Artificial intelligence for dementia prevention

Danielle Newby1 | Vasiliki Orgeta2 | Charles R. Marshall3,4 | Ilianna Lourida5,6 | Christopher P. Albertyn7 | Stefano Tamburin8 | Vanessa Raymont1 | Michele Veldsman9,10 | Ivan Koychev1 | Sarah Bauermeister1 | David Weisman11 | Isabelle F. Foote3,12 | Magda Bucholc13 | Anja K. Leist14 | Eugene Y. H. Tang5 | Xin You Tai15,16 | The Deep Dementia Phenotyping (DEMON) Network | David J. Llewellyn6,17 | Janice M. Ranson6

1Department of Psychiatry, Warneford Hospital, University of Oxford, Oxford, UK
2Division of Psychiatry, University College London, London, UK
3Preventive Neurology Unit, Wolfson Institute of Population Health, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK
4Department of Neurology, Royal London Hospital, London, UK
5Population Health Sciences Institute, Newcastle University, Newcastle, UK
6University of Exeter Medical School, Exeter, UK
7Department of Old Age Psychiatry, Institute of Psychiatry, Psychology and Neuroscience, King’s College London, London, UK
8Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona, Verona, Italy
9Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, UK
10Department of Experimental Psychology, University of Oxford, Oxford, UK
11Abington Neurological Associates, Abington, Pennsylvania, USA
12Institute for Behavioral Genetics, University of Colorado Boulder, Boulder, Colorado, USA
13Cognitive Analytics Research Lab, School of Computing, Engineering & Intelligent Systems, Ulster University, Derry, UK
14Department of Social Sciences, Institute for Research on Socio-Economic Inequality (IRSEI), University of Luxembourg, Esch-sur-Alzette, Luxembourg
15Nuffield Department of Clinical Neuroscience, University of Oxford, Oxford, UK
16Division of Clinical Neurology, John Radcliffe Hospital, Oxford University Hospitals Trust, Oxford, UK
17The Alan Turing Institute, London, UK

Correspondence
Danielle Newby, Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences (NDORMS), Centre for Statistics in Medicine, University of Oxford, Windmill Road, Oxford OX3 7LD, UK. Email: danielle.newby@ndorms.ox.ac.uk

Abstract
INTRODUCTION: A wide range of modifiable risk factors for dementia have been identified. Considerable debate remains about these risk factors, possible interactions between them or with genetic risk, and causality, and how they can help in clinical trial recruitment and drug development. Artificial intelligence (AI) and machine learning (ML) may refine understanding.

METHODS: ML approaches are being developed in dementia prevention. We discuss exemplar uses and evaluate the current applications and limitations in the dementia prevention field.
### RESULTS: Risk-profiling tools may help identify high-risk populations for clinical trials; however, their performance needs improvement. New risk-profiling and trial-recruitment tools underpinned by ML models may be effective in reducing costs and improving future trials. ML can inform drug-repurposing efforts and prioritization of disease-modifying therapeutics.

**DISCUSSION:** ML is not yet widely used but has considerable potential to enhance precision in dementia prevention.

**KEYWORDS**
- artificial intelligence
- dementia
- machine learning
- prevention
- risk prediction

**Highlights**
- Artificial intelligence (AI) is not widely used in the dementia prevention field.
- Risk-profiling tools are not used in clinical practice.
- Causal insights are needed to understand risk factors over the lifespan.
- AI will help personalize risk-management tools for dementia prevention.
- AI could target specific patient groups that will benefit most for clinical trials.

### 1 | INTRODUCTION TO DEMENTIA PREVENTION, ARTIFICIAL INTELLIGENCE, AND MACHINE LEARNING

The incidence of dementia has been reported to be decreasing in some high-income countries. However, the prevalence of dementia worldwide is predicted to triple over the next 30 years due to increases in life expectancy, particularly in low- and middle-income countries. It is now recognized that pathological changes begin years before the onset of clinical symptoms of Alzheimer’s disease (AD) and other dementia subtypes. This creates a window of opportunity for early identification and development of targeted interventions and treatments that could ultimately prevent subsequent dementia. Estimates show that delaying AD onset by 5 years would reduce prevalence by 40%, which would have a huge public health impact. In the absence of an effective cure, the identification of new strategies to prevent dementia cases is of critical importance. Thus dementia research has focused increasingly on the treatment of preclinical stages and the identification of modifiable risk factors.

### 1.1 | Dementia prevention research

Decades of primary prevention research have included the assessment and identification of biological, behavioral, environmental, and social factors that may increase or decrease the likelihood of developing dementia. Findings from epidemiological studies provide a growing list of modifiable risk and protective factors across the life course that can be targeted for prevention. Several interventional studies have tested the efficacy of prevention strategies in cognitively healthy individuals at risk of dementia (primary prevention) or to prevent further cognitive decline or progression to dementia in people with mild cognitive impairment (MCI) or other early symptoms (secondary prevention). Dementia prevention clinical trials targeting single or multiple risk factors have evaluated a range of pharmacological (cholinesterase inhibitors) and non-pharmacological interventions (e.g., exercise programs, cognitive training, multidomain). However, evidence that the treatment of these risk factors reduces the risk or progression of dementia remains weak. Factors hindering progress in this area may include the difficulty in identifying the age and characteristics of the ideal target population, a lack of understanding of the causal pathways to be targeted, and the limitations of traditional statistical methods in analyzing high-dimensional multimodal data (e.g., imaging, biomarker, genetics, and medical records) and non-linear relationships between variables. This is particularly important in dementia prevention research, where hundreds, or even thousands, of genetic and non-genetic features may predict disease risk and progression, and most of the proposed risk factors do not have a linear relationship with dementia risk throughout life.

### 1.2 | Artificial intelligence and machine learning

Artificial intelligence (AI) refers to the broader field of computer science that focuses on creating machines or systems capable of performing tasks that would typically require human intelligence. Machine learning (ML) is a subset of AI that focuses on designing algorithms and models that enable computers to learn from and make predictions or decisions based on data without being explicitly programmed. The application of methodologies incorporating
AI and ML is gaining momentum in dementia prevention research (Figure 1).

In this review, we focus mainly on the application of ML methods in dementia prevention research, since this area of AI holds promise for advancing the field. ML can be broadly split into unsupervised, supervised, and semi-supervised methods. Unsupervised methods do not distinguish between dependent and independent variables (i.e., there is not a specified dementia-related outcome being predicted); rather, their goal is to discover patterns and associations between the variables (i.e., risk factor and dementia) to potentially define possible groups and clusters hidden in the data. Unsupervised methods are applied under the assumption that similar outcomes (with similar variables) will be grouped together and reveal meaningful patterns not seen in the raw data.

Clustering methods such as multi-layer clustering, k-means clustering, hierarchical clustering, or Gaussian mixture models (GMMs) are unsupervised methods that have been applied in dementia prevention research, for example, to identify subgroups of individuals with MCI with markedly different prognostic cognitive trajectories. The results from such studies can be used to identify subgroups of individuals that are at high risk of dementia to better understand different disease trajectories and tailor interventions or prevention strategies accordingly. Unsupervised anomaly detection algorithms, including one-class support vector machines (SVMs), autoencoders (AEs), or isolation forests, have been utilized to identify unusual or atypical patterns in data. In the context of dementia prevention, these methods could help detect outliers or abnormal biomarker patterns that may be indicative of underlying pathology or increased risk. Other unsupervised methods include unsupervised latent variable models, such as latent Dirichlet allocation (LDA) or GMMs, which have been employed to identify latent structures or hidden variables in data, which can reveal hidden patterns or subgroup data sets that may be relevant for understanding dementia risk factors or disease progression.

With supervised methods the dependent variable is known (i.e., dementia is specified to be the outcome), and it is this information plus the input from the independent variables (i.e., the risk factors) that guide the supervised ML method to predict the dependent variable. These methods can be further split into regression or classification methods, where the dependent variable to be predicted is either numerical (continuous) or categorical. Supervised ML methods such as logistic regression, random forest, SVMs, gradient boosting, deep learning, and decision trees can be utilized for clinical risk prediction—for example, where dementia diagnostic status is known in combination with features such as demographics, imaging, biomarkers, genetics, comorbidities, symptoms, medication use, and other health indicators are used to build models useful for primary and secondary prevention. Supervised ML methods have also been developed to classify biomarker data associated with dementia and these models can also be trained on neuroimaging data such as magnetic resonance imaging (MRI) or positron emission tomography (PET) scans to classify brain images as healthy or indicative of dementia-related abnormalities. Similarly, classification models can be developed using genetic markers or biomarkers obtained from cerebrospinal fluid to distinguish between individuals with or without dementia-related pathology. Supervised methods can also be used to predict cognitive decline in individuals using longitudinal data. These models learn patterns in the data and can predict the future cognitive decline trajectory of individuals based on their baseline assessments and other risk factors.

Finally, semi-supervised methods fall in-between supervised and unsupervised ML and use both labeled (e.g., information on dementia/MCI diagnosis) and unlabeled data. Semi-supervised methods such as active learning use a labeled data set to train a model, which is then used to select the most informative or uncertain samples from the unlabeled data set. These selected samples are then labeled manually by experts, and the model is retrained using the newly labeled data. Active learning could be applied to various tasks in dementia prevention, such as identifying relevant risk factors or biomarkers, or targeted samples for clinical trials recruitment. Other methods include self-training where a model is initially trained on a small, labeled data set. The model is then used to make predictions on the unlabeled data, and the most confident predictions are added to the labeled data set. This process is repeated iteratively, gradually expanding the labeled data set and improving the model’s performance. Self-training can be applied to tasks such as dementia-risk prediction, where a limited number of labeled samples is available.
number of labeled instances are available, but a large amount of unlabeled data exists. Finally, co-training is a semi-supervised technique that utilizes multiple views or feature sets of the data to improve the model’s performance. The model is trained on different sets of features or representations, each with its own labeled and unlabeled data. The model is then iteratively trained on each view and shares information between the views during training. Co-training could be applied in dementia prevention by utilizing multiple modalities, such as genetic data, neuroimaging, and cognitive assessments, to improve the accuracy and robustness of predictive models. Semi-supervised methods use information from unlabeled data to improve the performance of models trained on labeled data. A more in-depth description of ML methodologies and applications to the broader dementia field can be found in the methods optimization paper also included in this special issue.

Traditional statistical approaches rely on using domain knowledge to create and fit mathematical models to represent the relationships between variables and provide a quantitative measure of confidence that describe associations. In contrast, ML techniques do not make a priori assumptions about the distribution from which the modeling sample is drawn, and hence they can be effective even in the presence of complex non-linear relationships and when the data are collected without a controlled experimental design. ML models are more flexible and scalable compared with conventional statistical approaches, as they can explore the structure of high-dimensional multimodal data, detect patterns, and generate insights with minimal human intervention. In this review, we provide a critical summary of the current state of the science, challenges, implications, and major opportunities for the application of ML to dementia prevention.

This review is one of a series of eight articles in a Special Issue on “Artificial Intelligence for Alzheimer’s Disease and Related Dementias” published in Alzheimer’s & Dementia. Together, this series provides a comprehensive overview of current applications of AI and ML to dementia, and future opportunities for innovation to accelerate research. Each review focuses on a different area of dementia research, including experimental models, drug discovery and trials optimization, genetics and omics, biomarkers, imaging, prevention (this article), applied models and digital health, and methods optimization.

2 | USE OF ML TO UNDERSTAND MODIFIABLE RISK FACTORS FOR DEMENTIA PREVENTION

In this section we describe the current state of the science relating to the identification of risk factors for dementia and risk-prediction modeling. We then describe the major challenges that are hindering...
progress and provide examples of ML applications to these problems. Finally, we discuss the opportunities of how ML methods could enhance and improve our understanding of risk factors by using different data types (multimodal data) and application of ML approaches from other disciplines.

2.1 | State of the science

2.1.1 | Identification of risk factors

Several lifestyle, psychosocial, and cardiometabolic risk factors play an important role in the development of dementia.\(^5^2\) The Lancet Commission on Dementia Prevention, Intervention, and Care, first published in 2017, proposed a life course perspective to better understand how, and at which time point over the course of the human lifespan, nine potentially modifiable risk factors increase dementia risk. Fewer years of education during early life were found to increase risk, whereas chronic hypertension, obesity, and age-related hearing loss were considered to increase risk during mid-life. The remaining five factors (current smoking, depression, physical inactivity, social isolation, and type 2 diabetes) were shown to exert their effect in later life.\(^5^3\) In the Lancet Commission’s updated 2020 report, three new risk factors were added, two mid-life risk factors (alcohol consumption and traumatic brain injury) and one late-life risk factor (air pollution).\(^8\)

Several additional systematic reviews\(^5^4–5^8\) have been conducted within the last 10 years that aimed to comprehensively summarize the emerging evidence relating to dementia risk factors (Figure 2). The main risk factors highlighted by these reviews include those covered by the Lancet Commission as well as the dementia prevention guidelines from The World Health Organization (WHO). Evidence for other risk factors such as hyperlipidemia, coronary heart disease, renal dysfunction, poor diet, and cognitive inactivity were less consistent across reviews. More recent systematic reviews, which focus upon individual risk factors, also suggest the importance of additional risk factors, such as stroke\(^5^9\) and delirium,\(^6^0\) and it is likely that the number of “established” risk factors will grow as this body of evidence consolidates.

However, when considering the relationship between a proposed risk or protective factor with dementia, it is vital to consider a lifespan perspective, owing to the changing importance of potentially modifiable risk factors across the life course and the fact that dementia pathology is believed to begin up decades before it is detectable by current diagnostic methods (known as the prodromal phase).\(^6^1–6^3\) For example, hypertension and obesity are both associated with increased future dementia risk during mid-life, but weight and blood pressure have been shown to decrease in later life in those with or developing dementia, indicating that in later life, changes in these risk factors are consequences of disease progression.\(^6^4,6^5\) ML methods that consider the changing importance of risk factors using longitudinal data could be used to understand different disease trajectories and disease heterogeneity.\(^6^6\)

2.1.2 | Risk-prediction modeling

The purpose of risk-prediction models is to identify individuals who are at greatest risk of developing a disease to inform appropriate risk-reduction interventions. Numerous risk-prediction models for dementia have been developed, some of which have been externally
validated. However, further work is required to assess their generalizability, cost-effectiveness, and predictive value before they can be recommended for use in clinical practice. These models tend to use a combination of demographic, cognitive, health, lifestyle, and genetic factors to estimate an individual’s risk of developing dementia using straightforward regression-based approaches. The Cardiovascular Risk Factors, Ageing and Dementia (CAIDE) model is one example, which uses age, education, and sex alongside the mid-life risk factors of body mass index, blood pressure, cholesterol, smoking history, and physical activity. The CAIDE was developed using logistic regression to predict dementia risk in 20 years with moderate levels of accuracy (area under the curve [AUC] 0.77; 95% confidence interval [CI] 0.71–0.83). A recent study using a variety of different ML algorithms (logistic regression, SVMs, random forest, and gradient-boosted trees) was used to predict 2-year dementia risk and had superior accuracy compared to common dementia risk models including the CAIDE. However, due to the long prodromal phase of dementia, the value of a prediction model that is optimized for prediction only 2 years prior to diagnosis is unlikely to be of primary preventative value, since high levels of neurodegeneration will already have occurred but it could help clinicians to identify patients who might be in the earlier stages of disease. The optimal period for dementia risk prediction will, therefore, require different risk-prediction tools targeting different stages of pathology and dementia progression. For example, for primary prevention, long-term risk profiling is useful because risk factors can potentially be modified, whereas for secondary prevention, shorter-term risk profiling is more appropriate to identify those at higher risk of disease progression.

A systematic review highlighted that many ML techniques (predominantly SVMs) used in prediction modeling can accurately predict the conversion from MCI to AD, with most studies using neuroimaging data. However, more recently, methods such as ensemble-based approaches and deep learning are being used to predict future risk decades in advance. In addition, ML can be used to assess the relative importance of multiple risk factors in determining future cognitive impairment and dementia. ML methods are also being developed that improve understanding for the non-ML expert and, therefore, have the potential to improve clinical detection and timely intervention.

### 2.2 Major current challenges

#### 2.2.1 Understanding risk-factor mechanisms and interactions that contribute to dementia risk

Improved understanding of the causal biological pathways that link risk factors to dementia is required to support the design of preventative interventions and the identification of treatment targets for preserving brain health during early disease stages. Risk factors are likely to exert their effects through a range of distinct biological pathways including those related to protein production and clearance, vascular health, inflammation, and cognitive and neural reserve. However, the complex etiology of these risk factors coupled with high levels of comorbidity make it hard to pinpoint what the most efficacious mechanistic targets might be, and increase the likelihood that the key target will change between individuals, thereby increasing the need for reliable precision medicine for dementia prevention.

Although the main risk factors for dementia are common and frequently coexist, studies typically focus on a single risk factor or a small group of phenotypically similar factors, leaving networks of interactions and synergistic effects unquantified. However, the direct relationships between large numbers of observed traits can be achieved using network analysis. For example, one study used weighted gene correlation network analysis to identify modules of metabolites that are associated with cognition in late life, some of which seemed to be influenced by risk factors such as education. This highlights the potential value of network-based approaches to identify complex biological pathways that act as mediators between modifiable risk factors and dementia, which could represent realistic therapeutic targets on which to intervene in a clinical trial setting. Integration of deep learning into network analysis, such as DeepPPI, has been shown to outperform traditional methods in predicting interactions within complex networks. A study that used unsupervised ML methods applied across multimodal health data successfully identified novel cardiometabolic markers linked to hypertension and poor metabolic health. Use of similar network-based methods for other dementia risk factors could, therefore, help to identify the mechanisms that underlie known epidemiologic associations.

Furthermore, we currently have a poor understanding of the shared biology between dementia risk factors and their impact on brain health, even though shared pathways across risk factors represent promising targets for intervention. However, a study that applied genomic structural equation modeling to AD and 12 modifiable risk factors found high levels of genetic overlap between most dementia risk factors, suggesting that there are overarching shared biological pathways that might increase AD risk. These findings demonstrate the potential merits of using latent variable modeling (LVM) to reduce data complexity and identify common mechanistic patterns between risk factors, but traditional LVM methods perform poorly with highly complex and nonlinear data. In contrast, LVM methods that are based on deep learning approaches, such as autoencoders (AEs), deep neural networks (DNNs), or generative adversarial networks (GANs), are better suited to these kinds of data and represent a promising avenue for more accurately defining the shared biology between dementia risk factors.
to shared causality, confounding, or reverse causality.\textsuperscript{84,85} Examples of this include possible confounding between education and socioeconomic status, which make it unclear which might be more important as a causal influence on dementia risk,\textsuperscript{86} and preclinical effects of AD on hearing and depression, which could indicate reverse causality.\textsuperscript{87,88}

A method that is used increasingly to assess the causality of risk factors is Mendelian randomization (MR), which uses genetic variants that are associated with an exposure as instrumental variables to assess the causal influence of the exposure on an outcome. A systematic review\textsuperscript{85} of dementia-related MR studies highlighted causal evidence for smoking quantity, vitamin D, homocysteine, systolic blood pressure, fasting glucose, insulin sensitivity, and high-density lipoprotein cholesterol with AD. Other MR studies have also shown causal relationships between cognitive ability, education, and dementia.\textsuperscript{89}

However, recent MR evidence has provided limited and inconclusive evidence of causal relationships of the modifiable risk factors presented by the Lancet Commission with dementia.\textsuperscript{90,91} This may be in part due to effects of survival bias, particularly for risk factors that are significantly associated with earlier mortality, as well as loss of power due to heterogeneity in both the dementia outcome measure and risk-factor exposures. A description of these issues relating to MR and the potential for ML to help alleviate some of these problems in dementia prevention research is discussed in detail in the genetics and omics article also included in this special issue.\textsuperscript{48}

However, these shortcomings necessitate the application and triangulation of other causal modeling methods to help us make more confident conclusions about whether a risk factor is causal for dementia, which we focus on in this review. Other methods for causal inference include pathway or mediation analysis. This type of analysis has been used to show a causal role for a range of health and lifestyle factors in mediating the link between socioeconomic status and dementia risk.\textsuperscript{92} Mediation analysis is demanding in terms of power and large sample sizes,\textsuperscript{93} but ML methods have the potential to help make this type of inference more feasible in dementia prevention research. For example, deep learning assisted methods have been developed recently to identify mediators in high-dimensional data, such as functional imaging or complex population survey data. These have been shown to be successful in identifying brain connectivity networks that mediate the relationship between cognitive traits\textsuperscript{94} and in assessing the causal link between health insurance coverage and general health in the United States, even in the presence of multiple confounders.\textsuperscript{95}

An alternative approach to inferring causality involves analyzing the timing and trajectory of risk factors: those with a causal effect are likely to show increasing levels of dementia risk with greater duration and intensity of exposure. This approach has been used to demonstrate that depression may be a dementia prodrome rather than a causal risk factor, as it is most associated with dementia risk when occurring for the first time in the period immediately prior to dementia diagnosis.\textsuperscript{96} Analysis of the timing and trajectories of risk factors requires either large longitudinal cohorts, such as in recent work showing the most important age range for a deleterious effect of hypertension\textsuperscript{97} or improved ability to analyze routinely collected electronic health records, which is likely to be facilitated by advances in text mining\textsuperscript{98} and ML.\textsuperscript{99,100} Recent examples of text mining of notes from health records using natural language processing\textsuperscript{101} allow for neural network ML architectures to be developed to extract clinical, lifestyle, and pharmacological data. These methods could be applied to analyze the trajectories of how risk factors influence dementia risk over time to support the design of preventative interventions that could be implemented across the life-course of an individual.

2.3 | Opportunities and future directions for ML to understand modifiable risk-factor research for dementia prevention

2.3.1 | Utilizing multimodal data to understand dementia

The variety of existing data modalities acquired in research and health care can help to provide a more complete picture about a complex disease, such as dementia, than any one type of data on its own. Dementia pathogenesis and clinical progression is hugely heterogeneous between individuals, so the collection of rich and deep phenotyped, high-dimensional multimodal data is crucial for successful dementia prevention. By integrating and analyzing multimodal data such as omics, histology, imaging, clinical, and digital data, this can unveil novel mechanistic insights that could be used to understand the condition and predict risk across the whole population or specific subgroups.\textsuperscript{102,103} This could then lead to a patient-specific approach leading to personalized preventative interventions and specific risk-management strategies for dementia prevention.\textsuperscript{104}

ML approaches are well suited to dealing with the high-dimensional nature of multimodal data.\textsuperscript{105,106} Multimodal patient data better reflect clinical practice, in which medical histories, fluid biomarkers, genetics, neuropsychology, and structural and functional imaging may all be collected during diagnostic assessments in secondary care. Indeed, several dementia risk scores, such as the CAIDE risk score\textsuperscript{69} described in Section 2.1.2, use this sort of information to identify modifiable risk factors that can be managed in mid-life to prevent or delay dementia in later life. Whether ML methods can improve dementia risk prediction has not yet been well studied. One study found a supervised ML model (Disease State Index) to achieve similar sensitivity and specificity to CAIDE in predicting dementia 10 years later.\textsuperscript{107} By using data already routinely collected in clinical settings, risk scores and ML algorithms could easily be clinically translatable. On the other hand, the integration of many different data modalities, for example, genetic, multiomic, and advanced neuroimaging features,\textsuperscript{108} may improve sensitivity and specificity, but this may be at the expense of clinical translatable and would require a higher level of resources, which are often not available to many populations.

Furthermore, although the use of ML approaches may enhance the analysis of high-dimensional multimodal data, this often leads to increasingly complex models becoming uninterpretable.\textsuperscript{109} A good
example of this is DNNs, which are powerful and flexible in practice, although incorporate hidden layers, which have led them to be described as a "black box" technique. These limitations should be thoughtfully considered during the development of any ML model aimed at assisting clinical decision-making. Explainable ML is a burgeoning field that aims to promote easy interpretation of ML models.\textsuperscript{110} For example, a recent study created a deep learning framework using multimodal data (magnetic resonance imaging [MRI] and clinical variables). The deep learning framework linked a fully convolutional network to a traditional multilayer perceptron to intuitively visualize and predict AD risk.\textsuperscript{111} The model was validated across three independent data sets and could be easily used by practicing neurologists, providing a clinically translatable and interpretable tool using widely collected multimodal data.

In any case, the optimal use of multimodal data for dementia prevention requires triangulation of heterogeneous data all bringing their own unique issues in terms of noise, missingness, and quality.\textsuperscript{112} Data labeling becomes challenging in multimodal data sets, with the degree of accuracy or objectivity varying between modalities. For example, some risk factors based on genetics or blood biomarkers may be deemed more objective than features that typically require manual labeling, such as intricate lesions, or that incorporate subjective decision-making, such as behavioral ratings.

Additional challenges, such as data linkage across several data sources, can pose a common barrier to ML because different data modalities vary widely in their availability, accessibility, and generalizability. There have been some recent initiatives to help facilitate cross-cohort access, such as the creation of Dementia Platforms UK Data Portal,\textsuperscript{113} however, to promote the application of ML methodologies creating a suitable research infrastructure that makes this feasible is key. Multimodal data also increase the likelihood of being able to identify an individual, as specific combinations of data may be truly unique, which raises ethical issues especially in relation to future dementia risk.\textsuperscript{114} Therefore, as ML becomes more routinely applied within dementia prevention research, it will become necessary to create relevant policy and legal regulations that seek to protect individuals while enabling research. Although the use of ML approaches may enhance the analysis of high-dimensional multimodal data, this often comes at a cost with increasing complex models becoming uninterpretable.\textsuperscript{109} A good example of this are DNNs, which are powerful and flexible in practice, although incorporate hidden layers, which have led them to be described as a "black box" technique. Explainable ML is a burgeoning field, aiming to allow interpretation of ML models and is further discussed in Section 4.\textsuperscript{110} For example, a recent study created a deep learning framework using multimodal data (MRI and clinical variables). The deep learning framework linked a fully convolutional network to a traditional multilayer perceptron to intuitively visualize and predict AD risk.\textsuperscript{111} The model was validated across three independent data sets and could be easily used by practicing neurologists, providing a clinically translatable and interpretable tool using widely collected multimodal data. Despite these challenges, multimodal data combined with ML approaches have considerable potential to provide further insights into the natural progression of dementia and its risk factors.

2.3.2 | Applying multidisciplinary approaches from other research fields

Applying multidisciplinary approaches from other fields can significantly enhance the use of ML for dementia prevention research by providing fresh perspectives, innovative methodologies, and a broader range of expertise.

For example, to improve our ability to recruit participants at risk of developing dementia, it is important to identify individuals at risk with high accuracy over different timescales (months, 2 years, and so on). ML methods have been shown to be effective in predicting short-term risk in other disciplines, such as the prediction of hypoxemia during surgery,\textsuperscript{115} and identifying optimal timing to refer patients with terminal respiratory failure for lung transplantation.\textsuperscript{116} Much longer timescales are required to utilize these ML methods for primary dementia prevention, and this could be a challenge. However, studies have shown ML to improve the prediction of risk of other long-term conditions and related outcomes such as cardiovascular disease,\textsuperscript{117,118} and suicide,\textsuperscript{119} highlighting the ML potential. ML approaches to identify risk markers or factors are already used successfully for analyses of dementia-related neuroimaging data.\textsuperscript{71,120} However, approaches that identify new risk markers/factors temporally preceding the health outcomes would be advantageous, due to the long prodromal period of dementia. Such methods have been used previously to detect early metabolite markers as risk factors for type 2 diabetes.\textsuperscript{121} Already known or new risk factors can then be evaluated with ML for causal inference,\textsuperscript{122} and there are examples of its use in prenatal and perinatal care.\textsuperscript{123} Identifying causal risk factors will allow for targeted risk-reduction interventions for primary dementia prevention, and there are examples of ML being used to target high-risk individuals for cancer screening\textsuperscript{124} and suicide prevention.\textsuperscript{125}

A shift in focus is required to embed causal applications of ML and other methods in dementia research, by combining domain knowledge to apply ML for causal inference applied in health care, development economics, and other fields.\textsuperscript{122} To understand how to prevent dementia, causal ML methodology embedding multidisciplinary approaches is needed, which, can lead to more robust estimates compared to traditional inferential statistics.\textsuperscript{122} A major issue is that many studies use predictive ML when researchers are really interested in understanding the underlying pathological and mechanistic processes.

The use of ML could also be used to directly evaluate and compare interventions, which would be beneficial for interventions for risk-factor management in dementia. The application of ML methods in this way has been used to compare therapies for depression using elastic net regularization,\textsuperscript{126} and methods such as Bayesian Additive Regression Trees (BART) and regression adjustment on multivariate spline of generalized propensity scores (RAMS), have been used evaluate different surgery types for treatment of non-small cell lung cancer.\textsuperscript{127}
3 | USE OF ML TO ENHANCE PREVENTION INTERVENTIONS AND TRIALS

In this section we describe what pharmacological or non-pharmacological interventions are available currently to prevent or delay dementia. Second, we state the main challenges in the field such as targeted clinical trial recruitment and testing suitable interventions. Finally, we use examples to highlight how ML can improve personalized medicine in terms of risk-factor management and predicting treatment response, as well as the identification of potential drug candidates for drug repurposing for dementia prevention.

3.1 | State of the science

3.1.1 | Pharmacological and non-pharmacological interventions for dementia

Pharmacological and non-pharmacological interventions will both be key to preventing dementia. However, problems with a lack of understanding of disease complexity, disease heterogeneity, and timing of interventions have resulted in limited progress. Recently, there has been some hope with the approvals of aducanumab and lecanemab by the US Food and Drug Administration (FDA) for treatment of AD in the United States, but there has been much criticism over their effectiveness and severe side effects.128–132

Most pharmacological interventions that are being tested currently in dementia clinical trials are disease-modifying therapies, which aim to slow or stop disease progression in early stages. As of January 2023, there are 141 drug candidates being tested in 187 ongoing AD trials. Of these candidates, 28% represent agents that are being repurposed from use for other diseases, aimed at various biological targets with diverse mechanisms of action. There are currently five primary prevention trials in Phase 3 and four prevention trials in Phase 2 enrolling participants with preclinical AD or patients with varying levels of MCI and AD severity.133

Regarding non-pharmacological interventions, multiple studies have reported physical activity, cognitive reserve,134 Mediterranean diet,135 antioxidants,136,137 and some vitamins (e.g., vitamin D138,139) as protective factors for the development of dementia. This evidence paved the way for randomized clinical trials to determine whether these non-pharmacological interventions may prevent cognitive decline in healthy older people,140 or those with MCI.141 The FINGER study reported that a multidomain intervention encompassing diet, exercise, cognition, and vascular risk management improved cognition regardless of participants’ characteristics in comparison to regular health advice.142 Therefore, combining non-pharmacological treatments may be more effective for primary prevention trials. The extension of the FINGER model over 25 countries worldwide will offer important information on the role of multidomain lifestyle interventions as effective preventative strategies for cognitive decline.143

3.2 | Remaining challenges of prevention interventions and trials

3.2.1 | Difficulties with representative targeted recruitment for clinical trials

The lack of representation of minority groups in clinical research remains a significant obstacle in medical research144 and is not isolated to the dementia field, which affects the usefulness of ML methods to generalize to different sub populations to be used for targeted recruitment to clinical trials. The lack of representation is due in part to the ethical challenges of recruiting people with dementia for secondary prevention clinical trials, with a tendency to paternalistically protect them from harms of participating in research.145 However, this has been addressed by embedding research in clinical care and streamlining proxy and advanced consent procedures. In addition, the continued lack of representative real-world data has meant dementia risk scores that could be used to identify those at risk of dementia for potential primary prevention trials in those with preclinical dementia are not robustly validated146 again leading to clinical trials that are unrepresentative,147 and based upon skewed biomarker data that hinder dementia prevention drug development.148 The lack of representation of minority groups often leads to bias in ML models, which are then unable to generalize to other populations to identify those who could be targeted and recruited for clinical trials. Further discussion of bias in ML can be found in Section 4 (Limitations of ML for dementia prevention).

3.2.2 | Problems with interventions targeting individual risk factors

Population-level interventions shift the distribution of health risk by addressing the underlying social, economic, and environmental conditions.149 The high prevalence of modifiable factors for dementia raises the question of whether population-based prevention strategies could reduce the prevalence of dementia. Studies have shown that population level interventions such as those focused on single risk factors, such as maintaining cardiovascular health, have been thought to have contributed to a modest reduction of dementia incidence in some countries.11 However, dementia risk factors vary among individuals due to genetic, lifestyle, and environmental factors, making it difficult to design a one-size-fits-all population intervention. Furthermore, there are complex interactions between risk factors, and many cluster together8,53: therefore, interventions that target more than one risk factor and studies that investigate additive or even synergistic effects of risk factors are likely to be more appropriate than single...
interventions for dementia prevention. The challenge of how to develop personalized risk profile management and interventions to address individual variation remains.

3.3 Opportunities and future directions for ML to enhance prevention interventions and trials

3.3.1 Personalizing preventative interventions

Prevention strategies should begin early and continue throughout life, with public health programs, individually tailored interventions, and the need to tackle inequality given that many risk factors are linked to socioeconomic inequalities. ML could be valuable in gathering information and learning from broad data sets, including from existing medical records, which could be used to model trajectories and develop algorithms for personalized risk planning and management.

Dementia prevention needs to be able to answer a range of questions, both in individuals with no cognitive impairment (preclinical cases), those with mild memory problems, those who convert from MCI to dementia, and those diagnosed with dementia. There would also be a benefit for ML approaches that can support the integration of biological, psychological, and social factors when approaching diagnosis, classification, prognosis, and tracking of disease progression and its management.

A recent example used transfer learning with ensemble learning algorithms to develop a personalized prediction model that allowed the early detection and prediction of dementia risk. It is important to note that this approach had the ability to visualize the interaction of the risk factors that drove the prediction, which will be vital for risk-factor management and therefore has promising clinical utility.

Understanding how aging impacts these ML models is vital, given that diseases that underpin dementia often start in mid-life, and, therefore, how modifiable risk factors interact with their progression may also vary with age.

Some dementia studies are already collecting large and phenotypically deep data sets involving multimodal, cognitive data, language, novel and validated biomarkers, genomics, medical comorbidities, wearable technologies, and imaging. The collection of diverse data such as gait and speech allow for the applications of methods such as SVMs, random forest, and k-nearest neighbors (KNN) algorithms to predict different types of MCI and to differentiate between cognitively normal, MCI, and AD patients.

3.3.2 Drug repurposing

Drug repurposing encompasses the identification of new uses for drugs beyond the scope of their original indication and can enhance traditional drug development efforts. Identification of repurposing candidates could accelerate the identification of new treatments to modify or slow the onset of dementia as has been the case for cancer, Parkinson's disease, HIV, and other conditions. For dementia prevention, there are numerous examples of drugs used to treat conditions such as hypertension, diabetes, and rheumatoid arthritis that are associated with a reduced dementia risk.

Drug repurposing has been largely opportunistic. However, with the large number of drugs currently on the market or under investigation, and the increasing wealth of data available, ML methods could be a powerful tool in drug repurposing. For example, one study identified drug candidates using ML methods such as logistic regression, SVMs, boosted random forest models, and neural networks applied to gene expression data sets from neuronal cells and postmortem brains of individuals with different stages of AD. From this they identified drug candidates that target Janus kinases, which play a role in immune response, as potential repurposing candidates for AD.

Other methods such as deep learning and network-based approaches will be vital to identify and test new potential drug-repurposing candidates for dementia prevention. There is potential for ML methods such as neural networks to identify combinations of drugs that target multiple risk factors simultaneously. It is notable that deep learning could also help elucidate why certain drugs reduce risk mechanistically, which will help with understanding the causes of dementia and identify potential new drug targets.

3.3.3 Predicting treatment response

Prediction of treatment response for an individual patient remains an important goal. With no specific biomarkers being used in routine clinical practice, selecting medications for dementia or other diseases would remain a largely trial-and-error process. ML algorithms that suggest appropriate treatments based on individual patient characteristics are useful, particularly when several treatment options of generally equivalent efficacy are available. Successful examples from other disease areas include the ongoing PETRUSHKA study, which builds on network meta-analytic approaches to estimate individual efficacy of antidepressants through aggregate and individual patient data. The aim of the project is to generate a clinical support system that incorporates patient-level data and patient preferences (i.e., relative importance of efficacy and adverse event risk) to generate a ranking list of treatment recommendations.

Further elaboration of this method used anonymized secondary care patient data obtained through UK-CRIS to train a recurrent neural network algorithm to predict the most effective treatment for cognitive impairment (i.e., cholinesterase inhibitor or memantine) in patients with dementia. Although in current clinical practice there is little to distinguish between available treatments, the results showed that patients who were prescribed medications according to what was predicted in the ML model had better cognitive performance after 2 years. The promise of these ML approaches and others in other disease areas such as cancer, epilepsy, and other mental health conditions make them a primary opportunity to optimize symptomatic and disease-modifying treatments for dementia.
4 | LIMITATIONS OF ML FOR DEMENTIA PREVENTION

Although this review highlights the ways in which ML has the potential to enhance our understanding of dementia prevention, there are also several limitations that are important to consider. First, many standard ML methods do not model causality and are purely associative. They are unable to account for reverse causation and are designed to maximize predictive power, and lack the interpretability that would allow for casual modeling, which is preferable for dementia prevention.\textsuperscript{179,180} ML models, like any statistical models, rely on the data and assumptions made during their development. ML models are trained on available data, and the relationships between variables in the data can reflect both direct causal relationships and correlations.\textsuperscript{179} If reverse causation is present in the data, ML models can potentially learn and capture those relationships as well as identify appropriate features and consider temporal relationships to minimize the potential influence of reverse causation.\textsuperscript{181} Broadly speaking, although ML models can assist in identifying potential causal relationships by revealing associations between variables and providing insights into their predictive power, establishing true causality often requires experimental design, domain expertise, and rigorous causal inference methods beyond the scope of ML techniques.\textsuperscript{179}

There are specific methods within the field of ML that can assist with causal inference.\textsuperscript{182,183} Structural causal models, also known as causal graphical models, aim to represent causal relationships among variables explicitly.\textsuperscript{184} These models can be learned from data or expert knowledge and can help elucidate the causal pathways in dementia prevention research. Techniques such as Bayesian networks and directed acyclic graphs\textsuperscript{185} have been adapted to consider counterfactual reasoning to represent and analyze causal relationships, and a recent study has shown causal Bayesian networks outperform associative algorithms to diagnose disease, particularly for rarer diseases.\textsuperscript{186} Other ML methods include causal forests, which combine elements of random forests with causal inference to estimate treatment effects and identify causal relationships.\textsuperscript{187} These methods aim to model heterogeneous treatment effects and provide insights into causal relationships while controlling for confounding variables. They can be useful in identifying factors or interventions that may have a causal impact on dementia outcomes.\textsuperscript{188,189} However, a limitation of causal ML for dementia is that the risk factors consist of intricate causal relationships involving different interactive processes, which change over time. Capturing such complexity within a causal model will be a challenge for dementia, and simplified models may not fully capture the underlying causal mechanisms. Furthermore, the lack of interpretability is a common issue and a major limitation for some ML methods such as deep learning, which may also limit clinical acceptability even with causal adaptations.\textsuperscript{190,192}

Second, another important limitation is biased ML models (algorithmic bias), which can arise when the ML models are trained on data that are of poor quality or unrepresentative, which can lead to further disadvantages in already under-represented and marginalized groups, thereby worsening existing inequity in health care.\textsuperscript{193}

This is particularly poignant for dementia prevention, as evidence suggests that dementia is more prevalent in minority groups who have a younger age of disease onset.\textsuperscript{194,195} However, most ML models are based on data from studies incorporating people of European ancestry; therefore, ML models can have poor predictability and generalizability to different subgroups in the population.\textsuperscript{196,197} However, with appropriate data-preprocessing techniques and hyperparameter tuning for ML methods such as deep learning, biases can be reduced, resulting in better predictive accuracy achieved for subgroups of the population.\textsuperscript{197}

Third, overfitting and underfitting are major issues for all ML models. Overfitting occurs when a model is overtrained to the training data, so it does not generalize to new data sets. Various methods can be implemented to avoid overfitting such as penalties for increasing model complexity and cross validation.\textsuperscript{198} Underfitting is the opposite of overfitting and where ML models cannot capture the underlying trends and variability of the data and lack predictive power.\textsuperscript{199} Therefore, the training and development of ML models often require a large data set, model optimization, and suitable model validation\textsuperscript{200}

Finally, due to the black-box nature of many ML models, there can be difficulty in the understanding and interpretation and therefore lack of acceptance and usage.\textsuperscript{201,202} Consequently, there is a need to provide understandable and transparent explanations for ML model decisions.\textsuperscript{203} ML methods and tools such as Shapley,\textsuperscript{204} Local Interpretable Model agnostic Explanations (LIME),\textsuperscript{205} and glass box methods such as Explainable Boosting Machine\textsuperscript{206} can begin to overcome challenges of lack of interpretability of ML models to help provide transparency, trust, and insights into the risk factors and indicators contributing to the development and progression of dementia as well as assisting in early detection, personalized interventions, and informed decision-making for individuals at risk.

5 | CONCLUSIONS

A wide range of potentially modifiable risk factors for dementia have been identified across the lifespan. However, the optimal combination of risk factors, possible interactions between them and with genetic risk, and their causal status remain largely unknown. Better risk-prediction tools are needed if personalized dementia prevention interventions are to be delivered in the future. Interpretable and causal ML models may refine our understanding of risk factors and could be used to reduce costs and improve the efficiency and power of future clinical trials. ML frameworks are also being developed to inform AD drug-repurposing efforts and the prioritization of potentially disease-modifying therapeutics. With the increasing complex multimodal becoming available, it is more important to combine different data to not only identify drug repurposing opportunities but also to understand the mechanisms of disease. It is vitally important that challenges relating to infrastructure, data accessibility, and ML acceptance are overcome. If we can overcome these challenges then the true potential of ML for dementia prevention will be realized, leading to precision dementia prevention.
AUTHOR CONTRIBUTIONS
D.N. contributed to the conception of the work, coordinating the writing team, and drafting and revision of the manuscript for intellectual content. V.O., C.R.M., C.P.A., S.T., V.R., M.V., I.K., S.B., D.W., I.F.F., M.B., A.K.L., E.Y.H.T., and X.Y.T. contributed to drafting and revision of the manuscript for intellectual content. J.M.R. and D.J.L. contributed to the conception of the work, conceived and organized the symposium from which this paper and others in the series originated, revised the manuscript for intellectual content, and harmonized the manuscript with other papers in the series. I.L. contributed to drafting and revision of the manuscript for intellectual content and harmonized the manuscript with other papers in the series. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS
With thanks to the Deep Dementia Phenotyping (DEMON) Network State of the Science symposium participants (in alphabetical order): Peter Bagshaw, Robin Borchert, Magda Bucholc, James Duce, Charlotte James, David Llewellyn, Donald Lyall, Sarah Marzi, Danielle Newby, Neil Oxtoby, Janice Ranson, Tim Rittman, Nathan Skene, Eugene Tang, Michele Veldsman, Laura Winchester, and Zhi Yao. This review was facilitated by the Alzheimer’s Association International Society to Advance Alzheimer’s research and Treatment (ISTAART), through the AI for Precision Dementia Medicine Professional Interest Area (PIA). The views and opinions expressed in this publication represent those of the authors and do not necessarily reflect those of the PIA membership, ISTAART, or the Alzheimer’s Association. This article was the product of a DEMON Network State of the Science symposium entitled “Harnessing Data Science and AI in Dementia Research” funded by Alzheimer’s Research UK. J.M.R. and D.J.L. are supported by Alzheimer’s Research UK and the Alan Turing Institute/Engineering and Physical Sciences Research Council (EP/N510129/1). D.J.L. also receives funding from the Medical Research Council (MR/X005674/1), National Institute for Health Research (NIHR) Applied Research Collaboration South West Peninsula, National Health and Medical Research Council (NHMRC), and National Institute on Aging/National Institutes of Health (RF1AG056564). This work was additionally supported by: European Research Council (grant agreement no. 803239 (A.K.L.), Barts Charity (C.R.M.), George Henry Woolfe Legacy Fund and the National Institute on Aging (RF1AG073593) (I.F.F.), E.Y.H.T. (National Institute for Health Research (NIHR) Clinical Lecturer) is funded by the NIHR and the views expressed in this publication are those of the author(s) and not necessarily those of the NIHR, NHS, or the UK Department of Health and Social Care. S.B. is supported by Dementias Platform UK (DPUK). The Medical Research Council supports DPUK through grant MR/T033377/1. M.B. is supported by Alzheimer’s Research UK, Economic and Social Research Council (ES/W010240/1), EU (Special EU Programmes body (SEUPBP)) INTER-REG (European Region Development Fund (ERDF)/SEUPBP), Health and Social Care Research and Development HSC R&D (COM/5750/23) and Dr George Moore Endowment for Data Science at Ulster University.

CONFLICT OF INTEREST STATEMENT
V.O. is a commissioner of the 2017 and the 2020 Lancet Commission on dementia prevention, intervention, and care. D.C.W. has served on advisory boards with Roche, Biogen, and Merck. V.R. has provided consultancy to Biogen. A.K.L. has served on advisory boards with Roche. DPuk is a DEMON partner. All other authors declare no conflicts of interest. Author disclosures are available in the Supporting information.

ORCID
Danielle Newby https://orcid.org/0000-0002-3001-1478

REFERENCES


SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.