Can patients with dementia be identified in primary care electronic medical records using natural language processing?

Laura C Maclagan, Mohamed Abdalla, Daniel A Harris, Therese A Stukel, Branson Chen, Elisa Candido, Richard H Swartz, Andrea Iaboni, R. Liisa Jaakkimainen, Susan E Bronskill

Technical Appendix 1. Text preprocessing, hyperparameter tuning and cross-validation methodological details for machine learning models to classify patients with dementia in primary care electronic medical records, Ontario, Canada

Supplementary Table 1. Health administrative databases used in the present study

Supplementary Table 2. Chronic conditions and their associated diagnostic codes included in the present study

Supplementary Table 3. Hyperparameters and preprocessing for machine learning models to classify patients with dementia in primary care electronic medical records, Ontario, Canada

Supplementary Table 4. Characteristics of all consult and progress notes in the primary care electronic medical record database prior to March 31st, 2016, overall and by dementia, Ontario, Canada

Supplementary Table 5. Dementia-related features used as inputs for the natural language processing algorithm that was applied to primary care electronic medical records, by thematic group

Supplementary Figure 1. Methodological schematic of the evaluation of machine learning model performance to classify patients with dementia by time before and after patients' index dates in primary care electronic medical records in Ontario, Canada

Technical Appendix 1. Text preprocessing, hyperparameter tuning and cross-validation methodological details for machine learning models to classify patients with dementia in primary care electronic medical records, Ontario, Canada

The following text preprocessing steps were completed on the EMR free-text clinical notes and feature list to improve standardization and reduce noise in the data:

a. a. All punctuation was replaced by spaces (e.g., "word1-word2" was converted to "word 1 word 2").

b. Words were partitioned using space separation (i.e., "the cat meowed" would be composed of tokens ["the", "cat", "meowed"]).

c. All words were converted to lower-case.

d. All words were stemmed using the base Snowball stemmer available in the NLTK python package.

Five-fold nested cross validation was used to train each model and test its performance [1]. The original data were resampled five times, creating five distinct sets of training (70%) and test (30%) dataset pairs. We ensured that the relative proportion of patients with dementia was approximately balanced across all training and test dataset pairs. The test datasets were withheld from the entire training process until after models were tuned.

Within each training dataset, the data were further split into training (90%) and validation (10%) partitions to conduct 10-fold cross-validation to tune each model's hyperparameter(s). The number of trees was tuned for gradient boosted models. Number of hidden layers was tuned for neural networks. We treated feature pre-processing as an additional tunable parameter. All other hyperparameters were held at their default values. The set of hyperparameters and other tuned parameters that yielded the best model performance, defined as maximizing the F1 score on the training and validation datasets, was applied to the corresponding test datasets. The complete set of hyperparameters and preprocessing options are presented in **Supplementary Table 3**.

REFERENCES

1. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 2006;7(1):91. doi: 10.1186/1471-2105-7-91.

Supplementary Table 1. Health administrative databases used in the present study

Dataset name	Dataset description
Canadian Institute of Health Information – Discharge Abstract Database (CIHI-DAD)	The CIHI-DAD is compiled by the Canadian Institute for Health Information (CIHI) and contains administrative, clinical (diagnoses and procedures/interventions), demographic, and administrative information for all admissions to acute care hospitals in Ontario.
Ontario Health Insurance Plan (OHIP)	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.
Ontario Drug Benefit (ODB)	The ODB database contains prescription medication claims for those covered under the provincial drug program, including: all residents of long-term care, those aged 65 years and older, patients receiving services under the Ontario Home Care program, those receiving social assistance, and residents eligible for specialized drug programs.
Continuing Care Reporting System (CCRS) – Long-Term Care	The CCRS-LTC database is compiled by the Canadian Institute for Health Information (CIHI) and comprises all mandatory clinical assessments of residents living in long-term care in Ontario.
Registered Persons Database (RPDB)	The RPDB is a population-based registry maintained by the Ontario Ministry of Health, used to derive basic sociodemographic characteristics (age, sex, postal code) and date of death (for deceased individuals). The RPDB also indicates the time periods for which an individual was eligible to receive publicly funded health insurance benefits.
Electronic Medical Records Primary Care Database (EMRPC)	The EMRPC includes electronic medical record data from a convenience sample of primary care physicians in Ontario, Canada. EMRPC includes all information contained in the patient EMR, such as clinical encounters, current and past medical history, family history of disease, risk factors, treatments (e.g., medications and immunizations), laboratory and diagnostic tests, physician billings, referral letters, and clinical notes.

Supplementary Table 2. Chronic conditions and their associated diagnostic codes included in the present study

Condition/diagnosis	ICD 9 / OHIP	ICD 10
Acute Myocardial Infarction (AMI)[1]	410	I21
Osteoarthritis	715	M15-M19
Other Arthritis (includes Synovitis, Fibrositis, Connective tissue disorders, Ankylosing spondylitis, Gout Traumatic arthritis, pyogenic arthritis, Joint derangement, Dupuytren's contracture, Other MSK disorders)	727, 729, 710, 720, 274, 716, 711, 718, 728, 739	M00-M03, M07, M10, M11-M14, M20-M25, M30-M36, M65-M79
Arthritis - Rheumatoid arthritis[2]	714	M05-M06
Asthma[3]	493	J45
Cancers (all)	140-239	C00-C26, C30-C44, C45-C97
Congestive Heart Failure[4]	428	I500, I501, I509
Chronic Obstructive Pulmonary Disease[5]	491, 492, 496	J41, J42, J43, J44
Coronary syndrome (excluding AMI)	411-414	I20, I22-I25
Diabetes[6]	250	E10, E11, E13, E14
Hypertension[7]	401, 402, 403, 404, 405	I10, I11, I12, I13, I15
Inflammatory Bowel Disease (Crohn's disease or Ulcerative Colitis)[8]	555, 556	K50, K51
Liver disease	5712, 5715, 5716	K703, K71.7, K74
Mood, anxiety, depression, and nonpsychotic disorders	296, 300, 309, 311	F30, F31, F32, F33, F34 (excl. F34.0), F38, F39, F40, F41, F42, F43.1, F43.2, F43.8, F44, F45.0, F45.1, F45.2, F48, F53.0, F68.0, F93.0, F99
Other mental illnesses	291, 292, 295, 297, 298, 299, 301, 302, 303, 304, 305, 306, 307, 313, 314, 315, 319	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
Osteoporosis	733	M81, M82
Stroke (excluding transient ischemic attack)	430, 431, 432, 434, 436	I60-I64
Traumatic brain injury[9]		S01.0-S01.9; S02.0, S02.1, S02.3; S02.7-S02.9; S04.0; S06.0-S06.9; S07.0; S07.1; S07.8; S07.9; S09.7-S09.9;T01.0; T02.0; T04.0; T06.0; T90.1; T90.2; T90.4; T90.5; T90.8; T90.9; ICD-9: 800.0-801.9, 803.0-804.9, 850.0- 854.1, 950.1-950.3, 959.0

AMI, Asthma, COPD, CHF, Dementia, Diabetes, Hypertension, Rheumatoid Arthritis, and Inflammatory Bowel disease are based on validated case algorithms (see Sources, respectively). All other conditions required at least one diagnosis recorded in acute care or two diagnoses recorded in physician billings within a two-year period[10]

REFERENCES

1. Austin PC, Daly PA, Tu JV. A multicenter study of the coding accuracy of hospital discharge administrative data for patients admitted to cardiac care units in Ontario. American Heart Journal 2002;144:290–6

2. Widdifield J, Bernatsky S, Paterson JM, Tu K, Ng R, Thorne JC, Pope JE, Bombardier C. Accuracy of Canadian health administrative databases in identifying patients with rheumatoid arthritis: a validation study using the medical records of rheumatologists. Arthritis Care Res 2013; 65(10): 1582-1591.

3. Gershon AS, Wang C, Guan J, Vasilevska-Ristovska J, Cicutto L, To T. Identifying patients with physician-diagnosed asthma in health administrative databases. Can Respir J 2009;16:183–8.

4. Schultz SE, Rothwell DM, Chen Z, Tu K. Identifying cases of congestive heart failure from administrative data: a validation study using primary care patient records. Chronic Diseases and Injuries in Canada 2013;33:160–6.

5. Gershon AS, Wang C, Guan J, Vasilevska-Ristovska J, Cicutto L, To T. Identifying Individuals with Physician Diagnosed COPD in Health Administrative Databases. Copd 2009;6:388–94.

6. Hux JE, Ivis F, Flintoft V, Bica A. Diabetes in Ontario: Determination of prevalence and incidence using a validated administrative data algorithm. Diabetes Care 2002;25:512–6.

7. Tu K, Campbell NR, Chen Z-L, Cauch-Dudek KJ, McAlister FA. Accuracy of administrative databases in identifying patients with hypertension. Open Med 2007;1:e18–26.

8. Benchimol EI, Guttmann A, Mack DR, Nguyen GC, Marshall JK, Gregor JC, Wong J, Forster AJ, Manuel DG. Validation of international algorithms to identify adults with inflammatory bowel disease in health administrative data from Ontario, Canada. J Clin Epidemiol. 2014 Aug;67(8):887-96. doi: 10.1016/j.jclinepi.2014.02.019.

9. Marr A, Coronado V, editors. Central Nervous System Injury Surveillance Data Submission Standards—2002. Atlanta (GA): Centers for Disease Control and Prevention, National Center for Injury Prevention and Control; 2004.

10. Mondor, L., Maxwell, C. J., Hogan, D. B., Bronskill, S. E., Gruneir, A., Lane, N. E., & Wodchis, W. P. (2017). Multimorbidity and healthcare utilization among home care clients with dementia in Ontario, Canada: a retrospective analysis of a population-based cohort. *PLoS medicine*, *14*(3), e1002249.

Supplementary Table 3. Hyperparameters and preprocessing for machine learning models to classify patients with dementia in primary care electronic medical records, Ontario, Canada

LASSO (L1 norm) and Ridge Regression (L2 norm)	
λ	**Adjustment for notes**
0.01	Yes
0.1	No
1.0	
10.0	

Gradient boosted models	
Number of trees	**Adjustment for notes**
100	Yes
150	No
200	

Neural network	
Hidden layer sizes	**Adjustment for notes**
100,	Yes
100,25	No
100,25,10	

Adjustment for notes: whether the counts for each term is divided by the number of notes.

C (LASSO and ridge): Each λ describes the inverse of regularization strength.

Number of trees (gradient boosted models): The number of boosting stages to perform. All other hyperparameters held at their default values.

Hidden Layer Sizes (neural network): The ith element represents the number of neurons in the ith hidden layer. All other hyperparameters held at their default values.

Supplementary Table 4. Characteristics of all consult and progress notes in the primary care electronic medical record database prior to March 31st, 2016, overall and by dementia, Ontario, Canada

Types of clinical notes	Overall Sample (N=44,674)	Dementia (N=526)	Without Dementia (N=44,148)*
Consult & progress notes	N_{notes}=5,663,327	N_{notes}=112,053	N_{notes}=5,551,274
Mean (SD)	126.8 (94.2)	213.0 (128.4)	125.7 (93.3)
Median (IQR)	102.0 (63.0-163.0)	179.5 (124.0-271.5)	101.0 (63.0-161.0)
Range	3-1,920	25-857	3-1,920
Consult notes only	N_{notes}=2,299,276	N_{notes}=43,585	N_{notes}=2,255,691
Mean (SD)	51.5 (41.3)	82.9 (53.4)	51.2 (41.0)
Median (IQR)	41.0 (23.0-68.0)	71.0 (46.0-108.8)	40.0 (23.0-67.0)
Range	1-708	5-444	1-708
Progress notes only	N_{notes}=3,364,051	N_{notes}=68,468	N_{notes}= 3,295,583
Mean (SD)	75.3 (61.0)	130.2 (86.5)	74.7 (60.4)
Median (IQR)	59.0 (35.0-96.0)	108.5 (71.0-167.0)	58.0 (35.0-95.0)
Range	2-1,243	13-597	2-1,243

Notes: N=Sample size; N_{notes} = sample of notes; SD=standard deviation; IQR=interquartile range
*Complete progress notes were available for the entire sample. Consult notes were available for 100% of patients with dementia and 99.9% (n=44,103) of patients without dementia

Supplementary Table 5. Dementia-related features used as inputs for natural language processing algorithm that was applied to primary care electronic medical records, by thematic group

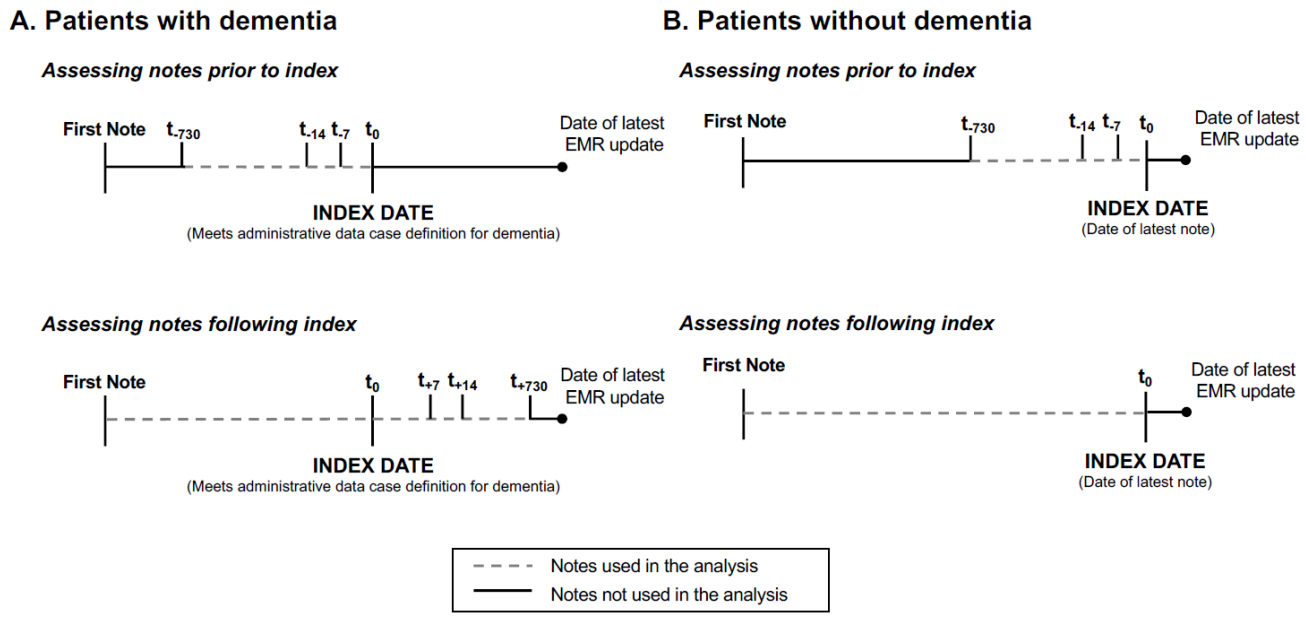
Dementia-related feature categories	Features (separated by commas)
Cognition	+6 item cognitive impairment test, 0 memory/recall, 6cit, 7-minute screen, abbreviated mental test, ad8, addenbrookes cognitive examination, agnosia, amnesia, amt, anosagnosia, aphasia, apraxia, attention, backwards, cannot [unable/does not] identify/recognize, can't follow, cdt, clock, clock drawing test, cmo, cognition, cognitive, confused [confusion], decision making, decreased/impaired cognition, difficulty [problems/ trouble/hard to] finding words [word finding], difficulty/unable to follow commands, directions, instructions, does not [unable/cannot] remember/recall/recognize/recognise, does not know [self/who/why/where], domain, drawing, eight-item informant interview to differentiate aging and dementia, executive dysfunction, executive function, fcsrt, follow, forgetful/forgets/forgetfulness, free and cued selective reminding test, general practitioner assessment of cognition, gpcog, inattentive, iqcode, lack of/poor insight, language, limited by cognition, making decisions, mci, memory, memory impairment screen, mic, mild, mild cognitive impairment, mini mental state exam, mini mental state examination, mini-cog, mis, mmse, moca, montreal cognitive assessment, poor [unable to] recall, poor [unreliable/unable] historian/history, question, recall, remember, rowland universal dementia assessment scale, rudas, sbt, scale, scored/score, short blessed test, short informant questionnaire on cognitive decline in the elderly, short portable mental status questionnaire, spmsq, standardized mini-cog instrument, thought process disorganized/illogical/ incoherent, unable [does not] understand, vascular cognitive impairment, visuospatial, word
Function	[getting/gets/got] lost, adl/ adls/adl's, ambulation, bank/banking, bath/bathe/bathing, bed, bladder, bowel/fecal, can't/cannot button, can't/cannot find, car, chair, continent/continence, daily activities/activities of daily living/ activity of daily living, dress, dressing, driveable, driving, feed, feeding, fire, food on stove, function, functional, gait, iadl/iadl's/iadls, incontinent/incontinence, independent, independent activities of daily, living/independent activity of daily living/, leave [leaving/leaves/left] food, left alone, losing/loses keys, losing/loses objects, meals, mobile, mobility, not independent/ no longer independent, personal support worker, plan, plan/planning, planning, psw, recipes, safety, self-care/ self care, shop/shopping, shower/showering, sit, slide, stand, toilet/toileting, transfer, urine/urinary, walker
Health System Use	acute, admission, alzheimer, assessment, atrophy, bipolar, blister, brain, care, ccac, cerebral, colitis, community care access centre, computed tomography, consult, contact, continue, contrast, ct, cxr, debridement, dementia, dermatology, device, diabetes, discharge,

	disease, disorder, disruptive behavior disorder, dni, dnr, emergency, geriatric, geriatrician, gtube, hematoma, hemorrhage, history, history , hospice, hospital, infection, influenza, injury, inr, international normalized ratio, intracranial, long-term care, ltc, magnetic resonance imaging, management, microangiopathic, mri, nph, nurse, nursing, nursing home, osteomyelitis, pacemaker, palliative, parkinson, peg, picc, placement, pneumonia, pressure, primary, psych, psychiatric, report, respite, rib, sacral, scan, services, slp, spine, staph, supplement, surgery, surgery, test, testing, trach, tract, transition, trauma, treat, tube, unit, visit, white matter disease, white matter hyperintensities
Dementia Medication	aricept, cholinesterase/ cholinesterase inhibitor/ che-i/ chei, donepezil, elexon, galantamine, memantine, reminyl, rivastigmine
Other Medication	glimepiride, nadolol, acarbose, accupril, ace, acebutolol , ace-iii, ace-inhibitor, ace-r, aliskiren, alpha-glucosidase inhibitor, altace, amaryl, amlodipine, ampicillin, angiotensin converting enzyme inhibitors, angiotensin ii receptor blocker, angiotensin receptor blockers, anti-anxiety, antibiotic, anti-cholinergic, antidepressant, anti-depressant, antipsychotic, anxiolytic, apixaban, arb, asa, aspirin, atacand, atenolol, ativan, atorvastatin, avapro, b12, benazepril, benzo, benzodiazepine, beta-blocker, betaloc, betapace, bid, biguanide, bisoprolol, blocadren, brevibloc, bumetanide, bumex, byetta, bystolic, calcium channel blockers, canagliflozin, candesartan, capoten, captopril, carvedilol, ccbs, cefepime, ceftriaxone, celexa, chlorthalidone, cilazapril, cipralex, citalopram, clonazepam, clopidogrel, coconut oil, coreg, corgard, coumadin, coversyl, crestor, dabigatran, dapagliflozin, depakote , diabeta, diamicron, dilaudid, direct oral anticoagulants, direct renin inhibitors, diuretics, doac, dose, dpp-4 inhibitor, ebixa, edecrin, edoxaban, eliquis, empagliflozin, enalapril, entresto, eprosartan, escitalopram, esmilol, ethacrynic acid, euglucon, exelon, exenatide, fluvastatin, forxiga, fosinopril , furosemide, gingko, glicazide, glimepiride, glipizide, glp-1 receptor agonist, glucobay, gluconorm, glucophage, glucotrol, glumetza, glyburide, haldol, haloperidol, hctz, heparin, hydrochlorothiazide, hydrochlorothiazide, imovane, indapamide, inderal, inhibace, insulin, invokana, irbesartan, janumet, januvia, jardiance, jentadueto, klonopin, komboglyze, labetalol, lantus, lasix, lescol, levatol, linagliptin, lipitor, liraglutide, lisinopril, lmwh, lopressor, lorazepam, losartan, lotensin, lovastatin, lozide, mavik, medication, meds, melatonin, metformin, metolazone, metoprolol, mevacor, monocor, monopril, morphine, nateglinide , nebivolol, nmda receptor antagonists/ n-methyl-d-aspartate receptor antagonists, noac, non-sulfonylurea insulin secretagogue, normodyne, novel oral anticoagulants, olanzapine, olmesartan, olmetec, omega-3, onglyza, oral hypoglycemic, oxazepam, paroxetine, paxil, penbutolol, perindopril, pindolol, plavix, pradaxa, pravachol, pravastatin, prinivil, prn, prophylaxis, propranolol, qhs, quetiapine, quinapril, ramipril, rasilez, repaglinide, restoril, risperidal, risperidone, rivaroxaban, rosuvastatin, route, saxagliptin,

	sectral, selective serotonin reuptake inhibitor, serax, seroquel, serotonin-norepinephrine reuptake inhibitor, sertraline, sglt2 inhibitors, simvastatin, sitagliptin, sleeping pill, snri, sotalol, ssri, starlix, statins, subcutaneously, sulfonylurea insulin secretagogue, sulfonylureas, telmisartan, temazepam, tenormin, tevetan, timolol, toprol-xl, trajenta, trandate, trandolapril, trazodone, twynsta, tylenol, valsarta, vasotec, victoza, viskazide, warfarin, xarelto, zaroxolyn, zestril, zocor, zoloft, zopiclone
Social	access, accompanied, address, adjustment, assist, assistance, attorney, call, caregiver, chaplain, client, community, dad, daughter, faith, family, father, finances, home, home care, house, husband, law, license, meditation, mom, mother, outreach, phone, poa, power of attorney, prayer, relation, religion, religious, residence, retirement, sdm, son, spiritual, spirituality, spouse, staff, substitute decision maker, support, team, wife, work, social
Symptoms	a&o x 1, 1-2, or 2, abdominal, aberrant behavior, abnormal behavior, accidental fall, acts out dreams, afraid/scared, aggression, aggressive, aggressive behavior, aggressive personality, agitate, agitated, agitated, behavior, agitated depression, agitation, alcohol, alert, anger, angry, anxious, apathy/apathetic, appetite, aspiration, attention seeking behavior, auditory hallucination, awake at night, bacteremia, balance, baseline, behavior, behavior abnormality, behavior changes, behavior disorder, behavior impaired, behavior issues, behavioral, behavioral abnormality, behavioral and psychological symptoms of dementia, behavioral changes, behavioral impaired, behavioral issues, behaviour, biting, blood, bpsd, burn, burn pots, change, change in behavior, chest, collar, combative, comfort, comfortable, command, complaining, compression, consistency, constant requests for attention/neediness, constipation/constipated, cough, crying, crying spells, cue, decline, decreased motivation [unmotivated], decubitus, delinquent behavior, delirious, delirium, delusions, deny, depressed/depression, destruction of property, diarrhea, diet, diff, difficulties, difficulty, difficulty [does not have capacity/unable], discussion, disinhibit/disinhibited, disoriented/not oriented, displace, distress, disturbance in physical behavior, dysphagia, edgy, emotional, emotional crisis, emotional instability, emotional issues, emotional lability, emotional stress, emotional upset, emotionally labile, euphoric/euphoria, excessive spitting, excessively loud voice volume, fall, fall down stairs, fall down steps, fall from bed, fall from chair, fall from height, fall from slipping, fall from stairs, fall from standing height, fall from stool, fall from toilet seat, fall from wheelchair, fall in bathtub, fall in home, fall in nursing home, fall in shower, fall on concrete, fall on ice, fall on snow, fall on stairs, fall on steps, fall risk, falling/fallen, falls, fearful [fear], fever, fidgets [fidgeting/ fidgety], finger pointing, flat/blunted/restricted affect, food, foot, foul language, fracture, gastronomy, gesturing, give, glucose, grabbing, hallucination, hallucinations/ hallucinating, hand , harm self, head, hearing things, heel, hitting, hitting self or others,

	hoarding, hold, hostile, hostile behavior, hostility, hurt self or others, hurting self or others, hypomanic behavior, illness, impairment, improve, impulsive, impulsive behavior , impulsivity, inappropriate behavior, increase, intermittent explosive outburst, irritability, irritability and anger, irritable, irritation, jittery, jumpiness, jumpy, kicking, labile, lack of cooperation, lack of patient cooperation, lesion, malnutrition, mania, manic behavior, mannerism, meal, mechanical fall, mental, misplace, misplacing items, mood, motor restlessness, mouth, movement, multiple complaints, multiple somatic complaints, nail biting, negative, negativism, nervous, nervousness, neurological, non-compliant, non-self-regulatory behavior, nonsensical [does not make sense], nutrition, obsessive compulsive, odd behavior, orientation, oriented, oriented to person/self only, outbursts, pacing, pai, paranoid, perseverative, perseveration, perseverating, physical sexual advances , picking at, poor [changing/deceased/ worsening] mental status, poor/impaired judgement, profanity/cursing/swearing, progress protein, psychomotor agitation, pulling at [on], puree, pushing, rash, refuse, repeat/repeating/repeated (questions, story, stories), repetitions questions, statements, repetitious mannerisms, repetitive, repetitive questions/sentences, repetitive verbilization/vocalization, resistance to change, resistant, resistiveness, responsive behaviour, restless, restless, restlessness, restlessness and agitation, restraints, rocking, sad/sadness/tearful/anhedonic/andedonia, scalp, scratching, screaming, seeing things, self-abusive, self-abusive behavior, self-cutting, self-destructive, self-destructive behavior, self-injurious behavior, self-injury, self-mutilation, sexual issues, shakes, shakiness, shaking, shaking all over, shouting, shoving, skin, slamming doors, sleep, smearing, solid, speech, spitting, spoke, stool, strange behavior, strange noises (unusual laughter, crying), stubbornness, sugar, sundowning [sundowns], supine, suspicious, swallow, symptom, tearing objects, tense, tensenless, tension, thin, threatened, thrive, throwing objects, time, tremor, ulcer, unable, unable to control behavior, unable to keep still, uncooperative, urinary, urinary tract infection, urine, uti, verbal sexual advances, verbally abusive behavior, violence, visual hallucination, wander, wandering, wanders, wanders at night, water, weight, weird behavior, weirdness, well, wind, yelling
Other	able, ago, alter, altered, america, application, apply, area, attempt, back, course, culture, daily, date, day, delayed, due, extend, face, failure, goal, health, hour, information, intake, leave, level, life, line, liquid, list, live , lives, living, measure, meet, normal, oral, outcome, past year, person/place and/or time, pna, provide, recent, recurrent, review, roll, says, secondary, seniors, seniors moments, site, source, stage, state, status, subjective, susceptible, term, thick, top, unable/ not able, unite, vac, year, nectar, objective

Supplementary Figure 1. Methodological schematic of the evaluation of machine learning model performance to classify patients with dementia by time before and after patients' index dates in primary care electronic medical records in Ontario, Canada

Panel A describes patients with dementia, where the study index date represents time zero (t_0). When assessing notes prior to index and following index, only notes indicated by the hashed lines were included. For example, when assessing performance using notes 7 days prior to index, only notes between 7 days prior to the index date and the index date (but not including the index date) were included.

Panel B describes patients without dementia, where the study index date (t_0) is date of the last note in the EMR database. When assessing notes prior to index and following index, only notes indicated by the hashed lines were included. For example, when assessing the seven days prior to index, only notes from 7 days prior to the latest note in the EMR database (not inclusive) were included. Given the index date for patients without dementia was the date of the last note in the EMR, we cannot look forward in the days following index, therefore all notes are included.