BMJ Open

Predicting dementia diagnosis from cognitive footprints in electronic health records: a case–control study protocol

Hao Luo ,[1,2] Kui Kai Lau,[3] Gloria H Y Wong ,[1] Wai-Chi Chan,[4] Henry K F Mak,[5] Qingpeng Zhang,[6] Martin Knapp,[7] Ian C K Wong[8,9]

To cite: Luo H, Lau KK, Wong GHY, *et al*. Predicting dementia diagnosis from cognitive footprints in electronic health records: a case–control study protocol. *BMJ Open* 2020;**10**:e043487. doi:10.1136/bmjopen-2020-043487

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-043487).

Received 05 August 2020
Revised 31 October 2020
Accepted 02 November 2020

Check for updates

© Author(s) (or their employer(s)) 2020. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to
Dr Hao Luo; haoluo@hku.hk

ABSTRACT

Introduction Dementia is a group of disabling disorders that can be devastating for persons living with it and for their families. Data-informed decision-making strategies to identify individuals at high risk of dementia are essential to facilitate large-scale prevention and early intervention. This population-based case–control study aims to develop and validate a clinical algorithm for predicting dementia diagnosis, based on the cognitive footprint in personal and medical history.

Methods and analysis We will use territory-wide electronic health records from the Clinical Data Analysis and Reporting System (CDARS) in Hong Kong between 1 January 2001 and 31 December 2018. All individuals who were at least 65 years old by the end of 2018 will be identified from CDARS. A random sample of control individuals who did not receive any diagnosis of dementia will be matched with those who *did* receive such a diagnosis by age, gender and index date with 1:1 ratio. Exposure to potential protective/risk factors will be included in both conventional logistic regression and machine-learning models. Established risk factors of interest will include diabetes mellitus, midlife hypertension, midlife obesity, depression, head injuries and low education. Exploratory risk factors will include vascular disease, infectious disease and medication. The prediction accuracy of several state-of-the-art machine-learning algorithms will be compared.

Ethics and dissemination This study was approved by Institutional Review Board of The University of Hong Kong/Hospital Authority Hong Kong West Cluster (UW 18-225). Patients' records are anonymised to protect privacy. Study results will be disseminated through peer-reviewed publications. Codes of the resulted dementia risk prediction algorithm will be made publicly available at the website of the Tools to Inform Policy: Chinese Communities' Action in Response to Dementia project (https://www.tip-card.hku.hk/).

Strengths and limitations of this study

- The study will employ population-representative longitudinal data retrieved from the Hong Kong territory-wide public healthcare system currently serving 7 million people. Findings are highly generalisable to the Hong Kong population.
- Flexible machine-learning models will be adopted to use the size and depth of information in the dataset, which allows the generation of novel hypotheses.
- Since the predictive model is developed from real world data rather than research cohorts, it allows direct application of the derived algorithm for early identification of high-risk cases and early primary/secondary intervention.
- Electronic health records like the Clinical Data Analysis and Reporting System inevitably lack details regarding certain risk factors (eg, socioeconomic status and lifestyle information), and information on underdiagnosed and misdiagnosed cases. Estimation of the effects of putative risk factors on dementia, and the predictive accuracy of the corresponding machine-learning model, may therefore be biased.

INTRODUCTION

Dementia is a group of disabling disorders that can be devastating for persons living with it and their families. At present, it is estimated that 50 million people globally have dementia, and the prevalence is expected to triple by 2050.[1] To date, no cure has been found for any type of dementia.[2] The WHO has identified developing effective prevention strategies as a public health priority, and several predictive models have been developed over the past 10 years.[3–6] The primary purpose of predictive algorithms such as a risk score is to identify individuals with high risk of dementia and to target corresponding preventive measures. Examining predictors generated by a predictive model can also deliver important information about modifiable risk factors to the general public. As shown in a very recent UK study, effective intentions for potentially modifiable risk factors of dementia would save £1863 billion annually in England, reduce dementia prevalence by 8.5% and produce gains in quality-adjusted life year.[7] In societies where the proportions of undiagnosed dementia are particularly high, risk-predictive algorithms

BMJ Open: first published as 10.1136/bmjopen-2020-043487 on 19 November 2020. Downloaded from http://bmjopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.

may even serve as a valuable tool to support early diagnosis of dementia.

Established risk factors and predictive models for dementia

Substantial progress has been made in investigating the aetiology of dementia. Other than dominant risk factors that cannot be altered (such as age, family history and heredity),[8 9] modifiable factors, such as less education, hypertension, hearing impairment, smoking, obesity, depression, physical inactivity, diabetes and low social contact have also been identified.[10 11] The very recent 2020 report of the Lancet Commission on dementia prevention, intervention and care added three more risk factors for dementia with newer, convincing evidence, including excessive alcohol consumption, traumatic brain injury and air pollution.[12] In addition, many medications are shown to have either adverse (eg, anticholinergics) or protective effects (eg, statins, antihypertensive agents and non-steroidal anti-inflammatory drugs) on cognition.[13–15] The Lancet 2020 report also recommended distinguishing medical conditions in midlife and late life as risk factors.[12] It is worth noting that more population-based studies with longer observational periods are still needed to establish causal links.

Risk scores, a widely used tool for predicting disease risk, have been developed for many adverse health outcomes.[16–18] The most highly cited dementia risk score was proposed by a Nordic team.[6] Their score included only seven factors: age, education, sex, systolic blood pressure, body mass index, total cholesterol and physical activity. The authors recognised the model's limitations and suggested that including more factors can improve prediction accuracy.

Choice of predictive models differs between research questions focusing on *prediction* and *effect*. Standard predictive models, represented by parametric models such as logistic regression and the Cox model, are typically interested in quantifying the effect of a predictor on the likelihood of developing dementia, while holding other relevant predictors constant.[6 19] This approach tends to use a simplified linear depiction of reality and emphasises clinical interpretability. When prediction becomes the more valued goal, flexible machine-learning procedures, which have the ability to discover interaction, non-linear and higher-order effects, have the advantage of generating more accurate estimators of the likelihood.[20 21] A few studies have used machine learning for building predictive and diagnostic models of dementia at different stages using clinical records including imaging data.[22–26] A very recent study used unsupervised machine learning and successfully identified high likelihood of dementia in population-based surveys even without cognitive and behavioural measures.[27]

Life course approach and the cognitive footprint of dementia

In recent years, consensus has been growing that dementia is caused by complex interactions among genetic and environmental factors across the lifespan. Important theoretical models adopting this life course perspective are represented by *cognitive reserve* and *cognitive debts*. Cognitive reserve theory suggests that 'individual differences in the cognitive processes or neural networks underlying task performance allow some people to cope better than others with brain damage'.[28] Educational attainment obtained early in life, occupational complexity during the working lifetime and leisure activities in later life are among the factors shown to increase this reserve.[29 30] In contrast, cognitive debt suggests that vulnerability to symptomatic Alzheimer's disease accumulates through engagement in certain cognitive processes that actively deplete the cognitive reserve. Suggested cognitive debt factors include depression, anxiety, sleep disorder, neuroticism, life stress and post-traumatic stress disorder.[31] Dementia might therefore be an outcome of a lifelong battle between reserve and debts.

Starting from the micronutrients and fat stores during fetal life to the management of health conditions in old age, exposure to risk factors at different stages of life may exert differential influence on the risk of dementia. Many life-course epidemiological studies have divided a person's life into several periods. Identified 'critical periods' include the prenatal period, childhood to adolescence, adulthood, midlife, the transition period (young old) and old age.[32 33] Adding a time dimension to the interaction between risk and protective factors may further complicate the picture.[34]

The cognitive footprint concept, drawing an analogy with the term 'carbon footprint' from the realm of environmental science, was suggested in 2015.[35] In line with the life course perspective, the basic idea is that a person's cognition will be affected by a range of activities and events, that is, footprints, through the life course. Education, infectious diseases, head injuries, exercise, drugs and toxicity can all have effects on cognition, including in later life. The cognitive footprint can either be negative as cognitive debts or positive as cognitive reserve. The original proposal of a cognitive footprint included consideration of the potential cognitive effects of medical and public health intervention and argued the possibility of modelling 'a cognitive footprint of interventions and policies to meet the global challenges of dementia'. To date, this theory has not been comprehensively tested, although a recent study conducted in the UK adopted the term 'cognitive footprint' for psychiatric and neurological conditions and compared the prevalence of cognitive impairment in adults with a history of mood disorder, schizophrenia, multiple sclerosis and Parkinson's disease.[36]

The cognitive footprint theory is theoretically plausible yet difficult to test, as it encompasses activities and events across the whole lifespan. In this project, we will develop and test a predictive algorithm of dementia based on the cognitive footprint theory by using a subset of the cognitive footprint—the cognitive footprint of medical history.

BMJ Open: first published as 10.1136/bmjopen-2020-043487 on 19 November 2020. Downloaded from http://bmjopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.

Electronic health records and machine-learning techniques

In recent years, digitally stored data have grown exponentially, amassing extensive information on personal medical history and laboratory test results.[37] Meanwhile, clinical big data analytics featured by machine-learning techniques are ever-evolving. However, electronic health records remain an underinvestigated source in terms of building predictive algorithms and addressing public health and clinical problems.

The public healthcare system in Hong Kong adopts electronic health records. The Clinical Data Analysis and Reporting System (CDARS) captures microlevel clinical data including medical history of relevant dementia risk factors. Our preliminary analysis of CDARS *inpatient* data between 2001 and 2010 identified a total of 30 419 patients with dementia diagnoses. Eighty per cent of these had one or more records before their first diagnosis of dementia, and more than 12% had more than 10 previous records available. In terms of comorbidities before or at the point of diagnosis of dementia, 40% patients had at least one diagnosis of unspecified essential hypertension, one-quarter had urethra and urinary tract disorders, 23% had cerebrovascular disease, and approximately one-fifth had pneumonia, diabetes and a history of falling. These initial results suggest that, although we cannot exhaust all possible factors to model the life-long cognitive footprint, a substantial number of factors can be measured or approximated.

Machine learning is a very broadly defined method that automates analytical model building. It covers any type of data-driven approach whose objective is learning from data, identifying patterns and making decisions with minimal human intervention. Newer methods from machine-learning literature, such as random forest and neural networks, have been introduced in medical studies for building predictive models.[20 38 39] The conventional modelling approach has relied heavily on parametric methods with predetermined predictors. This contrasts with machine-learning models which have the ability to learn and generate new evidence by examining the complex structure of a large database of existing clinical information. Considering the vast amount of clinical information in CDARS, machine learning is a valuable tool for deriving insights that can guide clinical decisions.

Combining the strength of the CDARS and modern machine-learning techniques, this study aims to develop and validate a dementia-predictive algorithm using machine learning. We hypothesise that the predictive and diagnostic accuracy of dementia can be significantly improved by applying super learning to a wider range of clinical records. Specifically, we aim to (1) identify important characteristics of patients (predictors) before their first diagnosis of dementia; (2) evaluate existing risk scores, developed from research cohorts, in terms of their predictive power of future dementia in a clinical population in Hong Kong; (3) test the theory of cognitive footprint by including relevant predictors from previous medical records and their interactions with the time dimension in the predictive model; and (4) develop a more flexible predictive model using machine-learning techniques to further improve the predictive accuracy of risk scores developed from conventional parametric models.

METHODS AND ANALYSIS

The study involves a descriptive analysis of the research cohort, a validation and benchmarking analysis of a standard predictive model using established risk factors, and an exploratory and validation analysis for developing the predictive algorithm using machine learning.

Data source and sample

The CDARS, a territory-wide database in Hong Kong, contains population-based electronic health records from the Hong Kong Hospital Authority. It is a decision supporting system for facilitating the retrieval of clinical data stored in multiple operation systems, including the Clinical Management System, for management decisions, clinical audit, planning and research. The CDARS hosts comprehensive data on basic demographic, treatment, diagnoses, prescriptions, laboratory test results and admission/discharge information that are entered by well-trained hospital staff. Data from the CDARS have been used in several earlier epidemiological studies on either the relationship between exposure and health outcomes or disease/medication trends and have proven to be reliable.[40–43] This case–control study will be nested within the CDARS data from 2001 to 2018.

To protect patient privacy, patients' records are pseudo-anonymised. Diagnoses are stored in CDARS through International Classification of Disease (ICD) codes. Many local studies validated the coding accuracy in CDARS and reported positive predictive values for different diseases ranging from 85.4% to 100%.[41 44–46] A unique pseudo-identification number is generated for each patient to enable data linkage and retrieval for further analysis.

To date, CDARS holds more than 11 million patient records with clinical details from 1993 onwards.[47] Our preliminary investigation of the data revealed that CDARS hosts 70 083 patient records with dementia diagnoses from 2001 to 2015, which is equivalent to an average of 4672 dementia diagnoses per year. Ninety-six per cent of these patients received their diagnosis after the age of 65 years. The headcount of dementia diagnosis by gender and age group is shown in table 1.

Case identification

A cohort of individuals who were at least 47 years of age at 1 January 2001, so that all included individuals will be at least 65 years old at the end of 2018, will be identified from CDARS. The inclusion criteria for the dementia group are: (1) the individual received the diagnosis of dementia when they were 65 years or older; (2) the diagnosis was made within the study period (1 January 2001 to 31 December 2018). The date of first dementia diagnosis

BMJ Open: first published as 10.1136/bmjopen-2020-043487 on 19 November 2020. Downloaded from http://bmjopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.

Table 1 Number of dementia diagnoses* per year, stratified by gender and age group

	Female				Male			
Year	<50	50–64	65+	Subtotal	<50	50–64	65+	Subtotal
2001	12	71	4157	4240	23	117	2284	2424
2002	6	44	3160	3210	10	103	1846	1959
2003	5	32	1208	1245	8	66	791	865
2004	19	48	2627	2694	26	81	1452	1559
2005	15	71	2655	2741	15	84	1444	1543
2006	9	47	2506	2562	13	79	1340	1432
2007	15	59	2414	2488	20	94	1432	1546
2008	8	56	2613	2677	10	79	1452	1541
2009	6	71	3146	3223	9	96	1889	1994
2010	13	92	3389	3494	16	116	2113	2245
2011	5	105	3305	3415	11	123	1888	2022
2012	7	88	3170	3265	12	122	1896	2030
2013	10	68	3044	3122	13	103	1769	1885
2014	11	75	2477	2563	14	113	1594	1721
2015	1	90	2579	2670	8	92	1608	1708
Total	142	1017	42 450	43 609	208	1468	24 798	26 474

*Individuals who have any diagnosis records of ICD-9-CM-290, 294.1, 294.2, 331.0, 331.1, 331.82.
ICD, International Classification of Disease.

will be defined as the index date. A random sample of control individuals who did not receive any diagnosis of dementia at any period (including the period before 1 January 2001) will be matched with study cases by age, gender and index date with 1:1 ratio.[48]

Based on the average number of diagnoses obtained from 2001 to 2015, we expect the number of cases with a dementia diagnosis will be about 84 099. Assuming 80% statistical power at the 5% level of significance, our cohort will be able to detect an OR of 1.20 and 1.48, respectively, for conditions with 0.5% and 0.1% background rate.

Patient and public involvement

The abstract of the protocol is written in laymen's term and a layman's summary of project completion report will be published at the official website of the Research Grant Council (Hong Kong). The resulted dementia risk prediction algorithm and significant factors identified in the model will be made publicly available at the website of the Tools to Inform Policy: Chinese Communities' Action in Response to Dementia project (https://www.tip-card.hku.hk/) to raise public awareness of risk factors of dementia.

Measures

Dependent variable

The dependent variable in this study is whether an individual has received a diagnosis of dementia of any kind, including Alzheimer's disease, vascular dementia, Lewy body dementia or other kinds of dementia. Individuals who have any diagnosis records of ICD-9-CM-290, 294.1, 294.2, 331.0, 331.1, 331.82 in CDARS will be coded as 1; the matched controls will be coded as 0.

Our primary aim is to predict dementia of any kind. As a secondary objective, various types of dementia, represented by Alzheimer's disease and vascular dementia, will also be investigated. Mild cognitive impairment (MCI) (ICD-331.83) is also considered to account for preclinical dementia. However, preliminary analysis of the 2001–2010 inpatient data identified a zero record of MCI. Hence, the prevalence rate of MCI will likely be too low to generate any significant findings.

Risk factors: age period at exposure

All relevant medical conditions from 1993 onwards will be identified in CDARS. Age at exposure, approximated by the date of record, will be classified into three groups: 21–45 for early adulthood, 46–64 for midlife, and 65 and above for old age. Except for education, childhood factors will not be considered in the current proposal since the study cohort needs to be at least 47 years old on 1 January 2001 and information about their childhood and adolescence is unlikely to have been accurately documented. All other factors will be broken down into more detailed categories based on exposure period. For example, diabetes will be recoded into three variables: diabetes diagnosed at early life (yes=1; no=0), diabetes diagnosed at midlife (yes=1; no=0) and diabetes diagnosed at late life (yes=1; no=0). The theoretical model—a cognitive footprint of medical history—is shown in figure 1.

BMJ Open: first published as 10.1136/bmjopen-2020-043487 on 19 November 2020. Downloaded from http://bmjopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.

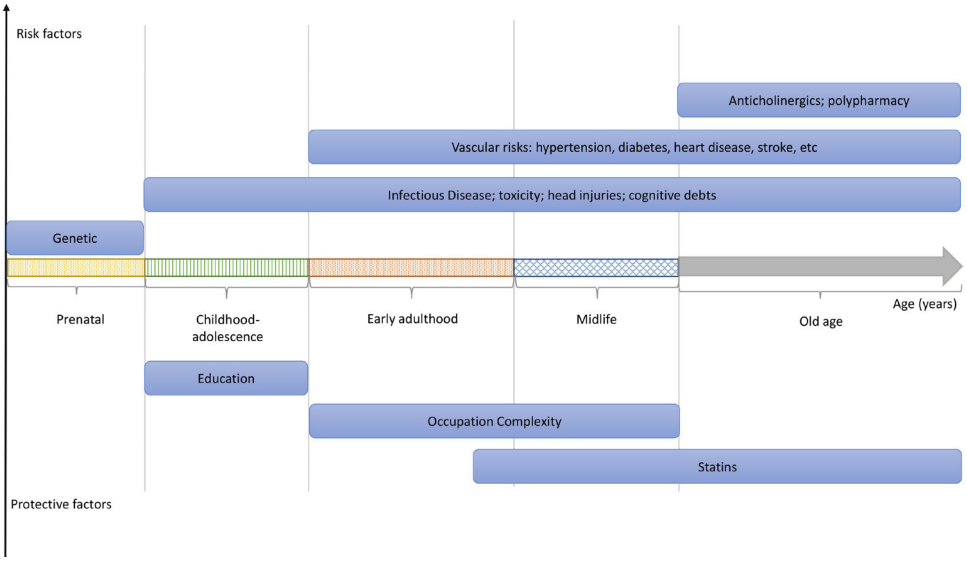

Figure 1 The theoretical model—a cognitive footprint of personal and medical history.

Established risk factors

The risk factors in this study will be divided into two general groups: established risk factors and exploratory risk factors. The established factors include diabetes mellitus (ICD-9-CM 250),[49] midlife hypertension (401),[49] midlife obesity (278), depression (296.2, 296.3, 300.4 and 311),[43] head injuries (800–804, 850–854 and 959.01)[50] and low education. In CDARS, educational level is recorded in five categories: less than primary, primary, secondary, tertiary education or above, and unknown. In this study, low education will be operationally defined as people who have less than primary or primary education. Since collecting information on educational level is not mandatory, a considerable percentage of missing values will be expected. We will perform sensitivity analyses using (1) a narrower definition of low education as people who have less than primary school education only and (2) the subsample of subjects with educational level information available to examine the robustness of the results. All these factors are measurable variables based on an influential review paper.[2]

Exploratory risk factors

Exploratory factors are selected based on the theory of cognitive footprint, which suggests that vascular disease, infectious disease, toxicity, nutrition and medication may all contribute significantly to the risk of dementia.[35] Infectious disease with ICD-9-CM codes from 001 to 139 will be merged into 16 wider categories-intestinal infectious diseases, tuberculosis, HIV and so on, according to the WHO classification. Toxicity includes poisoning by drugs, medicines and biological substances (ICD 960–979), as well as toxic effects of substances of a mainly non-medicinal source (ICD 980–989). Nutrition risk is measured by nutritional deficiencies (ICD 260–269). We also include hearing loss (ICD 389) based on more recent evidence.[10 51] Medication prescription will be identified in CDARS by British National Formulary (BNF) chapters.[52] Medication history of interest here is the prescription of antidepressants (BNF chapter 4.3), antipsychotics (4.2), lipid-regulating drugs including statins (2.12), and anti-hypertensive agents (2.5), diabetes medications (6.1) and polypharmacy. Polypharmacy is operationally defined as a medication count of five or more drugs.

All variables listed above are available in CDARS and can be retrieved electronically.[42 43]

Analytical plan

Data preparation and descriptive analysis

All data will be retrieved from the CDARS. Relevant variables for individuals with dementia diagnosis and their matched controls as listed in the Measures section will be retrieved for the identified cases. Comprehensive recoding processes will be carried out for all the risk factors. As missing values are presumably prevalent in the electronic health records, multiple imputations will be carried out using the MICE package in the open source software R.[53] Sensitivity analysis will be conducted in the later phases to compare the results with and without imputation.

The sample will be divided into two subsamples: a training set and a testing set. In the training set, 70% of individuals will be randomly selected from the dementia group and 70% from the control group. The remaining subjects will be assigned to the testing set. The validation set approach is chosen instead of cross-validation due to the large sample size and complex structure of the data.

We will descriptively present the clinical profiles of patients with and without dementia.

Differences in terms of risk established and exploratory risk factors will be compared using Student's t-test and χ^2 test. Characteristics of patients with different types of dementia will be compared using analysis of variances.

BMJ Open: first published as 10.1136/bmjopen-2020-043487 on 19 November 2020. Downloaded from http://bmjopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.

Benchmarking using established risk factors and multiple logistic regression model

Using the same simple technique adopted by several previous studies, a standard conditional logistic regression model will be fitted to the training sample using established risk factors only. We will use parameter estimates estimated from the training sample to compute estimated probabilities of developing dementia for individuals in the test set. The area under the receiver operating characteristic curve (AUC) and the c-statistics for the test sample will be calculated to evaluate the sensitivity and specificity.[54]

Developing the predictive algorithm using machine learning

This phase includes two steps. First, we will keep using the logistic regression model while adding exploratory risk factors based on the cognitive footprint of medical history. This step aims to examine the effect of exploratory predictors. Machine-learning techniques will be introduced in the next step.

Super learner

The concept of machine learning covers a broad range of algorithms. Given that there is rarely a single algorithm that universally outperforms others, it is often difficult to decide a specific machine-learning algorithm without adequate priori information about the data. In this project, a priori-specified ensembling machine-learning approach, super learning, will be implemented. Super learning combines multiple algorithms to a single algorithm and returns the best predictive model based on cross-validated test mean square error (MSE). It has optimality properties and was shown to be a powerful method in predicting mortality risk.[20] Technical details regarding super learning are published elsewhere.[55 56]

Specifically, more than 10 algorithms will be implemented in this super learning procedure, including generalised boosted regression, penalised regression, multivariate adaptive regression splines, random forest, support vector machine and neural network. The best algorithm will be selected based on the estimated MSE based on the 10-fold cross-validation. Estimation results obtained from the best algorithm will be applied to the testing set to predict group membership. Estimation outcomes, such as AUC values, sensitivity, specificity and c-statistics, obtained from conventional logistic models and machine-learning models will be compared and discussed. The SuperLearner package in R will be used to perform the machine-learning analysis.[57] The open source statistical software R will be used for the data analysis.[58]

LIMITATIONS

The proposed study has some limitations. First, a health registry database like the CDARS inevitably lacks details regarding relevant risk factors (eg, prenatal, childhood, adolescent and other early-life risk factors, socioeconomic status and lifestyle information). Findings regarding the relative importance of predictors included may be biased due to insufficient control of other putative factors, and the predictive accuracy for dementia may be compromised. Second, pieces of information on underdiagnosed and misdiagnosed cases are not available. Given the general undertreatment and underdiagnosis of dementia in Hong Kong and the possibility that mild cases of other conditions are managed in community outpatient clinics rather than public hospitals, the effects of risk factors on dementia may be overestimated as only severe cases were captured in electronic health records. Third, inference regarding the risk score or likelihood of dementia can only be made to clinical populations instead of the general population in Hong Kong. The effects of risk factors may therefore be underestimated since controls selected from this clinical population of people with complex medical needs are likely to carry a higher risk of dementia than the general population.[59] Fourth, as we are unable to have access to scores of the cognitive assessments, and as it appears that most clinicians may not be coding MCI cases, we will likely be picking up dementia cases which are already of moderate severity, leading to biased estimate of the effects and corresponding predictive accuracies generated from candidate machine-learning models. Despite these limitations, we believe it is important to evaluate the replicability of findings generated from research cohorts using real-life electronic health records. The purpose is to examine to what extent real-world diagnoses can predict dementia, irrespective of speculations about factors influencing these diagnoses.[60] Clinical algorithms and tools derived from real-life scenarios can be more easily translated and applied to assist clinical decision-making.

DATA STATEMENT

Patients' records that will be used in this study are required by law to be safety stored for privacy reasons. All data collected for this study will be anonymised. A designated server will be used to store the data and the server will be secured in a locked rack cabinet. This server will be backed up by another server with a similar level of security and the data stored inside will be encrypted. Only principal investigator (HL) and her delegates will have access to the servers. Technical appendix, statistical code and a synthetic dataset will be made available at the Hong Kong University website.

Author affiliations
[1]Department of Social Work and Social Administration, University of Hong Kong, Hong Kong, China
[2]Department of Computer Science, University of Hong Kong, Hong Kong, China
[3]Department of Medicine, University of Hong Kong, Hong Kong, China
[4]Department of Psychiatry, University of Hong Kong, Hong Kong, China
[5]Department of Diagnostic Radiology, University of Hong Kong, Hong Kong, China
[6]School of Data Science, City University of Hong Kong, Hong Kong, China
[7]Care Policy and Evaluation Centre (CPEC), The London School of Economics and Political Science, London, UK

BMJ Open: first published as 10.1136/bmjopen-2020-043487 on 19 November 2020. Downloaded from http://bmjopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.

[8]Centre for Safe Medication Practice and Research, Department of Pharmacology and Pharmacy, University of Hong Kong, Hong Kong, China
[9]Research Department of Practice and Policy, University College London School of Pharmacy, London, UK

Twitter Hao Luo @HaoLUO_hku and Gloria H Y Wong @GloW_hku

Acknowledgements The authors thank Kenneth KC Man and Celine SL Chui for their valuable comments on drafts for this protocol.

Contributors HL, GW and MK formulated the research questions. HL, KKL, W-CC, ICKW and GW designed the study. QZ and HKFM provided critiques of the study design. Analysis will be conducted by HL and QZ. HL drafted the protocol. All authors provided critiques of and reviewed the protocol.

Funding The work was supported by the Research Grant Council of Hong Kong under the Early Career Scheme 27110519.

Competing interests KKL has received grant support from Health and Medical Research Fund, Hong Kong Government Food & Health Bureau, Amgen, Boehringer Ingelheim, Eisai, Pfizer and Sanofi; as well as honorarium from Boehringer Ingelheim and Sanofi. All of which are not related to the current paper.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iDs
Hao Luo http://orcid.org/0000-0003-4261-3414
Gloria H Y Wong http://orcid.org/0000-0002-1331-942X

REFERENCES

1 World Health Organization. World health organization fact sheet - Dementia, 2019. Available: https://www.who.int/news-room/fact-sheets/detail/dementia
2 Barnes DE, Yaffe K. The projected effect of risk factor reduction on Alzheimer's disease prevalence. *Lancet Neurol* 2011;10:819–28.
3 WHO. *Dementia: a public health priority*. World Health Organization, 2012.
4 Exalto LG, Biessels GJ, Karter AJ, *et al*. Risk score for prediction of 10 year dementia risk in individuals with type 2 diabetes: a cohort study. *Lancet Diabet Endocrinol* 2013;1:183–90.
5 Exalto LG, Quesenberry CP, Barnes D, *et al*. Midlife risk score for the prediction of dementia four decades later. *Alzheimer's & Dementia* 2014;10:562–70.
6 Kivipelto M, Ngandu T, Laatikainen T, *et al*. Risk score for the prediction of dementia risk in 20 years among middle aged people: a longitudinal, population-based study. *Lancet Neurol* 2006;5:735–41.
7 Mukadam N, Anderson R, Knapp M, *et al*. Effective interventions for potentially modifiable late-onset dementia risk factors: their costs and cost-effectiveness. *Lancet Health Longevity* 2020.
8 Ferri CP, Prince M, Brayne C, *et al*. Global prevalence of dementia: a Delphi consensus study. *Lancet* 2006;366:2112–7.
9 Bird TD. Genetic factors in Alzheimer's disease. *New England J Med* 2005;352:862–4.
10 Livingston G, Sommerlad A, Orgeta V, *et al*. Dementia prevention, intervention, and care. *Lancet* 2017;390:2673–34.
11 Plassman BL, Havlik R, Steffens D, *et al*. Documented head injury in early adulthood and risk of Alzheimer's disease and other dementias. *Neurology* 2000;55:1158–66.
12 Livingston G, Huntley J, Sommerlad A, *et al*. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet* 2020;396:413–46.
13 Gray SL, Anderson ML, Dublin S, *et al*. Cumulative use of strong anticholinergics and incident dementia: a prospective cohort study. *JAMA internal medicine* 2015;175:401–7.
14 Lincoln P, Fenton K, Alessi C, *et al*. The Blackfriars consensus on brain health and dementia. *Lancet* 2014;383:1805–6.
15 Stoner CR, Knapp M, Luyten J, *et al*. *The cognitive footprint of medication: a review of cognitive assessments in clinical trials*, 2020.
16 Wilson PW, D'Agostino RB, Levy D, *et al*. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97:1837–47.
17 Pocock SJ, McCormack V, Gueyffier F, *et al*. A score for predicting risk of death from cardiovascular disease in adults with raised blood pressure, based on individual patient data from randomised controlled trials. *BMJ* 2001;323:75–81.
18 Lindström J, Tuomilehto J. The diabetes risk score. *Diabetes Care* 2003;26:725–31.
19 Kivipelto M, Helkala E-L, Laakso MP, *et al*. Midlife vascular risk factors and Alzheimer's disease in later life: longitudinal, population based study. *BMJ* 2001;322:1447–51.
20 Rose S. Mortality risk score prediction in an elderly population using machine learning. *Am J Epidemiol* 2013;177:443–52.
21 Austin PC, Tu JV, Ho JE, *et al*. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *J Clin Epidemiol* 2013;66:398–407.
22 Chen R, Herskovits EH. Machine-learning techniques for building a diagnostic model for very mild dementia. *Neuroimage* 2010;52:234–44.
23 Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia. Workshops at the twenty-seventh AAAI conference on artificial intelligence 2013.
24 Shankle WR, Mani S, Pazzani MJ, *et al*. *Detecting very early stages of dementia from normal aging with machine learning methods. Conference on artificial intelligence in medicine in Europe*. Springer, 1997.
25 Zhang Y, Dong Z, Phillips P, *et al*. Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning 2015;9:66.
26 Pellegrini E, Ballerini L, Hernandez MDCV, *et al*. Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review. *Alzheimers Dement* 2018;10:519–35.
27 Cleret de Langavant L, Bayen E, Yaffe K. Unsupervised machine learning to identify high likelihood of dementia in population-based surveys: development and validation study. *J Med Internet Res* 2018;20:e10493.
28 Stern Y. Cognitive reserve. *Neuropsychologia* 2009;47:2015–28.
29 Stern Y. Cognitive reserve and Alzheimer disease. *Alzheimer Dis Assoc Disord* 2006;20:112–7.
30 Smart EL, Gow AJ, Deary IJ. Occupational complexity and lifetime cognitive abilities. *Neurology* 2014;83:2285–91.
31 Marchant NL, Howard RJ. Cognitive debt and Alzheimer's disease. *J Alzheimer's Dis* 2015;44:755–70.
32 Muller M, Sigurdsson S, Kjartansson O, *et al*. Birth size and brain function 75 years later. *Pediatrics* 2014;134:761–70.
33 Fratiglioni L, Mangialasche F, Qiu C. Brain aging: lessons from community studies. *Nutr Rev* 2010;68:S119–27.
34 Whalley LJ, Dick FD, McNeill G. A life-course approach to the aetiology of late-onset dementias. *Lancet Neurol* 2006;5:87–96.
35 Rossor M, Knapp M. Can we model a cognitive footprint of interventions and policies to help to meet the global challenge of dementia? *Lancet* 2015;386:1008–10.
36 Cullen B, Smith DJ, Deary IJ, *et al*. The 'cognitive footprint'of psychiatric and neurological conditions: cross-sectional study in the UK Biobank cohort. *Acta Psychiatr Scand* 2017;135:593–605.
37 Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA* 2013;309:1351–2.
38 BioMed Central. *Modeling activation of inflammatory response system: a molecular-genetic neural network analysis. BMC proceedings*, 2007.
39 BioMed Central. *Classification of rheumatoid arthritis status with candidate gene and genome-wide single-nucleotide polymorphisms using random forests. BMC proceedings*, 2007.
40 Lao KSJ, Tam AWY, Wong ICK, *et al*. Prescribing trends and indications of antipsychotic medication in Hong Kong from 2004 to 2014: general and vulnerable patient groups. *Pharmacoepidemiol Drug Saf* 2017;26:1387–94.
41 Lau WCY, Chan EW, Cheung C-L, *et al*. Association between dabigatran vs warfarin and risk of osteoporotic fractures among patients with nonvalvular atrial fibrillation. *JAMA* 2017;317:1151–8.
42 Man KKC, Chan EW, Ip P, *et al*. Prenatal antidepressant use and risk of attention-deficit/hyperactivity disorder in offspring: population based cohort study. *BMJ* 2017;357:j2350.
43 Chai Y, Luo H, Wong GH, *et al*. Risk of self-harm after the diagnosis of psychiatric disorders in Hong Kong, 2000–10: a nested case-control study. *The Lancet Psychiatry* 2020;7:135–47.

BMJ Open: first published as 10.1136/bmjopen-2020-043487 on 19 November 2020. Downloaded from http://bmjopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.

44 Wong OF, PL H, Lam SK. Retrospective review of clinical presentations, microbiology, and outcomes of patients with psoas abscess. *Hong Kong Med J* 2013;19:416–23.
45 Wong AYS, Root A, Douglas IJ, *et al*. Cardiovascular outcomes associated with use of clarithromycin: population based study. *BMJ* 2016;352:h6926.
46 Chan EW, Lau WC, Leung WK, *et al*. Prevention of dabigatran-related gastrointestinal bleeding with gastroprotective agents: a population-based study. *Gastroenterology*2015;149:586–95.
47 Hospital Authority. Data collaboration lab 2020. Available: https://www3.ha.org.hk/Data/DCL/ProjectDataCatalogue
48 Grimes DA, Schulz KF. Compared to what? Finding controls for case-control studies. *Lancet* 2005;365:1429–33.
49 Yip TC-F, Wong GL-H, Chan HL-Y, *et al*. Elevated testosterone increases risk of hepatocellular carcinoma in men with chronic hepatitis B and diabetes mellitus. *J Gastroenterol Hepatol* 2020
50 Kleiven S, Peloso PM, von Holst H. The epidemiology of head injuries in Sweden from 1987 to 2000. *Inj Control Saf Promot* 2003;10:173–80.
51 Deal JA, Betz J, Yaffe K, *et al*. Hearing impairment and incident dementia and cognitive decline in older adults: the health ABC study. *J Gerontol A Biol Sci Med Sci* 2017;72:703–9.
52 Mehta D. *British national formulary*. Pharmaceutical Press, 2005.
53 Buuren Svan, Groothuis-Oudshoorn K. mice : Multivariate Imputation by Chained Equations in R. *J Stat Softw* 2011;45:1.
54 Friedman J, Hastie T, Tibshirani R. *The elements of statistical learning: Springer series in statistics New York*, 2001.
55 Van der Laan MJ, Polley EC, Hubbard AE, *et al*. Statistical applications in genetics and molecular biology 2007;6.
56 Polley EC, Van Der Laan MJ. *Super learner in prediction*, 2010.
57 Polley E, van der Laan MJ. SuperLearner: super learner prediction, 2013. Available: http://CRAN R-project org/package= SuperLearner R package version
58 R Core Team. *R: A Language and Environment for Statistical Computing [program]*. Vienna, Austria: R Foundation for Statistical Computing, 2017.
59 Chai Y, Luo H, Yip PSF. Prevalence and risk factors for repetition of non-fatal self-harm in Hong Kong, 2002–2016: a population-based cohort study. *Lancet Regional Health-Western Pacific* 2020;2:100027.
60 Luo H, Wong GHY, Chai Y. Risk of self-harm after diagnosis of psychiatric disorders - Authors' reply. *Lancet Psychiatry* 2020;7:305.

BMJ Open: first published as 10.1136/bmjopen-2020-043487 on 19 November 2020. Downloaded from http://bmjopen.bmj.com/ on April 19, 2024 by guest. Protected by copyright.