Title:

Describing a complex primary health care population in a learning health system to support future decision support and artificial intelligence initiatives

Authors:

Jacqueline K. Kueper MSc,[1,2] Jennifer Rayner PhD,[3,4] Merrick Zwarenstein MBBCh, PhD,[1,3] Daniel J. Lizotte PhD[1,2]

1. Department of Epidemiology and Biostatistics, Schulich School of Medicine & Dentistry, Western University
2. Department of Computer Science, Faculty of Science, Western University
3. Department of Family Medicine, Schulich School of Medicine & Dentistry, Western University
4. Alliance for Healthier Communities

Word Count: 3629

Figures: 4

Tables: 2

References: 50

Corresponding author:

Jacqueline K. Kueper

jkueper@uwo.ca

Abstract

Background: Learning health systems (LHS) use data to improve care. Descriptive epidemiology to reveal health states and needs of the LHS population is essential for informing LHS initiatives. To properly characterize complex populations, both simple statistical and artificial intelligence techniques, applied with an epidemiological lens, can be useful. We present the first large-scale description of the population served by one of the first primary care LHS in Canada.

Methods: We use electronic health record data from 2009-2019 to describe sociodemographic, clinical, and health care use characteristics of adult primary care clients served by the Alliance for Healthier Communities. In addition to simple summary statistics, we apply unsupervised leaning techniques to explore patterns of common condition co-occurrence, care provider teams, and care frequency.

Results: There are 221 047 eligible clients. Clients at community health centres that primarily serve those most at risk (homeless, mental health, addictions) and clients with multimorbidity tended to have more social determinants of health and chronic conditions. Most care is provided by physician and nursing providers, with heterogeneous combinations of other provider types. A subset of clients have many issues addressed within single-visits and there is notable within- and between-client variability in care frequency.

Conclusions: This population-level overview of clients served by the Alliance provides a foundation for future LHS initiatives. In addition to substantive findings, we demonstrate the use of methods from statistics and artificial intelligence to describe a complex primary care

population. We discuss implications for future initiatives, including development of decision support tools.

Keywords

Learning Health System, Primary Health Care, Descriptive Epidemiology, Artificial Intelligence/Machine Learning

Key messages

- The Alliance for Healthier Communities serves clients with barriers to care and complex health needs through Community Health Centres across Ontario; we provide the first large-scale description of sociodemographic, clinical, and health care use characteristics of their primary care population using electronic health record data.
- We demonstrate the use of both simple statistical techniques traditionally used in descriptive epidemiology and artificial intelligence techniques, applied with an epidemiological lens, to account for complexity.
- Implications for future learning health system initiatives, including the development of decision support tools, are discussed.

Introduction

The recognized potential for analysis of electronic health record (EHR) data to inform health care delivery led to the formalization of the concept of a Learning Health System (LHS) in 2007: a socio-technical system characterized by iterative cycles of data-to-knowledge-to-practice feedback.[1,2] LHS initiatives target quality improvement, research, or decision support; and usually rely on EHR data from the same population that the findings or end-product are intended to benefit.[2–5] These initiatives can support populations who have historically been excluded from medical research and clinical guideline development, such as those with complex health needs or barriers to participation.[6–9]

Primary care (PC), first contact care provided in a community setting over the life course, is inherently complex.[10,11] The Alliance for Healthier Communities provides team-based primary health care through 72 Community Health Centres (CHCs) across Ontario to clients who face barriers to care and challenges, such as poverty and mental illness, that increase their risk for poor health.[12–14] The Alliance officially adopted a LHS model in October 2020,[15,16] making them one of few documented PC LHSs in North America.[5]

A LHS may pursue multiple initiatives to inform and improve care delivery. 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,[17] and may be beneficial to add into these early stages of LHS 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.[18–20] In addition to proposed LHS benefits, descriptive studies can contribute towards closing the gap in understanding about the basic functions of PC in general.[21]

To properly understand complex EHR data, we propose using both simple statistical techniques traditionally used in descriptive epidemiology and more complex techniques from artificial intelligence, 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 PC system with its own EHR.

We present the first large-scale descriptive and exploratory study of ongoing primary care clients served by the Alliance using statistical and machine learning methodology. Our *objective* is to summarize sociodemographic, clinical, and health care use characteristics of this population. We use unsupervised learning techniques to identify patterns of multimorbidity, care provider teams, and care access frequency. Findings will provide a foundation for future Alliance LHS initiatives, 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 artificial intelligence to effectively describe a complex population and contribute to early stages of a LHS.

Methods

Study population and data source

We use a de-identified extract of the centralized, structured EHR database from all CHCs; clients have unique identifiers to allow tracking of care over time. Issues addressed during care are recorded using Electronic Nomenclature and Classification Of Disorders and Encounters for Family Medicine (ENCODE-FM)[22] and International Classification of Disease (ICD)-10 vocabularies.[23] PC EHRs represent an open cohort; **Supplementary Figure S1** 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 PC 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 follow RECORD reporting guidelines (**Supplementary Material A**).[24]

General analysis plan

Sociodemographic, clinical, and health care use characteristics are defined in **Supplementary Table S1**. Methods specific to each category are described below; we perform "table-based summaries" for all, whereby categorical variables are summarized by counts and percentages, and continuous variables by the range, median, mean, and standard deviation. Where specified, findings are stratified by client multimorbidity status (defined below) or CHC "urban at-risk" (UAR) status, which are CHCs located in major urban geographical areas and serve priority populations defined by homelessness and/or mental health and substance use challenges.[25] CHCs without UAR designation still focus on clients with barriers to care but may be in rural or urban settings and do not solely serve clients with the aforementioned complexities.[25]

Sociodemographic characteristics

We provide table-based summaries for select fields from the structured EHR client characteristic table and certain ENCODE-FM-derived variables. Missingness of the former occurs at the 1) CHC or provider level, whereby a client is not asked about the characteristic and 2) client level, whereby a client is asked and preferred to not respond. Results are presented overall and stratified by UAR and multimorbidity status.

Clinical characteristics

We describe 20 chronic conditions that define multimorbidity in PC research[26–28] and an additional four conditions of interest identified by Alliance stakeholders. For each condition, clients are assumed to receive related care upon the first record of a relevant code. Conditions are explored in single, composite, and pairwise manners.

Prevalence and incidence

To provide different perspectives on clinical complexity, we calculate two measures of prevalence and one measure of incidence for each of the 24 conditions. We also calculate prevalence of multimorbidity. Our primary multimorbidity definition, including for stratification, is presence of at least three of the 20 chronic conditions.[26–28] Multimorbidity of at least two conditions is also common and is presented for comparison.[27]

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, the number of clients who ever receive a condition indication is divided by an estimate of the average population size (technical details in *Supplementary Material B*). Sensitivity analyses

include 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 are added until no prevalence estimate is over 100%. Results are shown overall and UAR-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 have received care at a CHC. To calculate this, clients are separated 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, the number of clients who ever receive a condition indication is divided 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. Cumulative incidence curves are plotted using the R package *survival*.[29] To prioritize capture of incident condition-related care, clients with conditions recorded in 2009 are excluded 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 estimate an *Ising model* using R package *MRFcov*[30,31] for all conditions except Hepatitis C (Alliance-suggested condition that overlaps with one of the 20 chronic conditions). We convert coefficients, representing the strength of association between each condition pair adjusted for all other conditions, to odds ratios and interpret size using Chen et al. (2010) guidelines.[32] We also view the top frequency-based co-occurrences.

Health care use characteristics

We perform table-based summaries of provider and care access characteristics overall and stratified by UAR CHC, Rural Geography CHC, and client multimorbidity status.

Providers involved

To identify common care provider teams that clients are exposed to across their care histories, we use *non-negative matrix factorization (NMF)*[33] to identify frequently-occurring: 1) *"Ever-seen" teams* whereby dummy variables are used to indicate whether each provider type has ever been involved in care, and 2) *Relative "amount-seen" teams* based on volume of care whereby the number of events associated with each provider type are normalized within clients. For each version, analyses allowing 2,3,5,10, and 15 topics (provider teams) are run with the Python package *sklearn.decomposition.NMF* and the kullback-Leibler divergence distance metric.[34] Resulting topics are interpreted manually. Provider types are maintained as recorded in the EHR except "Other," "Unknown," and "Undefined" are combined. We also summarize the top frequency-based provider types involved in care and referrals. Eligible clients require at least one provider type indication in their EHR.

Care access patterns

Complexity of care is measured as the number of events (distinct issues addressed or types of care received) per visit (calendar day of access) to a CHC. *Care frequency* is measured as the number of calendar days at least one event is 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 perform *time series clustering* with the K Medoids algorithm and dynamic time warping distance metric[35] 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, R package *dtwclust*[36] is used to identify 2,3,4, and 5 clusters. Performance is assessed using the silhouette score and visual inspection.

Results

There are 221 047 eligible clients (Supplementary Material B), of whom 64 504 (29.18%) received care at least once in 2009, 141 627 (64.07%) in 2019, and 40 704 (18.4%) received care in both years.

Sociodemographic characteristics

Sociodemographic characteristics are described in **Table 1**, with remaining sub-strata in **Supplementary Table S2**. The UAR CHCs tend to provide care to clients who are more commonly male, English-speaking, and have lower levels of education, household income, immigration, stable housing, and/or food security. Clients with multimorbidity tend to be older and more commonly female, reside in rural locations, and have lower levels of education, immigration, stable residence, and/or food security.

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)
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.*

Clinical characteristics

Prevalence and incidence

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

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*

Observation-based period prevalence estimates tend to increase with length of observation; however, cumulative incidence plots for the 156 543 (70.82%) clients without care recorded in 2009 show the rate of condition indications notably decreases after the first year of observation. Sample plots are in **Figure 1**; all are in **Supplementary Figure S2 and S3**.

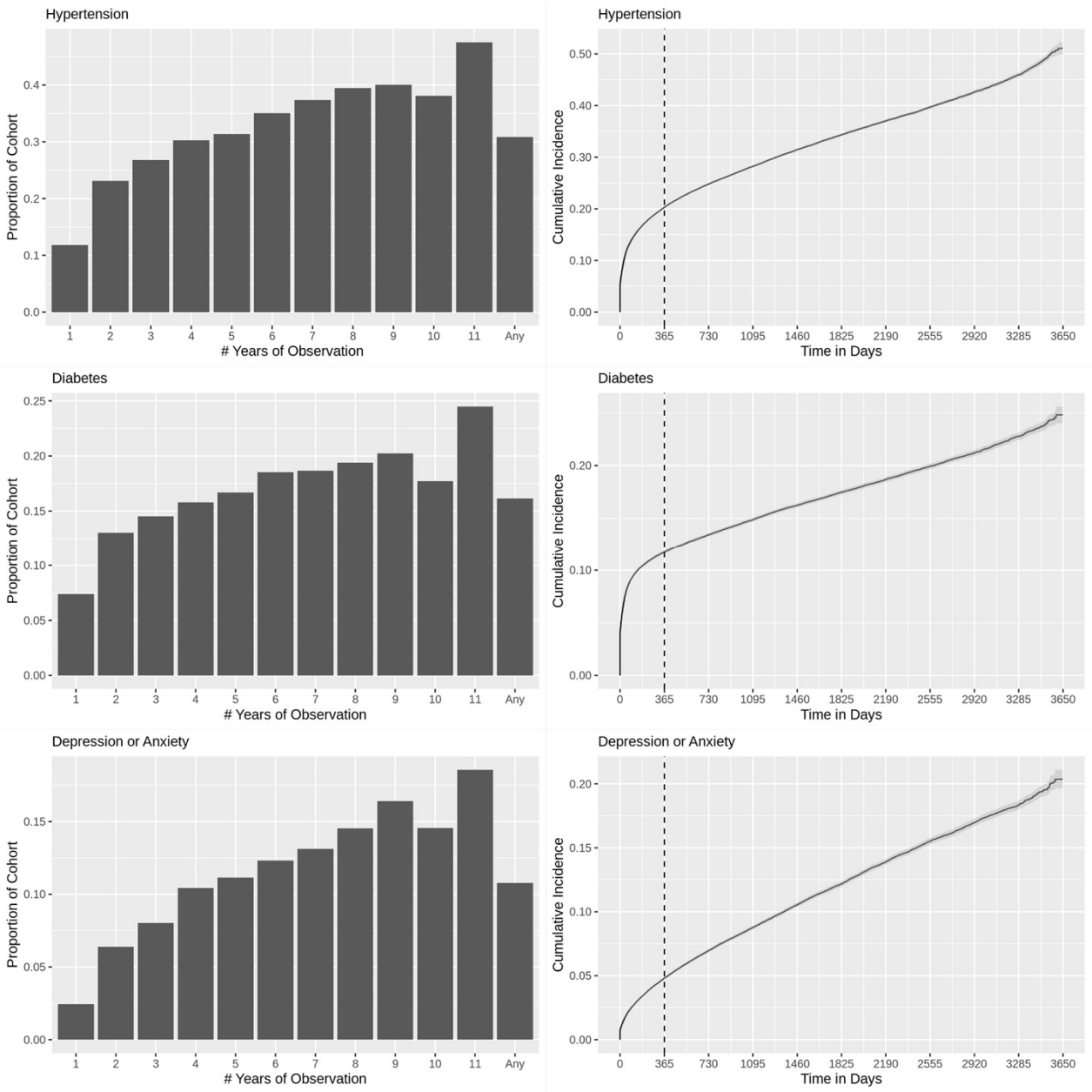

Figure 1: Example observation-based period prevalence and cumulative incidence plots. Left column: Observation-based period prevalence. Right column: Cumulative incidence by days of observation.

Condition co-occurrence patterns

Among the 103 172 (46.7%) clients with multimobidity of at least three chronic conditions, there are 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 range from -0.82 (Osteoporosis—Obesity) to 2.93 (Kidney disease or failure—Chronic urinary problem). There are 1 large, 5 medium, 40 small, and 207 very small associations based on odds ratio magnitude. The five largest positive associations are 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 are 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 with the highest marginal frequencies.

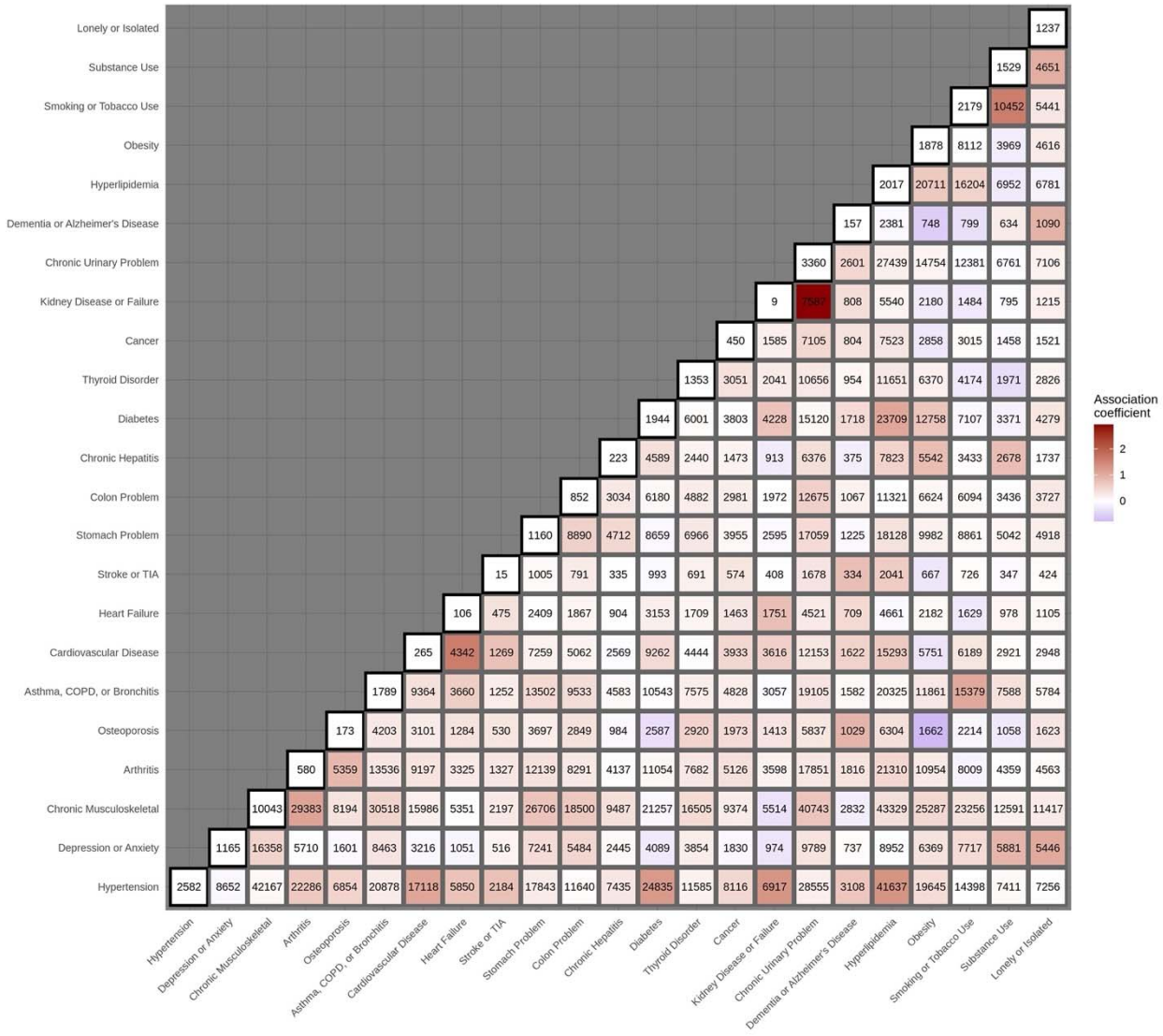

Figure 2: Condition co-occurrence patterns. Heatmap representing the results of the Ising model. Shading is relative to the edge weights or strength of condition co-occurrence. The numbers indicate raw counts in the data; diagonal counts represent clients who only have that single condition. *Legend:* TIA = Transient Ischemic Attack; COPD = Chronic Obstructive Pulmonary Disease.

Health care use characteristics

Table-based summaries of health care use characteristics are in **Supplementary Table S3**. In general, UAR and multimorbidity strata had higher health care use while rural geography CHCs were closer to the overall population.

Providers involved

There are 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 capture "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 increases, additional provider types emerge and then split apart to dominate separate topics. Exceptions are the high-weighted pairings of nurse and physician and of registered practical nurse and nurse practitioner. Overall, 18 of the 68 possible provider types emerge prominently in at least one topic; only one (respirologist) does not also appear in the amount-seen analysis.

The *amount-seen* provider team analysis has greater weight distributions between provider types within topics. For example, the first of the three-topic analysis has 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 have a non-zero weight for only one of the first two topics; in the amount-seen analysis more clients remain non-zero weight on only only one topic as the number of topics increase, e.g. 16.6% versus 2.5% at five topics. In general, results suggests most clients receive 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 referrals types (**Supplementary Table S4**), findings suggest referrals to dietitian/nutritionist are more common than to social worker, but frequent or longer-term care is more commonly provided by social workers.

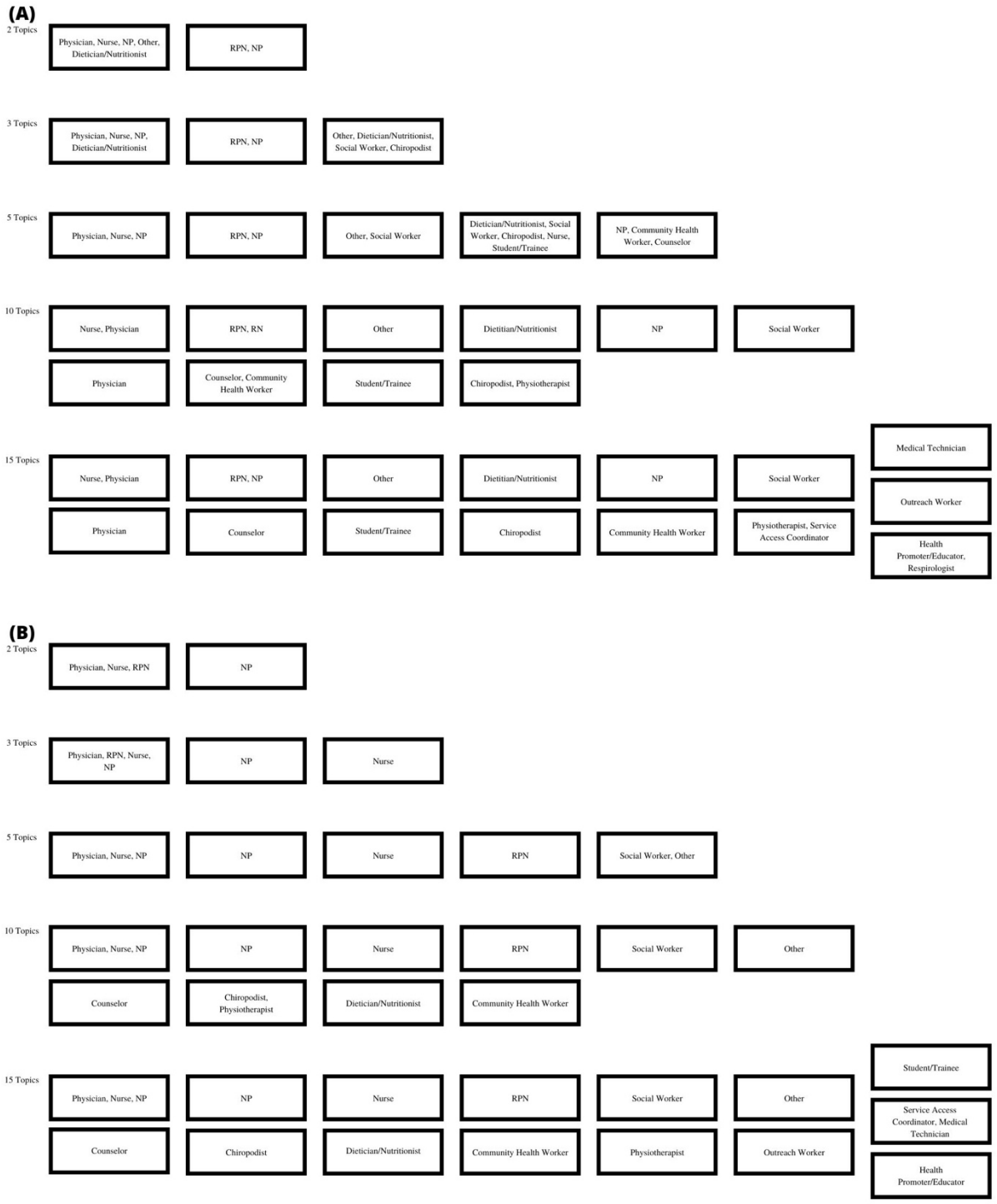

Figure 3: Common care provider teams. Boxes represent the topics resulting from the non-negative matrix factorization analysis for A) Ever-seen provider team analysis. B) Relative amount seen provider team analysis. Provider types are listed in order starting with the highest weighted provider; for any given topic, provider types with a weight less than three are not show. *Legend:* NP = Nurse Practitioner, RPN = Registered Practical Nurse.

Care access patterns

Complexity of care from a CHC-perspective is primarily low with 80.4% of client-visits associated with a single-issue and under 1% having over five issues addressed (higher intensity); however, from a client-perspective, 24 204 (11.0%) experience at least one visit with over five issues while 38 533 (17.4%) experience a maximum of one issue per visit across their care history. The mean *care access frequency* is 6 days per year (standard deviation=7.4). While 29 191 (13.2%) clients experience at least one year with over 25 days, 7455 (3.4%) average over 25 days per year across their entire care history. There are 8700 (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 includes 37 920 clients and 93 625 client-years of observation; the long-term cohort includes 42 855 clients and 387 035 client-years of observation. The silhouette score was always highest for two clusters (**Supplementary Table S5**). Visual inspection of plots (**Figure 4**) shows high variability within and between clients.

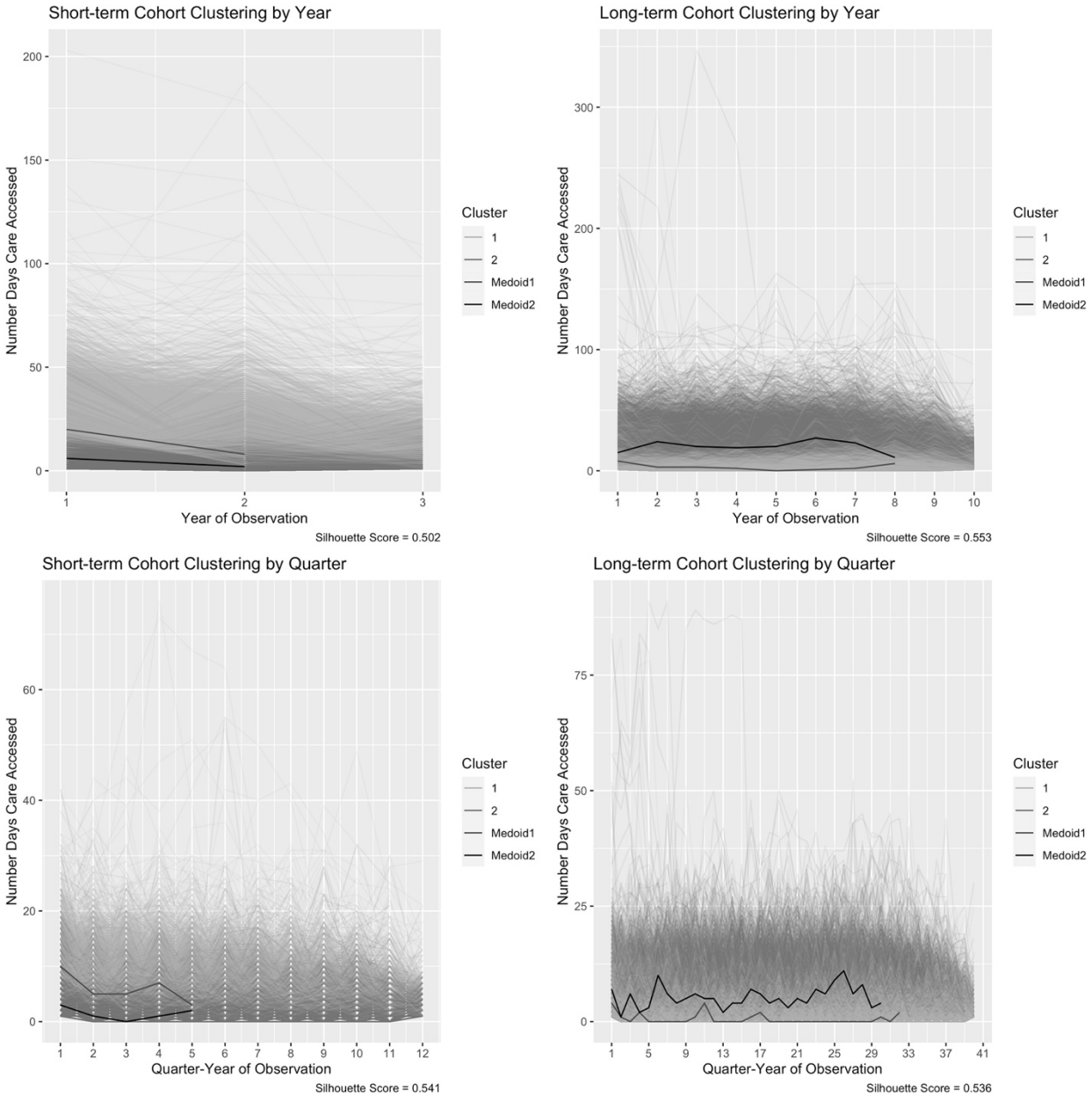

Figure 4: Care frequency clusters. Results from the four time series clustering analyses for each cohort and data-representation combination. Medoids are shown with raw time series data, separated by cluster number, for the number of clusters that resulted in the highest silhouette score (SS).

Discussion

We used statistical and artificial intelligence techniques to summarize sociodemographic, clinical, and health care use characteristics captured in the EHRs of ongoing PC clients served by

the Alliance. Substantive findings can motivate new topics for future LHS 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 LHS 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 were more prevalent in UAR CHC and multimorbidity strata, and there appears to be evidence for the healthy immigrant effect.[37] Completeness rates vary 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 1001 (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?"[38] 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.[39,40]

When assessing data quality and completeness, which is emphasized by machine learning for EHR guidelines,[2,4,19,41,42] the implications of pursuing LHS 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.[43] While these measures are not suitable for population level analyses, they should be considered for initiatives specific to the collecting CHCs.

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* estimates relate to a CHC-based perspective and are useful for long-term system-level planning, while the *observation-based period prevalence* 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.[44] For example, clients may not seek PC when they are healthy, hospitalized, or experiencing barriers to care.

The *cumulative incidence* plots 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 LHS 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. Although this care pattern is somewhat unique to PC settings, methods developed for related problems may be useful. For

example, accounting for variable lengths of stay in intensive care unit EHRs,[45] or handling cold-starts and sparse data for recommender systems.[46]

Condition co-occurrence patterns

There is a high prevalence of multimorbidity, but with so many different multimorbidity "compositions" it is hard to see how to make use of the category of multimorbidity. The *Ising model* demonstrates how to go beyond frequency-based comparisons and identify relationships between conditions irrespective of others, but again, this presents as a long tail problem, with very few combinations that are very prominent. PC 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 and so new techniques for providing evidence-based guidelines or recommendations for these vast numbers of combinations are needed and will be a subject of exploration.[47–50]

Health care use characteristics

Providers involved

While care for ongoing PC clients is typically led by physicians or nurse practitioners, CHCs include many provider types and LHS initiatives may choose to focus on particular provider type(s). The *NMF analyses* more easily identify reliable patterns of commonly seen provider types and teams than manually sifting through extensive count-based tables. 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.[33] For example, NMF-derived

403 topics could be used as inputs to a predictive model instead of separate variables to represent
404 each provider type or specific, manually selected combinations.
405

406 ### Care access patterns

407 *Complexity of care* from a CHC system-level perspective is primarily low intensity (few
408 problems addressed per visit). The subset of clients who experience higher care complexity do
409 not tend to also have high frequency of care. Sporadic visit patterns may be due to unstable living
410 arrangements or demanding life responsibilities; when there is uncertainty about when a client
411 will return, providers may pack together multiple types of care. The marginal distribution of *care*
412 *frequency* is right-skewed without a distinct break; most clients experience lower care frequency,
413 but higher frequencies are also observed. In contrast to expectations, we did not identify
414 consistent, distinct client groupings through the time-series clustering, e.g., to indicate a
415 subpopulation of "frequent visitors." This may be due to restrictions in the types of similarity that
416 dynamic time warping captures. Future analyses could try a different similarity metric or
417 including covariates to account for baseline variability.
418

419 ## Strengths and Limitations

420 Strengths include the deep interdisciplinary approach used to assess complex, longitudinal EHR
421 data. We used chronic condition definitions recommended for PC research;[26–28] although the
422 algorithms have not been validated for CHCs specifically. Our broad cohort definition supports a
423 high-level overview of the population, but may not be appropriate for specific research questions.
424

425 ## Conclusions

426 This study demonstrates the use of simple statistics and artificial intelligence techniques, applied
427 with an epidemiological lens, to describe EHR data from a budding LHS. Substantive findings
428 lay a foundation for future Alliance initiatives and may be informative for other organizations
429 serving complex PC populations.
430
431 Key suggestions for future LHS initiatives include the need to carefully deliberate the level of
432 analysis, or who a given initiative should be targeted at (e.g. population or specific CHCs, one or
433 many clinical presentations, all or subset of providers), and the associated implications for how
434 clients will be represented in the data. Representation will depend on analytical-, system-,
435 provider-, and client-level factors. Decision support initiatives need to consider heterogeneity in
436 conditions and care access patterns, including non-uniform risk of condition indications across
437 observation history.
438

439 Declarations

440 Ethics approval

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

442 Funding

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

445 Author contributions

446 JKK, JR, MZ, and DJL engaged in the conception and planning of the study. JKK conducted
447 analyses under the supervision of DJL. JKK drafted the manuscript and interpretation of findings.
448 JR, MZ, and DJL provided critical feedback. JKK is the guarantor.

Data availability

The data underlying this article were securely accessed from the Alliance for Healthier Communities. The data cannot be shared publicly due to their sensitive nature, as agreed upon in the ethics agreement.

Supplementary data

Supplementary data are available at *IJE online*. Supplementary material A includes the RECORD reporting guideline checklist. Supplementary material B includes additional tables, figures, and definitions.

Supplementary figure captions

Figure S1: 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 during (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. *Legend:* COPD = Chronic Obstructive Pulmonary Disease; TIA = Transient Ischemic Attack; AD = Alzheimer's Disease.

Figure S2: Observation-based period prevalence. 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.

Figure S3: Cumulative incidence 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).

Conflict of interest

None declared.

References

1. 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. doi:10.1002/lrh2.10020

2. Foley T, Horwitz L, Zahran R. *Realising the Potential of Learning Health Systems.* Newcastle University: The Learning Healthcare Project; 2021:101. https://learninghealthcareproject.org/wp-content/uploads/2021/05/LHS2021report.pdf. Accessed December 28, 2021.

3. Delaney BC, Peterson KA, Speedie S, Taweel A, Arvanitis TN, Hobbs FDR. Envisioning a Learning Health Care System: The Electronic Primary Care Research Network, a case study. *Ann Fam Med* 2012;**10**:54-59. doi: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. *Academic Medicine* 2021;**96**:1291-1299. doi:10.1097/ACM.0000000000004021

5. Nash DM, Bhimani Z, Rayner J, Zwarenstein M. Learning health systems in primary care: A systematic scoping review. *BMC Fam Prac* 2021;**22**:126. doi:10.1186/s12875-021-01483-z

6. 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. doi:10.1080/13557850701616987

7. 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. doi:10.2105/AJPH.2013.301706

8. 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. doi:10.2105/AJPH.2007.111138

9. 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. doi:10.1186/1471-2288-14-42

10. Starfield B. *Primary Care. Balancing Health Needs, Services, and Technology*. New York, NY: Oxford University Press, Inc.; 1998.

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

12. 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*. Toronto, Ont.: Institute for Clinical Evaluative Sciences; 2012. https://www.deslibris.ca/ID/232144. Accessed May 6, 2020.

13. Booth RG, Richard L, Li L, Shariff SZ, Rayner J. 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. doi:10.9778/cmajo.20190089

14. Albrecht D. Community health centres in Canada. *Leadersh Health Serv* 1998;**11**:5-10. doi:10.1108/13660759810202596

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

16. Alliance for Healthier Communities. *Towards a Learning Health System: Better Care Tomorrow When We Learn from Today*. Alliance for Healthier Communities; 2020:15.

https://www.allianceon.org/sites/default/files/documents/Learning%20Health%20System%20report%202020-10-20%20-%20FINAL_JR.pdf. Accessed November 23, 2020.

17. Cameron D, Jones IG. John Snow, the Broad Street Pump and Modern Epidemiology. *Int J Epidemol* 1983;**12**:393-396. doi:10.1093/ije/12.4.393

18. 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 Medl Inform Decis Mak* 2021;**21**:297. doi:10.1186/s12911-021-01669-6

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

20. 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. doi:10.23889/ijpds.v5i1.1123

21. 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. doi:10.1007/s11606-020-06560-0

22. ENCODE-FM. Electronic Nomenclature and Classification Of Disorders and Encounters for Family Medicine. *ENCODE-FM*. 2020. http://aix1.uottawa.ca/~fammed/fmcencod.htm. Accessed April 6, 2020.

23. Organization WH. ICD-10 Version:2019. *World Health Organization*. 2020. https://icd.who.int/browse10/2019/en. Accessed April 6, 2020.

24. 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. doi:10.1371/journal.pmed.1001885

25. 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,

553 Ontario: Institute for Clinical Evaluative Sciences in Ontario; 2015.
554 http://www.deslibris.ca/ID/248807. Accessed January 20, 2020.
555 26. Fortin M, Almirall J, Nicholson K. Development of a research tool to document self-
556 reported chronic conditions in primary care. *J Comorb* 2017;**7**:117-123.
557 doi:10.15256/joc.2017.7.122
558 27. Lee ES, Lee PSS, Xie Y, Ryan BL, Fortin M, Stewart M. The prevalence of
559 multimorbidity in primary care: A comparison of two definitions of multimorbidity with two
560 different lists of chronic conditions in Singapore. *BMC Public Health* 2021;**21**:1409.
561 doi:10.1186/s12889-021-11464-7
562 28. Lee YAJ, Xie Y, Lee PSS, Lee ES. Comparing the prevalence of multimorbidity using
563 different operational definitions in primary care in Singapore based on a cross-sectional study
564 using retrospective, large administrative data. *BMJ Open*. 2020;**10**:e039440.
565 doi:10.1136/bmjopen-2020-039440
566 29. Therneau T. A Package for Survival Analysis in R. 2021. https://CRAN.R-
567 project.org/package=survival.
568 30. Borkulo CD van, Borsboom D, Epskamp S, et al. A new method for constructing
569 networks from binary data. *Sci Rep* 2014;**4**:5918. doi:10.1038/srep05918
570 31. Clark NJ, Wells K, Lindberg O. Unravelling changing interspecific interactions across
571 environmental gradients using Markov random fields. *Ecology* 2018;**99**:1277-1283.
572 doi:10.1002/ecy.2221
573 32. Chen H, Cohen P, Chen S. How big is a big odds ratio? Interpreting the magnitudes of
574 odds ratios in epidemiological studies. *Commun Stat Simul Comput* 2010;**39**:860-864.
575 doi:10.1080/03610911003650383

33. Wang Y-X, Zhang Y-J. Nonnegative Matrix Factorization: A Comprehensive Review. *IEEE Trans Knowl Data Eng* 2013;**25**:1336-1353. doi:10.1109/TKDE.2012.51

34. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 2011;**12**:2825-2830. http://jmlr.org/papers/v12/pedregosa11a.html. Accessed January 1, 2022.

35. Aghabozorgi S, Seyed Shirkhorshidi A, Ying Wah T. Time-series clustering – A decade review. *Information Systems* 2015;**53**:16-38. doi:10.1016/j.is.2015.04.007

36. Montero P, Vilar JA. TSclust: An R Package for Time Series Clustering. *J Stat Softw* 2015;**62**:1-43. doi:10.18637/jss.v062.i01

37. 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. doi:10.1016/j.socscimed.2004.02.004

38. Haneuse S, Daniels M. A general framework for considering selection bias in EHR-based studies: What data are observed and why? *eGEMs* 2016;**4**. doi:10.13063/2327-9214.1203

39. 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. doi:10.1093/jamia/ocaa143

40. Zhao Y, Wood EP, Mirin N, Cook SH, Chunara R. Social determinants in machine learning cardiovascular disease prediction models: A systematic review. *Am J Prev Med* 2021;**0**:1-10. doi:10.1016/j.amepre.2021.04.016

41. 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. doi:10.1038/s41591-019-0548-6

598 42. Arbet J, Brokamp C, Meinzen-Derr J, Trinkley KE, Spratt HM. Lessons and tips for
599 designing a machine learning study using EHR data. *J Clin Transl Sci* 2020;**5**:1-10.
600 doi:10.1017/cts.2020.513
601 43. CIHI. Patient-reported outcome measures (PROMs). *Canadian Institute for Health*
602 *Information*. 2022. https://www.cihi.ca/en/patient-reported-outcome-measures-proms. Accessed
603 February 7, 2022.
604 44. Bagley SC, Altman RB. Computing disease incidence, prevalence and comorbidity from
605 electronic medical records. *J Biomed inform* 2016;**63**:108-111. doi:10.1016/j.jbi.2016.08.005
606 45. Zhang L, Chen X, Chen T, Wang Z, Mortazavi BJ. DynEHR: Dynamic adaptation of
607 models with data heterogeneity in electronic health records. In: *2021 IEEE EMBS International*
608 *Conference on Biomedical and Health Informatics (BHI)* 2021:1-4.
609 doi:10.1109/BHI50953.2021.9508558
610 46. Alyari F, Jafari Navimipour N. Recommender systems: A systematic review of the state
611 of the art literature and suggestions for future research. *Kybernetes* 2018;**47**:985-1017.
612 doi:10.1108/K-06-2017-0196
613 47. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research:
614 Application and impact of prognostic models in clinical practice. *BMJ* 2009;**338**:b606.
615 doi:10.1136/bmj.b606
616 48. O'Caoimh R, Cornally N, Weathers E, et al. Risk prediction in the community: A
617 systematic review of case-finding instruments that predict adverse healthcare outcomes in
618 community-dwelling older adults. *Maturitas* 2015;**82**:3-21. doi:10.1016/j.maturitas.2015.03.009
619 49. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in
620 developing risk prediction models with electronic health records data: A systematic review. *J Am*
621 *Med Inform Assoc* 2017;**24**:198-208. doi:10.1093/jamia/ocw042

622 50. Guthrie B, Boyd CM. Clinical guidelines in the context of aging and multimorbidity.

623 *Public Policy Aging Rep* 2018;**28**:143-149. doi:10.1093/ppar/pry038

624