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/).

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 are an important tool to assist with healthcare planning and allocation.
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), modifiable factors, such as less education, hypertension, hearing impairment, smoking, obesity, depression, physical inactivity, diabetes and low social contact have also been identified. 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. 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.

The Lancet 2020 report also recommended distinguishing medical conditions in midlife and late life as risk factors. 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. The most highly cited dementia risk score was proposed by a Nordic team. 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. 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.

A few studies have used machine learning for building predictive and diagnostic models of dementia at different stages using clinical records including imaging data. 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.

**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. 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. 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. 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. Adding a time dimension to the interaction between risk and protective factors may further complicate the picture.

The cognitive footprint concept, drawing an analogy with the term ‘carbon footprint’ from the realm of environmental science, was suggested in 2015. 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.

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.
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.35 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 percent 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, 25% 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 pseudonymised. 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%.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 4,672 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...
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. 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 recorded 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.

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

<table>
<thead>
<tr>
<th>Year</th>
<th>Female</th>
<th></th>
<th></th>
<th>Male</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;50</td>
<td>50–64</td>
<td>65+</td>
<td>Subtotal</td>
<td>&lt;50</td>
<td>50–64</td>
</tr>
<tr>
<td>2001</td>
<td>12</td>
<td>71</td>
<td>4157</td>
<td>4240</td>
<td>23</td>
<td>117</td>
</tr>
<tr>
<td>2002</td>
<td>6</td>
<td>44</td>
<td>3160</td>
<td>3210</td>
<td>10</td>
<td>103</td>
</tr>
<tr>
<td>2003</td>
<td>5</td>
<td>32</td>
<td>1208</td>
<td>1245</td>
<td>8</td>
<td>66</td>
</tr>
<tr>
<td>2004</td>
<td>19</td>
<td>48</td>
<td>2627</td>
<td>2694</td>
<td>26</td>
<td>81</td>
</tr>
<tr>
<td>2005</td>
<td>15</td>
<td>71</td>
<td>2655</td>
<td>2741</td>
<td>15</td>
<td>84</td>
</tr>
<tr>
<td>2006</td>
<td>9</td>
<td>47</td>
<td>2506</td>
<td>2562</td>
<td>13</td>
<td>79</td>
</tr>
<tr>
<td>2007</td>
<td>15</td>
<td>59</td>
<td>2414</td>
<td>2488</td>
<td>20</td>
<td>94</td>
</tr>
<tr>
<td>2008</td>
<td>8</td>
<td>56</td>
<td>2613</td>
<td>2677</td>
<td>10</td>
<td>79</td>
</tr>
<tr>
<td>2009</td>
<td>6</td>
<td>71</td>
<td>3146</td>
<td>3223</td>
<td>9</td>
<td>96</td>
</tr>
<tr>
<td>2010</td>
<td>13</td>
<td>92</td>
<td>3389</td>
<td>3494</td>
<td>16</td>
<td>116</td>
</tr>
<tr>
<td>2011</td>
<td>5</td>
<td>105</td>
<td>3305</td>
<td>3415</td>
<td>11</td>
<td>123</td>
</tr>
<tr>
<td>2012</td>
<td>7</td>
<td>88</td>
<td>3170</td>
<td>3265</td>
<td>12</td>
<td>122</td>
</tr>
<tr>
<td>2013</td>
<td>10</td>
<td>68</td>
<td>3044</td>
<td>3122</td>
<td>13</td>
<td>103</td>
</tr>
<tr>
<td>2014</td>
<td>11</td>
<td>75</td>
<td>2477</td>
<td>2563</td>
<td>14</td>
<td>113</td>
</tr>
<tr>
<td>2015</td>
<td>1</td>
<td>90</td>
<td>2579</td>
<td>2670</td>
<td>8</td>
<td>92</td>
</tr>
<tr>
<td>Total</td>
<td>142</td>
<td>1017</td>
<td>42450</td>
<td>43609</td>
<td>208</td>
<td>1468</td>
</tr>
</tbody>
</table>

*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.
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), midlife hypertension (401), midlife obesity (278), depression (296.2, 296.3, 300.4 and 311), head injuries (800–804, 850–854 and 959.01) 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 χ² test. Characteristics of patients with different types of dementia will be compared using analysis of variances.
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.55 56 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.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

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


