## Title
Temporal characterization of Alzheimer's Disease with sequences of clinical records.



## Authors
Estiri, Hossein
Azhir, Alaleh
Blacker, Deborah L
et al.

## Publication Date
2023-06-01

## DOI
10.1016/j.ebiom.2023.104629

Peer reviewed

# Temporal characterization of Alzheimer's Disease with sequences of clinical records

Hossein Estiri,[a,*] Alaleh Azhir,[a,b] Deborah L. Blacker,[c] Christine S. Ritchie,[a] Chirag J. Patel,[d] and Shawn N. Murphy[e]

[a]Department of Medicine, Massachusetts General Hospital, Boston, MA, USA
[b]Harvard Medical School, Harvard-MIT Program in Health Sciences and Technology, USA
[c]Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA
[d]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[e]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA

eBioMedicine
2023;92: 104629

Published Online xxx
https://doi.org/10.1016/j.ebiom.2023.104629

## Summary

**Background** Alzheimer's Disease (AD) is a complex clinical phenotype with unprecedented social and economic tolls on an ageing global population. Real-world data (RWD) from electronic health records (EHRs) offer opportunities to accelerate precision drug development and scale epidemiological research on AD. A precise characterization of AD cohorts is needed to address the noise abundant in RWD.

**Methods** We conducted a retrospective cohort study to develop and test computational models for AD cohort identification using clinical data from 8 Massachusetts healthcare systems. We mined temporal representations from EHR data using the transitive sequential pattern mining algorithm (tSPM) to train and validate our models. We then tested our models against a held-out test set from a review of medical records to adjudicate the presence of AD. We trained two classes of Machine Learning models, using Gradient Boosting Machine (GBM), to compare the utility of AD diagnosis records versus the tSPM temporal representations (comprising sequences of diagnosis and medication observations) from electronic medical records for characterizing AD cohorts.

**Findings** In a group of 4985 patients, we identified 219 tSPM temporal representations (i.e., transitive sequences) of medical records for constructing the best classification models. The models with sequential features improved AD classification by a magnitude of 3–16 percent over the use of AD diagnosis codes alone. The computed cohort included 663 patients, 35 of whom had no record of AD. Six groups of tSPM sequences were identified for characterizing the AD cohorts.

**Interpretation** We present sequential patterns of diagnosis and medication codes from electronic medical records, as digital markers of Alzheimer's Disease. Classification algorithms developed on sequential patterns can replace standard features from EHRs to enrich phenotype modelling.

**Funding** National Institutes of Health: the National Institute on Aging (RF1AG074372) and the National Institute of Allergy and Infectious Diseases (R01AI165535).



**Keywords:** Alzheimer's Disease; Temporal representation mining; Electronic health records; Cohort identification

## Introduction

Accurate characterization of Alzheimer's Disease (AD) is complex, often requiring neuro-cognitive, genetic, imaging markers and clinical judgment. These markers are seldom routinely collected in clinical care, which limits the scalability of models that rely on them and, thus, their utility in general practice. Despite the use of detailed biomarkers, current models of AD have produced variable and often moderate classification performance, with the area under the receiver operating characteristic (ROC) curves ranging from 0.52 to under 0.86.[1–5] Suboptimal characterization of AD cohorts can lead to the introduction of unnecessary noise, by including a slew of false positive patients to the cohort and excluding from the cohort those who were falsely identified as negatives. This can, for example, impede the recruitment process for clinical trials aimed at evaluating novel therapies. Further, problematic cohort

*Corresponding author. 399 Revolution Drive, Suite 790, Somerville, MA, 02145, USA.
*E-mail address:* hestiri@mgh.harvard.edu (H. Estiri).

## Research in context

### Evidence before this study

We searched PubMed on February 16, 2023, using the search terms ("Alzheimer's Disease" "cohort classification") OR ("Alzheimer's Disease" "cohort identification") OR ("Alzheimer's Disease" "cohort characterization") OR ("Alzheimer's Disease" "prediction"). This search returned 268 articles. Our synthesis of the previous efforts to create multifactorial models for Alzheimer's Disease cohort identification suggests that the evidence before this study either provided incremental predictive value and external validity or resulted in inconsistent disease prevalence estimates due to variable case definitions. More importantly, most current models require cognitive, genetic, and imaging markers data that are seldom routinely collected in clinical care, limiting the utility of such models in general practice.

### Added value of this study

This study offers a cohort characterization model for Alzheimer's Disease (AD) built on medications and diagnoses data that are widely available in a structured format in electronic health records (EHR). To train and validate the models in this study, we applied state-of-the-art sequential representation mining and dimensionality reduction algorithms that were customized for extracting signals from noise in noisy clinical data stored in electronic health records. In addition to the demonstrated improvements in classification performance achieved by using sequences of medication and diagnosis records, compared to stand-alone diagnosis records, the models developed in this study for AD are interpretable and enable clinical storytelling. Future modelling efforts to model Alzheimer's Disease should consider utilizing EHR data to scale utility for patient screening and early prediction and leverage sequential pairs of clinical records to minimize noise and incorporate time.

### Implications of all the available evidence

Our study, with its state-of-the-art classification performance and interpretable results, suggests that a standard Machine Learning applied to sequences of EHR data can produce scalable computational characterization of Alzheimer's Disease cohorts. Digital health tools incorporating these models can enable smart screening and targeted interventions. Scalable computational cohort characterization models can also facilitate identifying those who would benefit from effective disease-modifying therapies (DMTs) as they become available. Future work should focus on the validation of generalizability in different population settings, the further exploration of the utility of temporal sequences for identifying undiagnosed cases and early prediction and deriving care/treatment pathways for AD patients from a general baseline population.

identification undermines the validity of outcomes research, for instance, that are based on inaccurately designed case and control groups.

Widely available structured real-world data (RWD) stored in electronic health record (EHR) systems, which include a longitudinal profile of symptoms, diagnoses, medications, and clinical measurements recorded in clinical care, offer possibilities to enrich cohort identification for AD for accurate outcomes research. Using diagnosis records (mainly in International Classification of Diseases [ICD] codes) in RWD is the prevalent mechanism for defining cohorts of AD patients. However, cases of Alzheimer's Disease are often poorly documented and misclassified in the EHRs,[6,7] as providers often assign a non-specific dementia code, and patients sometimes receive conflicting diagnoses. Further, many cases remain undiagnosed, especially at the early stages of disease presentation, due to difficulties in recognizing the signs and symptoms of cognitive impairment during a brief visit.[8] As a result, the reliability of diagnosis codes for identifying Alzheimer's Disease patients is suboptimal. The sensitivity and positive predictive value (PPV) of AD diagnosis code in clinical data range from (sensitivity: 60%–80%) and (PPV: 57%–100%), depending on the clinical data type and diagnosis code.[9,10]

In this study, we adopt a temporal approach for modelling evolving phenotypes with EHR data by Estiri et al.[11–14] to develop and validate a computational cohort identification algorithm for Alzheimer's Disease. Our temporal approach can be characterized as the data-driven equivalent of a rule-based algorithm that requires extensive input from clinical experts. The proposed algorithm suggests modifications to the direct use of ICD codes—e.g., a given ICD code is only useful if it proceeds to or precedes a certain medication/diagnosis pair. We demonstrate that classifying true AD cases can be better achieved through computational models that utilize transitive sequences of medications and diagnosis records—rather than AD diagnosis codes—from EHRs. We compare models trained with sequential features of medications and diagnoses to similar algorithms trained using diagnosis codes for AD. Results demonstrate that the models developed with EHR sequences achieve superior classification performances and are consistent with interpretable stories of various sequences of symptoms, recognition, evaluation, and care that often occur. In addition to the hypothesis-generating value of the sequential representations as digital markers of disease, they enable temporal storytelling capabilities and explainable Artificial Intelligence (AI).

Digital health tools built with sequential representations of real-world clinical data stored in electronic health records systems offer significant opportunities for scaling cohort identification across the AD

continuum at a lower cost than models needing hard-to-collect information, and with higher precision than models relying on diagnosis codes for the phenotype.

## Methods

### Ethics

The use of data for this study was approved by the Mass General Brigham Institutional Review Board (protocol# 2017P000282) with a waiver of informed consent.

### Statistics

#### *Dataset*

We utilized structured medication and diagnosis data for patients who received care at eight healthcare facilities, two tertiary medical centres and four community hospitals, two speciality hospitals, and over 35 primary care centres within the Mass General Brigham (MGB) integrated healthcare system's footprint in the New England region. Our dataset included patients with samples in the MGB Biobank[15] until the end of 2020.

#### *Study cohort and classification tasks*

The study cohort comprised all patients with at least an encounter with an Alzheimer's Disease or a Dementia diagnosis code. We used diagnosis codes from both the 9th, and 10th revision of International Classification of Diseases (ICD) codes collected under AD and Dementia Phecodes,[16] which we augmented with historic local codes—the list of ICD codes are available in Table S1 in the Appendix. We identified two classification tasks. First, the goal was to identify patients with a true AD diagnosis, given any diagnosis code (AD, other dementing illness, or nonspecific dementia)—henceforth, we call this task and the related study cohort the "Dx AD/Dementia". Looking at AD diagnosis codes alone, we ran a second classification task on a subset of the first cohort (those with at least one specific AD diagnosis code) to identify patients with true AD diagnosis given a specific AD diagnosis code—henceforth, we call this task and the subsequent cohort "Dx AD".

#### *Chart reviews for gold-standard labels*

To evaluate classification models, we performed chart reviews on 150 patients from our patient pool (i.e., Dx AD/Dementia cohort). In the chart reviews, an expert clinician performed a manual review of patients' charts to adjudicate the presence (or lack thereof) of Alzheimer's Disease, regardless of the structured diagnosis codes—chart review criteria for adjudicating cases is available in Table S2 in the Appendix.

#### *Semi-supervised learning*

The data and modelling pipeline is illustrated in Fig. 1. We curated silver-standard labels (see Appendix), applied temporal representation mining and dimensionality reduction to engineer features, and trained and tested the two classification tasks through semi-supervised learning with structured electronic health records data. We describe each step below:

#### *Feature engineering and modelling*

We developed two sets of features. First, we applied the transitive Sequential pattern Mining (tSPM)[11,17] algorithm to mine the sequential representation of medication and diagnosis records from the study cohort's electronic health records. To mine tSPM representations, the algorithm temporally sorts the medication and diagnosis records from the EHR, selects the earliest record for each clinical observation, and mines all pairs of medication or diagnosis codes that are sequentially related (Fig. 1). This results in a large vector of tSPM sequential representation that is then fed into a high-throughput dimensionality reduction algorithm. Second, as the baseline model, we used a list of Alzheimer's Disease diagnosis ICD-9/10 codes defined in Phecodes. We controlled for age and sex in all models. Sex was self-reported by study participants.

Following Estiri, Vasey, and Murphy (2021), we applied the Minimize Sparsity, Maximize Relevance (MSMR) dimensionality reduction algorithm[11] to select a small subset of sequential temporal patterns that convey useful information for classification tasks. The resulting sequences were then fed into a Gradient Boosting Machine (GBM), using the R package gbm.[18] GBM, through boosting, applies a final step of feature selection to the final modelling features. This model also provides a relative tree-based feature importance metric, determined based on the cumulative use of a given sequential feature in each decision tree step across all trees used in the final boosted model. We used the feature importance score in each classification task to rank the final sequential features.

#### *Understanding the sequential features*

To understand the meaning and functions of the obtained sequences of EHR observations, we used visualization techniques and clinical expertise. We leveraged the final set of tSPM representations to develop a dashboard that temporally connects the sequences for creating pathway visualization. We also created doughnut charts to understand the marginal benefit of the sequential approach by comparing the positive predictive values of each element of a sequence and the full sequence. The positive predictive values (PPVs) are computed based on the true positive patients identified by the feature in the computed AD cohort. For example, in a feature sequentially entailing elements $a$ and $b$, $a \rightarrow b$, we provided positive predictive values for $a$ and $b$ individually and then $a \rightarrow b$. We colour-coded sequences in the dashboard to indicate whether a sequential feature is positively or negatively associated with true AD. To assess the contribution of the temporal direction in the $a \rightarrow b$, we also evaluate PPVs of $a\&b$ and $b \rightarrow a$.

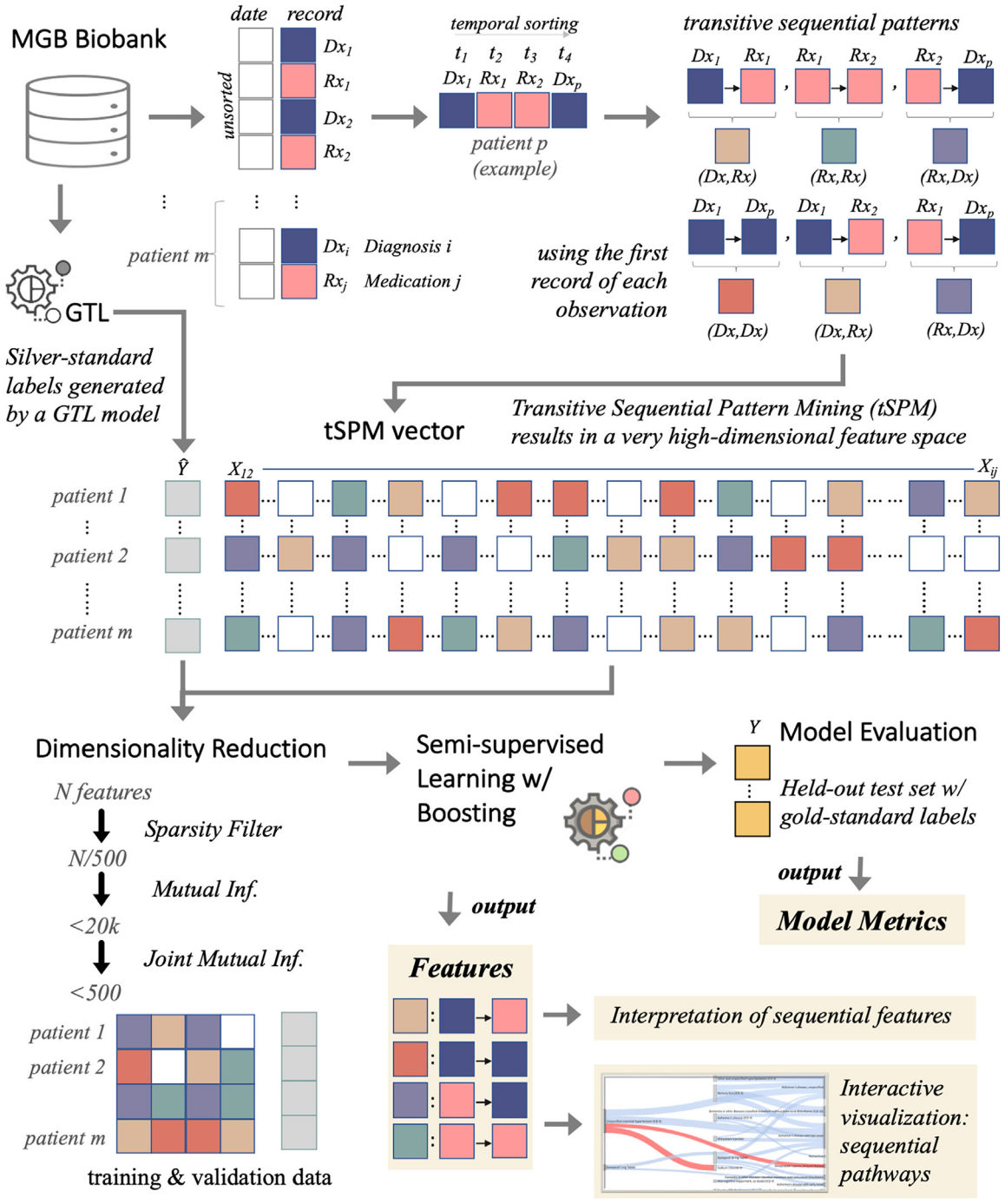


***Fig. 1:* The data pipeline for developing and testing AD classification models through semi-supervised learning with electronic health records data.** The pipeline encompasses mining transitive sequential patterns from clinical data, curating silver-standard labels, performing dimensionality reduction, modelling and model evaluations, and interpreting results in an interactive visualization dashboard.

*Model evaluations*

To evaluate the models, we used metrics of discrimination, namely areas under the receiver operating characteristics curves (ROC) and precision-recall curve receiver (PR) and calibration error, namely the Brier score and root mean square error RMSE. We computed these metrics for each classification task and model on the held-out test testing data curated through chart reviews.

### Role of funders
The funders had no role in study design, data collection, analysis, interpretation, or report writing.

### Data availability
Protected Health Information restrictions apply to the availability of the clinical data here, which were used under IRB approval for use only in the current study. As a result, this dataset is not publicly available. Qualified researchers affiliated with the Mass General Brigham (MGB) may apply for access to this data through the MGB Institutional Review Board.

## Results
4985 patients met our inclusion criteria for the overall cohort, based on having any diagnosis code (AD, other dementia type, or nonspecific dementia)—therefore, comprised our Dx AD/Dementia cohort for the first classification task. Nested within this cohort of patients, 1093 had a specific AD diagnosis record, constituting our Dx AD cohort for the second classification task. Aggregated demographic and clinical characteristics of these patients are provided in Table 1. In the gold-standard data used for testing the algorithms, 33 patients (of the 150 chart-reviewed) were labelled positive for AD. 56 (37.3%) of the chart-reviewed patients had a specific diagnosis code for AD, of whom 32 were labelled positive for AD—therefore, the positive predictive value (PPV) for AD diagnosis was 57.1 percent in the Dx AD cohort.

We mined 162,245,302 transitive sequences from 97,222 unique EHR records from the overall study population of 4985 patients. 12,501,934 of the sequences passed the sparsity filter in MSMR. Feature engineering resulted in a set of 219 features (Table S3 in the Appendix), which include age and sex as variables. Percentages have been rounded.

### Model performances
As demonstrated in Table 2 and Fig. 2, the model with tSPM sequences outperformed the model using AD diagnosis codes (Dx AD) in both discrimination and calibration metrics. On the discrimination metrics, the AUROCs and the precision-recall curves were between 2.88 and 9.84 percent improved in the sequential model. The calibration error metrics also showed improvement–between 3.24 and 30 percent in the models with EHR sequences, compared with the model with only AD diagnosis records (Dx AD).

Based on our trained tSPM model, 663 patients would be included in the AD cohort. 35 patients in the computed cohort did not have an AD diagnosis code, which suggests that our model is, to some extent, capable of detecting AD under coding. Compared to the two diagnosis code-based cohorts, patients in this computed cohort were slightly older, on average, had more diagnosis codes for AD and outpatient encounters, and had fewer inpatient encounters (Table 1). Among the demographic variables included in the models, age was the 30th and 10th important feature for classification tasks 1 (classifying true AD patients from patients with AD or Dementia diagnosis code) and 2 (classifying true AD patients from patients with an AD diagnosis code), respectively. Sex was not an important feature for classification in task 1 and marginally important (ranked 63) for task 2.

### Do the identified sequential features make clinical sense?
Among the 219 sequences we found for classifying Alzheimer's Disease, all diagnosis records that were the

| | Dx AD/Dementia cohort | Dx AD cohort[a] | Computed AD cohort[b] |
|---|---|---|---|
| Patients | 4985 | 1093 | 663 |
| Mean age | 72.58 | 77.93 | 79.94 |
| Mean Charlson score | 4.57 | 4.75 | 4.59 |
| Mean data depth (years) | 19.49 | 21.58 | 22.23 |
| Mean phenotype record | 3 | 12 | 17 |
| percent inpatient visits | 25 | 13 | 11 |
| percent outpatient visits | 35 | 46 | 49 |
| Unique EHR records | 97,222 | 54,882 | 43,360 |
| percent Female[a] | 47 | 52 | 52 |
| percent Male | 53 | 48 | 48 |
| percent African American | 5 | 3 | 3 |
| percent LatinX | 3 | 3 | 2 |
| percent White | 88 | 90 | 90 |

[a]Dx AD cohort is a subset of Dx AD/Dementia cohort who has at least a diagnosis record for AD. [b]Computed AD cohort comprises patients from the Dx AD/Dementia cohort who are likely to have AD based on the tSPM model's predicted probabilities for AD.

*Table 1:* **Demographic and clinical characteristics of the patient cohorts.**

| Classification task | 1: Dx AD/Dementia[a] | | | | 2: Dx AD[b] | | | |
|---|---|---|---|---|---|---|---|---|
| | Discrimination | | Calibration | | Discrimination | | Calibration | |
| | ROC | PR | Brier | RMSE | ROC | PR | Brier | RMSE |
| Dx AD[c] | 0.945 | 0.860 | 0.077 | 0.278 | 0.823 | 0.877 | 0.190 | 0.436 |
| tSPM[d] | 0.973 | 0.895 | 0.073 | 0.269 | 0.904 | 0.924 | 0.133 | 0.364 |
| Improvement | 2.88% | 3.91% | 5.19% | 3.24% | 9.84% | 5.36% | 30.0% | 16.51% |

[a]Classification task 1 classifies AD patients from patients with an AD or Dementia diagnosis code. [b]Classification task 2 classifies AD from patients with at least an AD diagnosis code. [c]GBM model trained with diagnosis codes for AD. [d]GBM model trained with tSPM sequences.

*Table 2:* **Model performance metrics.**

first element of the sequence (i.e., $a$ in $a \rightarrow b$) were ICD-9 codes, representing a possibly older record. Except for 2, all diagnosis records that were the later sequence element (i.e., $b$ in $a \rightarrow b$) were ICD-10 codes, representing a possibly newer record. We categorized the important sequences under the following groups. Visualizations in this section are provided from an interactive dashboard we developed to visualize sequential pathways and study positive predictive values for sequences and their elements (Fig. S1 and Appendix). The positive predictive values are computed based on the true positive patients identified by the feature in the computed AD cohort.

### A symptom in the past followed by an AD diagnosis or medication

The top 2 important sequences shared between the two classification tasks were sequences of an AD/Dementia symptom followed by an AD diagnosis or medication. An example of such sequences is memory loss (78,093–ICD 9 Diagnosis Code) followed by AD (late-onset G30.1 or unspecified G30.9) ICD-10 code. Memory loss is a common symptom of Alzheimer's Disease. Also, a past record of memory loss ICD-9 code followed by AD medications such as memantine and donepezil can positively indicate true Alzheimer's Disease. Our data shows that memory loss alone has a relatively low positive predictive value (~41%) for truly identifying AD (Fig. 3). The PPV for memantine (5 mg) and donepezil (10 mg) are 56% and 62%, respectively. However, when sequentially paired with memory loss, the respective positive predictive values increase to 71% and 72%.

### A risk factor in the past followed by an AD diagnosis

The 3rd most important sequential feature in classifying AD encompassed a potential risk factor in the past followed by an AD diagnosis. Notable in this category was

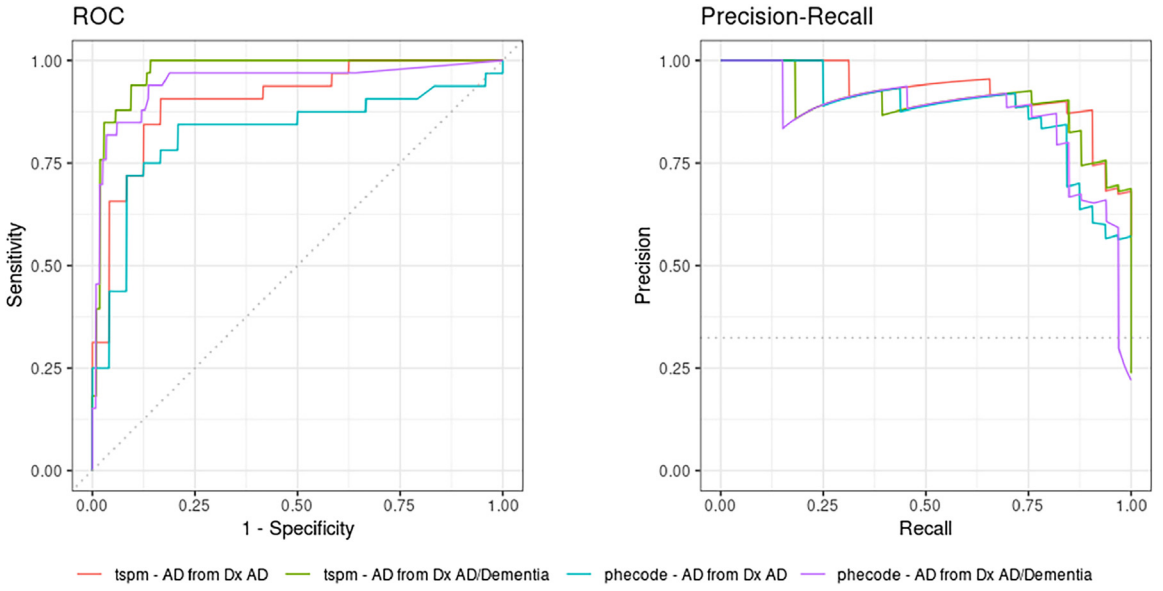


***Fig. 2:*** **The ROC and PR curves grouped by the task and features.** Classification task 1 identified AD patients from patients with an AD or Dementia diagnosis code, whereas classification task 2 identified AD from patients with at least an AD diagnosis code. Phecodes include all diagnosis codes for AD (Supplementary Table). tSPM models use sequential patterns of medications and diagnoses from the EHR data.

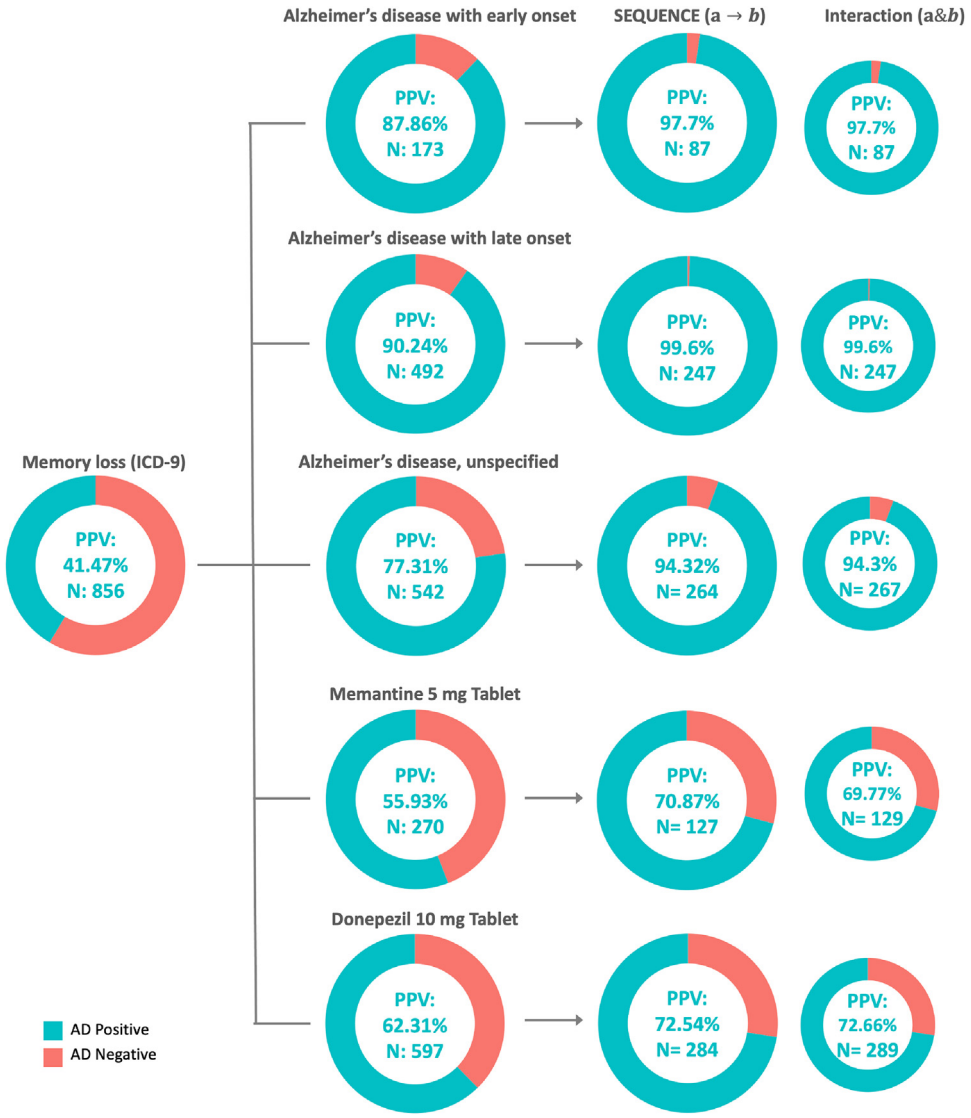


*Fig. 3:* **Comparing positive predictive values of the elements of a sequential feature with memory loss for classifying AD.**

unspecified essential hypertension (ICD-9 code 4019), followed by AD ICD-10 diagnosis code for AD or Dementia. Hypertension in midlife is particularly associated with an increased risk of developing dementia and Alzheimer's Disease.[19,20] Also in this category were past records of unspecified hyperlipidemia (ICD-9 code 2724) or hypercholesterolemia (ICD-9 code 2720), followed by AD diagnosis code (Fig. S2 and Appendix).

### Sequences involving AD medications

Some of the important sequential features we found for classifying AD included sequences of AD medications and diagnosis codes for AD or dementia (Fig. 4). Among medications, donepezil and memantine were primarily included. For instance, sequences of donepezil with AD or dementia diagnosis records and AD followed by memantine carry positive signals for classifying AD (Fig. 5).

Other important sequences of AD medication reflected changes in medications or in dosage. A highly important sequence, for example, indicated an increase in dosage for donepezil from 5 mg to 10 mg. Another sequence possibly reflected a change in the treatment plan from donepezil to memantine. Donepezil is a medication typically the first medication used to treat Alzheimer's Disease. In contrast, memantine is indicated for moderate to severe Alzheimer's Disease, though is sometimes used earlier if the patient cannot tolerate donepezil or requests it for other reasons.

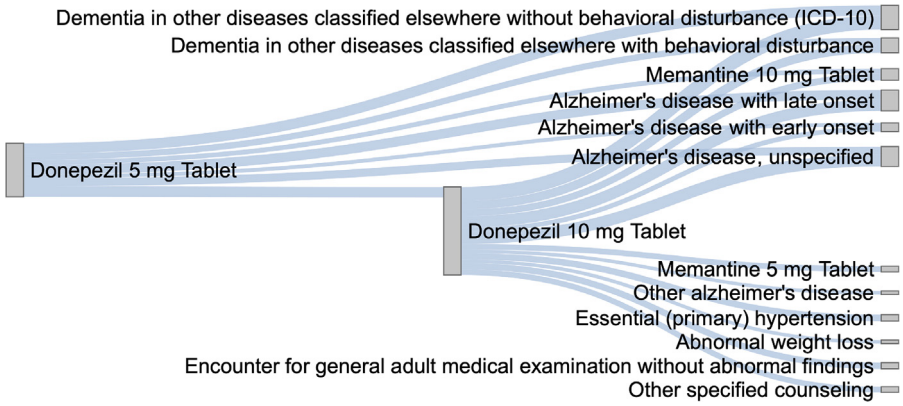


*Fig. 4:* **Sequential pathways that indicate true AD compiled by connecting identified sequences of AD medications and diagnoses in the EHR.** The visualization dashboard compiles sequential pathways by connecting identified sequences of EHR observations. The pathways presented here involve donepezil and memantine.

### Indirect medication relations to AD diagnosis

We found that the sequence of Sodium Chloride IV followed by Alzheimer's Disease with early onset is an important marker for classifying AD patients. The positive predictive value of the sequence is 96.72 percent, whereas PPV for Alzheimer's Disease with early onset and Sodium chloride IV alone is 87.86 and 9 percent, respectively. Sodium chloride IV solution is commonly used to treat a variety of conditions, including dehydration, electrolyte imbalances, and certain medical emergencies, and is. Several factors can also lead to dehydration in people with AD/dementia, including decreased thirst (or difficulty recognizing thirst), difficulty remembering to drink enough fluids or communicate their thirst, and due to AD or other medications that can lead to dehydration. Some medications used to

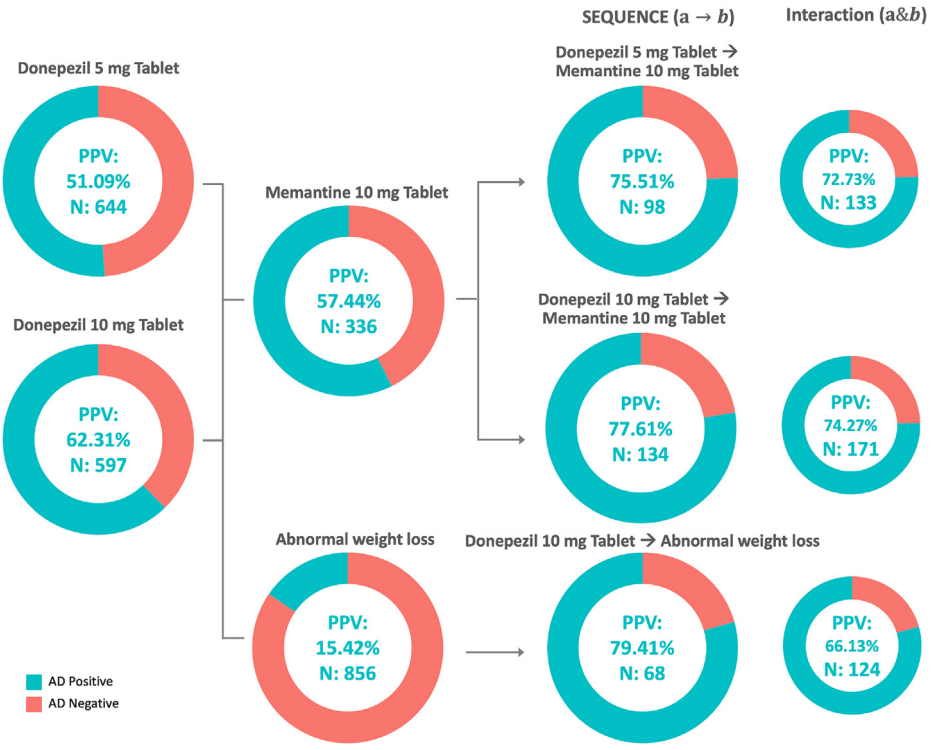


*Fig. 5:* **Comparing positive predictive values of the elements of a sequential feature with AD medications that can improve AD classification.**

treat AD, such as cholinesterase inhibitors, can cause side effects such as diarrhoea and increased urination, contributing to dehydration. Other AD medications, such as memantine, can cause side effects such as constipation, which can also contribute to dehydration if it leads to infrequent bowel movements.

### Comorbidities with shared roots or as side effects

We found that some sequences of comorbidities with a possible shared cause with AD or AD side effects can potentially indicate true AD. Examples included sequences of different joint pains followed by an AD diagnosis or medication as important digital markers for AD phenotyping (Fig. 6 and Fig. S3 in the Appendix). Pain in a joint can have a number of causes, including injuries, overuse, or underlying medical conditions, such as Rheumatoid Arthritis (RA), Osteoarthritis (OA), Peripheral arterial disease, and Compression fractures. Both RA and AD diseases are associated with older persons and genetic factors. Besides the inflammation associated with RA, reduced blood flow to vital body organs can increase the risk of developing dementia. Additionally, anti-rheumatic medications used by RA patients may increase the risk of developing dementia.[21]

Another sequence in this category is Anemia unspecified (ICD-9 code), followed by an AD or dementia diagnosis code. Less than 14% of patients in our cohort with an anaemia record are positive for AD. In this setting, an unspecified AD ICD-10 code has a PPV of 77.3 percent. When sequentially paired, Anemia followed by a non-specific AD has a PPV of 81.8 percent. When paired with a late-onset AD diagnosis code, the PPV goes up to 94.3. In both sequences, the positive predictive value for AD increases by about 5 percent. Several studies have found that older adults with anaemia may have an increased risk of developing Alzheimer's Disease.[22–24]

### Repeated AD diagnosis codes

Sequences that embody a temporal repetition of the diagnosis codes are also important in confirming the AD diagnosis. For example, one of the top features we found shows that if a patient has the ICD-9 code 3310 (Alzheimer's Disease) followed by an ICD-10 diagnosis


*Fig. 6:* **Comparing positive predictive values of the elements of a sequential feature with joint pain for classifying AD.**

code for AD, chances are higher that the patient truly has AD.

**Does the temporal direction of a sequence matter?** To assess the contribution of the temporal direction in the $a \rightarrow b$, we evaluated positive predictive values of $a\&b$ (i.e., interactions or joint observations) and $b \rightarrow a$ (i.e., reverse sequences). Out of the 219 sequences mined in this study, 107 reverse sequences were available in the data. That is, for about half of the sequences, the reverse sequence was not observed. As illustrated in Figs. 3, 5 and 6, we found that for all sequences identified by the algorithm, the positive predictive values were larger or equal to the PPVs obtained from the interactions or joint observations of the sequence elements. In many cases, the sequence reflected the interactions or joint observations between the two elements of the sequence ($a\&b$), especially when the reverse sequence was too sparse. Overall, the temporal direction mattered in identifying a sequence that contained a signal boost. For example, the sequence of donepezil 10 mg tablet preceding abnormal weight loss (Fig. 5) has a positive predictive value of 79.41%. The reverse sequence (i.e., abnormal weight loss → donepezil 10 mg tablet) and the joint observation of donepezil 10 mg tablet and abnormal weight loss had PPVs of 50% and 66.13%, respectively. Similarly, Alzheimer's Disease (ICD-9) followed by an ICD-9 diagnosis code for Dementia, unspecified, without behavioural disturbance had a PPV equal to 88.12%, while the PPVs for reverse sequence and the joint observation of the two records were respectively 72.37% and 81.36%.

## Discussion

Real-world clinical data stored in electronic health records systems offer significant opportunities for developing powerful tools to improve epidemiologic evidence on Alzheimer's Disease at a lower cost than face-to-face clinical studies. However, due to known data reliability issues, study cohorts from EHR data need to be carefully defined in order to minimize the introduction of additional and unnecessary noise in the study. This can be achieved through expert-driven algorithms, which are often costly,[25–27] or carefully developed computational models that can scale into digital health tools. Given the projected number of older adults that will develop AD in the future (as they age), scalable computational cohort characterization models can facilitate identifying those who would benefit from effective disease-modifying therapies (DMTs) as they become available.

AI-based digital health tools—such as one built upon the sequential models developed in this study—for improving AD cohort identification with widely available structured clinical data can accelerate precision drug development (e.g., through accelerating clinical trial recruitment) and improve outcomes research (e.g., via reducing the noise in identifying cases and control). However, prior research on the development of prognostic/diagnostic models of AD has had limited incremental predictive value and external validity or has required cognitive, genetic, and imaging markers that are not routinely collected in clinical care. Imprecise AD cohort identification can impede the recruitment process for clinical trials to evaluate novel therapies undermining the validity of outcomes research.

EHR observations reflect a complex set of processes that thwart their seamless translation into actionable knowledge. Namely, the raw EHR observation data may not be direct indicators of a patient's "true" health states at different time points but rather reflect the patients' interactions with the system, the clinical processes, and the recording processes. EHR observations are also recorded asynchronously across time (i.e., measured at different times and irregularly), which presents foundational challenges for directly applying standard temporal analysis methods. In this paper, we provided a way of identifying digital markers, via transitive sequential pattern mining (tSPM), to identify AD in EHR data. Our results demonstrated that, given the limited reliability of AD diagnosis codes, sequential pairs of clinical records stored in the EHRs can augment cohort identification of Alzheimer's Disease cohorts by a magnitude of 3–16 percent on a net improvement, over the use of AD diagnosis codes alone.

We categorized the sequences for identifying Alzheimer's Disease patients into six groups. Most important sequences represented (1) a symptom in the past followed by an AD diagnosis or medication (notably involving past records of memory loss and/or mild cognitive impairment), or (2) a risk factor in the past followed by an AD diagnosis, such as unspecified essential hypertension, unspecified hyperlipidemia or hypercholesterolemia. Another group of sequential patterns involving (3) sequences of AD medications, mainly donepezil and memantine, also carried positive signals for identifying AD. These sequences likely represented changes in the agent(s) prescribed due to the progression of AD or side effects. We also found sequences as (4) indirect digital markers for AD, such as indirect medication relations to AD diagnosis (e.g., supply of Fluorodeoxyglucose F18, Sodium Chloride IV, Midazolam followed by AD diagnosis). Sodium Chloride IV, PET scan (where Fluorodeoxyglucose F18 is needed), and Midazolam are all typically administered by a healthcare provider in a hospital or clinical setting. A fifth group of sequential features (5) represented comorbidities with possible shared roots or side effects. Sequences of different joint pains followed by an AD diagnosis or medication were important digital markers for AD phenotyping in this group. The final group of sequences was those that (6) represented two different records of AD diagnosis codes.

Similar to all other work based on observational data from electronic health records, this study has limitations. One of those limitations can be attributed to the data quality. We hypothesize that mining temporal representations from raw EHR data may alleviate some of the data quality issues, but we have not rested this hypothesis thoroughly. In addition, data quality issues, such as lack of longitudinal completeness, can be reflected in the sequential patterns mined. In addition, we used an already-enriched population to build our study on, which included all patients with at least a diagnosis record for either Dementia or Alzheimer's Disease. Therefore, the generalizability of the models developed here to the general population needs further study. Although the data used in this study originate from multiple institutions and care settings, the population distribution is geographically limited. Further evaluation is needed to understand possible fluctuations across geographic regions (and thus demographic characteristics) and care settings.

In addition to diagnosis and medication codes, EHR data contain more valuable information and data elements that can be leveraged within the sequencing framework. For instance, information about the context where an observation was recorded (e.g., primary care versus a new referral to speciality care, such as neurology or psychiatry) and the timing and patterns between clinical encounters. Additional data elements that can be incorporated into the sequencing framework include clinical notes, procedure codes, vitals, and laboratory tests from the EHR data and genomics from the biobank. Another limitation of this study is that performance comparisons to a comprehensive set of available benchmark AD algorithms (e.g., CALIBER)[28] are lacking.

In addition to the superior classification performance and the hypothesis-generating value of the sequential representations as digital markers of disease, they enable temporal storytelling and thus provide tools for explainable Artificial Intelligence (AI). As the clinical utility of complex AI algorithms has been under scrutiny,[29–31] explainability has become crucial for elevating clinical utility and impacting patient lives. As this method continues to develop, the classification pipeline, with the story-telling capabilities added through interpretive sequences, can be used in data-driven tools that could facilitate diagnostics and cohort characterization across the AD continuum and enhance public health surveillance, targeted screening, and individualize treatment, improve misdiagnosis rate, and enable more precise planning for health care and caregiving resources. Digital health tools built with tSPM representations on real-world clinical data stored in the EHRs offer significant opportunities for high-throughput precision cohort characterization across the AD continuum and at scale and at a lower cost.

#### Contributors

HE and SNM conceived the study design. SNM designed and oversaw the chart reviews. HE and AA developed and implemented the code for data mining, dimensionality reduction, classification tasks, and visualizations. HE and AA prepared the manuscript, tables, and figures. HE and SNM had access to all the data. HE, CJP, and SNM obtained funding. HE led the project administration. All authors reviewed and interpreted the results, commented on the paper, contributed to revisions, and read and approved the final version. HE had final responsibility for the decision to submit for publication.

#### Data sharing statement

Protected Health Information restrictions apply to the availability of the clinical data here, which were used under IRB approval for use only in the current study. As a result, this dataset is not publicly available. Qualified researchers affiliated with the Mass General Brigham (MGB) may apply for access to this data through the MGB Institutional Review Board. Computer code to perform the analyses in this study is available via: https://hestiri.github.io/mlho.

#### Declaration of interests

C Ritchie report grants from NIH and Retirement Research Foundation for other projects and being part of the steering committee of IMPACT Collaboratory as well as a board member of the International Neuropalliative Care Society. The other authors declare no competing interests.

#### Acknowledgment

The National Institute on Aging (RF1AG074372) and the National Institute of Allergy and Infectious Diseases (R01AI165535).

#### Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2023.104629.

#### References

1 Grassi M, Perna G, Caldirola D, Schruers K, Duara R, Loewenstein DA. A clinically-translatable machine learning algorithm for the prediction of Alzheimer's disease conversion in individuals with mild and premild cognitive impairment. *J Alzheimers Dis*. 2018;61:1555–1573.
2 Tjandra D, Migrino RQ, Giordani B, Wiens J. Cohort discovery and risk stratification for Alzheimer's disease: an electronic health record-based approach. *Alzheimers Dement*. 2020;6:e12035.
3 Barnes DE, Zhou J, Walker RL, et al. Development and validation of eRADAR: a tool using EHR data to detect unrecognized dementia. *J Am Geriatr Soc*. 2020;68:103–111.
4 Coley RY, Smith JJ, Karliner L, et al. External validation of the eRADAR risk score for detecting undiagnosed dementia in two real-world healthcare systems. *J Gen Intern Med*. 2022;38:351. https://doi.org/10.1007/s11606-022-07736-6. published online July 29.
5 Coley RY, Yates Coley R, Smith JJ, et al. eRADAR detects primary care patients at risk of having undiagnosed dementia in two real-world healthcare systems. *Alzheimers Dement*. 2022;18. https://doi.org/10.1002/alz.062967.
6 Chodosh J, Petitti DB, Elliott M, et al. Physician recognition of cognitive impairment: evaluating the need for improvement. *J Am Geriatr Soc*. 2004;52:1051–1059.
7 Cho K, Gagnon DR, Driver JA, et al. Dementia coding, workup, and treatment in the VA new England healthcare system. *Int J Alzheimers Dis*. 2014;2014:821894.
8 Bradford A, Kunik ME, Schulz P, Williams SP, Singh H. Missed and delayed diagnosis of dementia in primary care: prevalence and contributing factors. *Alzheimer Dis Assoc Disord*. 2009;23:306–314.
9 Solomon A, Ngandu T, Soininen H, Hallikainen MM, Kivipelto M, Laatikainen T. Validity of dementia and Alzheimer's disease diagnoses in Finnish national registers. *Alzheimers Dement*. 2014;10:303–309.
10 Wilkinson T, Ly A, Schnier C, et al. Identifying dementia cases with routinely collected health data: a systematic review. *Alzheimers Dement*. 2018;14:1038–1051.

11 Estiri H, Vasey S, Murphy SN. Transitive sequential pattern mining for discrete clinical data. *International conference on artificial*. 2020. https://link.springer.com/chapter/10.1007/978-3-030-59137-3_37.
12 Estiri H, Strasser ZH, Klann JG, et al. Transitive sequencing medical records for mining predictive and interpretable temporal representations. *Patterns*. 2020;1:100051. published online June 18.
13 Estiri H, Vasey S, Murphy SN. Generative transfer learning for measuring plausibility of EHR diagnosis records. *J Am Med Inform Assoc*. 2021;28:559–568.
14 Estiri H, Strasser ZH, Murphy SN. Individualized prediction of COVID-19 adverse outcomes with MLHO. *Sci Rep*. 2021;11:5322.
15 Karlson E, Boutin N, Hoffnagle A, Allen N. Building the partners HealthCare Biobank at partners personalized medicine: informed consent, return of research results, recruitment lessons and operational considerations. *J Pers Med*. 2016;6:2.
16 Wu P, Gifford A, Meng X, et al. Mapping ICD-10 and ICD-10-CM codes to Phecodes: workflow development and initial evaluation. *JMIR Med Inform*. 2019;7:e14325.
17 Estiri H, Strasser ZH, Murphy SN. High-throughput phenotyping with temporal sequences. *J Am Med Inform Assoc*. 2020;28:772. https://doi.org/10.1093/jamia/ocaa288. published online Dec 14.
18 Greenwell B, Boehmke B, Cunningham J, Developers GBM, Greenwell MB. Package 'gbm'. R package version. 2019;2. ftp://r-project.org/pub/R/web/packages/gbm/gbm.pdf.
19 Hu YH, Halstead MR, Bryan RN, et al. Association of early adulthood 25-year blood pressure trajectories with cerebral lesions and brain structure in midlife. *JAMA Netw Open*. 2022;5:e221175.
20 Liu L, Hayden KM, May NS, et al. Association between blood pressure levels and cognitive impairment in older women: a prospective analysis of the Women's Health Initiative Memory Study. *Lancet Healthy Longev*. 2022;3:e42–e53.
21 Sangha PS, Thakur M, Akhtar Z, Ramani S, Gyamfi RS. The link between rheumatoid arthritis and dementia: a review. *Cureus*. 2020;12:e7855.
22 Wolters FJ, Zonneveld HI, Licher S, et al. Hemoglobin and anemia in relation to dementia risk and accompanying changes on brain MRI. *Neurology*. 2019;93:e917–e926.
23 Hong CH, Falvey C, Harris TB, et al. Anemia and risk of dementia in older adults: findings from the Health ABC study. *Neurology*. 2013;81:528–533.
24 Jeong SM, Shin DW, Lee JE, Hyeon JH, Lee J, Kim S. Anemia is associated with incidence of dementia: a national health screening study in Korea involving 37,900 persons. *Alzheimers Res Ther*. 2017;9:94.
25 Agarwal V, Podchiyska T, Banda JM, et al. Learning statistical models of phenotypes using noisy labeled training data. *J Am Med Inform Assoc*. 2016;23:1166–1173.
26 Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep*. 2016;6:26094.
27 Yu S, Liao KP, Shaw SY, et al. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *J Am Med Inform Assoc*. 2015;22:993–1000.
28 Alexander N, Alexander DC, Barkhof F, Denaxas S. Identifying and evaluating clinical subtypes of Alzheimer's disease in care electronic health records using unsupervised machine learning. *BMC Med Inform Decis Mak*. 2021;21:343.
29 CONSORT-AI, SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med*. 2019;25:1467–1468.
30 Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak*. 2020;20:310.
31 Adadi A, Berrada M. Explainable AI for healthcare: from black box to interpretable models. In: *Embedded systems and artificial intelligence*. Singapore: Springer; 2020:327–337.