- and observation-based time

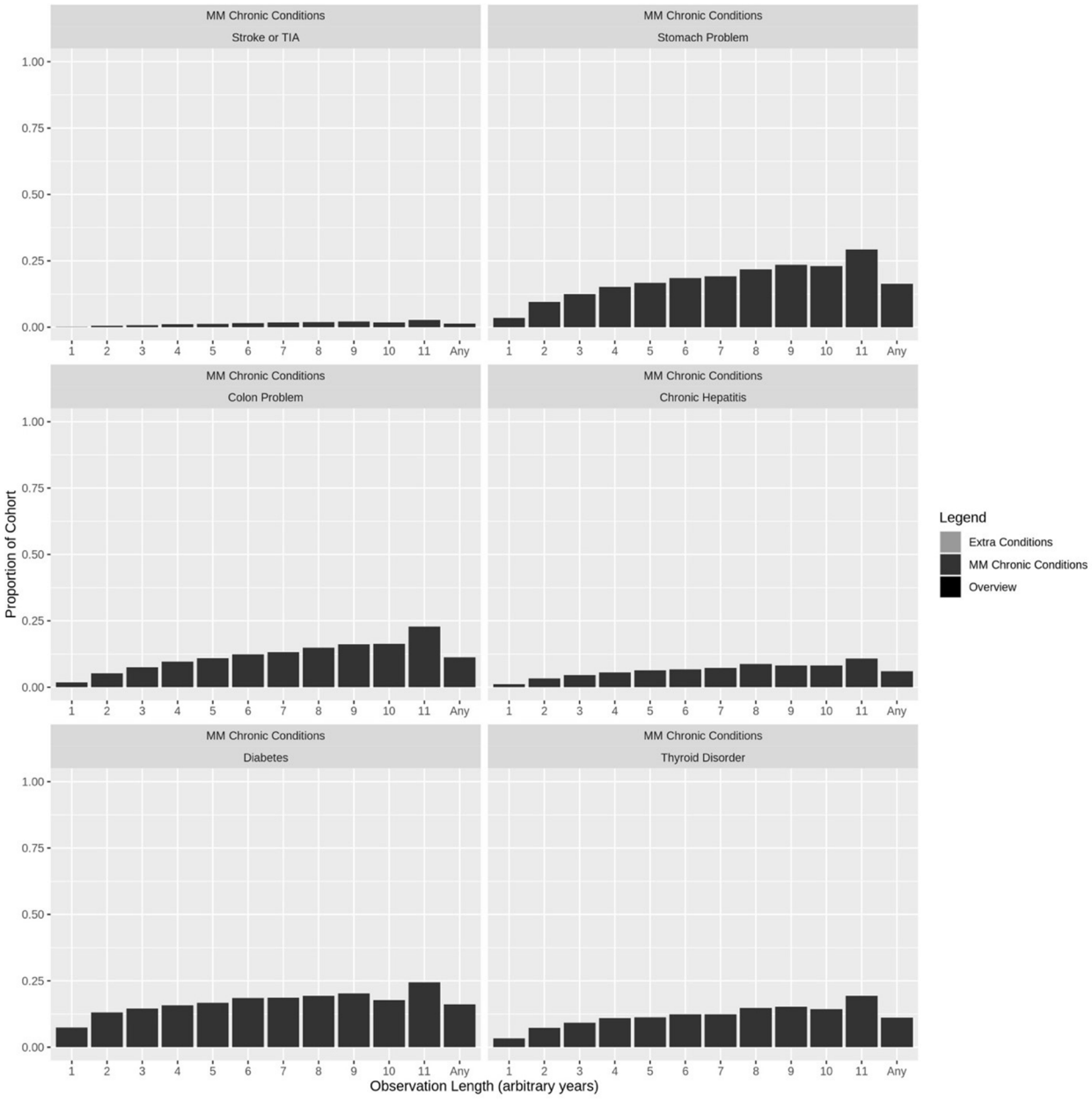


Active clients have at least one event during or after the year (calendar- or observation-based) of interest (gap years counted). The number of active observation years refers to the number of 365.25 day periods, counted from the first calendar date that an event was recorded for that client, that clients have at least one event recorded (gap years not counted). Length of observation refers to the number of years from the first to the last year that at least one event is recorded (gap years counted). Cumulative clients refers to the number of clients who have had at least one event during or before the year of interest.

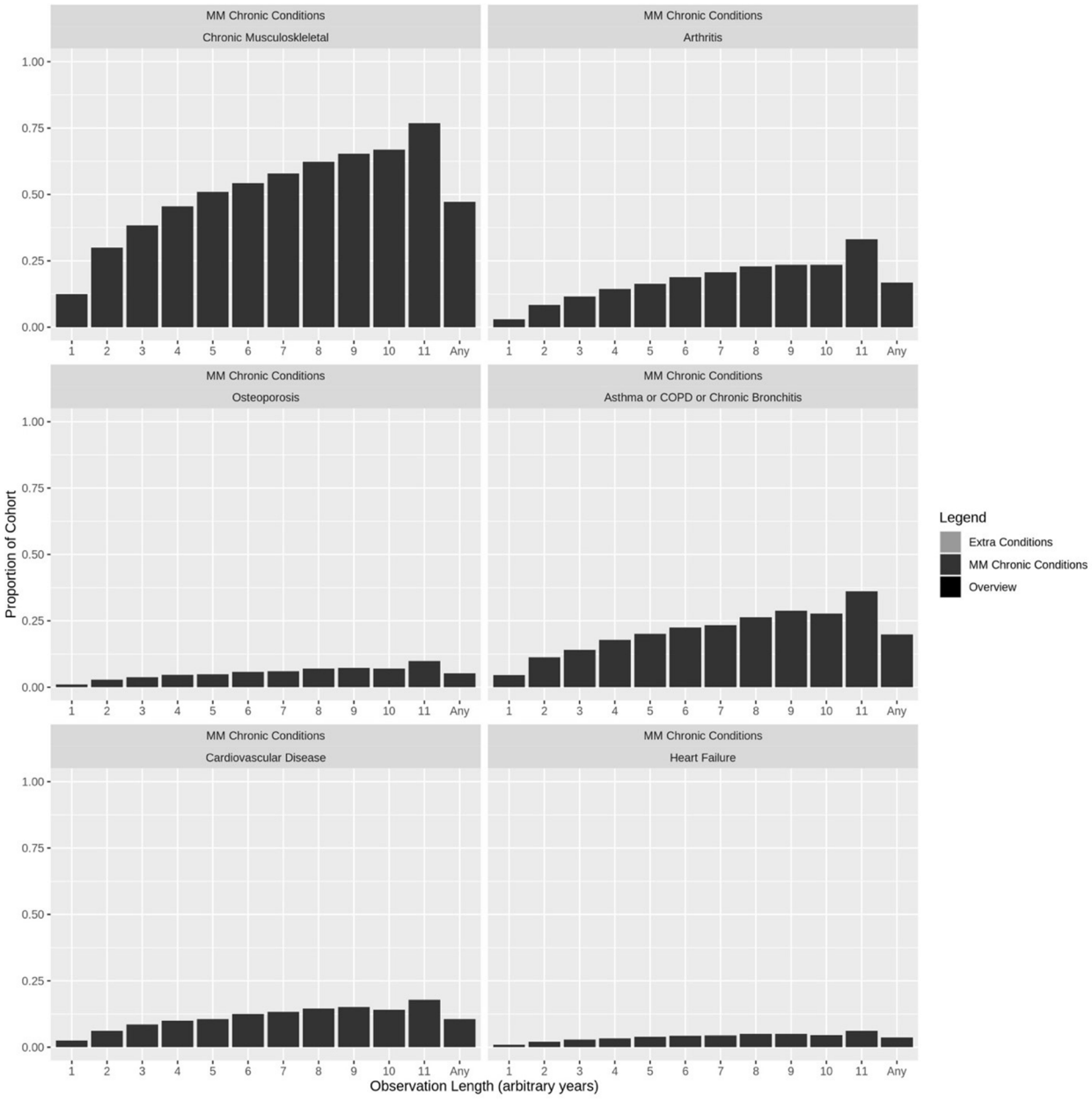
Supplementary Figure 2: Observation-based period prevalence



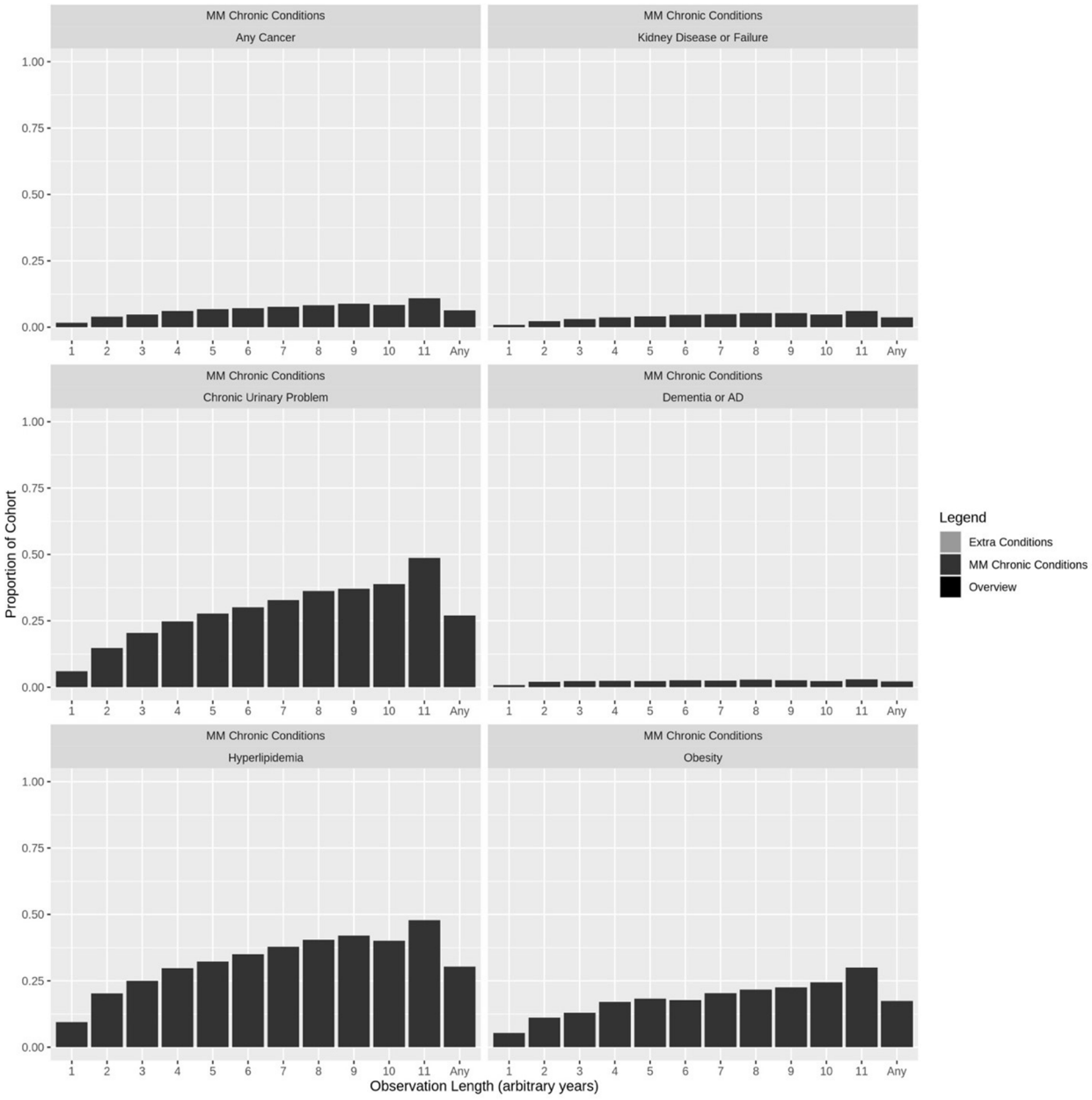
Supplementary Figure 2: Continued



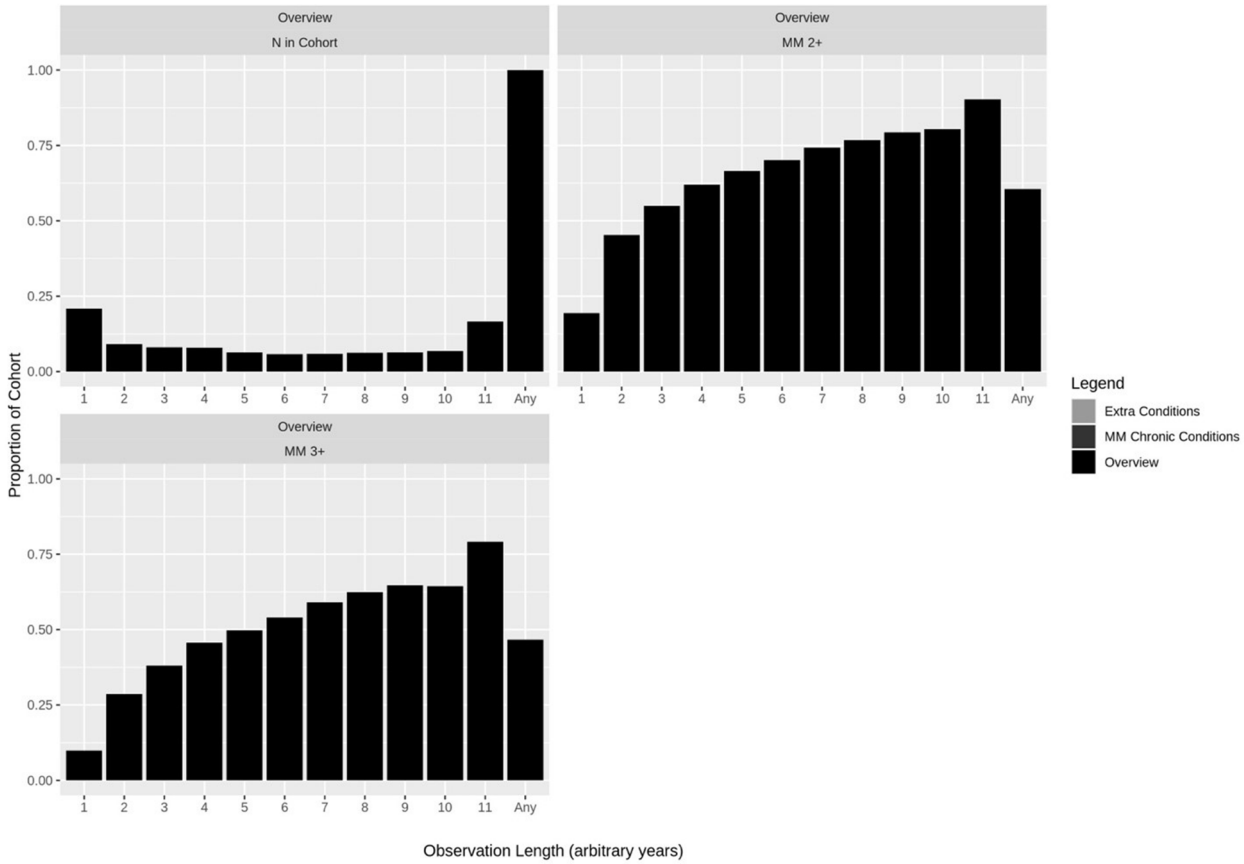
Supplementary Figure 2: Continued



Supplementary Figure 2: Continued

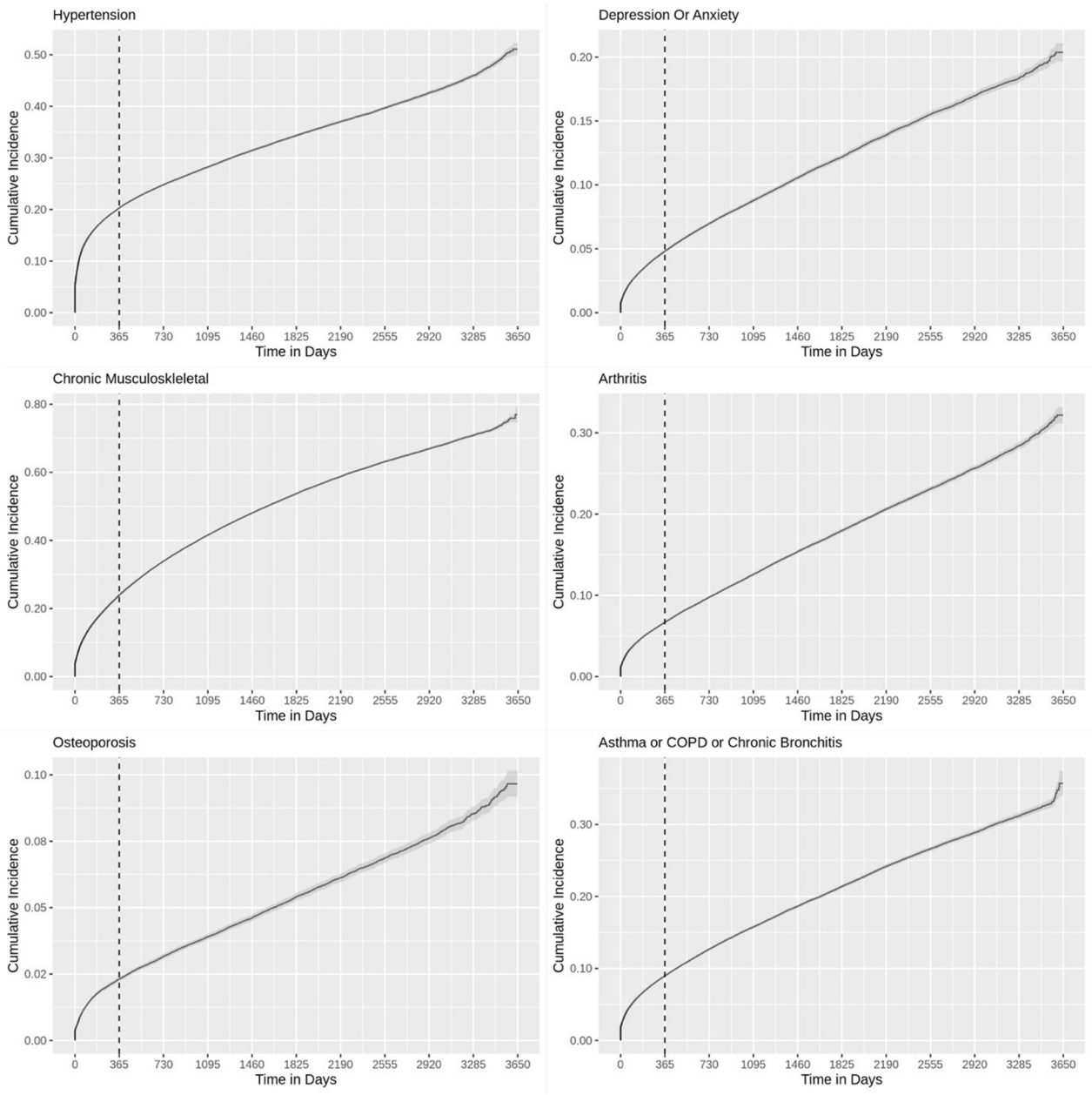


Supplementary Figure 2: Continued

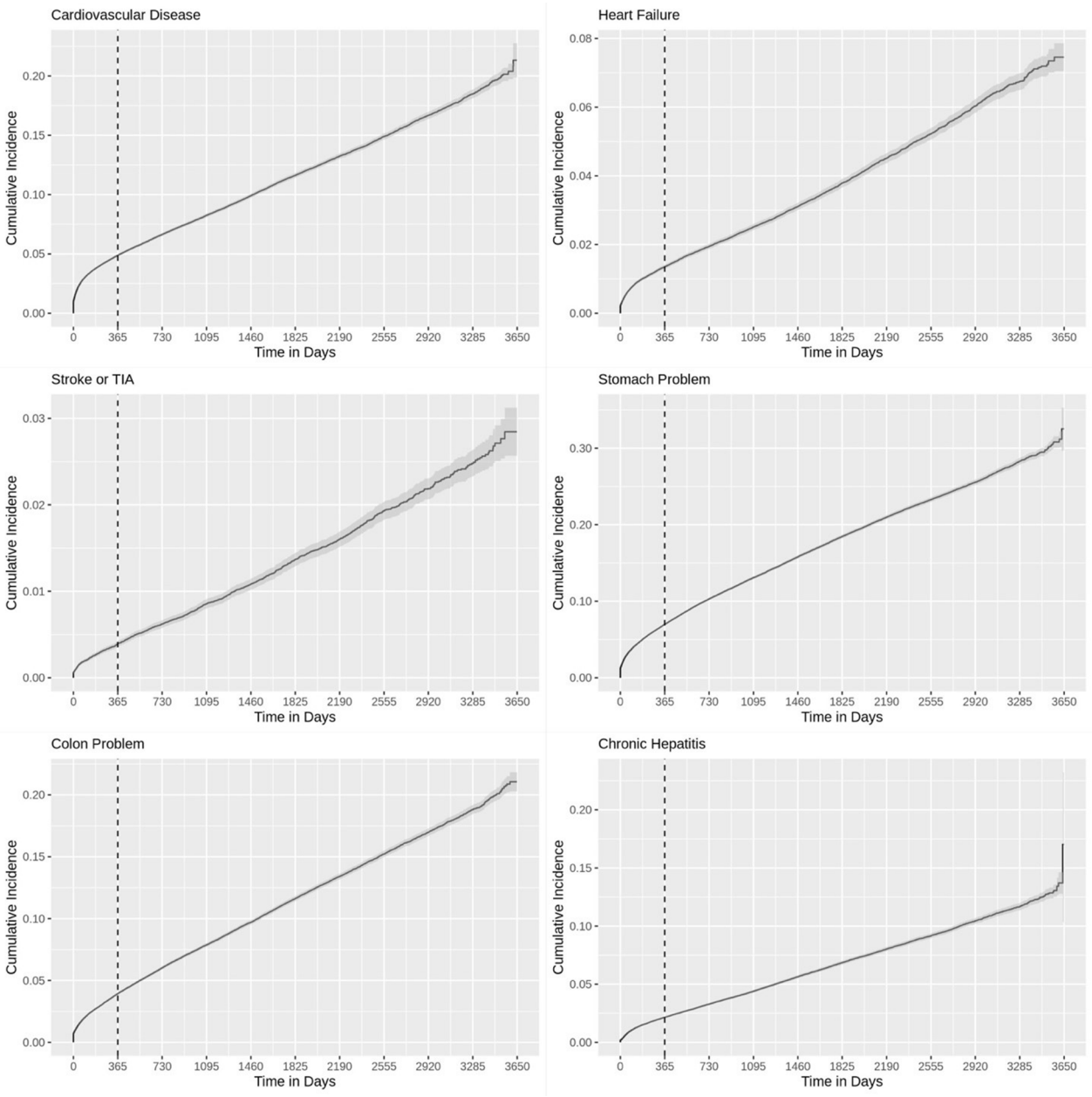


Each bar represents the proportion of clients within that observation-based cohort (years are arbitrary 365.25 day consecutive periods between the first and last recorded events) that have at least one indication of the condition of interest across their entire observation history. Conditions are grouped to represent 1) Extra conditions of interest to Alliance stakeholders, 2) 20 chronic conditions, which make up multimorbidity (MM) status, and 3) Overview indicators for the cohorts. *Legend:* COPD = Chronic Obstructive Pulmonary Disease; TIA = Transient Ischemic Attack; AD = Alzheimer's Disease.

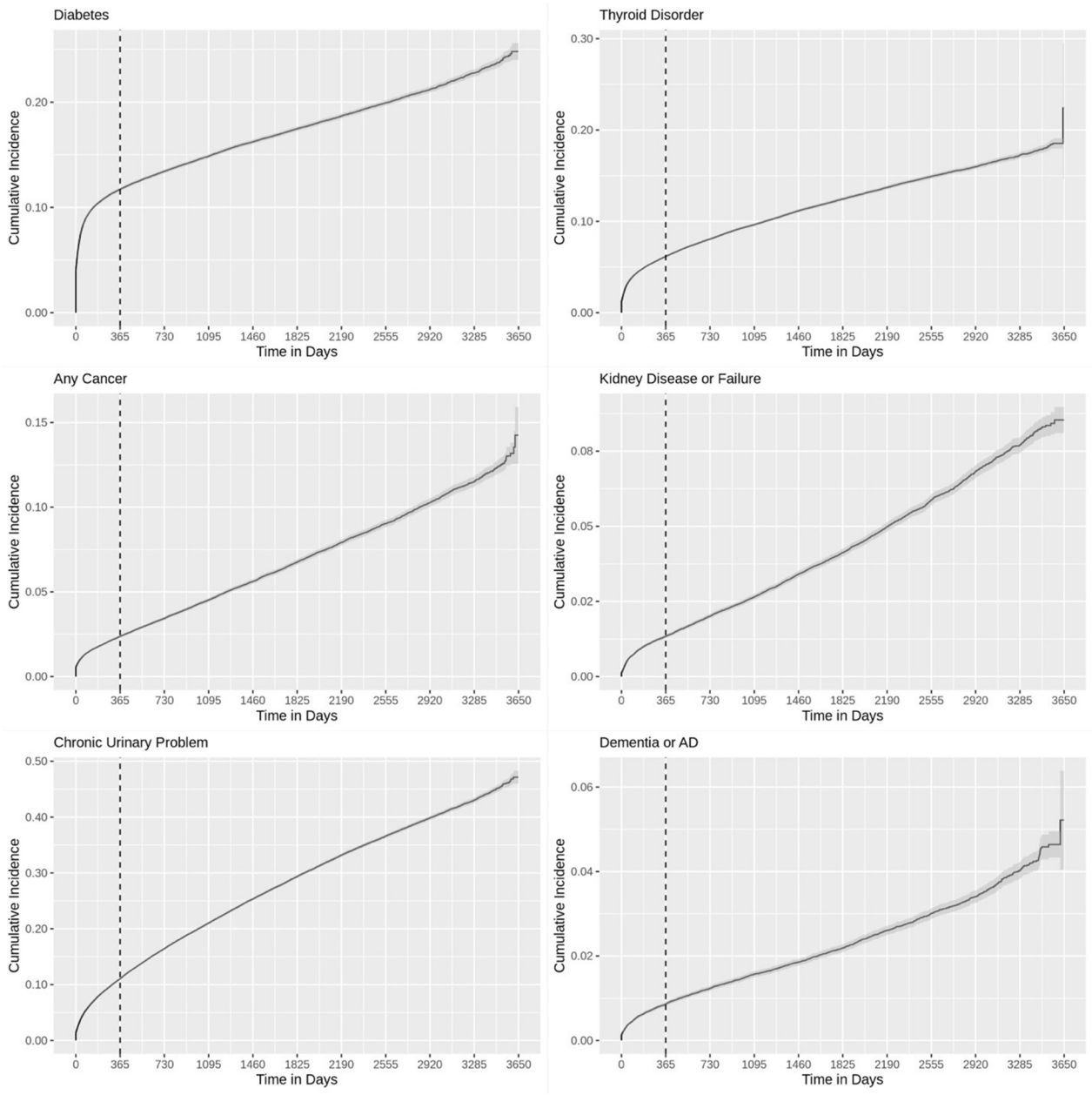
Supplementary Figure 3: Cumulative incidence



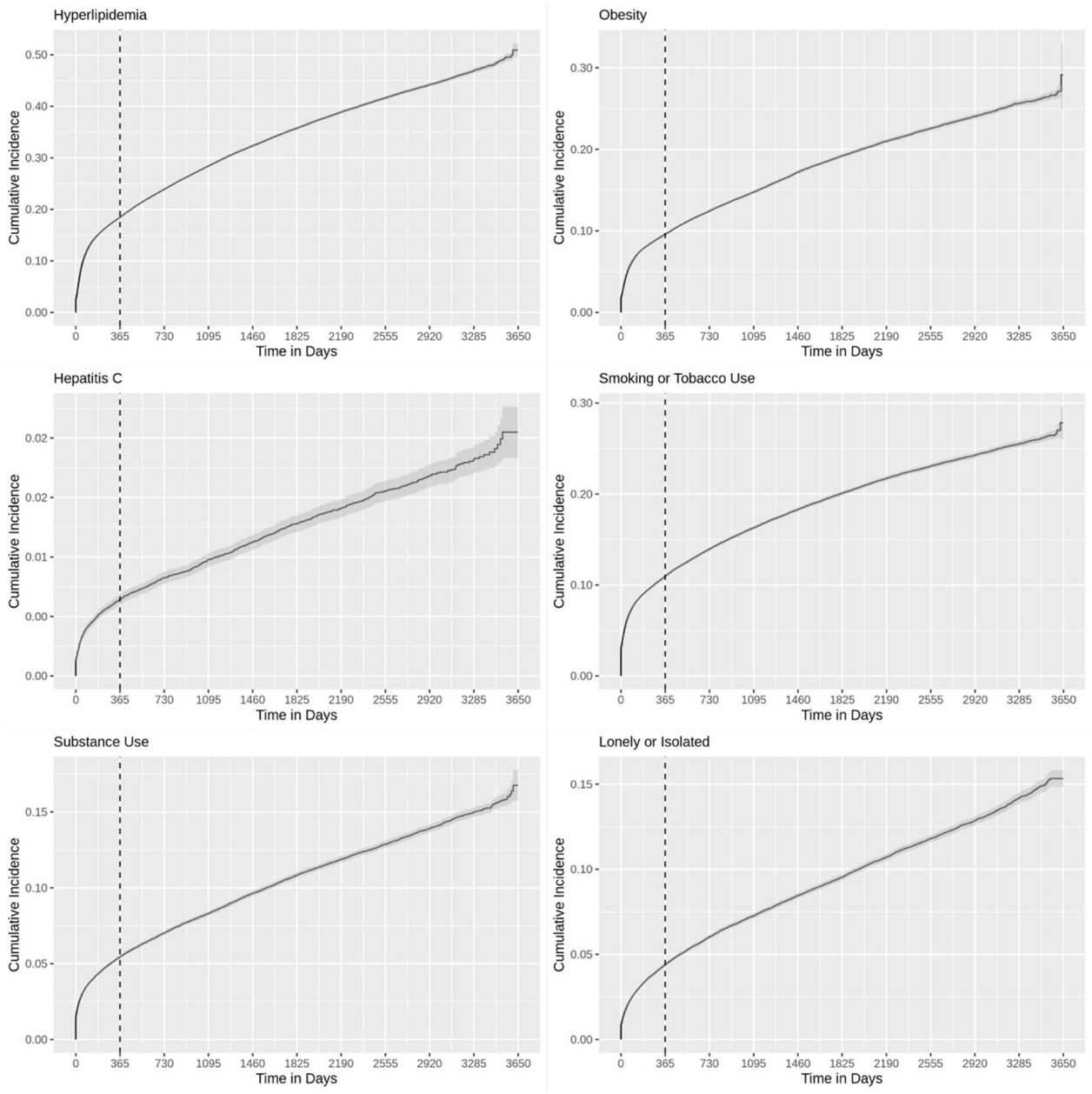
Supplementary Figure 3: Continued



Supplementary Figure 3: Continued



Supplementary Figure 3: Continued



Cumulative incidence plots by days of observation since the first recorded event. Clients eligible for this analysis must not have any care recorded in the first calendar-year of available data (2009).

Supplementary Appendix 2: The RECORD statement - checklist of items, extended from the STROBE statement, that should be reported in observational studies using routinely collected health data.

	Item No.	STROBE items	Location in manuscript where items are reported	RECORD items	Location in manuscript where items are reported
Title and abstract					
	1	(a) Indicate the study's design with a commonly used term in the title or the abstract (b) Provide in the abstract an informative and balanced summary of what was done and what was found	Abstract	RECORD 1.1: The type of data used should be specified in the title or abstract. When possible, the name of the databases used should be included. RECORD 1.2: If applicable, the geographic region and timeframe within which the study took place should be reported in the title or abstract. RECORD 1.3: If linkage between databases was conducted for the study, this should be clearly stated in the title or abstract.	Abstract
Introduction					
Background rationale	2	Explain the scientific background and rationale for the investigation being reported	Introduction		Introduction
Objectives	3	State specific objectives, including any prespecified hypotheses	Introduction		Introduction
Methods					
Study Design	4	Present key elements of study design early in the paper	Introduction, Methods – Study population and data source		Introduction, Methods – Study population and data source
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	Methods – Study population and data source; additional details specific to analyses are presented under the appropriate sub-heading		Methods – Study population and data source; additional details specific to analyses are presented under the appropriate sub-heading
Participants	6	*(a) Cohort study* - Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study* - Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study* - Give the eligibility criteria, and the sources and methods of selection of participants *(b) Cohort study* - For matched studies, give matching criteria and number of exposed and unexposed *Case-control study* - For matched studies, give matching criteria and the number of controls per case	Methods	RECORD 6.1: The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If this is not possible, an explanation should be provided. RECORD 6.2: Any validation studies of the codes or algorithms used to select the population should be referenced. If validation was conducted for this study and not published elsewhere, detailed methods and results should be provided. RECORD 6.3: If the study involved linkage of databases, consider use of a flow diagram or other graphical display to demonstrate the data linkage process, including the number of individuals with linked data at each stage.	Methods
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.	Methods, Supplementary Table 1	RECORD 7.1: A complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers should be provided. If these cannot be reported, an explanation should be provided.	Supplementary Table 1
Data sources/ measurement	8	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	Methods, Supplementary Table 1		Methods, Supplementary Table 1

Continued

Supplementary Appendix 2: Continued

	Item No.	STROBE items	Location in manuscript where items are reported	RECORD items	Location in manuscript where items are reported
Bias	9	Describe any efforts to address potential sources of bias	N/A		N/A
Study size	10	Explain how the study size was arrived at	Methods		Methods
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why	Methods, Supplementary Table 1		Methods, Supplementary Table 1
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding (b) Describe any methods used to examine subgroups and interactions (c) Explain how missing data were addressed (d) *Cohort study* - If applicable, explain how loss to follow-up was addressed *Case-control study* - If applicable, explain how matching of cases and controls was addressed *Cross-sectional study* - If applicable, describe analytical methods taking account of sampling strategy (e) Describe any sensitivity analyses	Methods		Methods
Data access and cleaning methods		..		RECORD 12.1: Authors should describe the extent to which the investigators had access to the database population used to create the study population. RECORD 12.2: Authors should provide information on the data cleaning methods used in the study.	Methods, Supplementary Table 1
Linkage		..		RECORD 12.3: State whether the study included person-level, institutional-level, or other data linkage across two or more databases. The methods of linkage and methods of linkage quality evaluation should be provided.	No linkage
Results					
Participants	13	(a) Report the numbers of individuals at each stage of the study (*e.g.*, numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed) (b) Give reasons for non-participation at each stage. (c) Consider use of a flow diagram	Supplementary Appendix 3	RECORD 13.1: Describe in detail the selection of the persons included in the study (*i.e.*, study population selection) including filtering based on data quality, data availability and linkage. The selection of included persons can be described in the text and/or by means of the study flow diagram.	Methods
Descriptive data	14	(a) Give characteristics of study participants (*e.g.*, demographic, clinical, social) and information on exposures and potential confounders (b) Indicate the number of participants with missing data for each variable of interest (c) *Cohort study* - summarise follow-up time (*e.g.*, average and total amount)	Results, Supplementary Appendix 3		

Continued

Supplementary Appendix 2: Continued

	Item No.	STROBE items	Location in manuscript where items are reported	RECORD items	Location in manuscript where items are reported
Outcome data	15	*Cohort study* - Report numbers of outcome events or summary measures over time *Case-control study* - Report numbers in each exposure category, or summary measures of exposure *Cross-sectional study* - Report numbers of outcome events or summary measures	Results, Supplementary Appendix 3		
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95% confidence interval). Make clear which confounders were adjusted for and why they were included (b) Report category boundaries when continuous variables were categorized (c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	Results, Supplementary Appendix 3		
Other analyses	17	Report other analyses done—e.g., analyses of subgroups and interactions, and sensitivity analyses	Results, Supplementary Appendix 1,3		
Discussion					
Key results	18	Summarise key results with reference to study objectives	Discussion		Discussion
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	Discussion	RECORD 19.1: Discuss the implications of using data that were not created or collected to answer the specific research question(s). Include discussion of misclassification bias, unmeasured confounding, missing data, and changing eligibility over time, as they pertain to the study being reported.	Discussion
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	Discussion		Discussion
Generalisability	21	Discuss the generalisability (external validity) of the study results	N/A		N/A
Other Information					
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	Declarations		Declarations
Accessibility of protocol, raw data, and programming code		..		RECORD 22.1: Authors should provide information on how to access any supplemental information such as the study protocol, raw data, or programming code.	Given the sensitive nature of the data, this information is not shared.

Appendix 3

Eligible Clients: Of the 881,129 adult clients in the Alliance EHR database in 2009-2019, 232,529 (26.4%) have ongoing primary care client indications, and 221,047 (25.1%) have at least one encounter in 2009-2019 (fully eligible).

Supplementary Table 1: Characteristic variable definitions

Variable Name	Definition	Source used to guide variable operationalization, if any
Sociodemographic Characteristics		
Age in 2015	2015 minus Year of Birth	
Geography	Geography of place of residence based on Forward Sortation Area: Rural if second digit is 0; Urban if any other valid digit; NA otherwise.	(1)
Sex	Categories as recorded in client characteristic table	
Gender	Collapsed client characteristic table categories	(2)
Sexual Orientation	Categories as recorded in client characteristic table	
Highest Level of Education Completed	Collapsed client characteristic table categories	(2) followed to the extent possible
Primary Spoken Language	Collapsed client characteristic table categories into the two official languages of Canada with remaining languages categorized as Other.	
Race and Ethnicity	Collapsed client characteristic table categories	(3,4)
Year Since Arrival in Canada	Cleaned free text entries from client characteristic table and collapsed into 5 years, 6 or more years, and None Recorded categories. None Recorded cannot differentiate between never-immigrated and never-asked.	
Household Income	Collapsed client characteristic table categories	Note: could not reliably follow guidelines in (2)
Household Composition	Categories as recorded in client characteristic table	
Stable Residence	Collapsed client characteristic table categories into Stable or Unstable (homeless, shelter, other temporary). Additional unstable residence situations were identified as the presence of at least one ENCODE-FM code: 8990, 9433, 9434, 9435, 9436, 9437, 9438, 9439, 9440, 9441, 9442, 9443, 9432, 8982, 8986, 9419, 9424, 8985, 9431, 9415, 9425, 9412, 9414, which were given priority.	List of codes are from an Alliance for Healthier Communities stakeholder
Food Insecurity	At least one ENCODE-FM code: 8972, 9782, 9802, 8971, 9568, 9805	List of codes are from an Alliance for Healthier Communities stakeholder
Clinical Characteristics		
Hypertension	At least one ICD-10 code: i10,i11,i12,i13,i14,i15	(5)
Depression or Anxiety	At least one ICD-10 code: f33,f40,f41	(5)
Chronic Musculoskeletal Conditions causing pain or limitation	At least one ICD-10 code: m40,m41,m42,m43,m44,m45, m46,m47,m48,m49,m50,m51,m52,m53,m54,m60, m61,m62, m63,m65,m66,m67,m68,m70,m71,m72,m73, m74,m75,m76,m77,m78,m79	(5)
Arthritis and/or Rheumatoid Arthritis	At least one ICD-10 code: m05.9,m13.0,m13.9, m15,m16,m17,m18,m19	(5)
Osteoporosis	At least one ICD-10 code: m81	(5)
Asthma, Chronic Obstructive Pulmonary Disease, or Chronic Bronchitis	At least one ICD-10 code: j40,j41,j42,j43,j44,j45,j46	(5)
Cardiovascular Disease (angina, myocardial infarction, atrial fibrillation, poor circulation in the lower limbs)	At least one ICD-10 code: i20,i25,i48,i70,i71,i72,i73, i74,i75,i76,i77,i78,i79	(5)

Continued

Supplementary Table 1: Continued

Variable Name	Definition	Source used to guide variable operationalization, if any
Heart Failure (including valve problems or replacement)	At least one ICD-10 code: i05,i06,i07,i08,i09,i34,i35, i36,i37,i38,i39,i42,i43,i50	(5)
Stroke and Transient Ischemic Attack	At least one ICD-10 code: g45,i62	(5)
Stomach Problem (irritable bowel, Chron's disease, ulcerative colitis, diverticulosis)	At least one ICD-10 code: k21,k25.7,k29.5	(5)
Colon Problem	At least one ICD-10 code: k50,k51,k52,k57,k58	(5)
Chronic Hepatitis	At least one ICD-10 code: k70,k71,k72,k73,k74, k75,k76,k77	(5)
Diabetes	At least one ICD-10 code: e10,e11,e12,e13,e14	(5)
Thyroid Disorder	At least one ICD-10 code: e00,e01,e02,e03,e04,e05, e06,e07	(5)
Any Cancer (including melanoma, but excluding other skin cancers)	At least one ICD-10 code: c00,c01,c02,c03,c04, c05,c06,c07,c08,c09,c10,c11,c12,c13,c14,c15,c16,c17,c18, c19,c20,c21,c22,c23,c24,c25,c26,c27,c28,c29,c30,c31,c32, c33,c34,c35,c36,c37,c38,c39,c40,c41,c42,c43,c44,c45,c46, c47,c48,c49,c50,c51,c52,c53,c54,c55,c56,c57,c58,c59,c60, c61,c62,c63,c64,c65,c66,c67,c68,c69,c70,c71,c72,c73,c74, c75,c76,c77,c78,c79,c80,c81,c82,c83,c84,c85,c86,c87,c88, c89,c90,c91,c92,c93,c94,c95,c96,c97	(5) modified by removing the 5 year restriction; taking any cancer indication within the 10 year period
Kidney Disease or Failure	At least one ICD-10 code: n18,n19	(5)
Chronic Urinary Problem	At least one ICD-10 code: n03,n11,n18,n20,n21,n22,n23,n25,n26,27,n28,n29, n30,n31,n32,n33,n34,n35,n36,n37,n38,n39,n40,n41, n42,n43,n44,n45,n46,n47,n48,n49,n50,n51+B38	(5)
Dementia or Alzheimer's Disease	At least one ICD-10 code: f00,f01,f02,f03	(5)
Hyperlipidemia (high cholesterol)	At least one ICD-10 code: e78	(5)
Obesity	At least one ICD-10 code: e66	(5) ICD-10 only; no BMI
Hepatitis C	At least one ICD-10 code: b18.2,b19.20,b19.21	(6)
Smoking or Tobacco Use	At least one ENCODE-FM code: 10072,5520, 679,9910,5339,5340,5341,5342,5343,5344,5345,5346, 5347,5348,5349	JKK selected relevant ENCODE-FM codes based on manual review
Substance Use	At least one ENCODE-FM code: 5304,10004,10005,5305,5306,5307,9754,5308,5309, 5310,5311,5312,5313,5314,5315,5316,5317,5318,5319, 5320,5321,5322,5323,5324,5325,5326,5327,5328,5329, 5330,5331,5332,5333,5334,5335,5336,5337,5338,5350, 5351,5352,5353,5354,5355,5356,5357,5358,5359,5360, 5361,5362,5363,5364,5365,5366,5367,5368,5369,5370, 5371,10007,5372,5373,5374,5375,5376,5377,5378,5379, 5380,5381,5382,5383,5384,5385,5386,5387,5388,5389, 5390,5391,5392,5393,5394,5395,5396,5397,5398,5399, 5400,9845,5401,9844,5401,5402,5403,5404,5405,5406, 5407,5408,5409,5410,5411,5412,5413,5414,5415,5416, 5417,5418,5419,5420,5421,5422,5423,5424,5425,5426, 5427,5428,5429,5430,5431,5432,5433,5434,5435,5436, 5437,5438,5439,5440,5441,5442,5443,5444,5445,5446, 5447,5448,5449,9277,9278,5450,5451,5452,5453,5454, 5455,5456,5457,5458,5459,5460,5461,5462,5463,5464, 5465,5466,5467,5468,5469,5470,5471,5472,5473,5474, 5475,5476,5477,5478,5479,5480,5481 or recorded in Disabilities Table	JKK selected relevant ENCODE-FM codes based on manual review

Continued

Supplementary Table 1: Continued

Variable Name	Definition	Source used to guide variable operationalization, if any
Lonely or Isolated	At least one ENCODE-FM code: 5138, 5139, 9265, 9267, 9268, 9512	List of codes are from an Alliance for Healthier Communities stakeholder
Health Care Use Characteristics		
# Years of Observation	Based on records in the service event table: Ceiling of number of days from first to last recorded event divided by 365.25	
# Provider Types Seen	Number of unique provider types recorded in providers involved table. Provider types were maintained as entered except Other, Unknown, and Undefined were collapsed	
# Internal Referrals	Number of records in the internal referrals table	
# External Referrals	Number of records in the external referrals table	
Avg. # Days per Year	Based on records in the service event table: Sum of unique calendar days with at least one event recorded divided by Number of Years of Observation	
Max # Days per Year	Based on records in the service event table: Maximum of number of unique calendar days care is accessed in a single calendar year	
Avg. # Events per Day	Based on records in the service event table: Sum of events divided by number of calendar days care is accessed at least once	
Max # Events per Day	Based on records in the service event table: Maximum number of events recorded in a single calendar day	

Legend: # = Number; Avg. = Average; CHC = Community Health Centre; ENCODE-FM = Electronic Nomenclature and Classification Of Disorders and Encounters for Family Medicine; ICD = International Classification of Disease; SD = Standard Deviation; UAR = Urban at Risk.

References:

(1) Canada Post. Addressing guidelines - Forward Sortation Area (FSA) [Internet]. Canada Post. 2022 [cited 2022 Feb 8]. Available from: https://www.canadapost-postescanada.ca/cpc/en/support/articles/addressing-guidelines/postal-codes.page.

(2) CIHI. In Pursuit of Health Equity: Defining Stratifiers for Measuring Health Inequality - A Focus on Age, Sex, Gender, Income, Education and Geographic Location. Ottawa, ON: Canadian Institute for Health Information; 2018 Apr. Available from: https://www.cihi.ca/sites/default/files/document/defining-stratifiers-measuring-health-inequalities-2018-en-web.pdf.

(3) CIHI. Proposed Standards for Race-Based and Indigenous Identity Data Collection and Health Reporting in Canada. Ottawa, ON: Canadian Institute for Health Information; 2020. Available from: https://www.cihi.ca/en/proposed-standards-for-race-based-and-indigenous-identity-data.

(4) Flanagin A, Frey T, Christiansen SL, AMA Manual of Style Committee. Updated guidance on the reporting of race and ethnicity in medical and science journals. JAMA. 2021 Aug 17;326(7):621–7.

(5) Fortin M, Almirall J, Nicholson K. Development of a research tool to document self-reported chronic conditions in primary care. J Comorb. 2017 Jan 1;7(1):117–23.

(6) Support Path. Hepatitis C ICD-10 Codes. Gilead Sciences; 2015 [cited 2020 Sep 25]. Available from: https://www.cvph.org/data/files/mysupportpath.pdf.

Supplementary Table 2: Sociodemographic characteristics with sub-strata

Characteristic	Values	All clients n (%)	UAR & MM n (%)	UAR & Non-MM n (%)	Non-UAR & MM n (%)	Non-UAR & Non-MM n (%)
Number of clients		221047	19237	16761	83935	101114
Age in 2015	25-34	55505 (25.11)	1864 (9.69)	6112 (36.47)	7482 (8.91)	40047 (39.61)
	35-44	45646 (20.65)	3154 (16.40)	4386 (26.17)	12388 (14.76)	25718 (25.43)
	45-54	44653 (20.20)	4784 (24.87)	3402 (20.30)	19198 (22.87)	17269 (17.08)
	55-64	37848 (17.12)	4935 (25.65)	1855 (11.07)	20643 (24.59)	10415 (10.30)
	65-74	23162 (10.48)	2952 (15.35)	692 (4.13)	14828 (17.67)	4690 (4.64)
	75+	14233 (6.44)	1548 (8.05)	314 (1.87)	9396 (11.19)	2975 (2.94)
Geography	Rural	49275 (22.29)	3479 (18.08)	2652 (15.82)	23339 (27.81)	19805 (19.59)
	Urban	167728 (75.88)	15291 (79.49)	13247 (79.03)	59720 (71.15)	79470 (78.59)
	Missing	4044 (1.83)	467 (2.43)	862 (5.14)	876 (1.04)	1839 (1.82)
Sex	Female	127070 (57.49)	10647 (55.35)	8052 (48.04)	49299 (58.73)	59072 (58.42)
	Male	93294 (42.21)	8561 (44.50)	8590 (51.25)	34563 (41.18)	41580 (41.12)
	Other	331 (0.15)	3 (0.02)	40 (0.24)	16 (0.02)	272 (0.27)
	Missing	352 (0.16)	26 (0.14)	79 (0.47)	57 (0.07)	190 (0.19)
Gender	Female	41352 (18.71)	3352 (17.42)	2157 (12.87)	18479 (22.02)	17364 (17.17)
	Gender diverse	340 (0.15)	52 (0.27)	60 (0.36)	92 (0.11)	136 (0.13)
	Male	29366 (13.28)	2425 (12.61)	2160 (12.89)	12308 (14.66)	12473 (12.34)
	Prefer not to answer	1001 (0.45)	37 (0.19)	14 (0.08)	339 (0.40)	611 (0.60)
	Missing	148988 (67.40)	13371 (69.51)	12370 (73.80)	52717 (62.81)	70530 (69.75)
Sexual Orientation	Bisexual	1578 (0.71)	141 (0.73)	144 (0.86)	549 (0.65)	744 (0.74)
	Gay	708 (0.32)	94 (0.49)	98 (0.58)	212 (0.25)	304 (0.30)
	Heterosexual	57065 (25.82)	4703 (24.45)	3744 (22.34)	24402 (29.07)	24216 (23.95)
	Lesbian	485 (0.22)	45 (0.23)	25 (0.15)	199 (0.24)	216 (0.21)
	Queer	323 (0.15)	14 (0.07)	20 (0.12)	77 (0.09)	212 (0.21)
	Two-Spirit	128 (0.06)	40 (0.21)	40 (0.24)	21 (0.03)	27 (0.03)
	Other	246 (0.11)	21 (0.11)	13 (0.08)	122 (0.15)	90 (0.09)
	Do not know	924 (0.42)	113 (0.59)	88 (0.53)	372 (0.44)	351 (0.35)
	Prefer not to answer	7561 (3.42)	565 (2.94)	312 (1.86)	3513 (4.19)	3171 (3.14)
	Missing	152029 (68.78)	13501 (70.18)	12277 (73.25)	54468 (64.89)	71783 (70.99)
Highest Level of Education	Post-secondary or equivalent	84888 (38.40)	6463 (33.60)	5593 (33.37)	29300 (34.91)	43532 (43.05)
	Secondary or equivalent	61831 (27.97)	6656 (34.60)	5127 (30.59)	25961 (30.93)	24087 (23.82)
	Less than high school	18941 (8.57)	1886 (9.80)	1380 (8.23)	8732 (10.40)	6943 (6.87)
	Other	8507 (3.85)	384 (2.00)	335 (2.00)	3694 (4.40)	4094 (4.05)
	Do not know	4860 (2.20)	734 (3.82)	584 (3.48)	1616 (1.93)	1926 (1.90)
	Prefer not to answer	2950 (1.33)	273 (1.42)	149 (0.89)	1312 (1.56)	1216 (1.20)
	Missing	39070 (17.67)	2841 (14.77)	3593 (21.44)	13320 (15.87)	19316 (19.10)
Primary Language	English	167163 (75.62)	17036 (88.56)	14622 (87.24)	62563 (74.54)	72942 (72.14)
	French	22547 (10.20)	554 (2.88)	390 (2.33)	10537 (12.55)	11066 (10.94)
	Other	26847 (12.15)	1473 (7.66)	1475 (8.80)	9237 (11.00)	14662 (14.50)
	Missing	4490 (2.03)	174 (0.90)	274 (1.63)	1598 (1.90)	2444 (2.42)
Race and Ethnicity	Black	8861 (4.01)	337 (1.75)	388 (2.31)	3420 (4.07)	4716 (4.66)
	East/SouthEast Asian	3739 (1.69)	248 (1.29)	236 (1.41)	1297 (1.55)	1958 (1.94)
	Indigenous	2944 (1.33)	838 (4.36)	739 (4.41)	803 (0.96)	564 (0.56)
	Latino	4350 (1.97)	102 (0.53)	104 (0.62)	1606 (1.91)	2538 (2.51)
	Middle Eastern	2046 (0.93)	149 (0.77)	195 (1.16)	689 (0.82)	1013 (1.00)
	Other	567 (0.26)	79 (0.41)	69 (0.41)	227 (0.27)	192 (0.19)
	South Asian	3597 (1.63)	232 (1.21)	91 (0.54)	1620 (1.93)	1654 (1.64)
	White	38464 (17.40)	2661 (13.83)	1870 (11.16)	18843 (22.45)	15090 (14.92)
	Do not know	838 (0.38)	91 (0.47)	60 (0.36)	396 (0.47)	291 (0.29)
	Prefer not to answer	2649 (1.20)	165 (0.86)	96 (0.57)	1348 (1.61)	1040 (1.03)
	Missing	152992 (69.21)	14335 (74.52)	12913 (77.04)	53686 (63.96)	72058 (71.26)
Years Since Arrival in Canada	0 to 5 years	13 654 (6.18)	315 (1.64)	876 (5.23)	2732 (3.25)	9731 (9.62)
	6+ years	51815 (23.44)	2863 (14.88)	2077 (12.39)	19859 (23.66)	27016 (26.72)
	None recorded	155578 (70.38)	16059 (83.48)	13808 (82.38)	61344 (73.09)	64367 (63.66)

Continued

Supplementary Table 2: Continued

Characteristic	Values	All clients n (%)	UAR & MM n (%)	UAR & Non-MM n (%)	Non-UAR & MM n (%)	Non-UAR & Non-MM n (%)
Household Income	$0 to $14,999	40519 (18.33)	4476 (23.27)	4253 (25.37)	13281 (15.82)	18509 (18.31)
	$15,000 to $24,999	21102 (9.55)	2095 (10.89)	1460 (8.71)	8986 (10.71)	8561 (8.47)
	$25,000 to $39,999	20877 (9.44)	1772 (9.21)	1216 (7.25)	8964 (10.68)	8925 (8.83)
	$40,000 to $59,999	17245 (7.80)	1455 (7.56)	966 (5.76)	7216 (8.60)	7608 (7.52)
	$60,000 or more	28494 (12.89)	2092 (10.87)	1770 (10.56)	10776 (12.84)	13856 (13.7)
	Do not know	15408 (6.97)	1301 (6.76)	1357 (8.10)	4963 (5.91)	7787 (7.70)
	Prefer not to answer	27621 (12.50)	2437 (12.67)	1693 (10.10)	12453 (14.84)	11038 (10.92)
	Missing	49781 (22.52)	3609 (18.76)	4046 (24.14)	17296 (20.61)	24830 (24.56)
Household Composition	Couple with children	53398 (24.16)	3280 (17.05)	3479 (20.76)	17433 (20.77)	29206 (28.88)
	Couple without child	39664 (17.94)	3907 (20.31)	2038 (12.16)	19043 (22.69)	14676 (14.51)
	Extended family	7632 (3.45)	578 (3.00)	545 (3.25)	3003 (3.58)	3506 (3.47)
	Grandparents with grandchild(ren)	1746 (0.79)	187 (0.97)	60 (0.36)	996 (1.19)	503 (0.50)
	Siblings	1622 (0.73)	140 (0.73)	110 (0.66)	529 (0.63)	843 (0.83)
	Single parent	14445 (6.53)	1344 (6.99)	1183 (7.06)	5004 (5.96)	6914 (6.84)
	Sole member	32782 (14.83)	4503 (23.41)	2942 (17.55)	14094 (16.79)	11243 (11.12)
	Unrelated housemates	8622 (3.90)	669 (3.48)	898 (5.36)	2180 (2.60)	4875 (4.82)
	Other	8913 (4.03)	788 (4.10)	688 (4.10)	3414 (4.07)	4023 (3.98)
	Do not know	2475 (1.12)	301 (1.56)	342 (2.04)	978 (1.17)	854 (0.84)
	Prefer not to answer	3727 (1.69)	262 (1.36)	229 (1.37)	1665 (1.98)	1571 (1.55)
	Missing	46021 (20.82)	3278 (17.04)	4247 (25.34)	15596 (18.58)	22900 (22.65)
Stable Residence	True	199349 (90.18)	14813 (77.00)	13414 (80.03)	75666 (90.15)	95456 (94.40)
Food Insecurity	True	10985 (4.97)	2066 (10.74)	881 (5.26)	5257 (6.26)	2781 (2.75)

Legend: CHC = Community Health Centre; MM = Multimorbidity; UAR = Urban at Risk.

Supplementary Table 3: Health care use characteristics

Measure	Value	All clients	UAR CHC	Rural CHC	Multimorbidity
# Years of Observation	Min, Median, Max	(1, 5, 11)	(1, 6, 11)	(1, 7, 11)	(1, 8, 11)
	Mean (SD)	5.6 (3.7)	6.1 (3.8)	6.7 (3.6)	7.4 (3.3)
11 Years of Observation	n (%)	36 724 (16.6)	7976 (22.2)	5374 (25)	29 062 (28.2)
# Provider Types Seen	Min, Median, Max	(0, 4, 19)	(0, 5, 19)	(0, 5, 14)	(0, 6, 19)
	Mean (SD)	4.5 (2.3)	5.1 (2.6)	4.8 (2.1)	5.8 (2.2)
# Internal Referrals	Min, Median, Max	(0, 0, 300)	(0, 1, 300)	(0, 0, 51)	(0, 1, 300)
	Mean (SD)	1.7 (4.3)	2.8 (7.0)	1.4 (2.8)	2.8 (5.7)
# External Referrals	Min, Median, Max	(0, 1, 309)	(0, 2, 309)	(0, 1, 46)	(0, 3, 309)
	Mean (SD)	2.9 (4.5)	3.8 (5.9)	2.5 (3.2)	4.8 (5.5)
Avg. # Days/Year	Min, Median, Max	(0.2, 6, 176.9)	(0.2, 6.9, 129.7)	(0.2, 6.2, 120.3)	(0.3, 9.2, 176.9)
	Mean (SD)	8 (7.4)	9.4 (8.9)	8 (6.7)	11.4 (8.4)
Max # Days/Year	Min, Median, Max	(1, 10, 349)	(1, 12, 245)	(1, 11, 349)	(1, 17, 349)
	Mean (SD)	13.7 (13)	16.8 (16.6)	14.2 (12.1)	20.3 (14.4)
Avg. # Events/Day	Min, Median, Max	(1, 1.2, 66)	(1, 1.2, 29)	(1, 1.2, 31)	(1, 1.2, 31)
	Mean (SD)	1.3 (0.5)	1.3 (0.3)	1.3 (0.3)	1.3 (0.3)
Max # Events/Day	Min, Median, Max	(1, 3, 635)	(1, 3, 635)	(1, 3, 224)	(1, 3, 635)
	Mean (SD)	3.9 (7.8)	4.1 (8.2)	3.8 (6.1)	5.5 (10.7)

Legend: # = Number; Avg. = Average; CHC = Community Health Centre; SD = Standard Deviation; UAR = Urban at Risk.

Supplementary Table 4: Provider type counts

Type	Provider type	Number of events	% of events
Provider Involved in Care	Physician	3693760	30.1
	Nurse Practitioner (RN-EC)	2608238	21.3
	Nurse	2475621	20.2
	Registered Practical Nurse (RPN)	990144	8.1
	Social worker	452641	3.7
	Other/Unknown/Undefined	448761	3.7
	Dietitian/Nutritionist	268395	2.2
	Chiropodist	259101	2.1
	Counselor	212799	1.7
	Physiotherapist	171291	1.4
Internal Referral	Other/Unknown/Undefined	100649	26.7
	Physician	73070	19.4
	Nurse Practitioner (RN-EC)	37333	9.9
	Dietitian/Nutritionist	30670	8.1
	Nurse	29326	7.8
	Social worker	28357	7.5
	Physiotherapist	11210	3.0
	Chiropractor	9881	2.6
	Chiropodist	9741	2.6
	Counselor	6068	1.6
External Referral	Other/Unknown/Undefined	183804	28.5
	Dermatologist	41388	6.4
	Surgeon - general	40736	6.3
	Gastroenterologist	33737	5.2
	Surgeon - specialty (eye, heart, brain, etc.)	29370	4.6
	Physiotherapist	27639	4.3
	Ear Nose Throat (E.N.T.) specialist	25791	4.0
	Urologist	22546	3.5
	Gynecologist	21701	3.4
	Cardiologist	20592	3.2

Supplementary Table 5: Time series clustering of care access frequency

Analysis	# Clusters (Silhouette Score)	Cluster ID	# Clients	% Clients	Medoid
Short Term by Year	K = 2 (SS = 0.502)	1	11 552	30.5	20, 8
		2	26 368	69.5	6, 2
	K = 3 (SS = 0.301)	1	15 067	39.7	12, 3
		2	16 791	44.3	4, 1
		3	6062	16.0	24, 9
	K = 4 (SS = 0.142)	1	12 931	34.1	5, 1
		2	13 063	34.4	12, 3
		3	5602	14.8	1, 2
		4	6324	16.7	25, 8
	K = 5 (SS = 0.211)	1	3639	9.6	31, 8
		2	11 533	30.4	8, 1
		3	11 155	29.4	3, 2
		4	7722	20.4	12, 5
		5	3871	10.2	17, 11
Short Term by Quarter	K = 2 (SS = 0.541)	1	6068	16.0	10, 5, 5, 7, 3
		2	31 852	84.0	3, 1, 0, 1, 2
	K = 3 (SS = 0.249)	1	10 780	28.4	6, 3, 1, 2, 1
		2	20 431	53.9	2, 0, 0, 0, 1
		3	6709	17.7	6, 1, 3, 4, 3
	K = 4 (SS = 0.044)	1	14 389	37.9	3, 0, 1, 1, 1
		2	8939	23.6	6, 1, 0, 1, 2
		3	9072	23.9	2, 1, 0, 1, 2
		4	5520	14.6	9, 4, 2, 5, 3
	K = 5 (SS = 0.121)	1	4163	11.0	11, 8, 4, 5, 2
		2	7084	18.7	3, 1, 0, 4, 1
		3	17 282	45.6	3, 1, 0, 1, 2
		4	6111	16.1	5, 2, 1, 0, 1
		5	3280	8.6	6, 0, 1, 6, 1
Long Term by Year	K = 2 (SS = 0.553)	1	34 265	80.0	8, 3, 3, 2, 0, 1, 2, 6
		2	8590	20.0	15, 24, 20, 19, 20, 27, 23, 11
	K = 3 (SS = 0.149)	1	15 831	36.9	9, 4, 8, 3, 3, 2, 5, 2
		2	10 557	24.6	24, 9, 13, 19, 12, 12, 16, 6
		3	16 467	38.4	4, 0, 0, 1, 0, 0, 1, 4
	K = 4 (SS = 0.155)	1	2402	5.6	18, 35, 34, 46, 34, 39, 27, 9
		2	23 637	55.2	8, 3, 2, 2, 4, 1, 2, 3
		3	8440	19.7	5, 0, 0, 0, 0, 1, 13, 3
		4	8376	19.5	20, 8, 10, 12, 16, 11, 19, 9
	K = 5 (SS = 0.136)	1	6206	14.5	17, 7, 1, 1, 2, 2, 4, 2
		2	5166	12.1	9, 13, 11, 11, 15, 21, 22, 9
		3	4716	11.0	27, 16, 11, 7, 10, 10, 13, 5
		4	11 254	26.3	6, 2, 6, 6, 7, 7, 12, 3
		5	15 513	36.2	3, 0, 0, 0, 0, 1, 5, 2
Long Term by Quarter	K = 2 (SS = 0.536)	1	36 775	85.8	4, 1, 0, 2, 0, 0, 0, 0, 0, 1, 4, 0, 0, 0, 0, 1, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 2
		2	6080	14.2	7, 1, 6, 2, 3, 10, 6, 4, 5, 6, 5, 5, 2, 4, 4, 7, 6, 4, 5, 3, 5, 4, 7, 6, 9, 11, 6, 8, 3, 4
	K = 3 (SS = 0.007)	1	16 528	38.6	1, 0, 5, 0, 1, 2, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 2, 1, 0, 1
		2	9729	22.7	5, 3, 1, 1, 1, 3, 2, 1, 3, 3, 8, 3, 1, 1, 2, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 3, 4, 7, 2
		3	16 598	38.7	3, 0, 1, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 3, 0, 1, 1, 4

Continued

Supplementary Table 5: Continued

Analysis	# Clusters (Silhouette Score)	Cluster ID	# Clients	% Clients	Medoid
	K = 4 (SS = 0.236)	1	998	2.3	7, 0, 3, 4, 1, 2, 2, 1, 1, 1, 1, 1, 1, 0, 1, 2, 1, 1, 0, 2, 2, 1, 6, 2, 3, 9, 20, 10, 13, 9, 12, 13, 2
		2	26 775	62.5	3, 0, 1, 0, 2, 0, 0, 0, 1, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 1, 1, 0, 1
		3	10 169	23.7	3, 2, 0, 0, 4, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 6, 2, 4, 2, 5, 1, 2
		4	4913	11.5	8, 5, 4, 3, 3, 2, 3, 4, 2, 4, 6, 4, 5, 4, 4, 5, 3, 7, 5, 4, 5, 4, 3, 3, 2, 4, 2, 7, 3
	K = 5 (SS = -0.031)	1	11 624	27.1	3, 2, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 2, 1, 0, 1, 0, 0, 0, 3, 1
		2	6981	16.3	2, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 3, 0, 1, 0, 1, 0, 0, 0, 2, 1
		3	6448	15.0	6, 8, 3, 1, 1, 2, 3, 3, 3, 2, 6, 4, 1, 1, 3, 1, 2, 2, 0, 2, 2, 1, 2, 2, 3, 2, 2, 1, 2
		4	5840	13.6	4, 3, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 2, 4, 9, 5, 2
		5	11 962	27.9	2, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 1, 0, 2, 0, 0, 1, 0, 0, 0, 1

Notes: Short-term cohort includes clients with 2–3 years of observation; long-term cohort includes clients with 8–10 years of observation. Year represents consecutive 365.25 day intervals; quarter-year represents consecutive 90.30 day intervals. Medoids are the time series of a real client, and represent the "middle" time series for a cluster, i.e., selected to minimize within- and maximize between-cluster distance. *Legend:* SS = Silhouette Score. K represents the number of clusters allowed to explain the data.

Eleven-year period prevalence technical details

Since not all clients receive care from CHCs 2009–2019, they are not all at-risk of condition indications in their electronic health record for the entire calendar-based period of observation. Thus, the denominator requires estimation of the average or mid-point size of the population. This is challenging given that primary care electronic health records represent an open cohort with no standard expectation for frequency of care, and the overall number of clients receiving care increases across calendar time (see Supplementary Figure 1). We used the following process to calculate 11-year period prevalence:

Numerator: number of clients with at least one relevant code at any point from 2009 through 2019.

Denominator: First, we calculated the median number of calendar-based years of observation across all eligible clients (i.e., median number of "at-risk" years): 5 years. Second, we calculated the number of clients who received any type of care at least once in each of the seven possible five-year intervals (2009-13; 2010-14; 2011-15; 2012-16; 2013-17; 2014-18; 2015-19), representing the size of the population within each of those five-year intervals. Finally, the median size of those seven cohorts was used as the denominator, representing the overall average size of the population across 11 years.

The same process was followed to get estimates for the entire eligible population and for the subset of clients who receive care from urban at risk community health centres.