

ARCH: Large-scale Knowledge Graph via Aggregated Narrative Codified Health Records Analysis

Ziming Gan^{1*}, Doudou Zhou^{2*}, Everett Rush³, Vidul A. Panickan^{4,5}, Yuk-Lam Ho⁵,
George Ostrouchov³, Zhiwei Xu⁶, Shuting Shen², Xin Xiong², Kimberly F. Greco², Chuan Hong⁷,
Clara-Lea Bonzel^{4,5}, Jun Wen⁴, Lauren Costa⁵, Tianrun Cai^{5,8}, Edmon Begoli³, Zongqi Xia⁹,
J. Michael Gaziano^{4,5,8}, Katherine P. Liao^{5,8}, Kelly Cho^{4,5,8†}, Tianxi Cai^{2,4,5}†, Junwei Lu^{2,5}†

¹University of Chicago, Chicago, IL, USA

²Harvard T.H. Chan School of Public Health, Boston, MA, USA

³Oak Ridge National Laboratory, Oak Ridge, TN USA

⁴Harvard Medical School, Boston, MA, USA

⁵VA Boston Healthcare System, Boston, MA, USA

⁶University of Michigan, Ann Arbor, MI, USA

⁷Duke University, Durham, NC, USA

⁸Brigham and Women's Hospital, Boston, MA, USA

⁹University of Pittsburgh, Pittsburgh, USA

Summary

Objective: Electronic health record (EHR) systems contain a wealth of clinical data stored as both codified data and free-text narrative notes, covering hundreds of thousands of clinical concepts available for research and clinical care. The complex, massive, heterogeneous, and noisy nature of EHR data imposes significant challenges for feature representation, information extraction, and uncertainty quantification. To address these challenges, we proposed an efficient **A**ggregated **n**a**R**rative **C**odified **H**ealth (ARCH) records analysis to generate a large-scale knowledge graph (KG) for a comprehensive set of EHR codified and narrative features.

Methods: The ARCH algorithm first derives embedding vectors from a co-occurrence matrix of all EHR concepts and then generates cosine similarities along with associated p -values to measure the strength of relatedness between clinical features with statistical certainty quantification. In the final step, ARCH performs a sparse embedding regression to remove indirect linkage between entity pairs. We validated the clinical utility of the ARCH knowledge graph, generated from 12.5 million patients in the Veterans Affairs (VA) healthcare system, through downstream tasks including detecting known relationships between entity pairs, predicting drug side effects, disease phenotyping, as well as sub-typing Alzheimer's disease patients.

Results: ARCH produces high-quality clinical embeddings and KG for over 60,000 EHR concepts, as visualized in the R-shiny powered web-API (<https://celehs.hms.harvard.edu/ARCH/>). The ARCH embeddings attained an average area under the ROC curve (AUC) of 0.926 and 0.861 for detecting pairs of similar EHR concepts when the concepts are mapped to codified data and to NLP data; and 0.810 (codified) and 0.843 (NLP) for detecting related pairs. Based on the p -values computed by ARCH, the sensitivity of detecting similar and related entity pairs

*Gan and Zhou contributed equally

†Cho, Cai and Lu contributed equally

25 are 0.906 and 0.888 under false discovery rate (FDR) control of 5%. For detecting drug side
26 effects, the cosine similarity based on the ARCH semantic representations achieved an AUC of
27 0.723 while the AUC improved to 0.826 after few-shot training via minimizing the loss function
28 on the training data set. Incorporating NLP data substantially improved the ability to detect
29 side effects in the EHR. For example, based on unsupervised ARCH embeddings, the power of
30 detecting drug-side effects pairs when using codified data only was 0.15, much lower than the
31 power of 0.51 when using both codified and NLP concepts. Compared to existing large-scale
32 representation learning methods including PubmedBERT, BioBERT and SAPBERT, ARCH
33 attains the most robust performance and substantially higher accuracy in detecting these rela-
34 tionships. Incorporating ARCH selected features in weakly supervised phenotyping algorithms
35 can improve the robustness of algorithm performance, especially for diseases that benefit from
36 NLP features as supporting evidence. For example, the phenotyping algorithm for depression
37 attained an AUC of 0.927 when using ARCH selected features but only 0.857 when using codified
38 features selected via the KESER network[1]. In addition, embeddings and knowledge graphs
39 generated from the ARCH network were able to cluster AD patients into two subgroups, where
40 the fast progression subgroup had a much higher mortality rate.

41
42 **Conclusions:** The proposed ARCH algorithm generates large-scale high-quality semantic rep-
43 resentations and knowledge graph for both codified and NLP EHR features, useful for a wide
44 range of predictive modeling tasks.

45 **Keywords:** Electronic health records, natural language processing, representation learning, knowl-
46 edge graph.

47 1 Introduction

48 The increasing adoption of electronic health record (EHR) systems has provided opportunities for
49 clinical studies and biomedical research ranging from patient phenotyping [2] and prediction of
50 medical events [3], to relationship extraction between medications and adverse drug effects [4].
51 EHR data often cover hundreds of thousands of unique clinical features from both codified data
52 and unstructured clinical narrative notes. With the goal of analyzing these two types of data
53 simultaneously, the main challenges lie in combining the codified and unstructured data efficiently,
54 representing their covered clinical features meaningfully, and quantifying statistically the presence-
55 absence as well as the strength of relationships between different features.

56 The goal of combining codified and unstructured data arises from the fact that both contain
57 clinically relevant and inextricably linked health information. Together, these complementary data
58 sources capture a more complete picture of a patient’s medical history. The codified data, also
59 referred to as structured data, typically consists of diagnostic codes, procedure codes, medication
60 prescriptions, and laboratory orders and results. The utilization of codified data is straightforward;
61 data entry is standardized and in the necessary format for analysis. For example, diagnostic codes
62 have been used to predict the risk of heart failure [5], and procedure and medication codes have
63 been used to predict childhood obesity [6]. Conversely, the utilization of unstructured free-text data
64 in clinical notes is less direct [7]. This textual data covers a broad range of clinical concepts that
65 need to be extracted via natural language processing (NLP). These NLP concepts include diseases
66 and syndromes, clinical attributes and findings, clinical drugs, as well as laboratory, diagnostic,
67 and therapeutic procedures, which can provide complementary information to the structured data.
68 The NLP concepts are also referred to as clinical concept of unique identifiers (CUIs) in the Unified
69 Medical Language System (UMLS) [8].

70 Many studies have shown that incorporating this textual information into analyses can enhance
71 model performance by significant margins [9, 10]. In many cases, relevant information is only
72 documented in clinical notes and not well codified. For instance, spontaneous reporting databases
73 for adverse drug events are underreported when assessed using codified data only [11] since over 90%
74 of adverse drug events are not codified [12]. As a result, it is necessary to utilize unstructured EHR
75 data for active pharmacovigilance [13, 14]. Furthermore, NLP concepts are particularly valueable
76 for capturing drug side effects, as a significant proportion of these effects, such as symptoms, cannot
77 be adequately represented by diagnostic codes. For example, healthcare-associated infection (HAI),
78 a potentially lethal condition, is widely underreported in the codified data but can be detected and
79 even predicted using manual annotation in EHRs [15].

80 Combining codified and unstructured data also yields benefits for disease phenotyping. In the
81 United States, a diagnosis code is required by the healthcare provider during the evaluation for
82 a condition. Even if the patient is ultimately diagnosed with a different condition, the initial
83 diagnosis code will remain in the patient’s record and may be misleading if viewed in isolation [16].
84 It has been shown that prediction models that combine unstructured clinical notes with codified
85 data outperform models that utilize either unstructured or codified data alone [17, 18]. The utility
86 of this approach is highlighted in the case of geriatric syndromes, which are associated with high
87 morbidity, mortality, and healthcare utilization but are not fully represented by diagnostic codes
88 found in major coding standards. Many impairments associated with geriatric syndromes, such
89 as walking difficulty and weight loss, are not fully captured in codified fields. However, a study
90 [19] demonstrates that incorporating unstructured data can increase the sensitivity of identifying
91 individuals with geriatric syndromes. The supplementation of codified data with data extracted
92 using NLP can achieve more accurate and comprehensive assessments of patient health, thereby
93 reducing disease misclassification.

94 Given a large number of codified and NLP concepts, understanding their relatedness to each
95 other can improve the efficiency of downstream predictive modeling tasks. To generate prior knowl-
96 edge on the relationship among the clinical codes and NLP concepts, a potential solution is to con-
97 struct a large-scale clinical knowledge graph (KG) on these concepts [20, 21, 1]. Representing EHR
98 concepts with low-dimensional semantic embedding, KG embedding provides a quantitative glimpse
99 into the degree of inter-relatedness of medical entities. Once high-quality embeddings of medical
100 concepts are learned, they can improve the efficiency of downstream applications in biomedical and
101 healthcare research including information retrieval [22, 23, 24], cohort selection [25, 26], and risk
102 prediction [27, 28, 29].

103 In recent years, word embedding techniques [30, 31, 32] in NLP have been successfully applied for
104 representing clinical concepts in a low-dimensional space. Many of these embeddings were derived
105 for specific downstream tasks such as clustering [33] and prediction [34, 35, 3, 36]. While these
106 embedding methods can be used to assess the relatedness of NLP concepts, they do not naturally
107 generate a sparse KG that clearly indicates whether a link exists between entities. In addition, while
108 KG representation techniques have been successfully used to analyze biomedical data including
109 biomedical text and codified EHR concepts [21, 37, 38, 1, 39], joint representation of large-scale
110 codified and NLP EHR concepts is currently lacking, as summarized in Table 1. Recently, Bai [40]
111 proposed to jointly learn vector representations of medical concepts and words using MIMIC-III
112 data [41]. However, their work was limited in two ways. First, they did not represent words in
113 the clinical notes as CUIs, thus limiting the reproducibility of these representations. Second, the
114 MIMIC-III data only contains 58,597 in-patient visits, which confines the model performance and
115 cannot infer broader information for outpatients. As a result, their embeddings cannot be used to
116 generate high-quality knowledge graphs capturing general clinical information. To the best of our
117 knowledge, there is no existing work that derives comprehensive embeddings for both codes and
118 CUIs from a comprehensive EHR with both inpatient and outpatient data.

119 Generating KG with a large number of entities, however, is challenging for several reasons.
120 First, an efficient computational algorithm is needed to embed all concepts when both the number
121 of concepts and the number of EHR records are large. Second, no existing KG embedding methods
122 provide statistical certainty on whether a link exists between two entities. Most existing KG predicts
123 links via a supervised fashion by optimizing prediction tasks using the labeled links between entity
124 pairs, leveraging existing knowledge of such links. While such supervised approaches can be used to
125 assist in KG generation from EHR, it would require mapping EHR codes and narrative concepts to
126 existing entity pairs, which itself is a challenging task. In addition, these methods necessitate the use
127 of “negative samples”, which represent unlinked entity pairs. Unfortunately, this type of data is not
128 readily available. Relying on the complement of positive samples as negative samples is considered
129 unreliable, as indicated by previous research [42, 43]. These prediction-based approaches also do
130 not provide statistical uncertainty on the existence of the link between an entity pair. Equipping
131 the KG with certain quantification enables us to generate a sparse network while controlling for
132 the false discovery rate (FDR).

133 In summary, there is a great unmet need for an approach that can integrate and summarize
134 these high dimensional and large-scale clinical data into a KG for studies. In this paper, we
135 will address this need by proposing an Aggregated naRrative Codified Health (ARCH) records
136 analysis which is an efficient statistical algorithm that can generate KG embeddings along with
137 uncertainty measures on the links. With pairwise co-occurrence counts of all EHR concepts and
138 a few simple summary statistics, the ARCH algorithm generates low-dimensional embeddings for
139 each concept and performs large-scale hypothesis testing based on the cosine similarity between
140 these embedding vectors. The connectivity of entity pairs is assessed jointly by controlling for a
141 target FDR. We validate the clinical utility of the ARCH KG, generated from EHR data from the

Method	Type	Number of concepts
Choi et al. [44] (claims)	CUI	15,905
Finlayson et al. [21] (notes)	CUI	28,394
De Vine et al. [24] (MedTrack)	CUI	59,266
Code2Vec [33]	Code	8,477
Med2Vec [34]	Code	28,840
KESER [1]	Code	14,718
MIKGI [45]	Code	13,261
ARCH	CUI&Code	51,423 CUIs + 9,586 Codes

Table 1: A summary of existing EHR-derived medical embeddings.

142 Veterans Affairs, along with semantic embeddings through downstream tasks including detecting
 143 known relationships between entity pairs, predicting drug side effects, disease phenotyping, as well
 144 as sub-typing Alzheimer’s disease (AD) patients.

145 2 Methods

146 2.1 Generative model for the knowledge graph

147 Suppose there are a total of d EHR codified and NLP concepts, indexed by $\mathcal{V} = \{1, \dots, d\}$. The
 148 semantic meaning of each concept is represented by a p -dimensional embedding vector \mathbf{V}_j for
 149 $j = 1, \dots, d$. These embeddings are generated from a latent Gaussian graphical model [46]: each
 150 column of $\mathbf{V} = (\mathbf{V}_1, \dots, \mathbf{V}_d)^\top \in \mathbb{R}^{d \times p}$ is independent and identically distributed from $N(0, \Theta^{-1})$
 151 where the precision Θ embeds the conditional dependency network of the d concepts, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$,
 152 with the vertex set \mathcal{V} representing all EHR concepts and the edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ characterizing the
 153 conditional dependency between the concepts. Our goal is to learn the KG \mathcal{G} with \mathcal{E} characterized
 154 by Θ in that $(j, k) \in \mathcal{E}$ if and only if $\Theta_{jk} \neq 0$ or equivalently \mathbf{V}_j is conditionally dependent on
 155 \mathbf{V}_k given all remaining embeddings. We aim to identify \mathcal{E} through testing the set of hypotheses
 156 $\mathbb{H} = \{H_{0,jk}, (j, k) \in \mathcal{V} \times \mathcal{V}\}$:

$$H_{0,jk} : (j, k) \notin \mathcal{E} \text{ v.s. } H_{1,jk} : (j, k) \in \mathcal{E}, \text{ for all } (j, k) \in \mathcal{V} \times \mathcal{V}. \quad (1)$$

To learn the representations \mathbf{V} and test \mathbb{H} , we assume that the observed clinical concepts in the EHR are generated from a latent Markov process driven by the embeddings sampled from the graphical model [47]. In specific, let w_t be the concept at time t and the occurrence probability of concept j is modeled by

$$\mathbb{P}(w_t = j \mid \mathbf{c}_t, \mathbf{V}) = \frac{\exp(\langle \mathbf{V}_j, \mathbf{c}_t \rangle)}{\sum_{k=1}^d \exp(\langle \mathbf{V}_k, \mathbf{c}_t \rangle)},$$

where the latent discourse vector \mathbf{c}_t represents the embedding of the topic at time t and is generated from an autoregressive (AR) model

$$\mathbf{c}_1 \sim N(0, \mathbf{I}_p/p) \text{ and } \mathbf{c}_t = \sqrt{\alpha} \mathbf{c}_{t-1} + \sqrt{1-\alpha} \boldsymbol{\epsilon}_t, \text{ with } \boldsymbol{\epsilon}_t \stackrel{\text{i.i.d.}}{\sim} N(0, \mathbf{I}_p/p) \text{ for } t \geq 2,$$

where $0 < \alpha < 1$ is the weight parameter. Figure 1 illustrates the generation process. The \mathbf{c}_t represents the latent topic vector at each time (e.g., phenotype, treatment, lab measurement, etc). For example, in the model part of Figure 1, \mathbf{c}_t is related to phenotype, and the probability of the concept “Alzheimer’s Disease” occurring at time t is larger as its embedding is closer to \mathbf{c}_t . At the time

$t+1$, \mathbf{c}_{t+1} becomes topic related to medicine and thus the concept of “Memantine” has larger occurrence probability. Under this model, the embedding inner product $\sigma_{jk} = \langle \mathbf{V}_j, \mathbf{V}_k \rangle \equiv \sum_{k=1}^p V_{jl}V_{kl}/p$ can be approximated by the population positive point-wise mutual information (PPMI) between concept j and k [48]:

$$\sigma_{jk} = \text{PPMI}(j, k) + O(1/d),$$

157 where $\text{PPMI}(j, k) = \max \left\{ 0, \log \frac{\mathbb{P}(j, k)}{\mathbb{P}(j)\mathbb{P}(k)} \right\}$, $\mathbb{P}(j, k)$ is the co-occurrence probability of the concept
 158 pair (j, k) and $\mathbb{P}(j)$ is the occurrence probability of the concept j . Therefore, when the number of
 159 concepts d has a larger order than the square root of the sample size used to estimate the PPMI,
 160 testing $H_{0,jk}$ can be achieved by testing $\text{PPMI}(j, k) = 0$ based on the estimated PPMI.

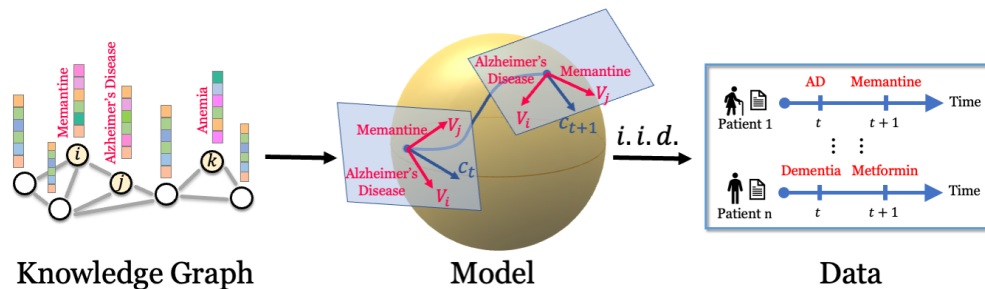


Figure 1: Data generation process of the EHR occurrence data. The embeddings of concepts are generated from a graphical model and the occurrence is then driven by a Markov process.

161 2.2 ARCH representation learning and graph recovery

162 For large-scale EHR datasets with a massive number of concepts and patient records, it is both
 163 statistically and computationally challenging to infer the network due to the latency and the large
 164 number of hypotheses involved. Our ARCH representation learning approach carries out the in-
 165 ference in two steps by (i) first screening for \mathcal{E} by identifying marginally dependent concept pairs
 166 with nonzero pointwise mutual information, and (ii) inferring about the Gaussian graphical model
 167 structure of Θ via sparse regression. In the first step of screening, we apply the SURE screening
 168 [49] by selecting pairs (j, k) with $\sigma_{jk} \neq 0$ after controlling for a desired FDR. In the second step,
 169 we further infer the edges from the network G via node-wise regression [50]. As the embedding
 170 vectors follow the Gaussian graphical model, the conditional distribution of embeddings is

$$\mathbf{V}_j \mid \{\mathbf{V}_k\}_{k \in \mathcal{C}_j} \sim N \left(- \sum_{k \in \mathcal{C}_j} \mathbf{V}_k (\Theta_{jj}^{-1} \Theta_{jk}), \Theta_{jj}^{-1} \mathbf{I}_p \right), \quad (2)$$

171 where \mathcal{C}_j is the set of concepts related to concept j obtained from the first prescreening step.

172 2.3 Pre-screening by PMI testing

To form a test statistic for $H_{0,jk} : \sigma_{jk} = 0$ and estimate \mathbf{V} , we first calculated the empirical PPMI
 as $\text{PPMI} = [\text{PPMI}(j, k)]$, with $\text{PPMI}(j, k) = \max \left\{ 0, \log \frac{\mathcal{C}(j,k)\mathcal{C}(\cdot,\cdot)}{\mathcal{C}(j,\cdot)\mathcal{C}(\cdot,k)} \right\}$, where $\mathcal{C}(j, \cdot)$ is the row sum
 of co-occurrence matrix $\mathcal{C}(j, k)$, and $\mathcal{C}(\cdot, \cdot)$ is the total sum of the co-occurrence. Details for the
 construction of $\mathcal{C}(\cdot, \cdot)$ is given in Section 3.1. We next took an SVD of the empirical PPMI matrix

as $\mathbb{P}\mathbb{P}\mathbb{M}\mathbb{I} = [\mathbb{P}\mathbb{P}\mathbb{M}\mathbb{I}(j, k)] = \mathbb{U}\text{diag}(\Lambda_1, \dots, \Lambda_d)\mathbb{U}^\top$, we can estimate \mathbf{V} and population PPMI matrix of d concepts respectively as

$$\begin{aligned}\tilde{\mathbf{V}} &= (\tilde{\mathbf{V}}_1^\top, \dots, \tilde{\mathbf{V}}_d^\top)^\top = \mathbb{U}^{(p)}\text{diag}(\Lambda_1^{\frac{1}{2}}, \dots, \Lambda_p^{\frac{1}{2}}), \\ \widehat{\mathbb{P}\mathbb{P}\mathbb{M}\mathbb{I}} &= \tilde{\mathbf{V}}\tilde{\mathbf{V}}^\top = \mathbb{U}^{(p)}\text{diag}(\Lambda_1, \dots, \Lambda_p)(\mathbb{U}^{(p)})^\top,\end{aligned}$$

173 where $\mathbb{U}^{(p)}$ being the first p singular vectors of $\mathbb{P}\mathbb{P}\mathbb{M}\mathbb{I}$ with positive eigenvalues. The dimension p
174 can be selected to optimize embedding quality similar to KESER [1] by maximizing the area under
175 the Receiver Operating Characteristics curve (AUC) of distinguishing those known relation pairs
176 from random pairs, where known relation pairs are curated from online sources, detailed in the
177 validation studies in Section 3.2.1.

178 The estimator $\widehat{\mathbb{P}\mathbb{P}\mathbb{M}\mathbb{I}}$ is close to the population PPMI matrix with a high approximation rate
179 and asymptotically normal, which allows us to approximate σ_{jk} with $\hat{\sigma}_{jk} = \tilde{\mathbf{V}}_j^\top \tilde{\mathbf{V}}_k$. Furthermore, we
180 may form test statistic $z_{jk} = \hat{\sigma}_{jk}/\hat{s}_{jk}$ to identify $\sigma_{jk} \neq 0$ since z_{jk} follows approximately standard
181 normal distribution under the null hypothesis [48], where \hat{s}_{jk} is an estimated standard error for
182 $\hat{\sigma}_{jk}$ detailed in Appendix S.1. To control for multiple comparisons, we performed the Benjamini-
183 Hochberg (BH) procedure under dependence and identified related concept pairs with z_{jk} higher
184 than a BH-controlled threshold as detailed in Appendix S.2.

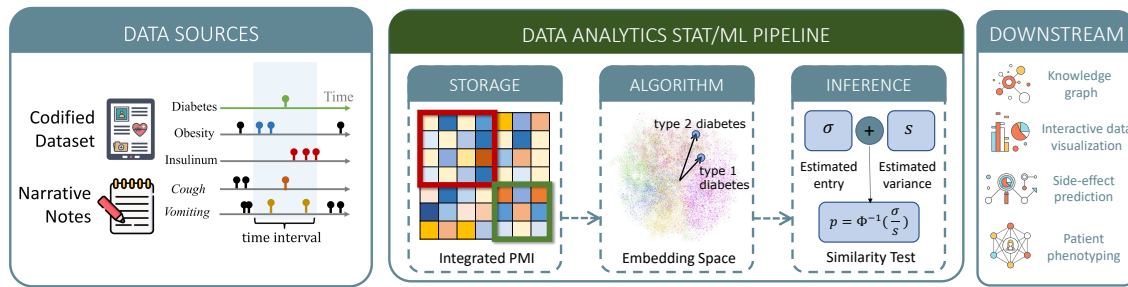


Figure 2: Data source, including codified data and narrative notes, and data analytics pipeline.

185 2.4 Sparse embedding regression

The FDR controlled testing procedure based on $\hat{\sigma}_{jk}$ could serve as a prescreening of related concepts from the large number of concept pairs. To further screen for the most relevant concepts to form $\mathcal{E}_j = \{k : \Theta_{jk} \neq 0\}$, we further performed a sparse regression of $\tilde{\mathbf{V}}_j$ against all embedding vectors identified as related to the concept j after initial screening, denoted by \mathbb{C}_j , to recover $\Theta_j, \Theta_{jj}^{-1}$ and hence its associated graph structure. Due to the potentially large number of elements identified in the pre-screening stage, we adopted an adaptive elastic-net penalized regression

$$\hat{\theta}_j(\gamma, \lambda) = \arg \min_{\theta} \left[\|\tilde{\mathbf{V}}_j - \tilde{\mathbf{V}}^\top[\mathbb{C}_j]\theta\|_2^2 + \lambda \sum_{l \in \mathbb{C}_j} \left\{ \frac{1-\gamma}{2} \left(\theta_l \frac{\hat{\sigma}_{jj}\sqrt{\hat{\sigma}_{ll}}}{\hat{\sigma}_{jl}} \right)^2 + \gamma \left| \theta_l \frac{\hat{\sigma}_{jj}\sqrt{\hat{\sigma}_{ll}}}{\hat{\sigma}_{jl}} \right| \right\} \right],$$

186 where $\tilde{\mathbf{V}}^\top[\mathbb{C}_j]$ is the submatrix of $\tilde{\mathbf{V}}$ corresponding to \mathbb{C}_j . The tuning parameters λ and γ control
187 the support of $\hat{\theta}_j$ and hence the network structure. We determined the optimal values for the
188 hyperparameters λ and γ for each target concept j by performing a grid search to balance the
189 external and internal validation losses. Specifically, we computed the average of the internal Akaike
190 information criterion (AIC) loss and an external validation loss, which was obtained using an
191 independent dataset $\tilde{\mathbf{V}}^*$, as detailed in Appendix S.3.

192 3 Validation of Real World EHR Trained ARCH Knowledge Graph

193 3.1 EHR data sources and preprocessing

194 We trained a large-scale ARCH KG using EHR data from the Veterans Affairs (VA) Corporate
195 Data Warehouse (CDW), integrating both codified and narrative data from 12.6 million patients
196 with at least 1 visit between 2000-2019. We gathered four domains of codified data including ICD
197 diagnosis codes, procedures, lab tests, and medication prescriptions. All raw codes are rolled up to
198 higher level codified concepts: ICD codes were aggregated into PheCodes using the ICD-to-PheCode
199 mapping from PheWAS catalog (<https://phewascatalog.org/phecodes>); procedure codes, including
200 CPT-4, HCPCS, ICD-9-PCS, ICD-10-PCS, were grouped into clinical classification software (CCS)
201 categories based on the CCS mapping*; laboratory codes were either mapped to LOINC codes
202 (<https://loinc.org/>) or manually annotated lab concepts; and medication codes were mapped to
203 RXNORM codes. All free text clinical notes were processed with the Narrative Information Linear
204 Extraction (NILE) NLP software [51], which maps clinical terms to CUIs in the UMLS. All codified
205 and NLP data were organized as triplets: (Patient id, date, concept). Using these processed data,
206 we created a co-occurrence matrix for all concept pairs by counting the number of co-occurrences
207 within a 30-day window across all patients. To reduce noise, we removed concepts that have less
208 than 3000 occurrences and concept pairs that have less than 1000 co-occurrences. Furthermore,
209 we removed all concepts that co-occur with more than 95% of other concepts as they tend to be
210 overly non-specific. This results in a total of over 61,000 concepts, out of which 51,423 are CUIs
211 and 9,586 are codified concepts.

212 3.2 Validation analyses

213 The ARCH KG was validated in four downstream tasks: (1) detecting known similar or related
214 clinical concepts; (2) detecting drug side effects; (3) disease phenotyping; and (4) profiling of pa-
215 tients with AD. For the detection of known relationships and drug side effects, we also compared
216 to embedding vectors from pretrained language model (PLM) embeddings based on Bidirectional
217 Encoder Representations from Transformer (BERT) [52], including Self-aligning pretrained BERT
218 (SAPBERT) [53], BERT for Biomedical Text Mining (BioBERT) [54], and BERT pretrained with
219 PubMed (PubmedBERT) [55]. BERT’s model architecture is a multi-layer bidirectional Trans-
220 former encoder, while BioBERT, PubMedBERT and SAPBERT are pretrained on different sources
221 based on BERT. BioBERT is pretrained on both general domain corpora and biomedical domain
222 corpora (PubMed abstracts and PMC full-text articles), PubMedBERT is pretrained purely with
223 in-domain text (PubMed text), and SAPBERT is pretrained on the biomedical KG of UMLS. The
224 language model based embeddings were obtained only based on the description of the EHR concepts
225 (e.g. preferred term for the CUI and code description).

226 3.2.1 Detecting known relationship pairs

227 We curated different categories of known relation pairs from online knowledge sources including
228 similar pairs and related pairs. Similar pairs of codified concepts were largely created based on
229 code hierarchies including the PheCode hierarchy. Since a majority of laboratory codes in the VA
230 are not mapped to LOINC codes, we augmented the LOINC hierarchy with manually annotated
231 similar pairs when assessing similar laboratory code pairs. Similar CUI pairs are extracted from
232 the relationship from the UMLS. We additionally evaluated the similarity between mapped CUI

*https://www.hcup-us.ahrq.gov/toolsoftware/ccs_svcsproc/ccssvcproc.jsp

↔ code pairs. We leveraged UMLS to obtain the mapping from different medical coding systems to concept unique identifiers [56]. For the related pairs, we first considered CUI-CUI pairs and used several categories of relationships given in the UMLS, including “may treat or may prevent”, “classifies”, “differential diagnosis”, “method of” and “causative”. For these CUI pairs, we map the disorder CUIs to PheCodes, drugs to the RxNorm, and procedures to CCS categories. These mapped code pairs are then further used to assess the ability to detect relatedness using codified data.

For each type of relationship, we calculated the cosine similarities of the embedding vectors of related pairs and those of randomly selected pairs to calculate AUC of the cosine similarities in distinguishing known pairs from random pairs. The random pairs were selected to match the semantic types of the related pairs. For example, when assessing “may treat or may prevent”, we restricted to disease-drug pairs. To reduce the noise of real data, we removed the features that have a pretty low frequency. Finally, we chose the dimension of embedding by optimizing the AUC. We performed ARCH testing procedure to determine whether a pair of entities are related with FDR chosen at 1%, 5%, and 10%, and reported the power of the ARCH procedure. Since no existing procedures are able to control FDR, we calculated the power of other algorithms in detecting known relationships by ranking entity pairs according to cosine similarity generated from their corresponding embeddings and then selecting the top M entity pairs as significant, where M is the number of entity pairs selected by ARCH. Among those M pairs, we calculated the proportion of those known to be related as their power.

3.2.2 Detecting drug side effect

The unintended effects or adverse events (AEs) of drugs threaten public health and patient safety [57]. However, the screening for and adjudication of AEs is costly and time-intensive and post-market drug retraction is expensive [58]. It is thus critical to predict the potential AEs of drugs prior to their widespread use. The ARCH KG provides semantic representations for both drugs and side effects, which can be subsequently modeled to identify potential side effects for a given drug. ARCH network includes both narrative and codified features, which can improve our ability to detect side effects that tend to be under-codified in the EHR. To develop and validate a side effect prediction model based on ARCH embeddings, we obtained labels from the Side Effect Resource (SIDER)[†] database of drugs and adverse drug reactions (ADRs) [59]. The SIDER database captures side-effect information from multiple data sources including placebo-controlled clinical trials, the FDA Adverse Event Reporting System (AERS), and biomedical literature. We followed the data cleaning procedure outlined in multimodal representation learning [60] and selected common AEs reported in more than 50 drugs. The AEs were mapped to both PheCodes and CUIs while the drugs, recorded as DrugBankID in SIDER, were mapped to RxNorm codes and CUIs. Following these steps using the VA data, we obtained 831 drugs and corresponding 4,010 AEs, which compose in total 128,220 drug-AE pairs. Similar to relation detection, we randomly sampled the same number of negative pairs from those drug-disease entity pairs that have not been reported as drug-AE pairs.

The AUC and power for detecting drug side effects based on ARCH embeddings or p -values as well as based on embeddings from existing language models were calculated similarly as those for the relation detection. Since the drug-AE pairs can exist in four forms: RxNorm-CUI pairs, CUI-CUI pairs, RxNorm-PheCode pairs, and CUI-PheCode pairs, we took the highest score among these four relationship pairs to represent the final score for each drug-AE pair. We also compared the score that uses all four forms of data to the score based on codified data only, i.e. RxNorm-PheCode

[†]<http://sideeffects.embl.de>

277 pairs, with respect to their power in detecting the drug-AE pairs. Since this KG representation
278 can be viewed as a pre-training step that can be further fine-tuned for the task of AE detection, we
279 further evaluated the quality of ARCH embeddings as well as embeddings from existing language
280 models based on the performance of a few-shot supervised model for this task. The fine-tuning
281 step employed a commonly used loss function [61] as detailed in Appendix S.5. We used 1% of the
282 positive and negative pairs to estimate model parameters, another 1% as validation data to select
283 optimal tuning parameters, and the remaining 98% pairs as a test data set for evaluation.

284 3.2.3 Disease phenotyping

285 A major bottleneck for conducting translational research studies with EHR is the lack of large-scale
286 precise data on disease outcomes needed for predictive modeling. For most conditions, ICD codes
287 do not accurately reflect the true disease status while manual annotation via chart review is not
288 scalable [62]. Recently, many unsupervised machine learning based phenotyping algorithms have
289 been shown to greatly improve the case definition over ICD codes [63, 64, 65, 62, 66]. However, most
290 of these algorithms require the specification of relevant features. Given the large number of potential
291 EHR features, automatically selecting features important for a disease of interest is an important
292 step to ensure the accuracy of the downstream modeling. We next illustrate how the ARCH network
293 can serve as an effective feature selection tool for EHR phenotyping and compare to the existing
294 KG based feature selection tool, KESER [1], which only identifies codified features. To compare
295 the performance of ARCH versus KESER, we employed the unsupervised PheNorm algorithm [65].
296 PheNorm can be viewed as weakly supervised in that it treats the counts of the PheCode and/or
297 CUI corresponding to the disease as “silver standard labels” to train an algorithm that combines
298 these key features with additional informative features including a measure of healthcare utilization
299 via drop-out training and mixture modeling. We compared PheNorm trained with ARCH selected
300 features, PheNorm trained with only KESER selected features, the MAP algorithm which only uses
301 counts of the main PheCode and CUI, and healthcare utilization [62], as well as two benchmark
302 methods that use the logarithm of the count of the main disease ICD code plus one (Main ICD
303 Only) and the logarithm of the count of the mention of the disease CUI plus one (Main NLP
304 Only) as the disease predictive scores, respectively. Since KESER only includes codified features
305 and MAP only uses the three key features, these comparisons also illustrate the value of other
306 informative features, particularly NLP features from free text, in improving the accuracy of the
307 algorithm. We trained these phenotyping algorithms using EHR data from 53, 549 MGB Biobank
308 participants for 8 conditions: coronary artery disease (CAD), Crohn’s disease (CD), rheumatoid
309 arthritis (RA), ulcerative colitis (UC), Congestive heart failure (CHF), type 1 diabetes mellitus
310 (T1DM), type 2 diabetes mellitus (T2DM) and depression. To evaluate their accuracy, the CAD,
311 CD, RA, UC, CHF, T2DM, T2DM and depression phenotyping algorithms were validated against
312 187, 138, 154, 127, 114, 540, 285 and 540 labeled observations curated via manual chart review,
313 and the AUCs were reported.

314 3.2.4 Profiling of AD patient via ARCH embeddings

315 Semantic representation of the EHR concepts can be linked with patient level EHR data to represent
316 patient clinical profile [67, 68, 69]. These patient embeddings can then be applied to perform
317 downstream tasks such as identifying “*patient like me*” [70] and mortality prediction [71]. However,
318 representing a patient’s clinical profile with respect to a specific condition, such as AD, requires the
319 knowledge of other EHR features relevant to AD progression as well as their relative importance
320 [72]. Our ARCH KG serves this purpose in that it can generate embeddings to represent an AD

321 patient. To demonstrate this, we used EHR data of 38,267 patients with AD diagnosis, collected
 322 from the University of Pittsburgh Medical Center (UPMC) over the period 2011-2021. We selected
 323 the AD relevant features and generate embeddings for the i th patient using the following term
 324 frequency-inverse document frequency (TF-IDF) procedure:

$$\mathbf{W}_i = \sum_{c \in \mathcal{V}_{AD}} \log\left(\frac{a_{ic}}{T_i} + 1\right) / \log(b_c + 1) \cdot \frac{\hat{\sigma}_{c,AD}}{\hat{\sigma}_{c,c} \sqrt{\hat{\sigma}_{AD,AD}}} \tilde{\mathbf{V}}_c, \quad (3)$$

325 where \mathcal{V}_{AD} is the feature set related to AD detected by ARCH, T_i is the follow-up time of the
 326 i th patient, $\tilde{\mathbf{V}}_c$ is the estimator of word representation for concept c , a_{ic} is the occurrence of the
 327 feature c in the EHR of the i th patient, b_c is the occurrence of feature c in all patients from VA
 328 between 2000-2019. Together, the PMI testing procedure and clinical embeddings can help us to
 329 generate patient embeddings that present phenotyping. As an illustration, we applied k -means
 330 algorithm to cluster patients into two groups using the patient embeddings. We analyzed the
 331 mortality risk of the two groups using the Kaplan-Meier (KM) curve of the time from first AD
 332 diagnosis to death. We characterized the between group differences in patient profile with respect
 333 to the distributions of AD related features selected via ARCH. For each AD related feature within
 334 each group, we compute its average intensity defined as the concept count normalized by total
 335 feature count within each patient. We summarize the group difference in patient profile based on
 336 the between-group differences in feature intensity.

337 4 Results

338 By optimizing the AUC of distinguishing known relation pairs from random pairs as detailed in
 339 Section 3.2.1, we set the dimension of embeddings as $r = 1500$ to optimize the embedding quality.
 340 We worked with 1500-dimensional embeddings on the following tasks.

341 4.1 Detecting known relationship pairs

	FDR	type	ARCH(p)	ARCH(c)	Pub	Bio	SAP
AUC		similar	0.873	0.871	0.670	0.589	0.735
		related	0.832	0.836	0.649	0.583	0.642
Power	0.1	similar	0.909	0.902	0.677	0.548	0.741
		related	0.892	0.884	0.701	0.594	0.672
	0.05	similar	0.906	0.898	0.668	0.536	0.733
		related	0.888	0.880	0.691	0.583	0.659
	0.01	similar	0.900	0.892	0.646	0.514	0.715
		related	0.880	0.871	0.670	0.559	0.638

Table 2: AUCs and power of detecting known similar pairs and related pairs with different algorithms with various target FDRs. Pub stands for PubmedBERT, Bio stands for BioBERT and SAP stands for SAPBERT.

342 The AUCs and power in detecting known relationships are summarized in Table 2 with details on
 343 the accuracy of detecting specific types of relationships given in Table 5 in Appendix S.6. The
 344 embeddings trained by ARCH achieved an AUC of 0.871 for detecting similar pairs and 0.836 for
 345 detecting related pairs, while pretrained language model derived embeddings including Pubmed-
 346 BERT, BioBERT and SAPBERT attained much lower AUCs ranging from 0.583 to 0.735. The

347 ARCH screening procedure attained power of 0.909 for similar pairs 0.892 for related pairs un-
348 der the target FDR 0.1, while the highest power among the three benchmarks was only 0.74 for
349 similar pairs and 0.70 for related pairs. Visualizations of the ARCH network can be found at
350 <https://celehs.hms.harvard.edu/ARCH/>, which enables users to visualize concepts relevant to a set
351 of target concepts.

352 4.2 Identifying drug side effects

	Method	FDR	ARCH(p)	ARCH(c)	Pub	Bio	SAP
AUC	Unsupervised		0.747	0.723	0.634	0.584	0.587
	Supervised		NA	0.826	0.651	0.657	0.686
Power	Unsupervised	0.1	0.522	0.374	0.277	0.199	0.235
		0.05	0.513	0.365	0.268	0.192	0.225
		0.01	0.493	0.346	0.250	0.179	0.209
	Supervised	0.1	NA	0.580	0.356	0.260	0.377
		0.05	NA	0.572	0.345	0.252	0.367
		0.01	NA	0.557	0.325	0.236	0.346

Table 3: AUCs and the sensitivities of different benchmark methods compared with ARCH for identifying drug side effects. The first and the third blocks show the performance of each method without supervision, while the second and the fourth blocks show the performance of the method with supervised learning using 1% drug-side effects pairs for training. Pub stands for PubmedBERT, Bio stands for BioBERT and SAP stands for SAPBERT.

353 Table 3 shows the AUC-ROC and power of ARCH embeddings, the pre-trained language model
354 embeddings, as well as the p -values from ARCH screening testing procedures in detecting drug side
355 effects. The unsupervised ARCH embeddings and the screening test p -values achieved substantially
356 a higher AUC of 0.723 and 0.747, compared to those from PLM which ranged from 0.584 to 0.634.
357 With few-shot supervised training, the ARCH embeddings attained an AUC of 0.826 while the
358 AUC of the fine-tuned PLM embeddings remained below 0.69. Comparing the power in detecting
359 drug side effects using codified data alone versus both codified and NLP data, we find that adding
360 NLP information greatly improved the ability to capture side effects for most drug classes as shown
361 in Figure 3. In Figure 4, we show most of the side effects of Levothyroxine and Hydrocodone can
362 be detected by ARCH while a significant fraction of the side effects can only be captured with the
363 help of NLP data. More examples of word-cloud figures are shown in Figure 10 in Appendix S.6.

364 4.3 Disease phenotyping

365 Figure 5 shows the AUCs of 8 phenotyping algorithms validated on labeled data from MGB.
366 PheNorm with ARCH selected features performs the best among all methods. The AUCs of the
367 PheNorm algorithms with features selected by ARCH exceeded 0.9 for all 8 diseases and on average
368 were 0.028 (p -value 3.30×10^{-5}), 0.067 (p -value 9.87×10^{-12}), 0.081 (p -value 3.29×10^{-11}), and
369 0.076 (p -value 1.06×10^{-11}) higher than that of PheNorm with KESER features, MAP, ICD only
370 and NLP only. The gain in performance is particularly noteworthy for conditions that benefit from
371 NLP features. For example, after applying ARCH in the feature selection step, the AUC of the
372 PheNorm algorithm for depression increased from 0.857 of KESER to 0.927 (p -value 2.47×10^{-4}).

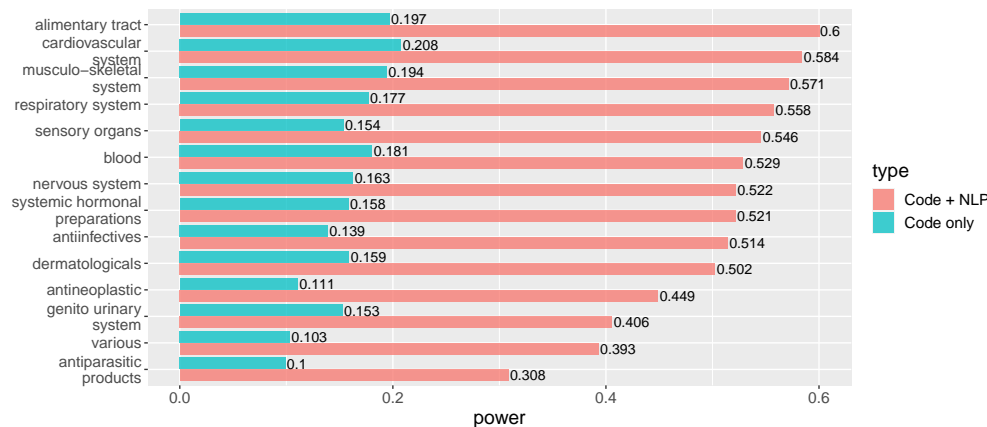


Figure 3: Sensitivity of detecting drug-side effects pairs with only codified data and that with both codified data and NLP with ARCH under target FDR 0.05.

373 4.4 Profiling of AD patient via ARCH embeddings

374 The AD cohort consists of about 64.7% female patients, 90.3% white and 7.6% black patients,
 375 with an average age of 82 years at first ICD code for AD and an average lifespan of 86 years.
 376 K-means clustering of the ARCH-based patient embeddings as detailed in Section 3.2.4 resulted in
 377 two subgroups: a fast progression group consisting of 12.3% the patients and a slow progression
 378 group formed by the remaining patients. As shown in Figure 6, the 5-year survival rate was 42.0%
 379 (95% CI: [38.6%, 45.7%]) and 80.9% (95% CI: [80.3%, 81.6%]) for the fast and slow progression
 380 groups, respectively.

381 Figure 7 highlights the top disease and drug features with the largest differences between the fast
 382 and slow progression groups. The phenotype features associated with faster progression are common
 383 phenotypes at the late stage of AD. Pneumonia is one of the two most serious medical conditions
 384 seen in late-stage AD patients [73]; hypovolemia and hypernatremia may be found in association
 385 with dehydration, which can occur in impaired late-stage AD patients who are dependent on others
 386 for fluid intake [74, 75, 76]. On the other hand, the features that appear more frequently in the
 387 slow progression group of patients, which are colored blue in the figure, are either common signs
 388 or possible causes of AD. Memory deficits begin from the early stage of AD [77], while vitamin
 389 deficiency and hypothyroidism are risk factors for AD [78, 79, 80]. As shown in the network
 390 of drug features and procedure features, the features ‘atorvastatin’, ‘metformin’, ‘escitalopram’,
 391 ‘melatonin’, among others, have been shown to moderate AD or slow down the progression of
 392 cognitive impairment in AD patients [81, 82, 83, 84]. Memantine, a type of N-methyl-D-aspartate
 393 receptor antagonist, is the only drug approved for use in moderate to severe AD under current AD
 394 treatment guideline [85, 86]; Rivastigmine and Donepezil are the drugs approved by FDA (Food
 395 and Drug Administration) for AD treatment besides Memantine and two accelerated approval
 396 drugs[‡]; all these three drugs are more common in the fast progression group of patients. With
 397 these references, the clustering of patients is practical and realistic, indicating the good quality of
 398 patient embedding based on the feature selection by ARCH.

[‡]<https://stanfordhealthcare.org/medical-conditions/brain-and-nerves/alzheimers-disease/treatments/medications.html>

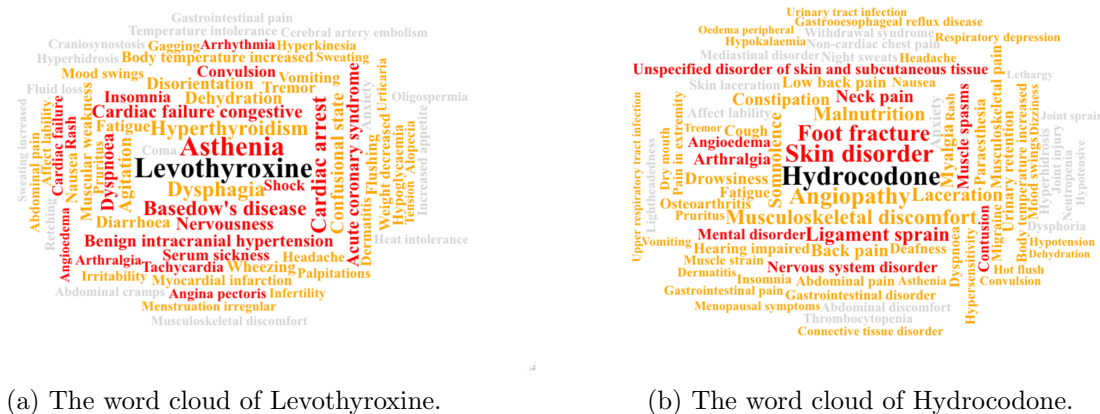


Figure 4: The word clouds of the side effects of two sample drugs - (a) Levothyroxine on the left and (b) Hydrocodone on the right. The surrounding words describe side effects. The words colored red are detected using codified data only while the words colored orange or red are detected by using both codified data and NLP codes. The words colored by grey are undetected. The size of the words is determined by the cosine similarity with the target drug code.

5 Discussion

Utilizing summary-level EHR data, the ARCH KG learning approach provides a highly scalable method for effectively representing codified and narrative EHR concepts on a large scale, while also recovering their network structure. The VA EHR-derived ARCH embeddings represent the first large-scale EHR embeddings to include both codified and NLP concepts, with the incorporation of NLP concepts proving particularly beneficial in real-world applications such as drug side effect monitoring and disease phenotyping. Additionally, the network structure derived from ARCH is constructed with a statistically guaranteed false discovery rate.

The versatility of the learned ARCH embeddings makes them ideal for a broad range of downstream tasks. These embeddings demonstrate greater robustness than existing PLM-based embeddings. Our semantic representation evaluations and drug side effect prediction studies show that

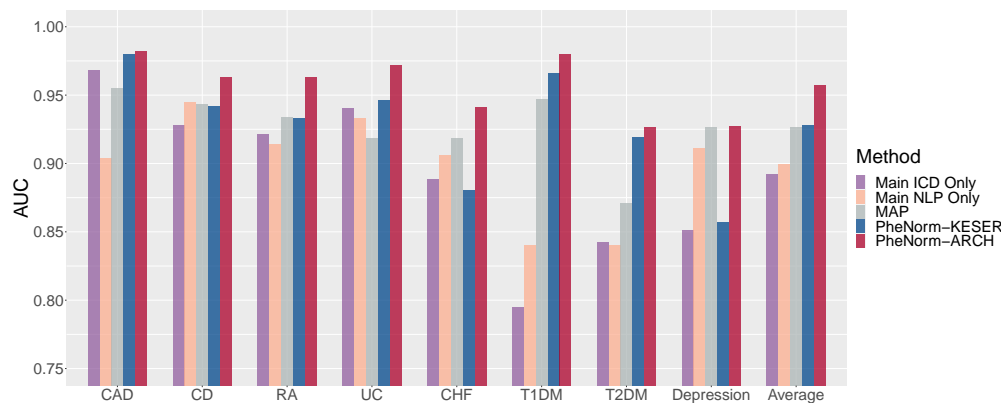


Figure 5: The AUC of different phenotyping algorithms trained with different feature sets across 8 diseases.

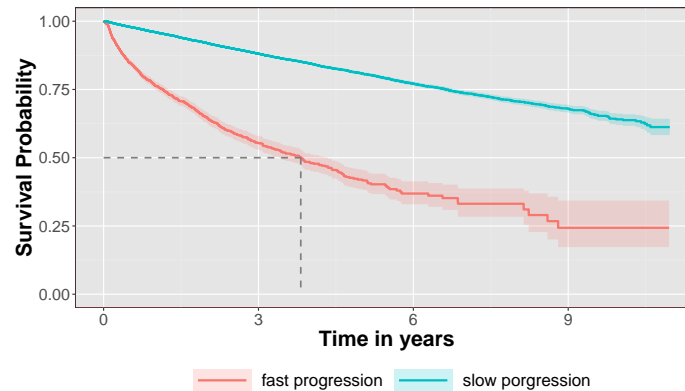


Figure 6: The KM survival curves for the fast and slow progression groups identified via k -means clustering of the ARCH patient level embeddings.



Figure 7: The word cloud of (a) phenotype features; and (b) drug features that drive the differences between the two subgroups. The size of the feature is determined by the between-group difference in the average intensity of such a feature. Red-colored features represent higher average intensity in the fast progression group and blue-colored features represent higher intensity in the slow progression group.

410 the ARCH embeddings can effectively capture the semantic relationships between EHR entities and
 411 concepts. Our results indicate that the ARCH embeddings with few shot training have the poten-
 412 tial to achieve high accuracy in KG-related tasks, such as entity matching and relation extraction.
 413 Additionally, the ARCH embeddings can serve as pre-trained representations of EHR concepts
 414 that can be linked to individual-level EHR data, further improving patient-level prediction tasks,
 415 as demonstrated in the AD patient profiling study. Joint representations of both codified and NLP
 416 data also enable more comprehensive multi-modal modeling of EHR data, significantly enhancing
 417 prediction performance for outcomes that require predictors that are not well-coded.

418 The use of summary-level data in learning the ARCH network creates an opportunity for col-
 419 laborative training of knowledge graphs across multiple institutions. This approach can enhance
 420 the quality of the trained representation and improve the portability of downstream prediction al-
 421 gorithms. However, co-training ARCH embeddings using multi-institutional data faces a challenge
 422 in dealing with coding differences between institutions. Even for institutions that have mapped
 423 their local EHR codes to a common ontology, such mappings are often incomplete. Future research
 424 needs to explore co-training knowledge graphs for overlapping yet non-identical EHR concepts from
 425 multiple institutions based on summary-level data.

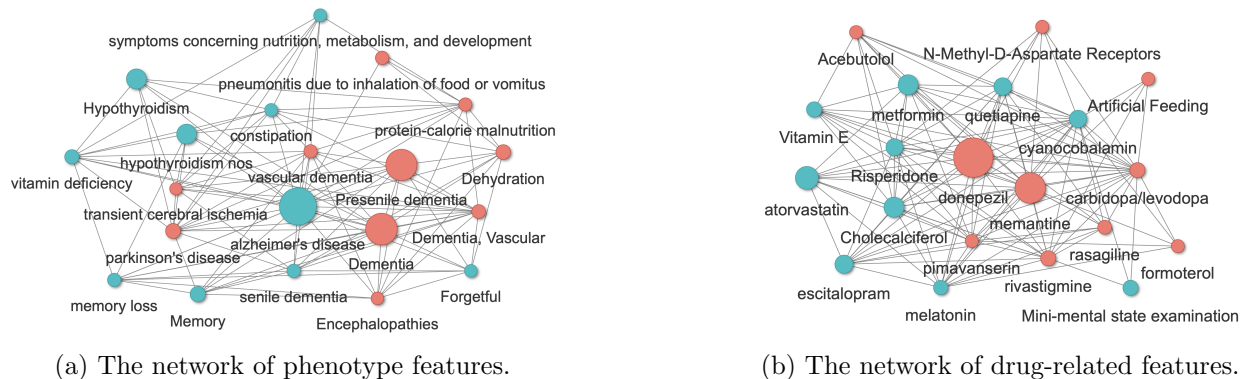


Figure 8: The network of (a) phenotype features; and (b) drug features that drive the differences between the two subgroups. The size of the feature is determined by the between-group difference in the average intensity of such a feature. Red-colored features represent higher average intensity in the fast progression group and blue-colored features represent higher intensity in the slow progression group.

426 Currently, the ARCH network relies solely on EHR occurrence patterns of concepts, disregarding
427 valuable information contained in their descriptions. Incorporating both occurrence patterns and
428 descriptions through language models is an intriguing avenue for further research in improving the
429 network.

430 6 Data Availability

431 The data that support the findings of this study are available from the Veterans Affairs (VA)
432 but restrictions apply to the availability of these data, which were used under license for the
433 current study, and so are not publicly available. Data are however available from the authors upon
434 reasonable request and with permission of VA.

435 Acknowledgements

436 We would like to acknowledge the invaluable contributions arising from the collaboration between
437 Veterans Affairs (VA) and the Department of Energy (DOE) which provided the computing infras-
438 tructure essential to develop and test these approaches at scale with nationwide VA EHR data. This
439 project was supported by the NIH grants 1OT2OD032581, R01 HL089778 and R01 LM013614, P30
440 AR072577, and the Million Veteran Program, Department of Veterans Affairs, Office of Research
441 and Development, Veterans Health Administration, and was supported by the award #MVP000.
442 This research used resources from the Knowledge Discovery Infrastructure at Oak Ridge National
443 Laboratory, which is supported by the Office of Science of the US Department of Energy under Con-
444 tract No. DE-AC05-00OR22725. This publication does not represent the views of the Department
445 of Veterans Affairs or the U.S. government.

446 Competing Interests

447 The authors declare that there are no competing interests.

448 **Author Contribution**

449 ZG: Methodology, Software, Writing – original draft. DZ: Methodology, Software, Writing – original
450 draft. ER: Resources. VAP: Data curation. YH: Data curation. GO: Resources. ZX: Methodology.
451 SS: Writing - review & editing. XX: Visualization. KFG: Writing - review & editing. CH: Writing
452 - review & editing. CB: Visualization. JW: Data curation. LC: Writing - review & editing. TC:
453 Writing - review & editing. EB: Writing - review & editing. ZX: Writing - review & editing. JMG:
454 Writing - review & editing. KPL: Writing - review & editing. KC: Conceptualization, Writing –
455 review & editing, Supervision. TC: Methodology, Conceptualization, Writing – review & editing,
456 Supervision. JL: Methodology, Conceptualization, Writing – review & editing, Supervision, Project
457 administration, Funding acquisition.

References

- 458
- 459 [1] Hong, C. *et al.* Clinical knowledge extraction via sparse embedding regression (KESER) with
460 multi-center large scale electronic health record data. *NPJ digital medicine* **4**, 1–11 (2021).
- 461 [2] Halpern, Y., Horng, S., Choi, Y. & Sontag, D. Electronic medical record phenotyping using
462 the anchor and learn framework. *Journal of the American Medical Informatics Association*
463 **23**, 731–740 (2016).
- 464 [3] Choi, E., Schuetz, A., Stewart, W. F. & Sun, J. Using recurrent neural network models for
465 early detection of heart failure onset. *Journal of the American Medical Informatics Association*
466 **24**, 361–370 (2017).
- 467 [4] Christopoulou, F., Tran, T. T., Sahu, S. K., Miwa, M. & Ananiadou, S. Adverse drug events
468 and medication relation extraction in electronic health records with ensemble deep learning
469 methods. *Journal of the American Medical Informatics Association* **27**, 39–46 (2020).
- 470 [5] Jin, B. *et al.* Predicting the risk of heart failure with ehr sequential data modeling. *IEEE*
471 *Access* **6**, 9256–9261 (2018).
- 472 [6] Gupta, M., Phan, T.-L. T., Bunnell, H. T. & Beheshti, R. Obesity Prediction with EHR Data:
473 A deep learning approach with interpretable elements. *ACM Transactions on Computing for*
474 *Healthcare (HEALTH)* **3**, 1–19 (2022).
- 475 [7] Birkhead, G. S., Klompas, M. & Shah, N. R. Uses of electronic health records for public health
476 surveillance to advance public health. *Annual Review of Public Health* **36**, 345–359 (2015).
- 477 [8] McInnes, B. T., Pedersen, T. & Carlis, J. Using UMLS Concept Unique Identifiers (CUIs) for
478 word sense disambiguation in the biomedical domain. In *AMIA Annual Symposium Proceed-*
479 *ings*, vol. 2007, 533–537 (American Medical Informatics Association, 2007).
- 480 [9] Ghassemi, M. *et al.* Unfolding physiological state: Mortality modelling in intensive care units.
481 In *Proceedings of the 20th ACM SIGKDD International Conference on knowledge Discovery*
482 *and Data Mining*, 75–84 (2014).
- 483 [10] Caballero Barajas, K. L. & Akella, R. Dynamically modeling patient’s health state from
484 electronic medical records: A time series approach. In *Proceedings of the 21th ACM SIGKDD*
485 *International Conference on Knowledge Discovery and Data Mining*, 69–78 (2015).
- 486 [11] Lopez-Gonzalez, E., Herdeiro, M. T. & Figueiras, A. Determinants of under-reporting of
487 adverse drug reactions. *Drug Safety* **32**, 19–31 (2009).
- 488 [12] Classen, D. C. *et al.* ‘Global trigger tool’ shows that adverse events in hospitals may be ten
489 times greater than previously measured. *Health Affairs* **30**, 581–589 (2011).
- 490 [13] Stang, P. E. *et al.* Advancing the science for active surveillance: rationale and design for
491 the observational medical outcomes partnership. *Annals of Internal Medicine* **153**, 600–606
492 (2010).
- 493 [14] LePendu, P., Iyer, S. V., Fairon, C. & Shah, N. H. Annotation analysis for testing drug safety
494 signals using unstructured clinical notes. *Journal of Biomedical Semantics* **3**, 1–12 (2012).

- 495 [15] Tayefi, M. *et al.* Challenges and opportunities beyond structured data in analysis of electronic
496 health records. *Wiley Interdisciplinary Reviews: Computational Statistics* **13**, e1549 (2021).
- 497 [16] Abhyankar, S., Demner-Fushman, D., Callaghan, F. M. & McDonald, C. J. Combining struc-
498 tured and unstructured data to identify a cohort of icu patients who received dialysis. *Journal*
499 *of the American Medical Informatics Association* **21**, 801–807 (2014).
- 500 [17] Zhang, D., Yin, C., Zeng, J., Yuan, X. & Zhang, P. Combining structured and unstructured
501 data for predictive models: a deep learning approach. *BMC Medical Informatics and Decision*
502 *Making* **20**, 280 (2020).
- 503 [18] Wang, Y. *et al.* Early detection of heart failure with varying prediction windows by struc-
504 tured and unstructured data in electronic health records. In *2015 37th Annual International*
505 *Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2530–2533
506 (2015).
- 507 [19] Kharrazi, H. *et al.* The value of unstructured electronic health record data in geriatric syndrome
508 case identification. *Journal of the American Geriatrics Society* **66**, 1499–1507 (2018).
- 509 [20] Bauer-Mehren, A. *et al.* Network analysis of unstructured ehr data for clinical research. *AMIA*
510 *Summits on Translational Science Proceedings* **2013**, 14–18 (2013).
- 511 [21] Finlayson, S. G., LePendou, P. & Shah, N. H. Building the graph of medicine from millions of
512 clinical narratives. *Scientific Data* **1**, 140032 (2014).
- 513 [22] Agarwal, P. & Searls, D. B. Can literature analysis identify innovation drivers in drug discov-
514 ery? *Nature Reviews Drug Discovery* **8**, 865–878 (2009).
- 515 [23] Cohen, T. & Widdows, D. Empirical distributional semantics: methods and biomedical appli-
516 cations. *Journal of Biomedical Informatics* **42**, 390–405 (2009).
- 517 [24] De Vine, L., Zuccon, G., Koopman, B., Sitbon, L. & Bruza, P. Medical semantic similarity
518 with a neural language model. In *Proceedings of the 23rd ACM International Conference on*
519 *Information and Knowledge Management*, 1819–1822 (2014).
- 520 [25] Glicksberg, B. S. *et al.* Automated disease cohort selection using word embeddings from
521 electronic health records. *Pacific Symposium on Biocomputing* 145–156 (2018).
- 522 [26] Segura-Bedmar, I. & Raez, P. Cohort selection for clinical trials using deep learning models.
523 *Journal of the American Medical Informatics Association* **26**, 1181–1188 (2019).
- 524 [27] Feng, Y. *et al.* Patient outcome prediction via convolutional neural networks based on multi-
525 granularity medical concept embedding. In *2017 IEEE International Conference on Bioinfor-*
526 *matics and Biomedicine (BIBM)*, 770–777 (IEEE, 2017).
- 527 [28] Choi, E., Xiao, C., Stewart, W. & Sun, J. Mime: Multilevel medical embedding of electronic
528 health records for predictive healthcare. *Advances in Neural Information Processing Systems*
529 **31** (2018).
- 530 [29] Li, Z., Roberts, K., Jiang, X. & Long, Q. Distributed learning from multiple ehr databases:
531 contextual embedding models for medical events. *Journal of Biomedical Informatics* **92**, 103138
532 (2019).

- 533 [30] Bengio, Y., Ducharme, R. & Vincent, P. A neural probabilistic language model. *Journal of*
534 *Machine Learning Research* **3**, 1137–1155 (2003).
- 535 [31] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations
536 of words and phrases and their compositionality. *Advances in Neural Information Processing*
537 *Systems* **26**, 3111–3119 (2013).
- 538 [32] Pennington, J., Socher, R. & Manning, C. D. Glove: Global vectors for word representation.
539 In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*
540 *(EMNLP)*, 1532–1543 (2014).
- 541 [33] Kartchner, D., Christensen, T., Humpherys, J. & Wade, S. Code2vec: Embedding and cluster-
542 ing medical diagnosis data. In *2017 IEEE International Conference on Healthcare Informatics*
543 *(ICHI)*, 386–390 (2017).
- 544 [34] Choi, E. *et al.* Multi-layer representation learning for medical concepts. In *Proceedings of*
545 *the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*,
546 1495–1504 (2016).
- 547 [35] Choi, E., Schuetz, A., Stewart, W. F. & Sun, J. Medical concept representation learning
548 from electronic health records and its application on heart failure prediction. *arXiv preprint*
549 *arXiv:1602.03686* (2016).
- 550 [36] Che, Z., Cheng, Y., Sun, Z. & Liu, Y. Exploiting convolutional neural network for risk
551 prediction with medical feature embedding. *arXiv preprint arXiv:1701.07474* (2017).
- 552 [37] Rossanez, A., Dos Reis, J. C., Torres, R. d. S. & de Ribaupierre, H. Kgen: a knowledge graph
553 generator from biomedical scientific literature. *BMC Medical Informatics and Decision Making*
554 **20**, 1–24 (2020).
- 555 [38] Harnoune, A. *et al.* Bert based clinical knowledge extraction for biomedical knowledge graph
556 construction and analysis. *Computer Methods and Programs in Biomedicine Update* **1**, 100042
557 (2021).
- 558 [39] Bonner, S. *et al.* A review of biomedical datasets relating to drug discovery: a knowledge
559 graph perspective. *Briefings in Bioinformatics* **23** (2022).
- 560 [40] Bai, T., Chanda, A. K., Egleston, B. L. & Vucetic, S. EHR phenotyping via jointly embedding
561 medical concepts and words into a unified vector space. *BMC Medical Informatics and Decision*
562 *Making* **18**, 15–25 (2018).
- 563 [41] Johnson, A. E. *et al.* MIMIC-III, a freely accessible critical care database. *Scientific Data* **3**,
564 1–9 (2016).
- 565 [42] Muñoz, E., Nováček, V. & Vandenbussche, P.-Y. Facilitating prediction of adverse drug reac-
566 tions by using knowledge graphs and multi-label learning models. *Briefings in Bioinformatics*
567 **20**, 190–202 (2019).
- 568 [43] Zhang, W., Chen, Y., Tu, S., Liu, F. & Qu, Q. Drug side effect prediction through linear
569 neighborhoods and multiple data source integration. In *2016 IEEE International Conference*
570 *on Bioinformatics and Biomedicine (BIBM)*, 427–434 (IEEE, 2016).

- 571 [44] Choi, Y., Chiu, C. Y.-I. & Sontag, D. Learning low-dimensional representations of medical
572 concepts. *AMIA Summits on Translational Science Proceedings* **2016**, 41–50 (2016).
- 573 [45] Zhou, D. *et al.* Multiview incomplete knowledge graph integration with application to cross-
574 institutional ehr data harmonization. *Journal of Biomedical Informatics* **133**, 104147 (2022).
- 575 [46] Koller, D. & Friedman, N. *Probabilistic graphical models: principles and techniques* (MIT
576 press, 2009).
- 577 [47] Arora, S., Li, Y., Liang, Y., Ma, T. & Risteski, A. A latent variable model approach to
578 pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*
579 **4**, 385–399 (2016).
- 580 [48] Xu, Z. *et al.* Codes clinical correlation test with inference on pmi matrix (2022). Preprint.
- 581 [49] Fan, J. & Lv, J. Sure independence screening for ultrahigh dimensional feature space. *Journal*
582 *of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 849–911 (2008).
- 583 [50] Zhou, S., Rütimann, P., Xu, M. & Bühlmann, P. High-dimensional covariance estimation
584 based on gaussian graphical models. *The Journal of Machine Learning Research* **12**, 2975–
585 3026 (2011).
- 586 [51] Yu, S., Cai, T. & Cai, T. Nile: fast natural language processing for electronic health records.
587 *arXiv preprint arXiv:1311.6063* (2013).
- 588 [52] Devlin, J., Chang, M., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional
589 transformers for language understanding. In *Proceedings of the 2019 Conference of the North*
590 *American Chapter of the Association for Computational Linguistics: Human Language Tech-*
591 *nologies, NAACL-HLT*, 4171–4186 (2019).
- 592 [53] Liu, F., Shareghi, E., Meng, Z., Basaldella, M. & Collier, N. Self-alignment pretraining for
593 biomedical entity representations. In *Proceedings of the 2021 Conference of the North American*
594 *Chapter of the Association for Computational Linguistics: Human Language Technologies*,
595 4228–4238 (2021).
- 596 [54] Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical
597 text mining. *Bioinformatics* **36**, 1234–1240 (2020).
- 598 [55] Gu, Y. *et al.* Domain-specific language model pretraining for biomedical natural language
599 processing. *ACM Transactions on Computing for Healthcare* **3**, 1–23 (2021).
- 600 [56] Bodenreider, O. The unified medical language system (UMLS): integrating biomedical termi-
601 nology. *Nucleic Acids Research* **32**, D267–D270 (2004).
- 602 [57] Zhang, T., Leng, J. & Liu, Y. Deep learning for drug–drug interaction extraction from the
603 literature: a review. *Briefings in Bioinformatics* **21**, 1609–1627 (2020).
- 604 [58] Timilsina, M., Tandan, M., d’Aquin, M. & Yang, H. Discovering links between side effects
605 and drugs using a diffusion based method. *Scientific Reports* **9**, 10436 (2019).
- 606 [59] Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The sider database of drugs and side effects.
607 *Nucleic Acids Research* **44**, D1075–D1079 (2016).

- 608 [60] Wen, J. *et al.* Multimodal representation learning for predicting molecule–disease relations.
609 *Bioinformatics* **39**, btad085 (2023).
- 610 [61] Yuan, Z. *et al.* Coder: Knowledge-infused cross-lingual medical term embedding for term
611 normalization. *Journal of Biomedical Informatics* 103983 (2022).
- 612 [62] Liao, K. P. *et al.* High-throughput multimodal automated phenotyping (MAP) with application
613 to PheWAS. *Journal of the American Medical Informatics Association* **26**, 1255–1262 (2019).
- 614 [63] Agarwal, V. *et al.* Learning statistical models of phenotypes using noisy labeled training data.
615 *Journal of the American Medical Informatics Association* **23**, 1166–1173 (2016).
- 616 [64] Levine, M. E., Albers, D. J. & Hripesak, G. Methodological variations in lagged regression for
617 detecting physiologic drug effects in ehr data. *Journal of Biomedical Informatics* **86**, 149–159
618 (2018).
- 619 [65] Yu, S. *et al.* Enabling phenotypic big data with phenorm. *Journal of the American Medical
620 Informatics Association* **25**, 54–60 (2018).
- 621 [66] Ahuja, Y. *et al.* surelda: A multidisease automated phenotyping method for the electronic
622 health record. *Journal of the American Medical Informatics Association* **27**, 1235–1243 (2020).
- 623 [67] Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep patient: an unsupervised representation
624 to predict the future of patients from the electronic health records. *Scientific Reports* **6**, 26094
625 (2016).
- 626 [68] Zhu, Z. *et al.* Measuring patient similarities via a deep architecture with medical concept
627 embedding. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 749–758
628 (2016).
- 629 [69] Dubois, S., Romano, N., Kale, D. C., Shah, N. & Jung, K. Learning effective representations
630 from clinical notes. *arXiv preprint arXiv:1705.07025* (2017).
- 631 [70] Sharafoddini, A., Dubin, J. A., Lee, J. *et al.* Patient similarity in prediction models based on
632 health data: a scoping review. *JMIR Medical Informatics* **5**, e6730 (2017).
- 633 [71] Allyn, J. *et al.* A comparison of a machine learning model with euroscore ii in predicting
634 mortality after elective cardiac surgery: a decision curve analysis. *PLoS one* **12**, e0169772
635 (2017).
- 636 [72] Lei, L. *et al.* An effective patient representation learning for time-series prediction tasks
637 based on EHRs. In *2018 IEEE International Conference on Bioinformatics and Biomedicine
638 (BIBM)*, 885–892 (2018).
- 639 [73] Kalia, M. Dysphagia and aspiration pneumonia in patients with Alzheimer’s disease.
640 *Metabolism* **52**, 36–38 (2003).
- 641 [74] Lauriola, M. *et al.* Neurocognitive disorders and dehydration in older patients: clinical expe-
642 rience supports the hydromolecular hypothesis of dementia. *Nutrients* **10**, 562 (2018).
- 643 [75] Farlow, M. R. Alzheimer’s disease. *Continuum: Lifelong Learning in Neurology* **13**, 39–68
644 (2007).

- 645 [76] Lee, T. J. & Kolasa, K. M. Feeding the person with late-stage Alzheimer’s disease. *Nutrition*
646 *Today* **46**, 75–79 (2011).
- 647 [77] Mimura, M. & Yano, M. Memory impairment and awareness of memory deficits in early-stage
648 Alzheimer’s disease. *Reviews in the Neurosciences* **17**, 253–266 (2006).
- 649 [78] Chai, B. *et al.* Vitamin D deficiency as a risk factor for dementia and Alzheimer’s disease: an
650 updated meta-analysis. *BMC Neurology* **19**, 1–11 (2019).
- 651 [79] Kim, J. H. *et al.* The association between thyroid diseases and Alzheimer’s disease in a national
652 health screening cohort in Korea. *Frontiers in Endocrinology* **13**, 815063 (2022).
- 653 [80] Hong, C. H. *et al.* Anemia and risk of dementia in older adults: findings from the health abc
654 study. *Neurology* **81**, 528–533 (2013).
- 655 [81] Sparks, D. L. *et al.* Atorvastatin for the treatment of mild to moderate Alzheimer disease:
656 preliminary results. *Archives of neurology* **62**, 753–757 (2005).
- 657 [82] Liao, W. *et al.* Deciphering the roles of metformin in Alzheimer’s disease: a snapshot. *Frontiers*
658 *in Pharmacology* **12**, 728315 (2022).
- 659 [83] Barak, Y., Plopsi, I., Tadger, S. & Paleacu, D. Escitalopram versus risperidone for the treat-
660 ment of behavioral and psychotic symptoms associated with Alzheimer’s disease: a randomized
661 double-blind pilot study. *International Psychogeriatrics* **23**, 1515–1519 (2011).
- 662 [84] Lin, L. *et al.* Melatonin in Alzheimer’s disease. *International Journal of Molecular Sciences*
663 **14**, 14575–14593 (2013).
- 664 [85] Liu, J., Chang, L., Song, Y., Li, H. & Wu, Y. The role of NMDA receptors in Alzheimer’s
665 disease. *Frontiers in Neuroscience* **13**, 43 (2019).
- 666 [86] Tariot, P. N. *et al.* Memantine treatment in patients with moderate to severe Alzheimer disease
667 already receiving donepezil: a randomized controlled trial. *Journal of the American Medical*
668 *Association* **291**, 317–324 (2004).