Predicting hospitalizations related to ambulatory care sensitive conditions with machine learning for population health planning: derivation and validation cohort study

Seung Eun Yi, Vinyas Harish, Jahir M. Gutierrez, Mathieu Ravaut, Kathy Kornas, Tristan Watson, Tomi Poutanen, Marzyeh Ghassemi, Maksims Volkovs, Laura Rosella

Supplementary Material

Supplementary Table 1

Description	**Description**
RPDB (Registered Persons Database)	The RPDB provides basic demographic information (age, sex, location of residence, date of birth, and date of death for deceased individuals) for those issued an Ontario health insurance number. The RPDB also indicates the time periods for which an individual was eligible to receive publicly funded health insurance benefits and the best known postal code for each registrant on July 1st of each year.
IRCC (Immigration, Refugees and Citizenship Canada's Permanent Resident Database)	The Ontario portion of the IRCC Permanent Resident Database includes immigration application records for people who initially applied to land in Ontario since 1985. The dataset contains permanent residents' demographic information such as country of citizenship, level of education, mother tongue, and landing date. New immigrants who are currently residing in Ontario but originally landed in another province are not captured in this dataset.
ON-MARG (Ontario Marginalization Index)	ONMARG is a geographically (census) based index developed to quantify the degree of marginalization occurring across the province of Ontario. It is comprised of four major dimensions thought to underlie the construct of marginalization: residential instability, material deprivation, dependency, and ethnic concentration. The dataset contains census divisions (CD), census tracts (CT), census subdivisions (CSD), consolidated municipal service manager areas (CMSM), public health units (PHU), local health integration networks (LHIN), sub-LHINs, and dissemination areas (DA).
CENSUS	The 2006, 2011, and 2016 Canadian census were used to capture information on area-level income and education.
Multimorbidity Dataset	This dataset was created combining multiple datasets such as OHIP, DAD, and ICES-derived chronic disease cohorts. It summarizes the diagnosis date of 18 chronic conditions for each patient. For more detail on the definition used to identify each chronic condition, see Table S2.
DAD (Discharge Abstract Database)	The DAD is compiled by the Canadian Institute for Health Information and contains administrative, clinical (diagnoses and procedures/interventions), demographic, and administrative information for all admissions to acute care hospitals, rehab, chronic, and day surgery institutions in Ontario. At ICES, consecutive DAD records are linked together to form 'episodes of care' among the hospitals to which patients have been transferred after their initial admission.
NACRS (National Ambulatory Care Reporting System)	The NACRS is compiled by the Canadian Institute for Health Information and contains administrative, clinical (diagnoses and procedures), demographic, and administrative information for all patient visits made to hospital- and community-based ambulatory care centres (emergency departments, day surgery units, hemodialysis units, and cancer care clinics). At ICES, NACRS records are linked with other data sources (DAD, OMHRS) to identify transitions to other care settings, such as inpatient acute care or psychiatric care.
DIN (Druglist File)	The DIN file contains a near exhaustive list of drug identification numbers used in Canada from 1990 forward. Contains information on drug and product names (generic and trade names), subclass information, PCG codes, Drug strength, Route of Administration, first and last dispensing dates from ODB.
ODB (Ontario Drug Benefit)	The ODB database contains prescription medication claims for those covered under the provincial drug program, mainly: those aged 65 years and older, nursing home residents, patients receiving services under the Ontario Home Care program, those receiving social assistance, and residents eligible for specialized drug programs. Main data elements include drug identifier, quantity, days supplied, date dispensed, cost, and patient, pharmacy and physician identifiers.
OHIP (Ontario Health Insurance Plan)	The OHIP claims database contains information on inpatient and outpatient services provided to Ontario residents eligible for the province's publicly funded health insurance system by fee-for-service health care practitioners (primarily physicians) and "shadow billings" for those paid through non-fee-for-service payment plans. The main data elements include patient and physician identifiers (encrypted), code for service provided, date of service, associated diagnosis, and fee paid.
OLIS (Ontario Laboratory Information System)	OLIS is a province-wide integrated repository of patients' lab test orders and results. Lab tests and results related to Hemoglobin A1C carried out from 2008 to 2015 were included for the study.

Supplementary Table 1. Dataset Description. Details of the datasets used are presented. Descriptions were adapted from the ICES data dictionary: https://datadictionary.ices.on.ca/Applications/DataDictionary/Default.aspx

Supplementary Table 2

Chronic condition	**Definition / Source**	**Lookback Window***
Asthma	ICES-derived cohort: Ontario Asthma Database (ASTHMA)	1992
Cancer	Ontario Cancer Registry (OCR)	1992
Congestive Heart Failure	ICES-derived cohort: Ontario Congestive Heart Failure Database (CHF)	1992
Chronic Obstructive Pulmonary Disorder	ICES-derived cohort: Ontario Chronic Obstructive Pulmonary Disease (COPD)	1992
Diabetes	ICES-derived cohort: Ontario Diabetes Database (ODD)	1992
Acute Myocardial Infarction	One hospitalization in DAD using the ICD codes (ICD9: 410, ICD10: I21)	1988 (DAD)
Rheumatoid Arthritis	ICES-derived cohort: Ontario Rheumatoid Arthritis Database (ORAD)	1992
Osteo- and other Arthritis	(1) One hospitalization in DAD; OR (2) Two or more OHIP physician billing claim within a two-year period using the ICD codes: ICD9/OHIP: 715, 710, 711, 716, 718, 720, 727, 728, 729, 739, 274 ICD10: M00-M03, M07, M10, M11-M14, M20-M25, M30-M36, M65-M79, M15-M19	1991 (OHIP) 1988 (DAD)
Crohn's Or Colitis	ICES-derived cohort: Ontario Chron's and Colitis Cohort Database (OCCC)	1992
Cardiac Arrhythmia	(1) One hospitalization in DAD; OR (2) Two or more OHIP physician billing claim within a two-year period using the ICD codes: ICD9 /OHIP: 427.3 (DAD) / 427 (OHIP) ICD 10: I48.0, I48.1	1991 (OHIP) 1988 (DAD)
Hypertension	ICES-derived cohort: Ontario Hypertension Database (HYPER)	1992
Chronic Coronary Syndrome	(1) One hospitalization in DAD; OR (2) Two or more OHIP physician billing claim within a two-year period using the ICD codes: ICD 9/OHIP: 411-414 ICD-10: I20, I22-I25	1991 (OHIP) 1988 (DAD)
Stroke (Excluding TIA)	(1) One hospitalization in DAD; OR (2) Two or more OHIP physician billing claim within a two-year period using the ICD codes: Any hospital admission with the following dx codes: ICD-9: 430, 431, 432, 434, 436 ICD-10: 160 (excl 160.8), 161, I62, 163 (excl 163.6), 164	1991 (OHIP) 1988 (DAD)
Osteoporosis	ICES-derived cohort: (1) One hospitalization in DAD; OR (2) Two or more OHIP physician billing claim within a two-year period using the ICD codes: ICD9/OHIP: 733 ICD10: M81, M82	1991 (OHIP) 1988 (DAD)

Chronic condition	Definition / Source	Lookback Window*
Mood Disorder (History of Mental Health-related Visit)	(1) One hospitalization in DAD/OMHRS; OR (2) Two or more OHIP physician billing claim within a two-year period using the ICD codes: From DAD var DX10CODE1 with any of the following ICD-10-CA codes: From OMHRS: - If var AXIS1_DSM4CODE_DISCH1 complete (i.e,. listed diagnosis from below present) use AXIS1_DSM4CODE_DISCH1 - No, use PROVDX1 - Exclude OMHRS admissions if AXIS1_DSM4CODE_DISCH1 in: (290.x OR 294.x). If AXIS1_DSM4CODE_DISCH1 missing, exclude if PROVDX1=2 - Include visits/admissions with suspect diagnoses (suspect = T). ICD9/OHIP: 311, 309, 300, 296 ICD10: F30—F34 (excl. F340), F38—F42, F431, F432, F438, F44, F450, F451, F452, F48, F530, F680, F930, F99	1991 (OHIP) 1988 (DAD) 2005 (OMHRS)
Other Mental Health Disorder (History of Mental Health-related Visit) - Note: This excludes dementia, deliberate self-harm codes, and mood disorder codes.	(1) One hospitalization in DAD/OMHRS; OR (2) Two or more OHIP physician billing claim within a two-year period using the ICD codes: From DAD var DX10CODE1 with any of the following ICD-10-CA codes: From OMHRS: - If var AXIS1_DSM4CODE_DISCH1 complete (i.e,. listed diagnosis from below present) use AXIS1_DSM4CODE_DISCH1 - No, use PROVDX1 - Exclude OMHRS admissions if AXIS1_DSM4CODE_DISCH1 in: (290.x OR 294.x). If AXIS1_DSM4CODE_DISCH1 missing, exclude if PROVDX1=2 - Include visits/admissions with suspect diagnoses (suspect = T). ICD9/OHIP: 291, 292, 295, 297, 298, 299, 301, 302, 303, 304, 305, 306, 307, 313, 314, 315, 319 ICD10: F04, F050, F058, F059, F060, F061, F062, F063, F064, F07, F08, F10, F11, F12, F13, F14, F15, F16, F17, F18, F19, F20 , F21, F22, F23, F24, F25, F26 F27 F28, F29, F340, F35, F36, F37, F430, F439, F453, F454, F458, F46, F47, F49, F50, F51, F52, F531, F538, F539, F54, F55, F56, F57, F58, F59, F60, F61, F62, F63, F64, F65, F66, F67, F681, F688, F69, F70, F71, F72, F73, F74 F75 F76 F77 F78, F79, F80, F81, F82, F83, F84, F85 F86 F87 F88, F89, F90, F91, F92, F931, F932, F933, F938, F939, F94, F95, F96, F97, F98	1991 (OHIP) 1988 (DAD) 2005 (OMHRS)

Chronic condition	**Definition / Source**	**Lookback Window***
Dementia	(1) One hospitalization in DAD; OR (2) Two or more OHIP physician billing claim within a two-year period using the ICD codes: ICD9/OHIP: 290, 331 (OHIP) / (DAD: 046.1, 290, 294, 331.0, 331.1, 331.5, 331.82) ICD10: F00, F01, F02, F03, G30	1991 (OHIP) 1988 (DAD)
Renal Failure	(1) One hospitalization in DAD; OR (2) Two or more OHIP physician billing claim within a two-year period using the ICD codes: ICD9/OHIP: 403,404,584,585,586,v451 ICD 10: N17, N18, N19, T82.4, Z49.2, Z99.2	1991 (OHIP) 1988 (DAD)

Supplementary Table 2. List of 18 chronic conditions used for the Multimorbidity Dataset.
∗ Lookback window refers to the time window used to extract the history of chronic conditions.

Supplementary Table 3

Type of Criteria	Criteria	Details
Inclusions	Hospitalization for an ambulatory care sensitive condition is identified as any most responsible diagnosis code of:	- **Chronic obstructive pulmonary disease (COPD):** Any most responsible diagnosis (MRDx) code of J41, J42, J43, J44, J47, or MRDx of acute lower respiratory infection (J10.0, J11.0, J12–J16, J18, J20, J21, J22), only when a *secondary diagnosis** of J44 is also present. *Secondary diagnosis refers to a diagnosis other than the most responsible one. - **Grand mal status and other epileptic convulsions:** G40, G41 - **Asthma:** J45. - **Diabetes:** E10.0, E10.1, E10.63, E10.64, E10.9, E11.0, E11.1, E11.63, E11.64, E11.9, E13.0, E13.1, E13.63, E13.64, E13.9, E14.0, E14.1, E14.63, E14.64, E14.90. - **Heart failure and pulmonary edema:** I50, J81. - **Hypertension:** I10.0, I10.1, I11. - **Angina:** I20, I23.82, I24.0, I24.8, I24.9.
	Admission to an acute care institution (Facility Type Code = 1).	
	Age at admission younger than 75.	
Exclusions	For heart failure and pulmonary edema, hypertension, and angina, exclude cases with cardiac procedures:	The full list of cardiac procedure codes for exclusion can be found in the CIHI Definition [2]. Codes may be coded in any position. Procedures coded as abandoned after onset (Intervention Status Attribute = A) are excluded.
	Records with missing sex.	
	Records with discharge as death (Discharge Disposition Code = 07, 72*, 73*, 74*).	
	Newborn, stillbirth or cadaveric donor records (Admission Category Code = N, R or S).	

Supplementary Table 3. Definition of Ambulatory Care Sensitive Conditions adapted from the criteria given by the Canadian Institute for Health Information [1]. All codes are in ICD-10-CA [2].

Supplementary Table 4

Feature Name	Dataset	Description
Age	RPDB	Age of the patient at the end of the observation window
Sex (female)	RPDB	Sex of the patient (binary)
Quarter of the year	-	Quarter of the year at the end of the observation window.
LHIN (1-14)	ON-MARG	Local Health Integrated Network the patient lives in. Takes on a binary value for each LHIN from 1 to 14. (14 features in total)
Living in rural areas	ON-MARG	Binary indicator of whether or not a patient lives in a rural area.
Marginalization Index - Dependency (1st quintile - 5th quintile)	ON-MARG	Index to quantify the dependency level of the neighborhood the patient lives in. Takes on a binary value for each index level from 1 to 5. Higher index indicates lower degree of marginalization. (5 features in total)
Marginalization Index - Ethnicity (1st quintile - 5th quintile)	ON-MARG	Index to quantify the ethnic marginalization of the neighborhood the patient lives in. Takes on a binary value for each index level from 1 to 5. Higher index indicates lower degree of marginalization. (5 features in total)
Marginalization Index - Instability (1st quintile - 5th quintile)	ON-MARG	Index to quantify the instability level of the neighborhood the patient lives in. Takes on a binary value for each index level from 1 to 5. Higher index indicates lower degree of marginalization. (5 features in total)
Marginalization Index - Deprivation (1st quintile - 5th quintile)	ON-MARG	Index to quantify the deprivation level of the neighborhood the patient lives in. Takes on a binary value for each index level from 1 to 5. Higher index indicates lower degree of marginalization. (5 features in total)
Income (1st quintile - 5th quintile)	ON-MARG	Index to quantify the income level of the neighborhood the patient lives in. Takes on a binary value for each index level from 1 to 5. Higher index indicates higher income. (5 features in total)
Education (1st quintile - 5th quintile)	ON-MARG	Index to quantify the education level of the neighborhood the patient lives in. Takes on a binary value for each index level from 1 to 5. Higher index indicates higher education level. (5 features in total)
Latitude	ON-MARG	Value in decimal degrees to a precision of 2 decimal places ($\sim 1km$).
Longitude	ON-MARG	Value in decimal degrees to a precision of 2 decimal places ($\sim 1km$).
Presence of arrythmia	MMB Macro	Presence of arrythmia throughout the observation window.
Presence of asthma	MMB Macro	Presence of asthma throughout the observation window.
Presence of asthma since quarter 7	MMB Macro	Presence of asthma starting from the 7th quarter of the observation window (at the latest).

Feature Name	Dataset	Description
Presence of chronic heart failure	MMB Macro	Presence of chronic heart failure throughout the observation window.
Presence of chronic heart failure since quarter 1	MMB Macro	Presence of chronic heart failure starting from the 1st quarter of the observation window (at the latest).
Presence of chronic heart failure since quarter 7	MMB Macro	Presence of chronic heart failure starting from the 7th quarter of the observation window (at the latest).
Presence of COPD	MMB Macro	Presence of chronic obstructive pulmonary disease throughout the observation window.
Presence of COPD since quarter 1	MMB Macro	Presence of COPD starting from the 1st quarter of the observation window (at the latest).
Presence of COPD since quarter 3	MMB Macro	Presence of COPD starting from the 3rd quarter of the observation window (at the latest).
Presence of COPD since quarter 4	MMB Macro	Presence of COPD starting from the 4th quarter of the observation window (at the latest).
Presence of COPD since quarter 5	MMB Macro	Presence of COPD starting from the 5th quarter of the observation window (at the latest).
Presence of COPD since quarter 6	MMB Macro	Presence of COPD starting from the 6th quarter of the observation window (at the latest).
Presence of COPD since quarter 7	MMB Macro	Presence of COPD starting from the 7th quarter of the observation window (at the latest).
Presence of coronary disease	MMB Macro	Presence of coronary disease throughout the observation window.
Presence of coronary disease since quarter 1	MMB Macro	Presence of coronary disease starting from the 1st quarter of the observation window (at the latest).
Presence of coronary disease since quarter 2	MMB Macro	Presence of coronary disease starting from the 2nd quarter of the observation window (at the latest).
Presence of coronary disease since quarter 7	MMB Macro	Presence of coronary disease starting from the 7th quarter of the observation window (at the latest).
Presence of diabetes	MMB Macro	Presence of diabetes throughout the observation window.
Presence of diabetes since quarter 1	MMB Macro	Presence of diabetes starting from the 1st quarter of the observation window (at the latest).
Presence of diabetes since quarter 2	MMB Macro	Presence of diabetes starting from the 2nd quarter of the observation window (at the latest).
Presence of diabetes since quarter 3	MMB Macro	Presence of diabetes starting from the 3rd quarter of the observation window (at the latest).
Presence of diabetes since quarter 4	MMB Macro	Presence of diabetes starting from the 4th quarter of the observation window (at the latest).
Presence of diabetes since quarter 5	MMB Macro	Presence of diabetes starting from the 5th quarter of the observation window (at the latest).
Presence of diabetes since quarter 7	MMB Macro	Presence of diabetes starting from the 7th quarter of the observation window (at the latest).
Presence of hypertension	MMB Macro	Presence of hypertension throughout the observation window.
Presence of hypertension since quarter 7	MMB Macro	Presence of hypertension starting from the 7th quarter of the observation window (at the latest).

Feature Name	Dataset	Description
Presence of mental disease	MMB Macro	Presence of mental disease throughout the observation window.
Presence of mental disease since quarter 1	MMB Macro	Presence of mental disease starting from the 1st quarter of the observation window (at the latest).
Presence of mental disease since quarter 7	MMB Macro	Presence of mental disease starting from the 7th quarter of the observation window (at the latest).
Presence of mood disorder	MMB Macro	Presence of mood disorder throughout the observation window.
Presence of renal failure	MMB Macro	Presence of renal failure throughout the observation window.
Presence of stroke since quarter 1	MMB Macro	Presence of stroke starting from the 1st quarter of the observation window (at the latest).
Number of ambulatory usage	NACRS	Number of ambulatory usage of the patient over the observation window
Time since last ambulatory usage	NACRS	If there is, time since the last ambulatory usage of the patient over the observation window
Presence of ACSC hospitalization	DAD, NACRS	Binary indicator of the presence of any ACSC-related hospitalization over the observation window
Time since last ACSC	DAD, NACRS	If there is, time since the last ACSC-related hospitalization over the observation window
Number of clinician visits	OHIP	Number of clinician visits of the patient over the observation window
Number of selective beta2-adrenergic agonists prescriptions	ODB	Number of selective beta2-adrenergic agonists prescriptions over the observation window
Time since last selective beta2-adrenergic agonists prescriptions	ODB	If there is, time since last selective beta2-adrenergic agonists prescription over the observation window
Number of beta-blockers prescriptions	ODB	Number of beta-blockers prescriptions over the observation window
Time since last beta-blockers prescriptions	ODB	If there is, time since last beta-blockers prescription over the observation window
Number of furosemide prescriptions	ODB	Number of furosemide prescriptions over the observation window
Number of albuterol sulfate prescriptions	ODB	Number of albuterol sulfate prescriptions over the observation window
Number of antilipemic statins prescriptions in quarter 8	ODB	Number of antilipemic statins prescriptions in the 8th quarter of the observation window
Time since last lab test	OLIS	If there is, time since last lab test the patient had over the observation window
Absence of LTC prescriptions in quarter 1	ODB	Binary indicator of LTC prescriptions over the first quarter of the observation window
Clinician visit feecode in quarter 8 - chest radiology	OHIP	Number of clinician visits related to chest radiology in the 8th quarter of the observation window
Number of calcium blockers prescriptions	ODB	Number of calcium blockers prescriptions over the observation window

Feature Name	Dataset	Description
Number of lab tests	OLIS	Total number lab tests the patient had over the observation window
Number of hospitalizations	DAD	Total number of hospitalizations the patient had over the observation window
Clinician visit location in quarter 4 - office	OHIP	Number of visits to a clinician in office during the 4th quarter of the observation window
Number of antilipemic statins prescriptions in quarter 5	ODB	Number of antilipemic statins prescriptions in the 5th quarter of the observation window
Absence of LTC prescriptions in quarter 5	ODB	Binary indicator of LTC prescriptions over the 5th quarter of the observation window
Absence of LTC prescriptions in quarter 6	ODB	Binary indicator of LTC prescriptions over the 6th quarter of the observation window
Absence of LTC prescriptions in quarter 8	ODB	Binary indicator of LTC prescriptions over the 8th quarter of the observation window
Number of antilipemic statins prescriptions in quarter 7	ODB	Number of antilipemic statins prescriptions in the 7th quarter of the observation window
Time since last furosemide prescription	ODB	If there is, the last time the patient was prescribed with furosemide during the observation window
Clinician visit feecode in quarter 7 - chest radiology	OHIP	Number of clinician visits related to chest radiology in the 7th quarter of the observation window
Clinician visit feecode in quarter 6 - chest radiology	OHIP	Number of clinician visits related to chest radiology in the 6th quarter of the observation window
Clinician visit feecode in quarter 5 - chest radiology	OHIP	Number of clinician visits related to chest radiology in the 5th quarter of the observation window
Time since last albuterol sulfate prescription	ODB	If there is, the last time the patient was prescribed with albuterol sulfate during the observation window
Number of macrolides prescriptions	ODB	Number of macrolides prescriptions over the observation window
Time since last calcium blocker prescription	ODB	If there is, the last time the patient was prescribed with calcium blocker during the observation window
Clinician visit feecode in quarter 4 - chest radiology	OHIP	Number of clinician visits related to chest radiology in the 4th quarter of the observation window
Clinician visit feecode in quarter 2 - chest radiology	OHIP	Number of clinician visits related to chest radiology in the 2nd quarter of the observation window
Clinician visit location in quarter 2 - office	OHIP	Number of visits to a clinician in office during the 2nd quarter of the observation window
Clinician visit feecode in quarter 1 - chest radiology	OHIP	Number of clinician visits related to chest radiology in the 1st quarter of the observation window
Clinician specialty in quarter 8 - emergency medicine	OHIP	Number of visits to a clinician specialized in emergency medicine during the 8th quarter of the observation window
Number of non-steroidal anti-inflammatory prescriptions	ODB	Number of non-steroidal anti-inflammatory prescriptions during the observation window
Time since last antilipemic statins prescription	ODB	If there is, the last time the patient was prescribed with antilipemic statins during the observation window

Feature Name	Dataset	Description
Emergency visit in quarter 8 - no blood transfusion	NACRS	No blood transfusion during emergency visits in the 8th quarter during the observation window
Time since last clinician visit	OHIP	If there is, the last time the patient goes for a clinician visit over the observation window
Number of antilipemic statins prescriptions in quarter 6	ODB	Number of antilipemic statins prescriptions in the 6th quarter of the observation window
Clinician visit location in quarter 6 - office	OHIP	Number of visits to a clinician in office during the 6th quarter of the observation window
Clinician specialty in quarter 7 - emergency medicine	OHIP	Number of visits to a clinician specialized in emergency medicine during the 7th quarter of the observation window
Clinician visit location in quarter 5 - office	OHIP	Number of visits to a clinician in office during the 5th quarter of the observation window
Presence of COPD-related ACSC	DAD, NACRS	Binary indicator of the presence of any COPD ACSC-related hospitalization over the observation window
Number of benzodiazepine prescriptions	ODB	Number of benzodiazepine prescriptions over the observation window
Number of fluoroquinolones prescriptions	ODB	Number of fluoroquinolones prescriptions over the observation window
Number of antilipemic statins prescriptions in quarter 3	ODB	Number of antilipemic statins prescriptions in the 3rd quarter of the observation window
Number of antilipemic statins prescriptions in quarter 1	ODB	Number of antilipemic statins prescriptions in the 1st quarter of the observation window
Clinician specialty in quarter 3 - internal medicine	OHIP	Number of visits to a clinician specialized in internal medicine during the 3rd quarter of the observation window
Time since last fluoroquinolones prescriptions	ODB	If there is, time since last fluoroquinolones prescription over the observation window
Clinician visit feecode in quarter 3 - chest radiology	OHIP	Number of clinician visits related to chest radiology in the 3rd quarter of the observation window
Time since last narcotics (opiate agonists) prescriptions	ODB	If there is, time since last narcotics (opiate agonists) prescription over the observation window
Time since last ACE inhibitors prescriptions	ODB	If there is, time since last ACE inhibitors prescription over the observation window
Number of ACE inhibitors prescriptions	ODB	Number of ACE inhibitors prescriptions over the observation window
Number of narcotics (opiate agonists) prescriptions	ODB	Number of narcotics (opiate agonists) prescriptions over the observation window
Clinician specialty in quarter 7 - dermatology	OHIP	Number of visits to a clinician specialized in dermatology during the 7th quarter of the observation window
Clinician specialty in quarter 6 - emergency medicine	OHIP	Number of visits to a clinician specialized in emergency medicine during the 6th quarter of the observation window

Supplementary Table 4. List of all features used in the XGBoost model and Linear Regression.

Supplementary Table 4 - Description

The features that were extracted from the data sources include patients' demographic and geographical information, drug prescription history, chronic conditions, clinician visits, hospital usage, as well as past history of ACSC and laboratory results.

Demographic information included the age of the patient, their sex, immigration status, and the date of arrival in Canada and the country of birth if applicable. Geographical information not only consisted of the address of the patient at the three postal code digit level (also called as a *forward sortation area* of *FSA*), but also of the aggregation of different socioeconomic status measures at the FSA-level. This included quintiles of area-level education or income, as well as marginalization indices measuring material deprivation, residential instability, dependency, and ethnic concentration.

Chronic conditions of the patient examine the presence of 18 comorbidities such as hypertension, diabetes, chronic obstructive pulmonary disease, and mood disorder. Drug prescription history has the information of the quantity and the name of medications that were prescribed to a given patient. It also has the information of whether or not the medication was prescribed in a long-term care facility. It is important to note that this contains the information of the medications that were dispensed but we cannot know if they were actually taken by the patient. These records are nevertheless a good proxy for the condition of the patient as well as an indicator of whether or not the patient is exposed to polypharmacy, the concurrent use of more than five medications. The physician visits and hospital usage information contained the type of physician or hospital visit and the fee code related to the visit. Finally, from hospitalization data, information on the presence and type of ACSC-related hospitalization was extracted for each patient.

To control for the diversity of the patients, we set a threshold to the frequency of each feature to avoid processing very rare features values that are not generalizable- for instance, the drug class information was processed only if at least 25% of the patients were prescribed medications of the same class. All categorical values were one-hot encoded. Demographic and geographical features were prepared as fixed attributes of the patient at the time of observation window. The other features were aggregated at a quarterly-level to account for the characteristics of the datasets being updated every three months. We also aggregated the latter at the observation window-level to obtain global health status of the patient, such as the total number of prescriptions of drug class A or the time since the last ACSC-related hospitalization.

The total number of features reached 2,082 after the initial preparation. In order to select the most important features for the model as well as to ensure its generalizability, we took a greedy approach to select a small subset of features that would ensure the performance of the model to match that of the model using all features. Starting from a subset of 50 most contributing and geographic features we wanted to keep in, other features were added to the subset only if it led to a visible increase in model performance when evaluated on the validation set. At the end of the process, we ended with 140 features in total.

Supplementary Method 1

Given the low incidence rate of ACSCs (1.83% in the training set), we had a very imbalanced dataset and while training, we undersampled negative data points (no ACSC hospitalization in prediction window) by selecting only one out of 8 negative samples, and kept all positive data points (ACSC hospitalization in prediction window). The validation and testing sets were left untouched (i.e. were not undersampled). The final predictions made by the model were calibrated to account for undersampling [3]. The model was trained with the following hyperparameters: a learning rate of 0.05, a maximum tree depth of 10, and both the fraction of columns to be randomly sampled and the subsample ratio of columns for each split set at 0.7. The alpha, gamma and lambda values were 0.3, 0.1 and 0.5 [4]. These were selected after a hyperparameter grid search, consisting in fixing ranges and increments for given hyperparameters and testing all combinations of values to find the optimal one [5] .

Supplementary Method 2

We trained a Logistic Regression (LR), a model that is widely used in developing healthcare risk prediction models. The LR model was trained using the same features as the XGBoost model. Feature values were normalized and scaled between 0 and 1, and depending on the type of features either 0 or 1 was imputed for missing values. For instance, if the feature referred to the incidence / number of events then 0 was imputed, and if the feature was time-related (lower value meant a more recent interaction) the value was set to 1.

As seen in Table S5, XGBoost model is able to predict the risk of ACSC-related hospitalizations with a higher AUC. While the XGBoost model is able to handle multiple types of variables without any feature engineering, LR requires feature normalization and all features were scaled between 0 and 1.

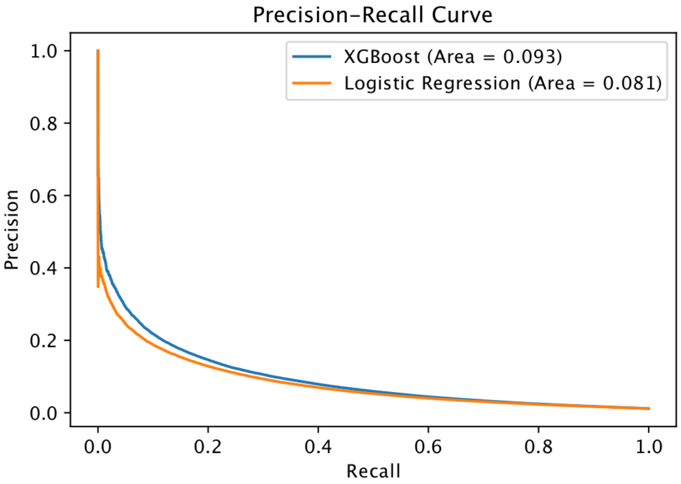
We compared the AUC value for the whole cohort of patients, where we saw a gain of 1.2 by using the XGBoost model. We reported the range of AUC as well as its average obtained by training the model 5 times with random restarts. We also compared the AUC values when looking at the “young” patients who newly qualified by turning 65 during the test set study period (target window between January 2016 and December 2017). This ensures that the algorithm will be able to correctly assess the risk of the patients who just turn 65 and are added to the patient group in the test set. XGBoost again shows an AUC gain of 1.4 compared to LR. We also compared the precision-recall curve of our XGBoost model to a logistic regression model. The rapid drop in precision is predictable due to the rarity of the ACSC-related outcomes.

Supplementary Table 5

	All Patients	New Patients
Logistic Regression	79.3 (79.2-79.5)	78.4 (78.2-78.6)
XGBoost	80.5 (80.4-80.5)	79.8 (79.6-79.9)

Supplementary Table 5. AUC values for XGBoost and LR on all patients and newly added patients to the test set.

Supplementary Figure 1



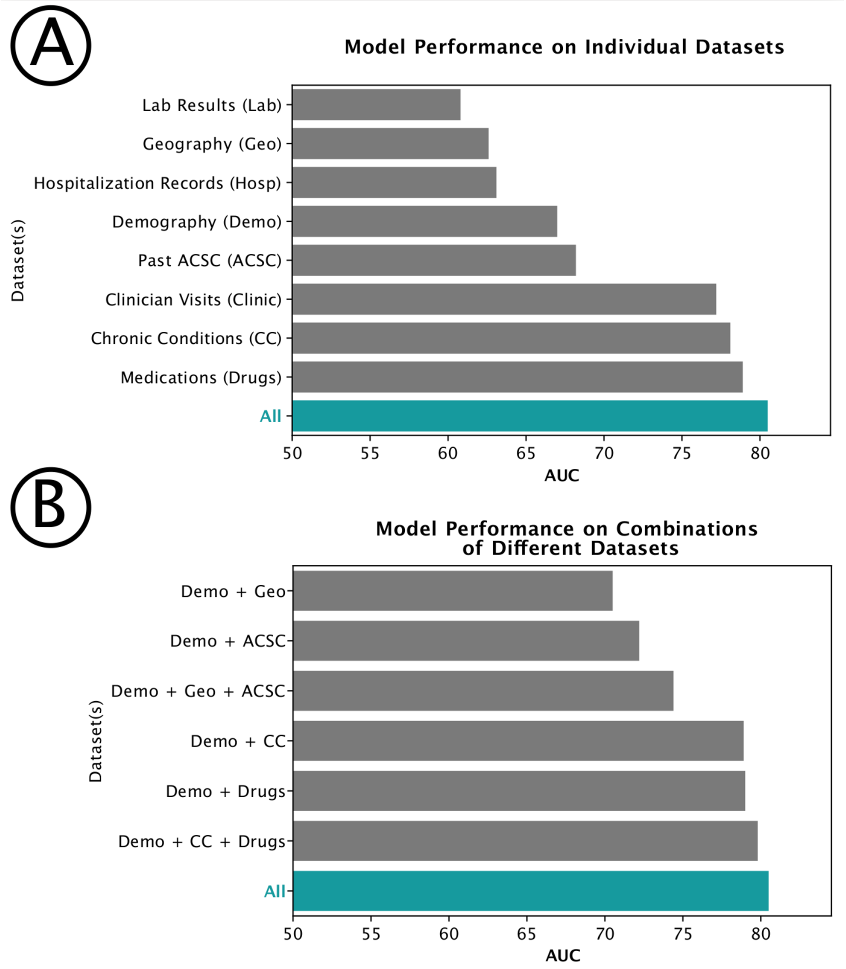
Supplementary Figure 1. Precision-Recall Curve comparing XGBoost model against Logistic Regression.

Supplementary Table 6

	Top 1%	Top 5%	Top 10%	All patients
Sex - Female (%)	50.2	48.7	48.7	52.3
Sex - Male (%)	49.8	51.3	51.3	47.7
Age (mean)	70.5	70.1	69.9	69.0
Immigrant (%)	3.61	4.46	4.92	10.9
Non-Immigrant (%)	96.4	95.5	95.1	89.1
No history of ACSC (%)	13.4	45.9	62.8	94.7
History of ACSC (%)	86.6	54.1	37.2	5.28
Lives in Rural Areas (%)	24.3	23.5	22.9	13.6
Lives in Urban Areas (%)	75.7	76.5	77.1	86.4
Education quantile (mean)	2.41	2.51	2.57	3.09
Income quantile (mean)	2.45	2.54	2.60	3.05
Number of events (median)	688	524	439	210

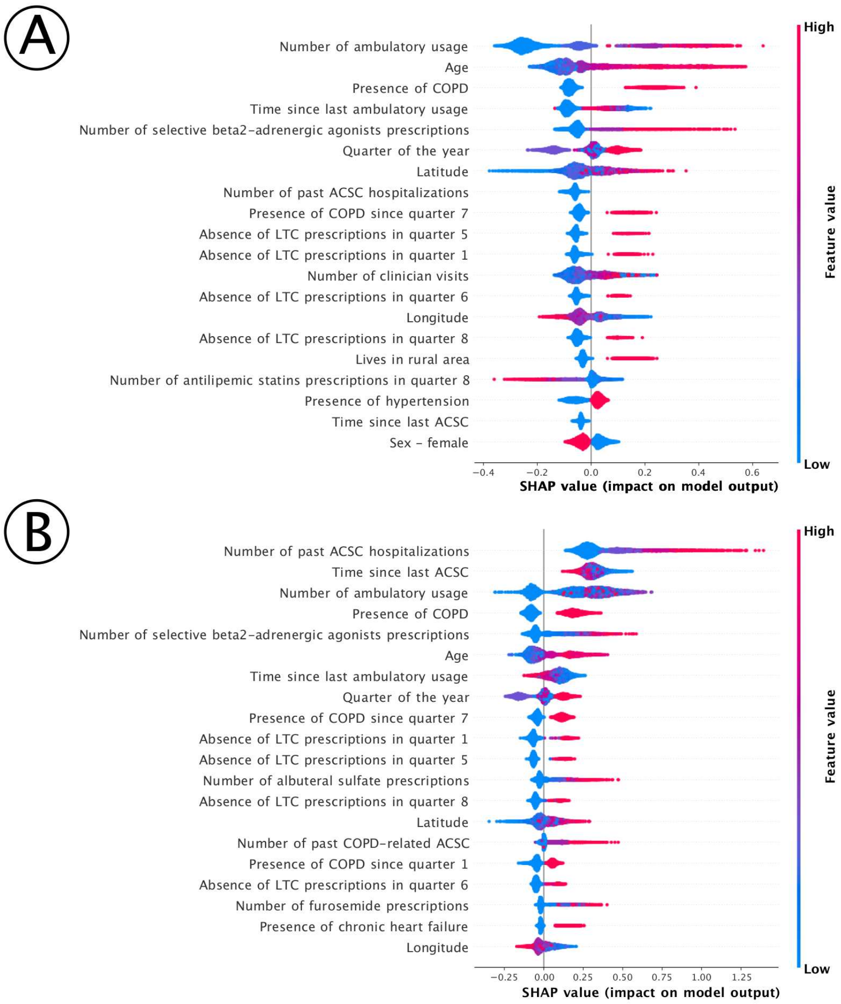
Supplementary Table 6. Baseline characteristics comparison for patients in different risk level groups, predicted by the model. For education and income quintiles, higher index refers to higher education level and income respectively, in the area a given patient lives in. The number of events refers to the number of any interaction a given patient had with the healthcare system - clinician visits, hospitalization, ambulatory usage, lab tests, and drug prescriptions.

Supplementary Figure 2



Supplementary Figure 2. Model performance on different features of subsets.

Supplementary Figure 3



Supplementary Figure 3. A) Feature importance for patients who do not have a history of an ACSC-related hospitalization. B) Feature importance for patients who have a history of one or more ACSC-related hospitalizations.

References

1 Ambulatory Care Sensitive Conditions. https://indicatorlibrary.cihi.ca/display/HSPIL/Ambulatory+Care+Sensitive+Conditions (accessed 6 Jan 2021).

2 Codes and classifications for clinical health data. https://www.cihi.ca/en/submit-data-and-view-standards/codes-and-classifications (accessed 6 Jan 2021).

3 Pozzolo AD, Caelen O, Johnson RA, *et al.* Calibrating Probability with Undersampling for Unbalanced Classification. 2015 IEEE Symposium Series on Computational Intelligence. 2015. doi:10.1109/ssci.2015.33

4 Python API Reference — xgboost 1.4.0-SNAPSHOT documentation. https://xgboost.readthedocs.io/en/latest/python/python_api.html (accessed 6 Jan 2021).

5 Larochelle H, Erhan D, Courville A, *et al.* An empirical evaluation of deep architectures on problems with many factors of variation. Proceedings of the 24th international conference on Machine learning - ICML ’07. 2007. doi:10.1145/1273496.1273556