

# Rapid Universal Early Screening for Alzheimer's Disease and Related Dementia via Pattern Discovery in Diagnostic History

Dmytro Onishchenko<sup>1</sup>, Sam Searle<sup>7</sup>, Kenneth Rockwood<sup>7</sup>, James A. Mastrianni<sup>5,6</sup> and Ishanu Chattopadhyay<sup>1,2,3,4★</sup>

<sup>1</sup>Department of Medicine, University of Chicago, Chicago, IL USA

<sup>2</sup>Committee on Genetics, Genomics & Systems Biology, University of Chicago, Chicago, IL USA

<sup>3</sup>Committee on Quantitative Methods in Social, Behavioral, and Health Sciences, University of Chicago, Chicago, IL USA

<sup>4</sup>Center for Health Statistics, Department of Medicine, University of Chicago, Chicago, IL USA

<sup>5</sup>Department of Neurology, University of Chicago, Chicago, IL USA

<sup>6</sup>Committee on Neurobiology, University of Chicago, Chicago, IL USA

<sup>7</sup>Division of Geriatric Medicine, Department of Medicine, Department of Community Health and Epidemiology, School of Health Administration, Halifax, NS Canada

★To whom correspondence should be addressed: e-mail: [ishanu@u-chicago.edu](mailto:ishanu@u-chicago.edu).

## SUMMARY

**Alzheimer's disease (AD) is a progressive, incurable and ultimately fatal neurodegenerative condition. In this study, we introduce the Zero-burden Co-morbid Risk (ZCoR) score to screen for the future risk of AD and related dementia (ADRD) 1 – 10 years before a clinical diagnosis. Requiring no new bloodwork or cognitive tests, ZCoR leverages uncharted comorbidity patterns, to potentially enable near-instantaneous universal point-of-care screening of entire patient populations. In validation, ZCoR ( $n = 729,018$ ) achieves out-of-sample AUC > 90% for predicting a diagnosis immediately after screening, an AUC > 87% for a diagnosis made one year earlier than in current practice, and maintaining over > 80% AUC for predictions made a decade earlier, irrespective of sex. We achieve high predictability in patients lacking any of the currently suspected risk factors; demonstrating effectiveness in cohorts at higher risk of missed diagnoses. Additionally, ZCoR can target mild cognitive impairment (MCI) with performance at par with questionnaire-based assessments (AUC 88 – 90%), maintaining high effectiveness (AUC  $\approx$  80%) for predicting impairment upto 3 years into the future. Powered by stochastic learning algorithms that enhance standard machine learning, ZCoR enables discovery in electronic health record databases, can reduce ADRD and MCI diagnostic delays, and the impact of socio-economic and demographic variables, with immediate impact on patient outcomes.**

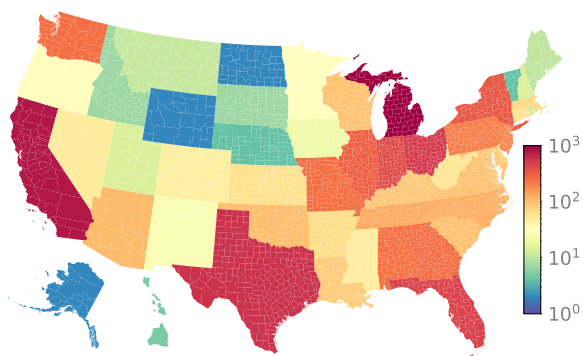
## INTRODUCTION

**D**EMENTIA is an acquired loss of cognition in one or more domains including learning and memory, social cognition, language, executive function, complex attention, and perceptual motor function, severe enough to significantly diminish social or occupational function<sup>1</sup>. Affecting approximately 47 million people worldwide<sup>1</sup>, including over 5.5 million in the United States<sup>2,3</sup>; and projected to be over 81 million worldwide by 2040<sup>4</sup>, there is an immediate need to find effective interventions, and screening tools that enable them.

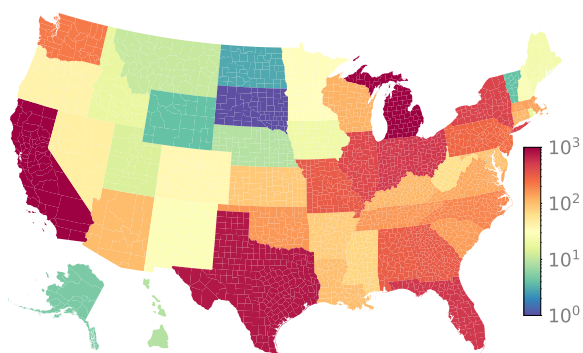
The most common cause of dementia, contributing to 60%–80% of cases, is believed to be Alzheimer's disease (AD), a progressive, fatal, and currently incurable neurodegenerative condition<sup>5</sup>. Based on patient years lived with disability plus years lost to premature mortality, AD ranked as the 6<sup>th</sup> most burdensome disease or injury in the US in 2016, up from 12<sup>th</sup> in 1990<sup>6</sup>, and was implicated in over 250,000 deaths in 2018<sup>5</sup>.

AD-related neuropathology appears to progress over years or decades, independently of the clinical course, suggesting lengthy asymptomatic, subclinical, and/or subtly symptomatic periods before a clinical diagnosis<sup>7</sup>.

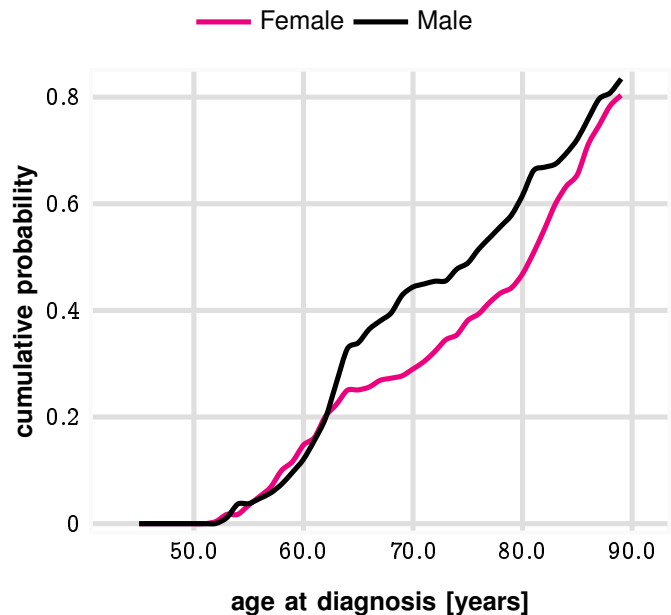
**a.** Prevalence in Truven Dataset:  
Male



**b.** Prevalence in Truven Dataset:  
Female



**C.** Cumulative probability of ADRD diagnosis in Truven dataset



**Fig. 1: Descriptive statistics of database.** Panels a and b show geographic locations of study participants. Panel c illustrates the distribution of patient ages at documented diagnosis, showing that risk starts increasing from around 60 years, matching known ADRD onset age characteristics<sup>3</sup>. Also, panel c illustrates that the empirical risk per patient is higher for males in the Truven dataset, although there are more females with ADRD in total (See Tables I and II), and more female patients in general in the database in the relevant age groups. The age-stratified prevalence in the Truven dataset align closely with prevalence numbers reported for the US in 2020<sup>20</sup>.

Accurate screening for both current and future cases on the Alzheimer's clinical spectrum may be expected to lead to earlier detection of AD biomarkers, neuropathology and incipient cognitive impairment or dementia. In turn, accelerating diagnosis may provide several important benefits for patients, caregivers, healthcare providers, and society<sup>1-3,8-11</sup>: first, pharmacologic and non-pharmacologic interventions may be applied to slow progression of cognitive impairment, while cognition is relatively preserved. Second, use of tailored education strategies may be facilitated to promote cognitively-impaired patients' adherence to and safe use of complex treatment regimens. Third, patient capacity for financial, legal, and health care decision-making may be optimized when it can occur as early as possible in the syndrome's course<sup>9</sup>. And finally, patient access to clinical trials of drugs targeting cognition and dementia may be fostered, and study samples may be enriched, accelerating progress and decreasing costs of such investigations.

However, accurate screening for ADRD is limited by the current diagnostic/prognostic modalities. Imaging or cerebrospinal fluid testing for evidence of beta-amyloid plaques and neurofibrillary tau deposits, the basis for an AD classification<sup>7</sup> and predictors of potential cognitive worsening, is expensive, invasive, and sometimes inaccessible. Although measurement of phosphorylated tau in plasma has shown promise as a specific marker and prognostic factor for ADRD<sup>12-16</sup>, this method is as yet unavailable in everyday practice, entails an invasive blood draw, and if widely used, may in aggregate prove costly. Neuropsychological testing instruments such as the Montreal Cognitive Assessment (MOCA)<sup>17,18</sup> have good diagnostic accuracy and some prognostic utility in identifying mild cognitive impairment (MCI) and mild AD<sup>19</sup>, but their time requirements, even when measured in minutes, may add appreciably to length-of-visit, and hence pose challenges in primary care settings<sup>9,10</sup>. Moreover, these instruments require validation when used in additional locales or languages, and efficacy in predicting future diagnoses might be suspect.

Analysis of routinely-collected health care data in past medical encounters may offer a passive, non-invasive,

TABLE I: Inclusion/Exclusion, Positive/Control Criteria & Cohort Definitions

	Definitions
Inclusion/Exclusion Criteria	Age 50+ years
	Has medical history for $\geq 3$ yrs
Cohort Definition	<b>Positive Cohort:</b> Patients either with at least one target code for ADRD from Tab. III (Case Dx), or with at least one of the target diagnostic codes or a prescription of an ADRD drug (See Tab. IV, Case Dx/Rx)
	<b>Control Cohort:</b> Patients lacking any target diagnostic code (Case Dx), or additionally any ADRD related prescription (Case Dx/Rx)

CONSORT Diagram

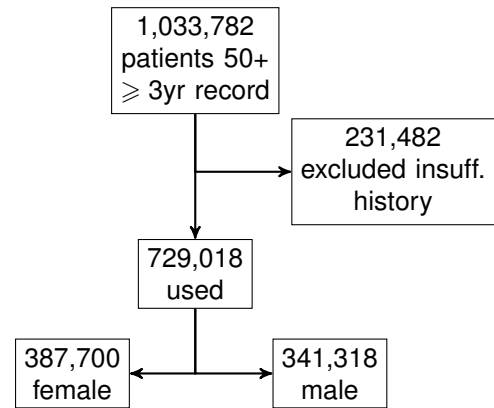


TABLE II: Cohort Sizes

	Male				Female			
	age range	n	n <sub>positive</sub>	n <sub>control</sub>	age range	n	n <sub>positive</sub>	n <sub>control</sub>
all patients	65-74	45943	1498	44445	65-74	47392	1546	45846
	75-84	16911	2870	14041	75-84	17303	3203	14100
	85+	7837	2383	5454	85+	10727	3753	6974
	total	341318	10397	330921	total	387700	12599	375101
low-risk §	65-74	3841	108	3733	65-74	5330	101	5229
	75-84	963	170	793	75-84	1165	213	952
	85+	328	106	222	85+	411	170	241
	total	51499	797	50702	total	78152	1019	77133
high-risk §	65-74	42102	1390	40712	65-74	42062	1445	40617
	75-84	15948	2700	13248	75-84	16138	2990	13148
	85+	7509	2277	5232	85+	10316	3583	6733
	total	289819	9600	280219	total	309548	11580	297968

§See Tab. I for cohort definitions, and SI-Table I for diagnostic codes defining the high-risk cohort.

inexpensive, fast and accessible solution, to accurately discover elevated risk of ADRD<sup>21</sup>. The multifactorial etiologies of ADRD imply that numerous risk factors are associated with these syndromes<sup>21</sup>, and administrative claims and hospital databases, due to their large or even vast scope, offer sufficient statistical power to discover exquisitely-detailed algorithms to pinpoint potential cases. Notably, administrative claims and hospital databases may be especially amenable to exploration in heretofore unprecedented depth of associations of ADRD and comorbidities. Observational studies already have suggested that such associations encompass a large number and variety of disorders covering much of the human disease spectrum<sup>22</sup>; for example, non-neuropsychiatric chronic conditions such as diabetes, hypertension, hypercholesterolemia, obesity, sleep apnea, thyroid disorders, osteoporosis, and glaucoma have been linked to ADRD<sup>22,23</sup>. The validated or suspected associations of ADRD with both “intuitive” categories of comorbidity, e.g., neurological, psychiatric, and cardiovascular disorders, and with non-intuitive categories, e.g., metabolic, endocrine, ophthalmologic disorders, and more recently infections<sup>24,25</sup>, provide rationale for us to seek to leverage comorbid diagnoses to quantify ADRD risk.

Despite extensive documentation of co-morbidities, a reliable risk estimator — purely from ICD code-based co-morbidity patterns without any pre-selection of diagnostic codes already known to be a ADRD co-morbidity — is under-explored. The heterogeneity of brain aging processes and ADRD presentation<sup>26</sup>, make such an endeavor challenging. Here we report the Zero Burden Co-morbid Risk Score (ZCoR) for ADRD, developed and validated using 729,018 unique patients drawn from across the United States, which reliably identifies patients up to 10 years before a contemporary documented clinical diagnosis.

ZCoR is a 701-feature digital signature distilled automatically from past diagnostic code sequences. We make no preselection of codes or ADRD-related risk factors, and require no new blood-work, laboratory tests, familial history or other patient-specific information that might preclude applicability at the point of care. Yet, we achieve an out-of-sample AUC exceeding 87% for either sex (predictions for one year earlier), and  $\approx 80\%$  (prediction

---

made a decade earlier). Importantly, our predictive performance matches the highest AUC achieved by MOCA (92.1%<sup>19,27</sup>) when making predictions just before a clinical diagnosis ( $\geq 91\%$ , see Table. IX). Our underlying algorithms are fundamentally novel, designed to learn from sparse, noisy categorical diagnostic sequences, with demonstrable non-trivial performance boost over standard tools used in recent studies.

Additionally, ZCoR for ADRD is sex-stratified, with separate signatures generated for males and females, in line with the growing appreciation of sex as a key contributor to the phenotypic heterogeneity of ADRD<sup>28,29</sup>. Predictability of eventual dementia in the presence of specific high-risk conditions such as type 2 diabetes has been studied before<sup>30-32</sup>. However, the complex pathobiology of ADRD implies that “low-risk” patients without any of the known risks, might still develop ADRD. Lacking the known flags, such a low-risk cohort is at a much higher risk of a missed or a delayed diagnosis. We show that ZCoR maintains high predictive performance in such patient groups.

The effectiveness of diverse cognitive assessment tools, often combined with pre-selected risk factors, has been recently surveyed<sup>33</sup>, recording AUCs between 55%-89%, sometimes for diagnoses 10 – 20 years into the future<sup>34,35</sup>. Nevertheless, substantial resource burden - often requiring detailed neurological, cognitive and psychiatric consults - limits applicability of such tools at the point-of-care, which were never conceived of for universal screening. Similarly, recent advances with machine learning<sup>26</sup> have often focused on classification of brain imaging data, which, while effective, and backed by well-understood mechanistic models, do not mitigate the barrier to universal adoption.

Thus, our key contribution in this study is to potentially alleviate obstacles to universal testing of the older population. The necessity of such universal screening tools has been well-recognized<sup>36,37</sup>, with recent attempts at developing electronic health record (EHR)-based digital signatures to assess future ADRD risk. While two other digital signatures<sup>36,37</sup> have, to our knowledge, been reported since 2020, ZCoR demonstrates significantly better performance. More importantly, Boustani *et al.* used both structured and unstructured data including clinical notes processed for specific AD related keywords, and and Park *et al.* makes use of laboratory test results (*e.g.* blood hemoglobin) which might not be available for every patient at the point-of-care; in contrast, ZCoR exclusively uses data already present in patient records, which would typically vary from one patient to another, with no *a priori* fixed “demand” on any specific item of clinical, familial, demographic or lifestyle information. Thus, ZCoR can be applied almost universally, passively, and nearly instantaneously at the point-of-care.

EHR-based screening has also been explored to a limited extent for MCI<sup>38</sup>, leveraging patterns extracted from clinical notes. When retrained to detect MCI, ZCoR-MCI is demonstrated to perform significantly better compared to reported results (both at both at the time of screening, and for predictions made years into the future), while, as before, using only ICD codes from past encounters.

## RESULTS

### Patient Selection

Our patient data comes from the IBM MarketScan<sup>®</sup> Commercial Claims and Encounters Database for the years 2003-2018<sup>39</sup> (previously Truven Health Analytics, and referred to as the “Truven dataset”). This US national database merges data contributed by over 150 insurance carriers and large self-insurance companies, and comprises over seven billion time-stamped diagnosis codes. The database tracks over 87 million patients for 1 to 15 years, reflecting a substantial cross-section of the US population. We select our cohort(s) in accordance with the inclusion/exclusion criteria described in Table I, ensuring that selected patients have at least three years of medical history recorded in the dataset. The geographical distribution of the patients in our selected cohort(s) is illustrated in Fig. 1a-b. Fig. 1c illustrates the age distribution at the time of ADRD diagnosis, which is consistent with the reported onset age characteristics for ADRD (mid-sixties<sup>3</sup>). Notably, the cumulative risk of onset (number of ADRD cases normalized by the total number of patients in the given age category) increases with age, as shown in Fig. 1c.

Predicting future ADRD diagnosis is modeled as a binary classification problem: we classify time-stamped sequences of diagnostic codes into positive and control categories, where the “positive” category refers to patients diagnosed with ADRD at 1 year from the point of screening, as identified by one or more ICD codes from Table III appearing in record (referred to as the Dx problem definition in Fig. 2 and Tables VII and VIII) or the prescription of AD-related medication<sup>37,40</sup> (Table IV, donepezil, galantamine, memantine or rivastigmine, referred to as the “Dx/Rx problem definition” in Table VII and VIII). The breakdown of diagnostic codes used for different sex and problem-definition combinations are shown in Table V.

We also consider screening up to  $M$  years before the actual diagnosis, considering values of  $M = 0, \dots, 10$ ,



TABLE III: ADRD ICD diagnostic codes

ICD code	description
290.0	Senile dementia uncomplicated
290.1	Presenile dementia
290.11	Presenile dementia w delirium
290.12	Presenile dementia w delusion
290.13	Presenile dementia w depression
290.2	Senile dementia w delusion
290.21	Senile dementia w depressive
290.3	Senile dementia w delirium
290.8	Senile psychosis NEC
290.9	Senile psychot condition NOS
293.9	Transient mental disease NOS
294.2	Demn NOS w/o behavior dstrb
294.21	Demn NOS w behavior distrb
294.8	Mental disorder NEC other disease
294.9	Mental disorder NOS other disease
331.0	Alzheimer's disease
F00 F00.0 F00.1 F00.2 F00.2 F00.9	Dementia in Alzheimer's disease
F03.9	Unspecified dementia without behavioral disturbance
F03.90	Unspecified dementia without behavioral disturbance
F03.91	Unspecified dementia with behavioral disturbance
F05	Delirium due to known physiological condition
F06.8	Other specified mental disorders due to known physiological condition
G30	Alzheimer's disease with early onset
G30.0	Alzheimer's disease with early onset
G30.1	Alzheimer's disease with late onset
G30.8	Other Alzheimer's disease
G30.9	Alzheimer's disease unspecified

TABLE IV: ADRD common pre-  
scriptions active ingredients

Drugs
Donepezil Hydrochloride
Galantamine Hydrobromide
Memantine Hydrochloride
Rivastigmine
Rivastigmine Tartrate

TABLE V: Number of diagnostic codes used

target case	gender	Number of codes	Number of unique codes
Dx	Male	6729970	18895
Dx	Female	9693456	19998
Dx/Rx	Male	6721267	18887
Dx/Rx	Female	9682339	19988

i.e., which relate to predicting ADRD immediately before a clinical diagnosis to up to a decade in the future. The control cohort comprises patients who never develop ADRD, i.e., do not have target codes, and are never prescribed related medication. Due to our requirement of minimum 3 years of medical history of record implies the absence of a diagnosis for at least  $M + 2$  years in the future from the time of screening. We base our predictions on the past 2 years of diagnostic history. Overall we analyze  $n = 729,018$  patients, with 22,996 patients in the positive group and 706,022 patients in the control group (See CONSORT diagram in Fig. 2c), considering approximately 42 million diagnostic codes, with altogether over 46K unique codes for both sexes.

We do not pre-select any diagnostic code based on its suspected comorbidity with ADRD. To investigate if our performance changes substantially for “high-risk” patients identified based on known co-morbidities including obesity, type II diabetes mellitus, hypertension, atherosclerosis, atrial fibrillation, dyslipidemia, depression, alcohol abuse, and pneumonia, we separately test our performance in high-risk and low-risk sub-cohorts. The high-risk sub-cohort comprises patients with one or more of the diagnoses enumerated in SI-Table I, which identify the top known co-morbidities<sup>22,41</sup>. The low-risk sub-cohort comprises patients who are not at high-risk as specified by the previous condition. Results in the low-risk sub-cohort is of particular significance; these patients are at a higher risk of missed or delayed diagnosis.

Determining the optimal set of target codes for making clinically useful predictions is challenging; too wide a definition makes predictions non-specific, while selecting too few erodes statistical power. The selection of target codes in Table III closely follows Park *et al.*<sup>37</sup> to enable a direct performance comparison. We also consider an expanded set of targets, including vascular dementia, frontotemporal dementia, vascular cognitive impairment, dementia with Lewy Bodies, and major neurocognitive disorder, with no significant performance variation (See Results).

The EHR codeset we use to ascertain ADRD-related disorders is intentionally broad, and not meant to diagnose a specific pathology, but predict risk of general dementia. This comports with our aim of developing a universal

TABLE VI: Feature Definitions (Total number of features used: 701)

feature name	explanation	$\mathbb{N}_{\text{features}}$
<b>feature-phenotype</b> scores relative to phenotype score	Mean p-score of <b>feature-phenotype</b> codes within sequence divided by general p-score of <b>feature-phenotype</b>	45
<b>feature-phenotype</b> scores relative to whole score	Mean p-score of <b>feature-phenotype</b> codes within sequence divided by mean p-score of all codes in the record	45
aggregation score	aggregation of the p-scores in the record	13
high scores proportion	proportion of codes with very high p-scores among all codes in the record	1
low scores proportion	proportion of codes with very low p-scores among all codes in the record	1
dynamics of mean score	mean p-score of second half of the record divided by mean p-score of first half of the record	1
dynamics of geometric mean score	geometric mean p-score of second half of the record divided by mean p-score of first half of the record	1
dynamics of st.dev score	standard deviation of p-scores of second half of the record divided by standard deviation of p-scores of first half of the record	1
dynamics of score range	range of p-scores of second half of the record divided by range of p-scores of first half of the record	1
dynamics of score skew	skew of p-scores of second half of the record divided by skew of p-scores of first half of the record	1
aggregation relative to phn score	aggregation of all <b>feature-phenotype</b> 's mean scores divided by corresponding general p-score of <b>feature-phenotype</b>	9
aggregation relative to whole score	aggregation of all <b>feature-phenotype</b> 's mean scores divided by mean p-score of all codes in the record	9
<b>feature-phenotype</b> proportion	Ratio of number of weeks with the codes of a given phenotype to the total number of weeks in sequence	45
<b>feature-phenotype</b> prevalence	Ratio of number of weeks with the codes of a given phenotype to the number of weeks with any diagnosis code recorded	45
<b>feature-phenotype</b> first incident	Time interval from observation date to the first phenotype code, normalized by record length	45
<b>feature-phenotype</b> last incident	Time interval from observation date to the last phenotype code, normalized by record length	45
<b>feature-phenotype</b> mean position	Mean time position of phenotype codes in the record, normalized by record length	45
<b>feature-phenotype</b> streak	Length of the longest uninterrupted subsequence of weeks with the codes of a given phenotype recorded	45
<b>feature-phenotype</b> code prevalence	Ratio of number of codes of a given phenotype to the total number of codes in sequence	45
<b>feature-phenotype</b> code density	Ratio of number of codes of a given phenotype to the total number of weeks in sequence	45
Max/Mean/Std/Range intermission	Maximum/Mean/Standard Deviation/Range of the lengths of subsequences of consequent weeks with codes	4
Max/Mean/Std cluster	Maximum/Mean/Standard Deviation of the lengths of subsequences of consequent weeks without codes	3
Max/Std/Range prevalence	Maximum/Standard Deviation/Range of the phenotype prevalences	3
Max/Std/Range code prevalence	Maximum/Standard Deviation/Range of the Ratio of number of codes of a given phenotype to the total number of codes in sequence	3
Max/Std/Range code density	Maximum/Standard Deviation/Range of the Ratio of number of codes of a given phenotype to the total number of weeks in sequence	3
Density of DX Record	Proportion of weeks in a record observed where at least one DX code was recorded	1
<b>feature-phenotype</b>	Sequence Likelihood Defect for a given phenotype	45
<b>feature-phenotype</b> neg <b>log-likelihood</b>	negative log-likelihood score for a given phenotype	45
<b>feature-phenotype</b> pos <b>log-likelihood</b>	positive log-likelihood score for a given phenotype	45
<b>feature-phenotype</b> <b>log-likelihood</b> ratio	Ratio of positive to negative log-likelihood score for a given phenotype	45
Mean $\Delta$ ‡	Mean negative Sequence Likelihood Defect	1
Geometric Mean $\Delta$ ‡	Geometric Mean negative Sequence Likelihood Defect	1
Range $\Delta$ ‡	Range of Sequence Likelihood Defect	1
Std. deviation $\Delta$ ‡	Standard Deviation of Sequence Likelihood Defect	1
Mean neg <b>log-likelihood</b>	Mean negative log-likelihood score	1
Geometric Mean pos <b>log-likelihood</b>	Geometric Mean negative log-likelihood score	1
Range neg <b>log-likelihood</b>	Range of negative log-likelihood score	1
Std. deviation neg <b>log-likelihood</b>	Standard Deviation of negative log-likelihood score	1
Mean pos <b>log-likelihood</b>	Mean positive log-likelihood score	1
Geometric Mean pos <b>log-likelihood</b>	Geometric Mean positive log-likelihood score	1
Range pos <b>log-likelihood</b>	Range of positive log-likelihood score	1
Std. deviation pos <b>log-likelihood</b>	Standard Deviation of positive log-likelihood score	1
Mean <b>log-likelihood</b> ratio	Mean log-likelihood score ratio	1
Geometric Mean <b>log-likelihood</b> ratio	Geometric Mean log-likelihood score ratio	1
Range <b>log-likelihood</b> ratio	Range of log-likelihood score ratio	1
Std. deviation <b>log-likelihood</b> ratio	Standard Deviation of log-likelihood score ratio	1

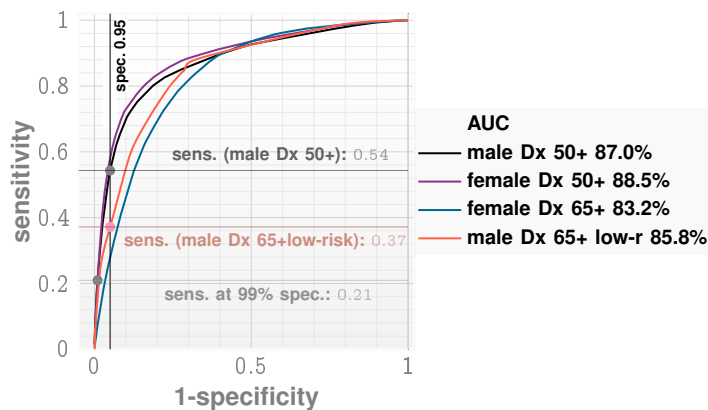
\*feature: ICD disease categories, or sets of diagnostic codes tracked

† $\Delta$ : Sequence Likelihood Defect (See Methods)

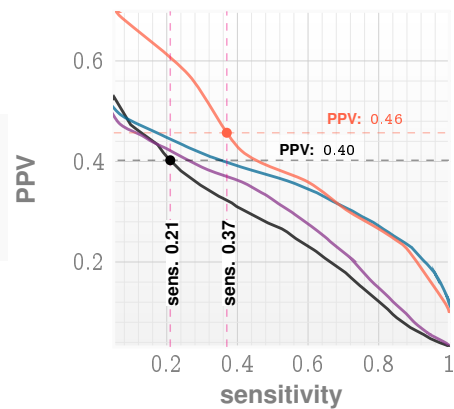
‡ neg log-likelihood: log-likelihood of observed sequence generated by model inferred from control (See Methods)

# pos log-likelihood: log-likelihood of observed sequence generated by model inferred from positive (See Methods)

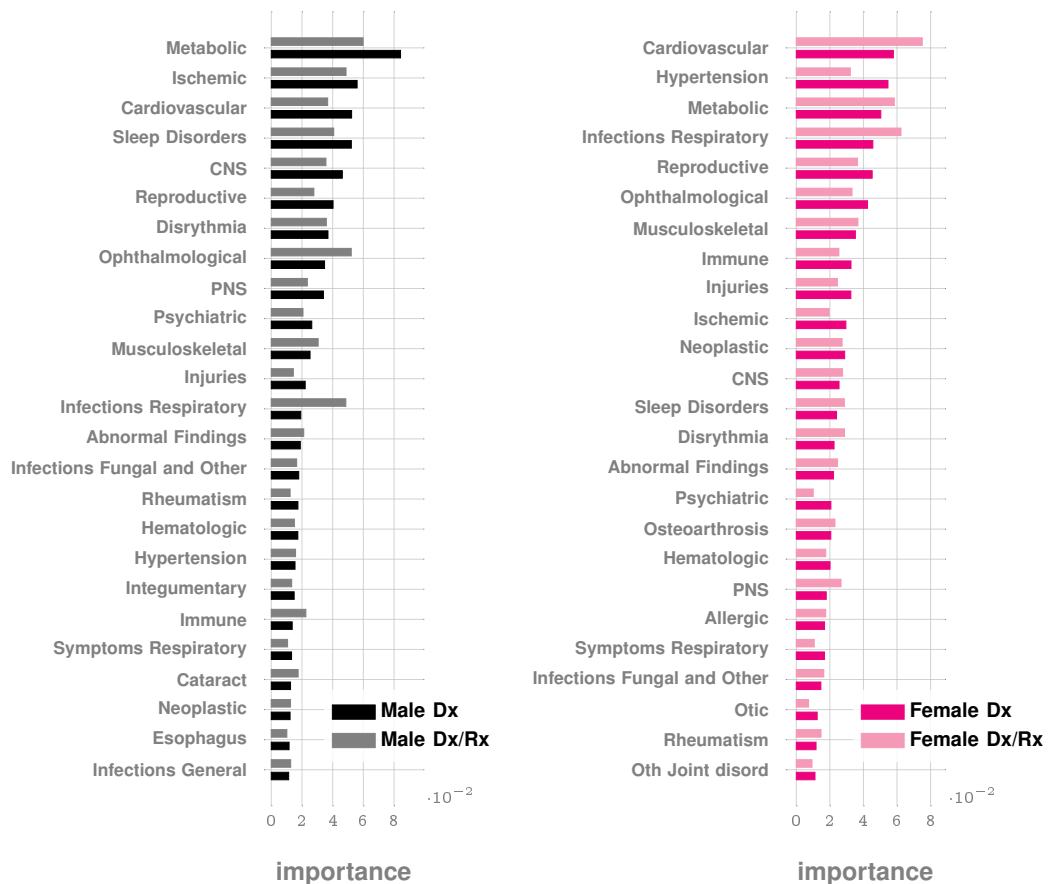
### a. Receiver Operator Characteristic curves



### b. Precision-Recall curves



### c. Feature importances for broad categories of co-morbidities



**Fig. 2: Predictive performance of ZCoR for ADRD diagnosis 1 year in the future.** Panels a and b show the out-of-sample ROC and precision-recall curves for diagnosis 1 year from the point of screening. We achieve AUCs > 88% for male and > 86% for females in the age group 50+, for the diagnostic criteria based on ICD codes (See description of diagnostic criteria considered in Table I), with sensitivities at 58% (females) and 54% (females) at 95% specificity. See Tables VII and VIII for performance within 65+ cohort, and within the low-risk and high-risk cohorts in each age strata. Panel c shows the top 20 comorbidity categories sorted in the order of inferred importance in estimating risk, where categories for mental and cognitive disorders have been removed to highlight the role of other physiological co-morbidities. Importantly, the comorbidities modulate risk differentially by sex, although the patterns are broadly similar, e.g., metabolic, cardiovascular, ophthalmological, ischemic categories appear in both males and females, with slightly altered ranking. Infections and immunologic disorders appear with high importance.

screening tool, as opposed to a diagnostic instrument, that triggers more detailed neurological assessment.

In addition to ADRD, we also investigate the ability of our basic approach to predict MCI (both at the time of screening, and as a prediction of a future diagnosis), identified by the appearance of ICD codes 331.83 (ICD9) and G31.84 (ICD10). We evaluate the performance of ZCoR-MCI with a retrained pipeline, which, in out-of-sample validation, shows significant improvement over reported literature.

---

## Feature Importance & Comorbidity Spectra

The aggregate importance of the ZCoR features (See Fig. 2c), estimated as the mean change in the raw risk via random perturbations in the feature values, illustrates that metabolic and cardiovascular disorders are the most important diagnostic category modulating risk.

Additionally, we compute the statistically significant log-odds ratio of specific ICD codes occurring in the true positive vs the true negative patient sets. We call these the “comorbidity spectra” (See Figs. 3 and 4). These spectra are based on individual codes, as opposed to the aggregated feature importances shown in Fig. 2c. Clearly, every disorder listed in the co-morbid spectra does not all appear in a single patient; the codes with high log-odds ratio are significantly more likely in the positive cohort. The comorbidity spectra, so named because of disease category-specific color coding, offers unique insight into the predictive co-morbidity burden of ADRD.

## Outcomes

In this study we demonstrate the following key results: 1) high out-of-sample predictive performance for identifying a ADRD diagnosis 1 year into future via leveraging subtle comorbidity patterns recorded in the medical history of individual patients (Tables VII-VIII), 2) high predictive performance for diagnosis up to 10 years into the future with sufficiently slow loss of predictive performance to remain clinically useful (Table IX), significantly outperforming recent results (Tables X-XI), 3) effective performance for both low-risk and high-risk cohorts (Tables VII-VIII, see relevant rows). Here, the high-risk cohort comprises patients with commonly surveilled for ADRD co-morbidities. Additionally, 4) maintain high performance for expanded target definitions which include vascular dementia, frontotemporal dementia, vascular cognitive impairment, dementia with Lewy Bodies, and major neurocognitive disorders (Table XII). And finally, 5) high predictive performance to screen for MCI for current and future diagnosis (Table XIII).

Pertaining to our main prediction results for 1-3, Fig. 2a-b illustrate the ROC and the precision-recall curves respectively (for screening one year before current diagnosis), shown separately for males and females. As noted in the panel legends, our out-of-sample predictive performance is  $> 88\%$  AUC for females (age 50+) and  $> 86\%$  for males (age 50+), with  $> 50\%$  sensitivity at 95% specificity (53% for males and 57% for females). At 99% specificity, we obtain a PPV of 42% for females (50+) and 40 – 41% for males (50+) respectively. At these values we obtain an accuracy of  $\approx 96 - 97\%$  (Table VII), which indicates the overall fraction of correct predictions. The PPV achieved by ZCoR at maximum accuracy is 54 – 55% for females (50+) and 51 – 53% for males (50+), with a corresponding NPV of 97%. The corresponding results for age 65+ are tabulated in Table VIII.

Thus, to summarize: our predictive pipeline detects about 53-57 out of every 100 patients who get a diagnosis in 1 year, if we operate at 95% specificity. If we wish to operate at the higher specificity of 99%, then out of 100 positive flags, we have about 41-42 true positives. The accuracy metric indicates that we correctly identify the risk status (positive or control) for 96-97 out of 100 patients, irrespective of sex, highlighting the potentially high clinical significance of ZCoR, as a universal screening tool to identify patients for diagnostic workup and/or intensified surveillance.

From the inferred relative importance of the co-morbidity categories (See Fig. 2d-e), we conclude, that metabolic and ischemic diseases, cardiovascular abnormalities, sleep disorders, nervous system disorders, and diseases of the eye are important modulators of risk. Infections also feature in the top 20 co-morbidities shown in these panels. Importantly while there are sex differences, the overall pattern of the relative importance ranking remains substantially sex-invariant. With some exceptions, many of these patterns are not particularly surprising; the contribution of this study is to bring them together systematically to realize an accurate risk estimate via the ZCoR score.

As expected, our predictive performance degrades as we predict earlier (See Table IX, and inset). Importantly, the degradation is slow enough that we can use ZCoR with acceptable reliability up to 10 years into the future, and significantly outperform reported results.

Understanding the seat of this predictive power is important. The feature importances discussed earlier (Fig. 2c) identify the relative impact of broad disease categories. Importantly, to evaluate the feature importance of a specific diagnostic category, we sum the importance of all features based on that category, not just the presence or absence of individual diagnoses. The latter aspect, *i.e.*, the risk burden from the presence of specific codes, is investigated via the co-morbidity spectra for out-of-sample patients, shown separately in Figs. 3 and 4 for males and females, and the two target definitions (Dx and Dx/Rx). We find that the important co-morbidities are diverse, vary with the sex of the patients, but are clearly dominated by mental disorders, circulatory disorders, injuries, and a range of disorders categorized broadly as “ill-defined symptoms” in the ICD framework. Again,

TABLE VII: Detailed ZCoR performance for patients aged 50+, predictions made 1 year before diagnosis

sex	definition	cohort	sens.	PPV	acc	PPV <sup>†</sup>	NPV <sup>†</sup>	spec.	auc
Female	Dx/Rx	all patients	0.57	0.29	0.94	0.55	0.97	95%	0.884 ± 0.008
Female	Dx	all patients	0.58	0.29	0.94	0.54	0.97	95%	0.885 ± 0.004
Female	Dx/Rx	all patients	0.22	0.42	0.96	0.55	0.97	99%	0.884 ± 0.008
Female	Dx	all patients	0.21	0.42	0.96	0.54	0.97	99%	0.885 ± 0.004
Male	Dx/Rx	all patients	0.53	0.25	0.94	0.51	0.97	95%	0.866 ± 0.006
Male	Dx	all patients	0.54	0.27	0.94	0.53	0.97	95%	0.870 ± 0.011
Male	Dx/Rx	all patients	0.21	0.41	0.97	0.51	0.97	99%	0.866 ± 0.006
Male	Dx	all patients	0.21	0.40	0.97	0.53	0.97	99%	0.870 ± 0.011
Female	Dx/Rx	high-risk	0.55	0.28	0.94	0.54	0.97	95%	0.883 ± 0.008
Female	Dx	high-risk	0.55	0.28	0.94	0.50	0.97	95%	0.883 ± 0.005
Female	Dx/Rx	high-risk	0.20	0.41	0.96	0.54	0.97	99%	0.883 ± 0.008
Female	Dx	high-risk	0.19	0.40	0.96	0.50	0.97	99%	0.883 ± 0.005
Male	Dx/Rx	high-risk	0.52	0.25	0.94	0.58	0.97	95%	0.867 ± 0.008
Male	Dx	high-risk	0.53	0.25	0.94	0.50	0.97	95%	0.871 ± 0.011
Male	Dx/Rx	high-risk	0.21	0.40	0.97	0.58	0.97	99%	0.867 ± 0.008
Male	Dx	high-risk	0.20	0.39	0.97	0.50	0.97	99%	0.871 ± 0.011
Female	Dx/Rx	low-risk	0.59	0.28	0.94	0.64	0.98	95%	0.830 ± 0.039
Female	Dx	low-risk	0.54	0.28	0.94	0.62	0.98	95%	0.833 ± 0.039
Female	Dx/Rx	low-risk	0.41	0.58	0.97	0.64	0.98	99%	0.830 ± 0.039
Female	Dx	low-risk	0.37	0.56	0.97	0.62	0.98	99%	0.833 ± 0.039
Male	Dx/Rx	low-risk	0.54	0.26	0.94	0.56	0.97	95%	0.816 ± 0.039
Male	Dx	low-risk	0.59	0.28	0.94	0.68	0.97	95%	0.826 ± 0.029
Male	Dx/Rx	low-risk	0.28	0.48	0.97	0.56	0.97	99%	0.816 ± 0.039
Male	Dx	low-risk	0.31	0.51	0.97	0.68	0.97	99%	0.826 ± 0.029

\*Calculated at 95% specificity

†Maximum PPV at observed prevalence, and NPV at maximum PPV

TABLE VIII: Detailed ZCoR performance for patients aged 65+, predictions made 1 year before diagnosis

sex	definition	cohort	sens.	PPV	acc	PPV <sup>†</sup>	NPV <sup>†</sup>	spec.	auc
Female	Dx/Rx	all patients	0.29	0.43	0.87	0.56	0.89	95%	0.836 ± 0.007
Female	Dx	all patients	0.28	0.42	0.87	0.54	0.89	95%	0.832 ± 0.010
Female	Dx/Rx	all patients	0.09	0.54	0.89	0.56	0.89	99%	0.836 ± 0.007
Female	Dx	all patients	0.07	0.49	0.88	0.54	0.89	99%	0.832 ± 0.010
Male	Dx/Rx	all patients	0.31	0.41	0.89	0.54	0.91	95%	0.827 ± 0.011
Male	Dx	all patients	0.31	0.40	0.89	0.54	0.91	95%	0.829 ± 0.013
Male	Dx/Rx	all patients	0.10	0.53	0.90	0.54	0.91	99%	0.827 ± 0.011
Male	Dx	all patients	0.09	0.51	0.90	0.54	0.91	99%	0.829 ± 0.013
Female	Dx/Rx	high-risk	0.26	0.43	0.87	0.57	0.89	95%	0.831 ± 0.007
Female	Dx	high-risk	0.27	0.42	0.87	0.57	0.89	95%	0.827 ± 0.009
Female	Dx/Rx	high-risk	0.09	0.54	0.88	0.57	0.89	99%	0.831 ± 0.007
Female	Dx	high-risk	0.07	0.48	0.88	0.57	0.89	99%	0.827 ± 0.009
Male	Dx/Rx	high-risk	0.31	0.41	0.89	0.53	0.91	95%	0.824 ± 0.013
Male	Dx	high-risk	0.30	0.40	0.89	0.52	0.91	95%	0.827 ± 0.014
Male	Dx/Rx	high-risk	0.10	0.53	0.90	0.53	0.91	99%	0.824 ± 0.013
Male	Dx	high-risk	0.09	0.51	0.90	0.52	0.91	99%	0.827 ± 0.014
Female	Dx/Rx	low-risk	0.52	0.58	0.90	0.64	0.92	95%	0.888 ± 0.025
Female	Dx	low-risk	0.45	0.54	0.89	0.55	0.93	95%	0.869 ± 0.033
Female	Dx/Rx	low-risk	0.17	0.71	0.89	0.64	0.92	99%	0.888 ± 0.025
Female	Dx	low-risk	0.10	0.90	0.89	0.55	0.93	99%	0.869 ± 0.033
Male	Dx/Rx	low-risk	0.41	0.48	0.90	0.61	0.92	95%	0.866 ± 0.050
Male	Dx	low-risk	0.37	0.46	0.89	0.57	0.92	95%	0.858 ± 0.037
Male	Dx/Rx	low-risk	0.15	0.62	0.91	0.61	0.92	99%	0.866 ± 0.050
Male	Dx	low-risk	0.14	0.64	0.91	0.57	0.92	99%	0.858 ± 0.037

\*Calculated at 95% specificity

†Maximum PPV at observed prevalence, and NPV at maximum PPV

while many of these patterns are known at the population level, design of the personalized ZCoR score is not immediately obvious.

We include predictive performance in conventional high-risk (defined in SI-Table I) and low-risk cohorts in Table VII and VIII, showing that our performance in the high-risk sub-cohort is comparable with that in the full cohort. The AUCs in the low-risk cohort are somewhat lower (> 81% for males 50+ and > 83% for females 50+ respectively), albeit high enough to be clinically effective: we have a maximum PPV of 62 – 68%, and a sensitivity of 54 – 59% at specificity of 95% for 50+ patients who get diagnosed 1 year in the future (See

TABLE IX: Long-range ZCoR AUC estimates (95% confidence bounds) for target set listed in Table III

years to diagnosis	AUC Female	AUC Male	AUC Female Dx/Rx*	AUC Male Dx/Rx
0	0.912 ± 0.015	0.913 ± 0.015	0.909 ± 0.015	0.918 ± 0.015
1	0.885 ± 0.015	0.871 ± 0.015	0.875 ± 0.015	0.867 ± 0.015
2	0.872 ± 0.017	0.858 ± 0.017	0.867 ± 0.017	0.858 ± 0.017
3	0.858 ± 0.019	0.850 ± 0.018	0.852 ± 0.019	0.855 ± 0.019
4	0.863 ± 0.021	0.842 ± 0.020	0.843 ± 0.021	0.860 ± 0.021
5	0.850 ± 0.023	0.841 ± 0.022	0.840 ± 0.024	0.846 ± 0.023
6	0.840 ± 0.026	0.831 ± 0.025	0.832 ± 0.027	0.835 ± 0.025
7	0.830 ± 0.031	0.810 ± 0.028	0.839 ± 0.031	0.835 ± 0.028
8	0.815 ± 0.036	0.815 ± 0.033	0.828 ± 0.037	0.823 ± 0.034
9	0.799 ± 0.046	0.809 ± 0.041	0.811 ± 0.049	0.807 ± 0.042
10	0.841 ± 0.059	0.784 ± 0.049	0.834 ± 0.062	0.810 ± 0.058

\* Dx/Rx refers to diagnosis inferred from either codes or AD-related prescriptions

INSET. ZCoR AUC over time

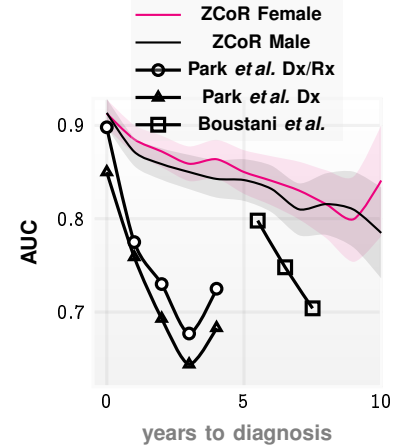


TABLE X: Comparison of AUC achieved in out-of-sample data between ZCoR and Park *et al.*<sup>37</sup>

Year to diagnosis	Park <i>et al.</i> (Dx/Rx)	Park <i>et al.</i> (Dx)	ZCoR (Dx)	ZCoR (Dx/Rx)	$\Delta(\text{Dx})\%^\ddagger$	$\Delta(\text{Dx/Rx})\%^\dagger$
0	0.90	0.85	0.91	0.92	7.4726	2.3315
1	0.78	0.76	0.89	0.88	16.621	12.933
2	0.73	0.69	0.87	0.87	25.851	18.793
3	0.68	0.64	0.86	0.86	33.373	26.404
4	0.72	0.68	0.86	0.86	26.467	18.639

TABLE XI: Comparison of AUC achieved in out-of-sample data between ZCoR and Boustani *et al.*<sup>36</sup>

year to diagnosis	Boustani <i>et al.</i> (Dx)	ZCoR (Dx)	$\Delta(\text{Dx})\%^\ddagger$
1-10	0.80	0.85	6.48
3-10	0.75	0.84	11.9
5-10	0.70	0.83	17.8

$\ddagger$  Percentage outperformance of ZCoR with the Dx target definition

$\dagger$  Percentage outperformance of ZCoR with the Dx/Rx target definition

Tables VII and VIII).

Additionally, Table XII shows that an expanded target definition (described in Methods) yields no significant change in predictive performance. Finally, our results on MCI prediction are shown in Table XIII (carried out with a retrained pipeline targeting MCI), which illustrates an average AUC between 88-90% at the point of screening, degrading to under  $\approx 80\%$  for predictions made 3 years into the future. Notably, the performance at the point of screening is at par with MOCA ( $0.9 - 0.91 \pm 0.015$  vs  $.921^{19}$ ), while being significantly superior to a recently published EHR-based approach using standard machine learning algorithms<sup>42</sup>.

## DISCUSSION

We report the development and validation of the ZCoR automated universal screening tool for AD/DRD, leveraging previously-uncharted co-morbidity patterns discovered from individual longitudinal diagnostic history. Across sexes, ZCoR accurately preempts AD/DRD cases up to 10 years before a clinical diagnosis is first documented. The broad co-morbidity categories that we infer to be important (Fig. 2c) include metabolic, cardiovascular, ischemic, ophthalmological, and sleep disorders. Diseases of the nervous system, unrelated to AD/DRD, infections, and immunologic disorders also appear in the list of top risk-modulating co-morbidities. Importantly, the co-morbidities modulate risk differentially by sex (Fig. 2c), although the patterns are broadly similar, with slightly altered ranking.

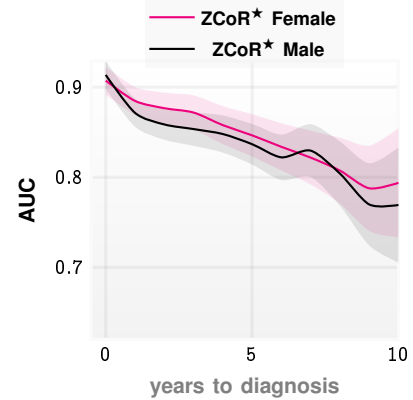
TABLE XII: Long-range ZCoR<sup>★</sup> AUC estimates (95% confidence bounds) for expanded target set (adding vascular dementia, frontotemporal dementia, vascular cognitive impairment, dementia with Lewy Bodies, and major neurocognitive disorder to Table III).

years to diagnosis	AUC Female	AUC Male
0	0.907 ± 0.155	0.913 ± 0.155
1	0.884 ± 0.155	0.871 ± 0.155
2	0.876 ± 0.173	0.858 ± 0.170
3	0.871 ± 0.191	0.853 ± 0.187
4	0.857 ± 0.210	0.847 ± 0.206
5	0.846 ± 0.236	0.836 ± 0.226
6	0.833 ± 0.262	0.822 ± 0.253
7	0.821 ± 0.299	0.829 ± 0.294
8	0.807 ± 0.365	0.804 ± 0.358
9	0.787 ± 0.473	0.769 ± 0.456
10	0.793 ± 0.606	0.769 ± 0.640

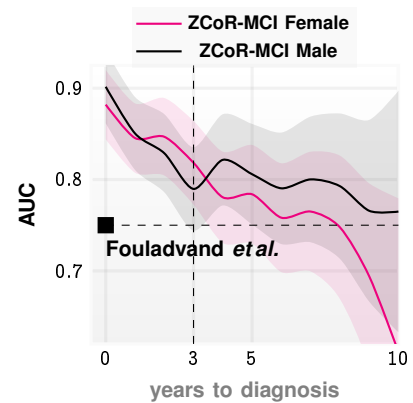
TABLE XIII: Long-range ZCoR-MCI AUC estimates (95% confidence bounds) with MCI identified via ICD codes 331.83 (ICD9) and G31.84 (ICD10).

years to diagnosis	AUC Female	AUC Male
0	0.882 ± 0.038	0.901 ± 0.040
1	0.845 ± 0.038	0.850 ± 0.040
2	0.846 ± 0.042	0.828 ± 0.043
3	0.818 ± 0.046	0.789 ± 0.046
4	0.780 ± 0.050	0.821 ± 0.050
5	0.783 ± 0.054	0.805 ± 0.055
6	0.758 ± 0.060	0.790 ± 0.061
7	0.764 ± 0.066	0.800 ± 0.069
8	0.747 ± 0.079	0.792 ± 0.080
9	0.694 ± 0.102	0.765 ± 0.100
10	0.611 ± 0.168	0.764 ± 0.132

INSET. ZCoR<sup>★</sup> AUC over time



INSET. ZCoR-MCI AUC over time



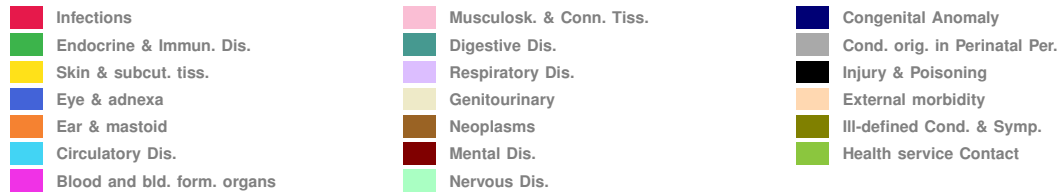
Focusing on the presence/absence of individual diagnostic codes modulating ADRD risk in the co-morbidity spectra (Figs. 3 and 4), we find circulatory disorders are generally over-represented, along with injuries, and conditions related to age-related cognitive decline. Many of these patterns are unsurprising: AD is an amnesic syndrome, injuries might indicate neuropathies from known AD co-morbidities such as diabetes or stroke, and cerebrovascular diseases might signal vascular dementia. Other prominent codes such as ataxia and psychiatric signs were recently associated with specific biomarkers implicated in autosomal dominant early-onset Alzheimer’s disease<sup>43,44</sup>. Appearance of other codes are more surprising, e.g. dysphagia or swallowing impairment is usually noted in the late stages of AD. However, recent studies have documented changes in cortical control of swallowing beginning before dysphagia becomes apparent in dementia patients<sup>45,46</sup>. Thus, the important illness categories that we find to be associated with ADRD in either sex align with suspected or documented links in statistical<sup>29,36,37</sup> and observational studies<sup>22,23</sup>, lending credence to ZCoR rationale and accuracy. Also lending such credence is the score’s incorporation, via sex-stratification, of differences between males and females in ADRD risk factors, natural history, and symptoms<sup>28,29,47–51</sup>.

We find that with increasing patient age, it becomes more difficult to distinguish age related cognitive decline from ADRD (SI-Fig. 1), suggesting that ADRD comorbidities have confounding overlaps with conditions that arise more frequently as patients get older.

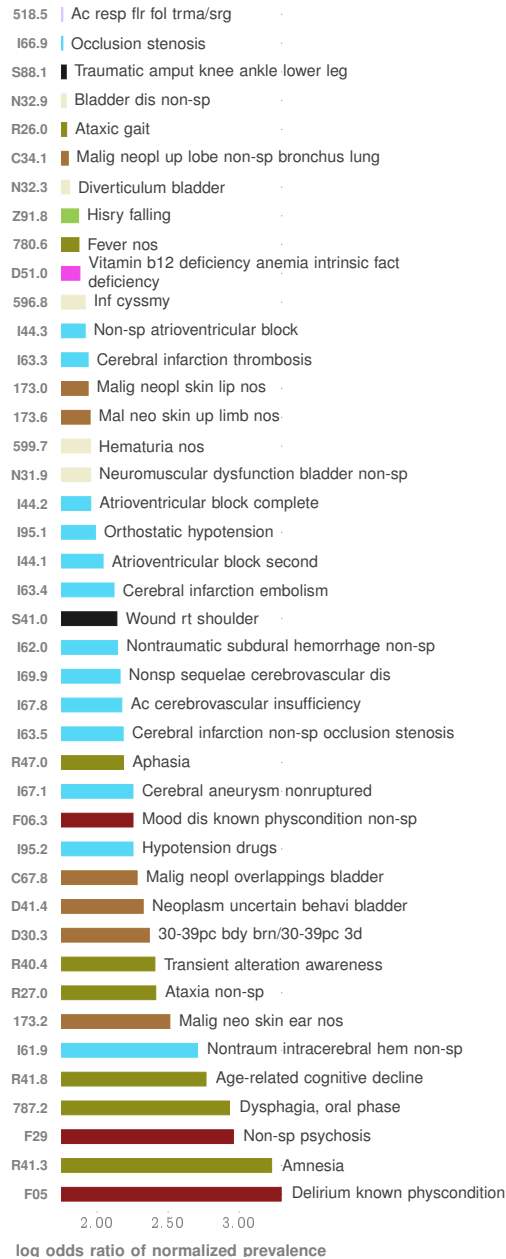
To our knowledge, ZCoR is one of three digital signatures for ADRD reported since 2020, joining those of Boustani *et al.*, developed utilizing data from the Indiana Network for Patient Care<sup>36</sup>, and of Park *et al.*, developed utilizing data from the Korean National Health Insurance Service<sup>37</sup>. Although the respective reported prognostic time-frames are not fully comparable, our digital signature appeared to achieve the best performance of the three (Table IX inset, and Tables X and XI). Notably, the AUC of ZCoR for ADRD at 10 years before documented diagnosis surpassed the AUCs of the Boustani *et al.*<sup>36</sup> signature for the 1-10 year, 3-10 year, or 5-10 year before diagnosis time-frames by 6.5%, 11.9%, and 17.8% respectively, while leveraging diagnostic histories of 1,400% more patients ( $\approx 50K$  vs  $\approx 700K$  for ZCoR). Also for each prediction time-point made 0 through 4 years before documented diagnosis, the AUCs of ZCoR exceeded those of the Park *et al.*<sup>37</sup> signature by 2-7% (0



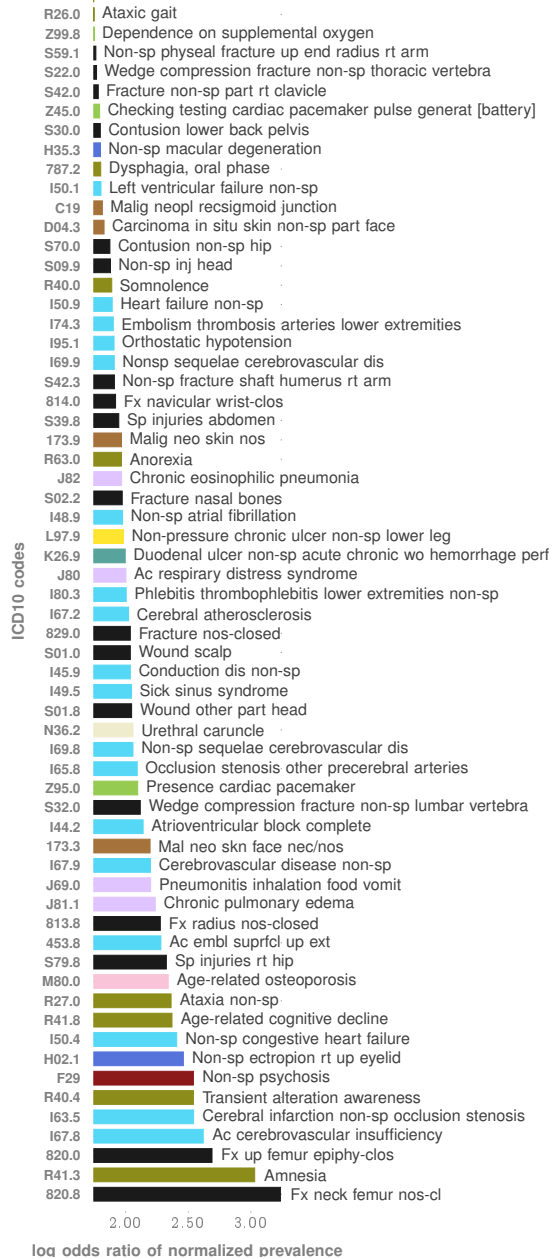
## ICD Class



### a. Male Alzheimer's Dis.



### b. Female Alzheimer's Dis.

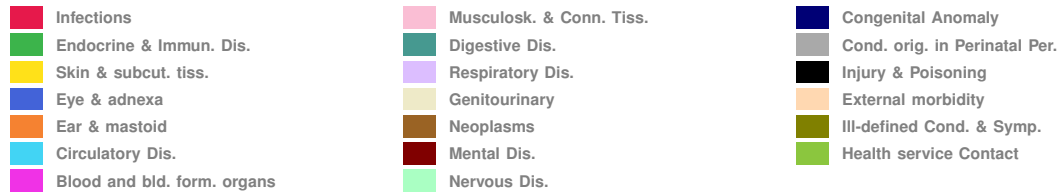


**Fig. 3: Co-morbidity Spectrum for the Dx/Rx case.** Disorders that increase the odds of the patient being a “true positive” vs a “true negative”, where diagnosis is determined using either ICD codes (See Table III) or ADRD-related medications (See Table IV) in history. Such disorders (ranked according to the log-odds ratio) are more likely to be found in patients who are in the positive cohort. Comparing **panel a** with **panel b**, we note that these odds change from males to females, but as expected the patterns are broadly similar, with over-representation of circulatory disorders.

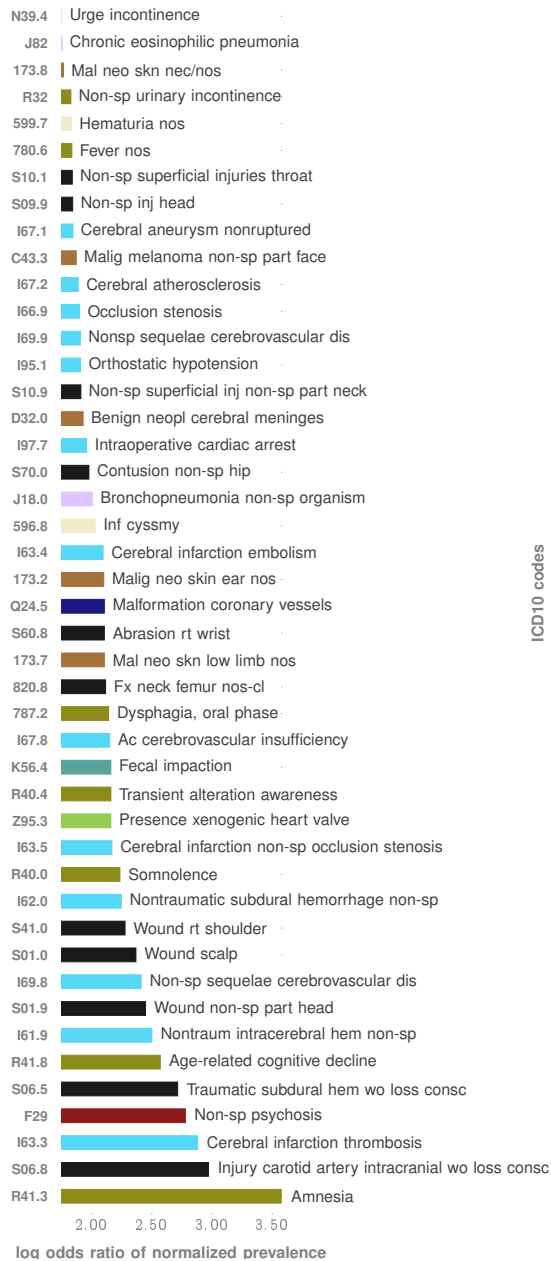
year), 12.9-16.6% (1 year), 18.8-25.8% (2 years), 26.4-33.3% (3 years) and 18.6-26.5% (4 years), while using 1,750% more patients ( $\approx 40K$  vs  $\approx 700K$  for ZCoR).

A number of methodological differences contribute to these respective performances. First, ZCoR uses sophisticated pattern discovery on patient history, and is not limited by known risk factors and co-morbidities, allowing for high performance on low-risk and high-risk cohorts alike. More importantly, our new stochastic inference

## ICD Class



### a. Male Alzheimer's Dis.



### b. Female Alzheimer's Dis.



**Fig. 4: Co-morbidity Spectrum for the Dx case.** Disorders that increase the odds of the patient being a “true positive” vs a “true negative”, where diagnosis is determined using ICD codes (See Table III). Such disorders (ranked according to the log-odds ratio) are more likely to be found in patients who are in the positive cohort. Comparing **panel a** with **panel b**, we note that these odds change from males to females, but as expected the patterns are broadly similar, with over-representation of circulatory disorders.

algorithms are designed to leverage longitudinal patterns, and are not limited to using indicator variables, i.e., simply the presence or absence of specific historical codes. Thus we are able to substantially leverage the emergent dependencies and temporal ordering of patterns emergent across the human disease spectrum.

Additionally, ZCoR for ADRD is stratified by sex. Sex-stratification of AD risk has recently found support in the literature<sup>29</sup>. Finally, our algorithm is derived using a cohort roughly 10-18-fold larger than those of Boustani

---

*et al.* or Park *et al.* (729,018 versus 40,736 and 71,466 respectively), allowing our algorithms to capitalize on significantly larger quantities of data.

Beyond predictive performance, ZCoR addresses the barrier to universal testing. With no specific data demands (we use what we have on the individual patient file), and designed to operate on existing electronic healthcare architectures, the digital signature operates non-invasively, inexpensively, and nearly instantaneously, and is potentially very widely, if not universally accessible at least in developed countries using EHR. Unlike that of Boustani *et al.*, (but like that of Park *et al.*) who use expert opinion-generated variables in the first phase of their digital signature development, our algorithm is completely data-driven. Also, unlike Boustani *et al.* (but like Park *et al.*), we considered only structured data, *i.e.*, ICD codes, and not clinical notes. While clinical notes might reveal substantially more information, such insights most relevant to ADRD might not be available before a neurology consult. And unlike Park *et al.*, we do not use laboratory tests such as hemoglobin levels, which might not be available for every patient in primary care.

We envision three main potential applications of ZCoR. First, the score can serve in primary care or specialist settings (*e.g.*, neurology, gerontology) as a screening tool for future incident overt cases, with the potential diagnostic, therapeutic, psychosocial, caregiver-related, and research benefits noted in the introduction to this paper. ZCoR could, for example, be routinely deployed, alone or along with a brief, validated neuropsychological instrument, as recommended by the American Academy of Neurology<sup>52</sup>, in the cognitive screening mandated since 2011 as part of the Medicare annual wellness visit<sup>3</sup>. Alternatively, especially given the variable clinical natural history of such patients<sup>53</sup>, ZCoR could be employed in individuals with subjective memory decline but largely-intact cognition and function, or in those with incipient MCI, *e.g.*, worsening but still personally-appropriate serial neuropsychological test scores, who have not undergone biofluid or imaging assessment for ADRD-related or other dementia-related pathology. Notably, from pharmacoeconomic, practical, and psychosocial standpoints, use of ZCoR for “long-range” clinical prognostication may be compatible with the up-to-decades-long, pre-clinical progression of beta-amyloid and tau neuropathology in AD: even 10 years before overt cognitive impairment, biofluid testing or imaging performed due to ZCoR high-risk status is likely to be informative<sup>7</sup>, and the ZCoR classification, actionable. A second potential ZCoR application could be screening for undiagnosed prevalent cases of ADRD in primary care settings. Considering the estimated 45%–80% of dementia cases in older adults that go undiagnosed in the US<sup>54</sup>, availability of a non-invasive, inexpensive, near-instantaneous, and almost universally-accessible tool could revolutionize detection of such patients. Third, ZCoR could be applied in scientific research regarding ADRD natural history and prevention. Beyond enrichment of trials of prophylactic interventions against cognitive impairment, ZCoR opens intriguing avenues of investigation, *e.g.*, examination of the roles of previously-underrecognized comorbidity classes with important associations with ADRD, *e.g.*, musculoskeletal disorders in males, respiratory infections in females, reproductive or ophthalmological disorders in both sexes. More precise understanding of the particular diseases that indeed are associated with ADRD will facilitate assessment of intriguing hypotheses such as inflammation serving as a key link between comorbidities and ADRD genetic features and phenotype<sup>55</sup>.

## Limitations & Conclusion

Our key limitations arise from potential diagnostic mis-codings, and the current imprecision in Alzheimer’s-related nomenclature<sup>7</sup>. Coupled with the high prevalence of undiagnosed dementia, mis-coding could lead to our ADRD signature deriving from data of only a fraction, albeit a substantial fraction, of our true cases. This situation might pose a particular peril under our Dx target definition, which considers only diagnostic codes. Mitigating this concern is the vast size of our control groups ( $n=375,101$  females,  $n=330,921$  males), implying that “non-signal” from large numbers of “true controls” is likely to overwhelm “buried ADRD signal” from “false controls”. An additional possible concern related to mis-coding would be inclusion of non-ADRD age-related dementia cases among the ADRD group. However, given the “mixed” picture of dementia afflicting many patients with ADRD<sup>1,7</sup>, tracking the characteristics of patients with non-ADRD cognitive impairment is also pertinent.

The performance of ZCoR might be further enhanced with the inclusion of treatment-related factors, *e.g.*, medications, along with comorbidities. As noted, however, using only diagnostic codes may increase the availability of data inputs for ZCoR in everyday practice, and hence the tool’s scalability to routine settings. Moreover, comorbidity codes may be viewed to at least some extent as surrogates capturing the effects of medications that might influence Alzheimer neuropathology, *e.g.*, statins or anti-diabetic agents.

Predictive screening for ADRD raises some ethical concerns. In particular, early detection of progressive, not-yet-well-manageable brain disorders that have major effects on capacity, autonomy, and healthcare and other resource utilization, poses potential risks stemming from the possibility of stigmatization and discrimination<sup>8</sup>. It will be necessary to further explore these and other potential harms of early recognition of Alzheimer cognitive

---

impairment, and to seek their amelioration through legal and public health policy changes<sup>3,8</sup>.

It is important to note that however strong its predictive performance, ZCoR is a screening tool, not a diagnostic tool, and by itself certainly does not establish an ADRD diagnosis. ZCoR prediction of high-risk for ADRD should lead to diagnostic testing, e.g., application of cognitive tests or imaging, and to intensified surveillance, as indicated. ZCoR results also can inform discussion with, and planning by, patients and their significant others.

In conclusion, ZCoR opens potentially new avenues in identification of and intervention against cognitive impairment, in neurocognitive research, and in designing effective caregiver support. Moving forward, we will focus on: 1) prospective validation of ZCoR; 2) assessment of the effects of ZCoR use on patient and caregiver quality-of-life, patient cognition and function, and healthcare utilization; 3) correlation with ADRD clinical and neuropathological biomarkers such as neuropsychological and functional test results and biofluid and imaging findings related to beta-amyloid, tau, and neurodegeneration; 4) comparison of ZCoR prospective performance in different racial groups and ethnicities, including examination of the signature's ability to reduce disparities in the rate of diagnosis. The impact of ZCoR on the accuracy and speed of diagnosis, on health care resource utilization, and eventually, on patient and caregiver outcomes, warrant prospective study.

## ACKNOWLEDGEMENTS

Robert J. Marlowe, Spencer-Fontayne Corporation, Jersey City, NJ, USA provided editorial support, literature research, and writing assistance. This work is funded in part by the Defense Advanced Research Projects Agency, project number HR00111890043/P00004, National Institute on Aging, award number AG066619, and the Alzheimer's Association, award number 852418. The claims made here do not necessarily reflect the position of the funding organizations.

## AUTHOR CONTRIBUTIONS

DO implemented the algorithm and ran validation tests. DO and IC carried out mathematical modeling, and algorithm design. DO, SS, KR, JM and IC interpreted results. JM and IC guided the research. DO, KR, JM and IC wrote the paper. IC procured funding for the study.

## DECLARATION OF INTERESTS

IC is a founder and shareholder of Zero Burden Laboratories, Inc. He has not drawn any salary from the company. IC has received funding from the Alzheimer's Association, United States Department of Defense, the National Institutes of Health, and the Neubauer Collegium for Culture and Society. DO is a founder and shareholder of Zero Burden Laboratories. He has not drawn any salary from the company.

## STAR METHODS

### Resource Availability

#### *Materials availability*

This study did not generate or use new unique reagents.

#### *Data and code availability*

- **Data:** The Truven database used in this study is not in the public domain, and may be procured under license from <https://www.ibm.com/watson-health/about/truven-health-analytics>. A small de-identified set of patient diagnostic history is made available for testing purposes, and is publicly available as of the date of publication, as noted in the Key Resources Table. Description of ICD codes are available at <https://www.cdc.gov/nchs/icd/icd10cm.htm>, and <https://www.icd10data.com/>. A comprehensive interface for looking up ICD-10-CM code descriptions is provided by the National Library of Medicine, and may be accessed at <https://clinicaltables.nlm.nih.gov/apidoc/icd10cm/v3/doc.html>.
- **Code:** Working modules have been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table. Complete pseudocode is made available in the Supplementary Information text (Algorithms 1, and 2 in Supplementary Information).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

---

## Lead Contact

Further information and requests for resources and software should be directed to and will be fulfilled by the lead contact, Ishanu Chattopadhyay (ishanu@uchicago.edu).

## Method Details

### Summary of Modeling Steps

Individual diagnostic histories can have long-term memory<sup>56</sup>, implying that the order, frequency, and comorbid interactions between diseases are important for assessing the future risk of our target phenotype. The large number of possible ICD codes, along with the sparsity of codes per patient (approximately one entry every 100 steps on the diagnostic time series) makes this a difficult learning problem.

### Step 1: Partitioning The Human Disease Spectrum

We begin by partitioning the human disease spectrum into 45 non-overlapping categories. Each category is defined by a set of diagnostic codes from the International Classification of Diseases, Ninth Revision (ICD9) (See Table SI-II for description of the categories used in this study).

For this study, we ended up using 6462501 and 9426722 diagnostic codes for males and females respectively (17501 and 18633 unique codes) spanning both ICD9 and ICD10 protocols (using ICD10 General Equivalence Mappings (GEMS)<sup>57</sup> equivalents where necessary), from a total 729,018 patients. Transforming the diagnostic histories to report only the broad categories reduces the number of distinct codes that the pipeline needs to handle, thus improving statistical power.

Our categories largely align with the top-level ICD9 categories, with small adjustments, *e.g.* bringing all infections under one category irrespective of the pathogen or the target organ. We do not pre-select the phenotypes; we want our algorithm to seek out the important patterns without any manual curation of the input data.

For each patient, the past medical history is a sequence  $(t_1, x_1), \dots, (t_m, x_m)$ , where  $t_i$  are timestamps and  $x_i$  are ICD9 codes diagnosed at time  $t_i$ . We map individual patient history to a three-alphabet categorical time series  $z^k$  corresponding to the disease category  $k$ , as follows. For each week  $i$ , we have:

$$z_i^k = \begin{cases} 0 & \text{if no diagnosis codes in week } i \\ 1 & \text{if there exists a diagnosis of category } k \text{ in week } i \\ 2 & \text{otherwise} \end{cases} \quad (1)$$

The time-series  $z^k$  is observed in the inference period. Thus, each patient is represented by 43 mapped trinary series.

We refer to these individual diagnostic categories as “phenotypes”, since they are observable characteristics of the patients. Each patient is represented by 45 sparse stochastic time series of events, which are compressed into specialized Hidden Markov Models known as Probabilistic Finite Automata (PFSA)<sup>58,59</sup>. These models are inferred separately for each phenotype, for each sex, and for the control and the positive cohorts, to identify the distinctive average patterns emerging at the population level. We infer  $45 \times 2 \times 2 = 180$  PFSA models in total in this study. Our inference algorithm for these models does not presuppose a fixed structure, and is able to work with non-synchronized and variable-length data streams. Variation of these inferred models between the positive and control groups delineate the estimated risk of an AD RD diagnosis at the population level. Given these models, and the history of a specific patient, we can then quantify the likelihood of this patient’s particular history being generated by the control PFSA models as opposed to the positive models. We refer to this likelihood difference as the sequence likelihood defect (SLD)<sup>60</sup>, which is the one of the key informative features in our approach. The SLD is a novel concept, involving the generalization of the notion of Kullback-Liebler divergence<sup>61</sup> between probability distributions to a generalized divergence between possibly non-iid stochastic processes (See Step 2 below). SLD-based features allow the ZCoR measure to factor in complex longitudinal, *i.e.*, temporal patterns beyond simply the presence/absence of comorbidities.

### Inference & Event Periods

We train our predictive pipeline with all diagnostic codes that are recorded in the past 2 years from the point at which a prediction is made. This period from which we use data to train our pipeline is called the “inference window”. Our aim is to make predictions on the occurrence of the target diagnostic codes at 1 year from the end of the inference window. For patients in the control cohort, we make sure that no target code appears for 2 years after the end of the inference window. Additionally, when making predictions further into the future, we always make sure that the control group has no target codes for 1 year after the predicted time of diagnosis,

i.e., if we are making a prediction of a diagnosis  $m$  years in future, then control group patients are chosen to have no diagnosis in at least next  $m + 1$  years.

### Step 2: Model Inference & The Sequence Likelihood Defect $\Delta$

The mapped series, disease-category, and ADRD diagnosis-status are considered to be independent sample paths, and we want to explicitly model these systems as specialized HMMs (PFSA). We model the positive and the control cohorts and each disease category separately, ending up with a total of 86 HMMs at the population level (43 categories, 2 ADRD status categories: positive and control). Each of these inferred models is a PFSA; a directed graph with probability-weighted edges, and acts as an optimal generator of the stochastic process driving the sequential appearance of the three letters (as defined by Eq. (1)) corresponding to disease category, and ADRD status-type (See “**Probabilistic Finite State Automata Inference**” for background on PFSA inference).

To reliably infer the ADRD status-type of a new patient, i.e., the likelihood of a diagnostic sequence being generated by the corresponding ADRD status-type model, we generalize the notion of Kullback-Leibler (KL) divergence<sup>61,62</sup> between probability distributions to a divergence  $\mathcal{D}_{KL}(G||H)$  between ergodic stationary categorical stochastic processes<sup>63</sup>  $G, H$  as:

$$\mathcal{D}_{KL}(G||H) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{x:|x|=n} p_G(x) \log \frac{p_G(x)}{p_H(x)} \quad (2)$$

where  $|x|$  is the sequence length, and  $p_G(x), p_H(x)$  are the probabilities of sequence  $x$  being generated by the processes  $G, H$  respectively. Defining the log-likelihood of  $x$  being generated by a process  $G$  as :

$$L(x, G) = -\frac{1}{|x|} \log p_G(x) \quad (3)$$

The cohort-type for an observed sequence  $x$  — which is actually generated by the hidden process  $G$  — can be formally inferred from observations based on the following provable relationships (See Theorems 1 and 2):

$$\lim_{|x| \rightarrow \infty} L(x, G) = \mathcal{H}(G) \quad (4a)$$

$$\lim_{|x| \rightarrow \infty} L(x, H) = \mathcal{H}(G) + \mathcal{D}_{KL}(G||H) \quad (4b)$$

where  $\mathcal{H}(\cdot)$  is the entropy rate of a process<sup>61</sup>. Importantly, Eq. (4) shows that the computed likelihood has an additional non-negative contribution from the divergence term when we choose the incorrect generative process. Thus, if a patient is eventually going to be diagnosed with ADRD, then we expect that the disease-specific mapped series corresponding to her diagnostic history be modeled by the PFSA in the positive cohort. Denoting the PFSA corresponding to disease category  $j$  for positive and control cohorts as  $G_+^j, G_0^j$  respectively, we can compute the *sequence likelihood defect* (SLD,  $\Delta^j$ ) as:

$$\Delta^j \triangleq L(G_0^j, x) - L(G_+^j, x) \rightarrow \mathcal{D}_{KL}(G_0^j||G_+^j) \quad (5)$$

With the inferred PFSA models and the individual diagnostic history, we estimate the SLD measure on the right-hand side of Eqn. (5). The higher this likelihood defect, the higher the similarity of diagnosis history to that of women with ADRD.

### Step 3: Risk Estimation Pipeline With Semi-supervised & Supervised Learning Modules

The risk estimation pipeline operates on patient specific information limited to the available diagnostic history in the inference period, and produces an estimate of the relative risk of ADRD, with an associated confidence value. To learn the parameters and associated model structures of this pipeline, we transform the patient specific data to a set of engineered features, and the feature vectors realized on the positive and control sets are used to train a gradient-boosting classifier<sup>64</sup>. The complete list of 701 features used is provided in Tab. VI.

We need two training sets: one to infer the models, and one to train the classifier with features derived from the inferred models. Thus, we do a random 3-way split of the set of unique patients into *feature-engineering* (25%), *training* (25%) and *test* (50%) sets. We use the feature-engineering set of ids first to infer our PFSA models (*unsupervised model inference in each category*), which then allows us to train the gradient-boosting classifier using the training set and PFSA models (*classical supervised learning*), and we finally execute out-of-sample validation on the test set. Fig. 2c in the main text shows the top 20 features ranked in order of their relative importance (relative loss in performance when dropped out of the analysis).

In addition to the phenotype specific specialized Markov models, we use a range of engineered features reflecting various aspects of diagnostic histories:

*Prevalence scores (p-scores)*: The p-scores focus on individual diagnostic codes, and we create a dictionary of the ratio of relative prevalence of each code (relative to the set of all codes present) in the positive category (for each sex) to the control category. This is the second hyper-training step. In the later steps of the pipeline,

---

we use dictionary look ups to map codes to their p-scores, and also their aggregate measures such as mean, median, and variance to train a downstream LGBM.

*Rare scores:* These scores consist of a subset of p-scores which correspond to codes with particularly high and low relative prevalences (p-score  $> 2$  or  $< .5$ ). Thus, this feature category depends on the p-score dictionary generated in the second hyper-training step.

*Sequence scores:* Sequence scores are relatively straight-forward statistical measures such as mean, median, variance, time since last occurrence *etc.* on the trinary phenotype-specific sex-stratified histories. No hyper-training is required for the generation of the sequence features.

Thus we require three splits of the training dataset. The first split is used to carry out hyper-training of the PFSA models and the p-score dictionary. The second split is used to train the score-category specific LGBMs, one for each feature category. And the third split is used to train the final LGBM that takes inputs from the outputs of the four LGBMs in the previous layer. The network layout is shown in SI-Fig. 2.

### **Validation**

In validation, or actual prediction of patient fate, we use the trinary mapping, generate the features using the PFSA models and the p-score dictionary, and calculate the raw-risk via the trained LGBM network. The relative score is then obtained by a choice of the operating point reflecting the specificity/sensitivity trade-off discussed before.

### **Data Splits: Training and Validation Hold-out**

All eligible patients are randomly split into the Training set ( $\approx 75\%$  of data) and the Test set ( $\approx 25\%$  of data). The Training set is then split into 3 subsets: 1) The hyper-training set (SI-Fig. 2A) is used to train PFSA models p-score dictionary, 2) the second split (referred to as the pre-aggregation split, SI-Fig. 2B) is used to train the four feature-category specific LGBMs, and 3) the final split (referred to as the aggregation split, SI-Fig. 2C) is used to train the aggregate LGBM which uses outputs from the trained LGBMs in the previous layer. This trained pipeline is then validated on the held out validation split ( $\approx 25\%$  of data).

### **Generating PFSA Models From Set of Input Streams with Variable Input Lengths**

Our PFSA reconstruction algorithm<sup>59</sup> is distinct from standard HMM learning. We do not need to pre-specify structures, or the number of states in the algorithm, and all model parameters are inferred directly from data. Additionally, we can operate either with 1) a single input stream, or 2) a set of input streams of possibly varying lengths which are assumed to be different and independent sample paths from the unknown stochastic generator we are trying to infer. At an intuitive level, we use the input data to infer the length of histories one must remember to estimate the current state, and predict futures for the process being modeled. Thus, we do not step through the symbol streams with a pre-specified model structure, and avoid the need to have equal-length inputs. More details of the algorithm are provided in the next section.

The ability to model a set of input streams of varying lengths is particularly important, since medical histories of different patients are typically of different lengths.

### **Probabilistic Finite State Automata Inference**

Software for PFSA inference is mader available at <https://pypi.org/project/zedsuite/>.

Let  $\Sigma$  be a finite alphabet of symbols with size  $|\Sigma|$ . The set of sequences of length  $d$  over  $\Sigma$  is denoted by  $\Sigma^d$ . The set of finite but unbounded sequences over  $\Sigma$  is denoted by  $\Sigma^*$ , the Kleene star operation<sup>65</sup>, i.e.  $\Sigma^* = \bigcup_{d=0}^{\infty} \Sigma^d$ . We use lower case Greek, for example  $\sigma$  or  $\tau$ , for symbols in  $\Sigma$ , and lower case Latin, for example  $x$  or  $y$ , for sequences of symbols, i.e.  $x = \sigma_1\sigma_2 \dots \sigma_n$ . We use  $|x|$  to denote the length of  $x$ . The empty sequence is denoted by  $\lambda$ .

We denote the set of strictly infinite sequences over  $\Sigma$  by  $\Sigma^\omega$ , and the set of strictly infinite sequences having  $x$  as prefix by  $x\Sigma^\omega$ . Let  $\mathcal{S} = \{x\Sigma^\omega : x \in \Sigma^*\} \cup \{\emptyset\}$ , we can verify that  $\mathcal{S}$  is a semiring<sup>66</sup> over  $\Sigma^\omega$ . We use  $\mathcal{F}$  to denote the sigma algebra generated by  $\mathcal{S}$ .

**Definition 1** (Stochastic Process over  $\Sigma$ ). *A stochastic process over a finite alphabet  $\Sigma$  is a collection of  $\Sigma$ -valued random variables  $\{X_t\}_{t \in \mathbb{N}}$  indexed by positive integers<sup>67</sup>.*

We are specifically interested in processes in which the  $X_t$ s are not necessarily independently distributed.



**Definition 2** (Sequence-Induced Measure and Derivative). For a process  $\mathcal{P}$ , let  $\Pr_{\mathcal{P}}(x)$  or simply  $\Pr(x)$  denote the probability  $\mathcal{P}$  producing a sample path prefixed by  $x$ . The **measure**  $\mu_x$  **induced by a sequence**  $x \in \Sigma^*$  is the extension<sup>66</sup> to  $\mathcal{F}$  of the premeasure defined on the semiring  $\mathcal{S}$  given by

$$\forall x, y \in \Sigma^*, \mu_x(y\Sigma^\omega) \triangleq \frac{\Pr(xy)}{\Pr(x)}, \text{ if } \Pr(x) > 0 \quad (6)$$

For any  $d \in \mathbb{N}$ , the  **$d$ -th order derivative** of a sequence  $x$ , written as  $\phi_x^d$ , is defined to be the marginal distribution of  $\mu_x$  on  $\Sigma^d$ , with the entry indexed by  $y$  denoted by  $\phi_x^d(y)$ . The first-order derivative is called the **symbolic derivative** and is denoted by  $\phi_x$  for short.

**Definition 3** (Probabilistic Nerode Equivalence and Causal States<sup>68</sup>). For any pair of sequences  $x, y \in \Sigma^*$ ,  $x$  is equivalent to  $y$ , written as  $x \sim y$ , if and only if either  $\Pr(x) = \Pr(y) = 0$ , or  $\mu_x = \mu_y$ . The equivalence class of a sequence  $x$  is denoted by  $[x]$  and is called a **causal state**<sup>69</sup>. The cardinality of the set of causal states is called the **probabilistic Nerode index**, or the Nerode index for simplicity.

We can see from the definition that causal states captures how the history of a process influences its future. Since the probabilistic Nerode equivalence is right invariant, it gives rise naturally to a automaton structure introduced below.

**Definition 4** (Probabilistic Finite-State Automaton (PFSA)). A PFSA  $G$  is defined by a quadruple  $(Q, \Sigma, \delta, \tilde{\pi})$ , where  $Q$  is a finite set,  $\Sigma$  is a finite alphabet,  $\delta: Q \times \Sigma \rightarrow \Sigma$  is called the transition map, and  $\tilde{\pi}: Q \rightarrow \mathbf{P}_\Sigma$ , where  $\mathbf{P}_\Sigma$  is the space of probability distributions over  $\Sigma$ , is called the transition probability. The entry of  $\tilde{\pi}(q)$  indexed by  $\sigma$  is denoted by  $\tilde{\pi}(q, \sigma)$ .

**Definition 5** (Transition and Observation Matrices). The transition matrix  $\Pi$  is the  $|Q| \times |Q|$  matrix with the entry indexed by  $q, q'$ , written as  $\pi_{q, q'}$ , satisfying

$$\pi_{q, q'} \triangleq \sum_{\{\sigma \in \Sigma \mid \delta(q, \sigma) = q'\}} \tilde{\pi}(q, \sigma) \quad (7)$$

and the observation matrix  $\tilde{\Pi}$  is a  $|Q| \times |\Sigma|$  matrix with the entry indexed by  $q, \sigma$  equaling  $\tilde{\pi}(q, \sigma)$ .

We note that both  $\Pi$  and  $\tilde{\Pi}$  are stochastic, i.e. non-negative with rows summing up to 1.

**Definition 6** (Extension of  $\delta$  and  $\tilde{\pi}$  to  $\Sigma^*$ ). For any  $x = \sigma_1 \dots \sigma_k$ ,  $\delta(q, x)$  is defined recursively by

$$\delta(q, x) \triangleq \delta(\delta(q, \sigma_1 \dots \sigma_{k-1}), \sigma_k) \quad (8)$$

with  $\delta(q, \lambda) = q$ , and  $\tilde{\pi}(q, x)$  is defined recursively by

$$\tilde{\pi}(q, x) \triangleq \prod_{i=1}^k \tilde{\pi}(\delta(q, \sigma_1 \dots \sigma_{i-1}), \sigma_i) \quad (9)$$

with  $\tilde{\pi}(q, \lambda) = 1$ .

**Definition 7** (Strongly Connected PFSA). We say a PFSA is strongly connected if the underlying directed graph is strongly connected<sup>70</sup>. More precisely, a PFSA  $G = (Q, \Sigma, \delta, \tilde{\pi})$  is strongly connected if for any pair of distinct states  $q$  and  $q' \in Q$ , there is an  $x \in \Sigma^*$  such that  $\delta(q, x) = q'$ .

We assume all PFSA in the discussions in the sequel are strongly connected if not specified otherwise. For strongly connected PFSA  $G$ , there is a unique probability distribution over  $Q$  that satisfies  $\mathbf{v}^T \Pi = \mathbf{v}^T$ . This is the **stationary distribution**<sup>71,72</sup> of  $G$  and is denoted as  $\wp_G$ , or  $\wp$  if  $G$  is understood.

**Definition 8** ( $\Gamma$ -Expression). We can encode the information contained in  $\delta$  and  $\tilde{\pi}$  by a set of  $|Q| \times |Q|$  matrices  $\Gamma = \{\Gamma_\sigma \mid \sigma \in \Sigma\}$ , where

$$\Gamma_\sigma \Big|_{q, q'} \triangleq \begin{cases} \tilde{\pi}(q, \sigma) & \text{if } \delta(q, \sigma) = q', \\ 0 & \text{if otherwise.} \end{cases} \quad (10)$$

$\Gamma_\sigma$  is called **event-specific transition matrix**, with the event being that  $\sigma$  is current the output.  $\Gamma_\sigma$  can also be extended to arbitrary  $x \in \Sigma^*$  by defining  $\Gamma_x = \prod_{i=1}^k \Gamma_{\sigma_i}$  with  $\Gamma_\lambda = I$ .

**Definition 9** (Sequence-Induced Distribution on States). For a PFSA  $G = (Q, \Sigma, \delta, \tilde{\pi})$  and a distribution  $\wp_0$  on  $Q$ , the **distribution on  $Q$  induced by a sequence**  $x$  is given by  $\wp_{G, \wp_0}^T(x) = \left[ \wp_0^T \Gamma_x \right]$  with  $\wp_{G, \wp_0}(\lambda) = \wp_0$ . The entry indexed by  $q \in Q$  of the vector  $\wp_{G, \wp_0}(x)$  is written as  $\wp_{G, \wp_0}(x, q)$ . When  $\wp_0 = \wp_G$ , the stationary distribution of  $G$ , we write  $\wp_{G, \wp_0}(x)$  as  $\wp_G(x)$ , or simply as  $\wp(x)$ , if  $G$  is understood.

**Definition 10** (Stochastic Process Generated by a PFSA). Let  $G = (Q, \Sigma, \delta, \tilde{\pi})$  be a PFSA and let  $\wp_0$  be a distribution on  $Q$ , the  $\Sigma$ -valued stochastic process  $\{X_t\}_{t \in \mathbb{N}}$  generated by  $G$  and  $\wp_0$  satisfies that  $X_1$  follows the distribution  $\wp_0$  and  $X_{t+1}$  follows the distribution  $\wp_{G, \wp_0}(X_1 \dots X_t)$  for  $t \in \mathbb{N}$ .

For the rest of this paper, we will assume  $\varrho_0 = \varrho_G$  if not specified otherwise. We can show that, when initialized with  $\varrho_G$ , the process generated by a PFSA  $G$  is stationary and ergodic. We also note that, for the process generated by  $G$ , we have  $\phi_x = \varrho_G(x)^T \tilde{\Pi}$ . Since  $\varrho_G(\lambda) = \varrho_G$ , the symbolic derivative of the empty sequence  $\phi_\lambda$  is the stationary distribution on the symbols.

**Definition 11** (Synchronizable PFSA and Synchronizing Sequence). A **synchronizing sequence** is a finite sequence that sends an arbitrary state of the PFSA to a fixed state<sup>73</sup>. To be more precise, let  $G = (Q, \Sigma, \delta, \tilde{\Pi})$  be a PFSA, we say a sequence  $x \in \Sigma^*$  is a synchronizing sequence to a state  $q \in Q$  if  $\delta(q', x) = q$  for all  $q' \in Q$ . A PFSA is **synchronizable** if it has at least one synchronizing sequence. Given a sample path generated by a PFSA, we say the PFSA is **synchronized** if a synchronizing sequence transpires in the sample path.

**Definition 12** (Equivalence and Irreducibility). Two PFSA  $G$  and  $H$  are **equivalent** if they generate the same stochastic process. A PFSA  $G$  is said to be **irreducible**, if there is not another PFSA with smaller state set that is equivalent to  $G$ .

**Definition 13**. Consider a PFSA  $G$  over state set  $Q$ . For a given  $\varepsilon > 0$ , we say a sequence  $x$  is a  $\varepsilon$ -synchronizing sequence to a state  $q \in Q$  if

$$\|\varrho_G(x) - \mathbf{e}_q\|_\infty \leq \varepsilon. \quad (11)$$

While there exists PFSA that is not synchronizable, we can show that an irreducible PFSA always has an  $\varepsilon$ -synchronizing sequence for some state  $q$  for arbitrarily small  $\varepsilon > 0$ . Moreover, we can show that as length increases, sequences produced by PFSA become uniformly  $\varepsilon$ -synchronizing. These two are the underpinning properties for the inference algorithm of PFSA (See Alg. 1), because they imply that  $\phi_x$  can be used to approximate  $\tilde{\pi}(q)$  if  $x$  are properly prefixed and long enough.

**Definition 14** (Joint  $\varepsilon$ -Synchronizing Sequence). Let  $G$  and  $H$  be two PFSA over state sets  $Q_G$  and  $Q_H$ , respectively. For a fixed  $\varepsilon$ , a sequence  $x$  is said to be **jointly  $\varepsilon$ -synchronizing** to  $(q, r) \in Q_G \times Q_H$  if  $x$  is  $\varepsilon$ -synchronizing to  $q$  and to  $r$  simultaneously. We define

$$\Sigma_{\varepsilon, (q, r)}^d \triangleq \{x \in \Sigma^d : x \text{ jointly } \varepsilon\text{-synchronizing to } (q, r)\} \quad (12)$$

**Definition 15** (Joint Pair of States). Let  $G$  and  $H$  be two PFSA over state sets  $Q_G$  and  $Q_H$ , respectively. Define

$$p_G(q, r) \triangleq \lim_{d \rightarrow \infty} p_G(\Sigma_{\varepsilon, (q, r)}^d) \quad (13)$$

A pair of states  $(q, r) \in Q_G \times Q_H$  is called a **G-joint pair** of states if  $p_G(q, r) > 0$ . We also define

$$Q_G \triangleq \{(q, r) \in Q_G \times Q_H : (q, r) \text{ is a G-joint pair}\} \quad (14)$$

The inference algorithm for PFSA is called **GenESeSS** for Generator Extraction Using Self-similar Semantics. With an input sequence  $x$  and a hyperparameter  $\varepsilon$ , **GenESeSS** outputs a PFSA in the following three steps: 1) approximate an almost synchronizing sequence; 2) identify the transition structure of the PFSA; 3) calculate the transition probabilities of the PFSA. See Alg. 1<sup>59</sup> for details.

#### Theoretical Development of Sequence Likelihood Defect

**Definition 16** (Entropy Rate and KL Divergence). By **entropy rate** of a PFSA, we mean the entropy rate of the stochastic process generated by the PFSA<sup>74</sup>. Similarly, by **KL divergence** of two PFSA, we mean the KL divergence between the two processes generated by them<sup>75</sup>. More precisely, we have

$$\mathcal{H}(G) = - \lim_{d \rightarrow \infty} \frac{1}{d} \sum_{x \in \Sigma^d} p(x) \log p(x) \quad (15)$$

and the KL divergence

$$\mathcal{D}_{KL}(G \parallel H) = \lim_{d \rightarrow \infty} \frac{1}{d} \sum_{x \in \Sigma^d} p_G(x) \log \frac{p_G(x)}{p_H(x)} \quad (16)$$

whenever the limits exist.

**Theorem 1** (Closed-form Formula for Entropy Rate and KL Divergence). The entropy rate of a PFSA  $G = (\Sigma, Q, \delta, \tilde{\Pi})$  is given by

$$\mathcal{H}(G) = \sum_{q \in Q} \varrho_G(q) \cdot h(\tilde{\pi}(q)) \quad (17)$$

where  $h(\mathbf{v})$  is the based-2 entropy of the probability vector  $\mathbf{v}$ .

Consider two PFSA  $G = (Q_G, \Sigma, \delta_G, \tilde{\Pi}_G)$  and  $H = (Q_H, \Sigma, \delta_H, \tilde{\Pi}_H)$  with  $\mu_G$  being absolutely continuous with

respect to  $\mu_H$ . Let  $Q_c$  be the set of  $G$ -joint pairs of states, we have

$$\mathcal{D}_{KL}(G \parallel H) = \sum_{(q,r) \in Q_c} p_G(q,r) \mathcal{D}_{KL}(\tilde{\pi}_G(q) \parallel \tilde{\pi}_H(r)) \quad (18)$$

**Definition 17** (Log-likelihood). Let  $x \in \Sigma^d$ , the log-likelihood<sup>74</sup> of a PFSA  $G$  generating  $x$  is given by

$$L(x, G) = -\frac{1}{d} \log p_G(x) \quad (19)$$

The calculation of log-likelihood is detailed in Alg. 2.

**Theorem 2** (Convergence of log-likelihood). Let  $G$  and  $H$  be two reduced PFSA, and let  $x \in \Sigma^d$  be a sequence generated by  $G$ . Then we have

$$L(x, H) \rightarrow \mathcal{H}(G) + \mathcal{D}_{KL}(G \parallel H) \quad (20)$$

in probability as  $d \rightarrow \infty$ .

*Proof.* We first notice that

$$\sum_{x \in \Sigma^d} p_G(x) \log \frac{p_G(x)}{p_H(x)} = \sum_{x \in \Sigma^{d-1}} \sum_{\sigma \in \Sigma} p_G(x) \wp_G(x) \tilde{\Pi}_G \Big|_{\sigma} \log \frac{p_G(x) \wp_G(x) \tilde{\Pi}_G \Big|_{\sigma}}{p_H(x) \wp_H(x) \tilde{\Pi}_H \Big|_{\sigma}} \quad (21)$$

$$= \sum_{x \in \Sigma^{d-1}} p_G(x) \log \frac{p_G(x)}{p_H(x)} + \underbrace{\sum_{x \in \Sigma^{d-1}} p_G(x) \sum_{\sigma \in \Sigma} \wp_G(x) \tilde{\Pi}_G \Big|_{\sigma} \log \frac{\wp_G(x) \tilde{\Pi}_G \Big|_{\sigma}}{\wp_H(x) \tilde{\Pi}_H \Big|_{\sigma}}}_{D_d} \quad (22)$$

By induction, we have  $\mathcal{D}_{KL}(G \parallel H) = \lim_{d \rightarrow \infty} \frac{1}{d} \sum_{i=1}^d D_i$ , and hence by Cesàro summation theorem<sup>76</sup>, we have  $\mathcal{D}_{KL}(G \parallel H) = \lim_{d \rightarrow \infty} D_d$ . Let  $x = \sigma_1 \sigma_2 \dots \sigma_n$  be a sequence generated by  $G$ . Let  $x^{[i-1]}$  is the truncation of  $x$  at the  $(i-1)$ -th symbols, we have

$$-\frac{1}{n} \sum_{i=1}^n \log \wp_H(x^{[i-1]}) \tilde{\Pi}_H \Big|_{\sigma_i} = \underbrace{\frac{1}{n} \sum_{i=1}^n \log \frac{\wp_G(x^{[i-1]}) \tilde{\Pi}_G \Big|_{\sigma_i}}{\wp_H(x^{[i-1]}) \tilde{\Pi}_H \Big|_{\sigma_i}}}_{A_{x,n}} - \underbrace{\frac{1}{n} \sum_{i=1}^n \log \wp_G(x^{[i-1]}) \tilde{\Pi}_G \Big|_{\sigma_i}}_{B_{x,n}} \quad (23)$$

Since the stochastic process  $G$  generates is ergodic, we have

$$\lim_{n \rightarrow \infty} A_{x,n} = \lim_{d \rightarrow \infty} D_d = \mathcal{D}_{KL}(G \parallel H) \quad (24)$$

and  $\lim_{n \rightarrow \infty} B_{x,n} = \mathcal{H}(G)$ .  $\square$

## Quantification & Statistical Analysis

### Raw Risk & Relative Risk

We choose a decision threshold on the raw risk computed by our pipeline to make crisp predictions, i.e., if the raw risk is greater than this calibrated threshold, then the individual patient is predicted to be in the positive category.

### Threshold Selection on ROC Curve

In situations where the number of negatives vastly outnumber the number of positives (which is the case in our problem), it is better to base this trade-off on a measure that is independent of the number of true negatives. The two popular measures considered in the literature are accuracy and the F1-score:

$$\text{accuracy} = \frac{t_p + t_n}{t_p + f_p + f_n + t_n} \quad (25)$$

$$F1 = \frac{2t_p}{2t_p + f_p + f_n} \quad (26)$$

The F1-score is the same as accuracy where the number of true negatives is the same as the number of true positives, thus partially correcting for the class imbalance.

The selection of the threshold may also be dictated by the current practice of ensuring high specificities in screening tests. Thus, the most relevant clinically operating point is either the one corresponding to 95% specificity, which is highlighted in Fig. 2a.

### Performance Measurement

We measure our performance using standard metrics including the AUC, sensitivity, specificity, the positive predictive value (PPV), and the negative predictive value (NPV). We also report accuracy (acc, See Tables VII and VIII), which is the probability of a correct prediction (positive or control), and variation of AUC for predicting ADRD into the future up to 10 years (See Table IX).

Ninety-five percent confidence intervals (95% CIs) on ROC curves and AUCs were obtained via bootstrapping, and AUC p-values were obtained using the Mann-Whitney U-test statistic.

### Note on Receiver Operating Characteristics (ROC) and Precision-recall Curves

The ROC curve is a plot between the False Positive rate (FPR) and the True Positive Rate (TPR), and the area under the ROC curve (AUC) is often used as a measure of classifier performance. For the same of completeness, we introduce the relevant definitions:

In the following  $P$  denotes the total number of positive samples (number of patients who are eventually diagnosed), and  $N$  denotes the total number of negative samples (number of patients in the control group).

**Definition 18.** True positive rate, true negative rate, false positive rate, positive predictive value (PPV), and prevalence ( $\rho$ ) are defined as:

$$\text{sensitivity, or TPR} = \frac{t_p}{P} = \frac{t_p}{t_p + f_n} \quad (27)$$

$$\text{specificity, or TNR} = \frac{t_n}{N} = \frac{t_n}{t_n + f_p} \quad (28)$$

$$\text{FPR} = 1 - \text{TNR} \quad (29)$$

$$\text{precision, or PPV} = \frac{t_p}{t_p + f_p} \quad (30)$$

$$\rho = \frac{P}{N + P} \quad (31)$$

where as before  $t_p, t_n, f_p, f_n$  are true positives, true negatives, false positives, and false negatives respectively.

Denoting sensitivity by  $s$ , and specificity by  $c$ , it follows that:

$$\text{PPV} = \frac{t_p/P}{t_p/P + (f_p/N)(N/P)} = \frac{\text{TPR}}{\text{TPR} + ((N - t_n)/N)(N/P)} \quad (32)$$

$$\Rightarrow \text{PPV} = \frac{s}{s + (1 - c)(\frac{1}{\rho} - 1)} \quad (33)$$

Thus, we note that for a fixed specificity and sensitivity, the PPV depends on prevalence. Indeed, it is clear from the above argument that PPV decreases with decreasing prevalence, and vice versa.

### Effect of Class Imbalance

ROC curves are generally assumed to be robust to class imbalance. Note that if we assume that patient outcomes are independent (which is well-justified in the case of a non-communicable condition, particularly in large databases), then  $t_p$  should scale linearly with the total number of positives  $P$ , implying:

$$\text{TPR} = \frac{t_p}{P} = \frac{t'_p}{P'} \quad (34)$$

implying that with different sizes of the set of positive samples (or negative samples), the ROC curve remains unchanged. In particular, note that even if the prevalence is very small (say 0.01%), we cannot cheat to boost the AUC by labeling all predictions as negative, or stating that risk is always zero: in that case, our  $P$  is very small, but our  $t_p = 0$  strictly, implying that our  $\text{TPR} = 0$ , thus leading to a zero AUC. We can cheat to boost the accuracy (See the previous section), but not the AUC.

Note that while relative class sizes or imbalance does not affect the ROC (under the assumption that true positives and true negatives scale with the number of positives and negatives), very small absolute sample sizes might still result in poor performance of the model.

The precision-recall curves do get affected by class imbalance, or the prevalence, as shown by Eq (33). However, in diagnostic analysis, they are important since we are generally less interested in the number of true negatives; the ratio of false positives to the total number of positive recommendations by the algorithm is much more relevant, i.e., the PPV or the precision.

---

## REFERENCES

- [1] Arvanitakis, Z., Shah, R. C. & Bennett, D. A. Diagnosis and management of dementia. *Jama* **322**, 1589–1599 (2019).
- [2] Moyer, V. A. Screening for cognitive impairment in older adults: Us preventive services task force recommendation statement. *Annals of internal medicine* **160**, 791–797 (2014).
- [3] Owens, D. K. *et al.* Screening for cognitive impairment in older adults: Us preventive services task force recommendation statement. *Jama* **323**, 757–763 (2020).
- [4] Chu, L. *et al.* Alzheimer's disease: early diagnosis and treatment. *Hong Kong Med J* **18**, 228–237 (2012).
- [5] Association, A. *et al.* Alzheimer's disease facts and figures. *Alzheimer's & dementia* **13**, 325–73 (2017).
- [6] Murray, C. J. *et al.* The state of us health, 1990-2010: burden of diseases, injuries, and risk factors. *Jama* **310**, 591–606 (2013).
- [7] Jack Jr, C. R. *et al.* NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia* **14**, 535–562 (2018).
- [8] Ahlgrim, N. S., Garza, K., Hoffman, C. & Rommelfanger, K. S. Prodromes and preclinical detection of brain diseases: Surveying the ethical landscape of predicting brain health. *Eneuro* **6** (2019).
- [9] Patnode, C. D. *et al.* Screening for cognitive impairment in older adults: updated evidence report and systematic review for the us preventive services task force. *Jama* **323**, 764–785 (2020).
- [10] Borson, S. *et al.* Implementing routine cognitive screening of older adults in primary care: process and impact on physician behavior. *Journal of general internal medicine* **22**, 811–817 (2007).
- [11] Davidson, M. & Thibaut, F. Is dementia a preventable disease? *Dialogues in clinical neuroscience* **21**, 3 (2019).
- [12] Mattsson-Carlgen, N. *et al.* Longitudinal plasma p-tau217 is increased in early stages of alzheimer's disease. *Brain* **143**, 3234–3241 (2020).
- [13] Moscoso, A. *et al.* Longitudinal associations of blood phosphorylated tau181 and neurofilament light chain with neurodegeneration in alzheimer disease. *JAMA Neurology* .
- [14] Karikari, T. K. *et al.* Diagnostic performance and prediction of clinical progression of plasma phospho-tau181 in the Alzheimer's Disease Neuroimaging Initiative. *Molecular Psychiatry* 1–14 (2020).
- [15] Hall, S. *et al.* Plasma phospho-tau identifies Alzheimer's co-pathology in patients with lewy body disease. *Movement Disorders* (2020).
- [16] Janelidze, S. *et al.* Plasma p-tau181 in alzheimer's disease: relationship to other biomarkers, differential diagnosis, neuropathology and longitudinal progression to alzheimer's dementia. *Nature medicine* **26**, 379–386 (2020).
- [17] Bezdicek, O. *et al.* Determining a short form montreal cognitive assessment (s-moca) czech version: validity in mild cognitive impairment parkinson's disease and cross-cultural comparison. *Assessment* **27**, 1960–1970 (2020).
- [18] Lab, S. R. A. Montreal cognitive assessment in rehabmeasures database. <https://www.sralab.org/rehabilitation-measures/montreal-cognitive-assessment> (2020). (Accessed on 02/14/2021).
- [19] Nasreddine, Z. S. *et al.* The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society* **53**, 695–699 (2005).
- [20] Zhao, L. Alzheimer's disease facts and figures. *Alzheimers Dement* **16**, 391–460 (2020).
- [21] Wilkinson, T. *et al.* Identifying dementia cases with routinely collected health data: a systematic review. *Alzheimer's & Dementia* **14**, 1038–1051 (2018).
- [22] Duthie, A., Chew, D. & Soiza, R. Non-psychiatric comorbidity associated with Alzheimer's disease. *QJM: An International Journal of Medicine* **104**, 913–920 (2011).
- [23] Anstey, K. J. *et al.* Future directions for dementia risk reduction and prevention research: An international research network on dementia prevention consensus. *Journal of Alzheimer's Disease* 1–10 (2020).
- [24] Sipilä, P. N. *et al.* Hospital-treated infectious diseases and the risk of dementia: multicohort study with replication in the uk biobank. *medRxiv* (2020).
- [25] Muzambi, R. *et al.* Assessment of common infections and incident dementia using UK primary and secondary care data: a historical cohort study. *The Lancet Healthy Longevity* (2021).
- [26] Eavani, H. *et al.* Heterogeneity of structural and functional imaging patterns of advanced brain aging revealed via machine learning methods. *Neurobiology of aging* **71**, 41–50 (2018).
- [27] Montreal Cognitive Assessment Webpage. Normative test. <https://www.mocatest.org/moca-clinic-data/>. (Accessed on 05/04/2021).
- [28] Ferretti, M. T. *et al.* Sex and gender differences in alzheimer's disease: current challenges and implications for clinical practice: position paper of the dementia and cognitive disorders panel of the european academy of Neurology. *European journal of Neurology* **27**, 928–943 (2020).
- [29] Choi, J., Kwon, L.-N., Lim, H. & Chun, H.-W. Gender-based analysis of risk factors for dementia using

- senior cohort. *International journal of environmental research and public health* **17**, 7274 (2020).
- [30] Exalto, L. G. *et al.* Risk score for prediction of 10 year dementia risk in individuals with type 2 diabetes: a cohort study. *The Lancet Diabetes & Endocrinology* **1**, 183–190 (2013).
- [31] Reitz, C. *et al.* A summary risk score for the prediction of alzheimer disease in elderly persons. *Archives of Neurology* **67**, 835–841 (2010).
- [32] Barnes, D. E. *et al.* Development and validation of a brief dementia screening indicator for primary care. *Alzheimer's & Dementia* **10**, 656–665 (2014).
- [33] Tang, E. Y. *et al.* Current developments in dementia risk prediction modelling: an updated systematic review. *PloS one* **10**, e0136181 (2015).
- [34] Chary, E. *et al.* Short-versus long-term prediction of dementia among subjects with low and high educational levels. *Alzheimer's & Dementia* **9**, 562–571 (2013).
- [35] Ohara, T. *et al.* Apolipoprotein genotype for prediction of Alzheimer's disease in older japanese: the hisayama study. *Journal of the American Geriatrics Society* **59**, 1074–1079 (2011).
- [36] Boustani, M. *et al.* Passive digital signature for early identification of Alzheimer's disease and related dementia. *Journal of the American Geriatrics Society* **68**, 511–518 (2020).
- [37] Park, J. H. *et al.* Machine learning prediction of incidence of alzheimer's disease using large-scale administrative health data. *NPJ digital medicine* **3**, 1–7 (2020).
- [38] So, A., Hooshyar, D., Park, K. W. & Lim, H. S. Early diagnosis of dementia from clinical data by machine learning techniques. *Applied Sciences* **7**, 651 (2017).
- [39] Hansen, L. The truver health marketscan databases for life sciences researchers. *Truver Health Analytics IBM Watson Health* (2017).
- [40] FDA. Fda-approved treatments for Alzheimer's disease or treatments for Alzheimer's disease. <https://www.alz.org/media/documents/fda-approved-treatments-alzheimers-ts.pdf> (2019). (Accessed on 02/14/2021).
- [41] Tortajada-Soler, M. *et al.* Prevalence of comorbidities in individuals diagnosed and undiagnosed with alzheimer's disease in león, spain and a proposal for contingency procedures to follow in the case of emergencies involving people with alzheimer's disease. *International journal of environmental research and public health* **17**, 3398 (2020).
- [42] Fouladvand, S. *et al.* Deep learning prediction of mild cognitive impairment using electronic health records. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 799–806 (IEEE, 2019).
- [43] Anheim, M. *et al.* Ataxic variant of alzheimer's disease caused by pro117ala psen1 mutation. *Journal of Neurology, Neurosurgery & Psychiatry* **78**, 1414–1415 (2007).
- [44] Piccini, A. *et al.* Association of a presenilin 1 s170f mutation with a novel alzheimer disease molecular phenotype. *Archives of Neurology* **64**, 738–745 (2007).
- [45] Humbert, I. A. *et al.* Early deficits in cortical control of swallowing in Alzheimer's disease. *Journal of Alzheimer's disease* **19**, 1185–1197 (2010).
- [46] Kai, K. *et al.* Relationship between eating disturbance and dementia severity in patients with alzheimer's disease. *PloS one* **10**, e0133666 (2015).
- [47] Gannon, O., Robison, L., Custozzo, A. & Zuloaga, K. Sex differences in risk factors for vascular contributions to cognitive impairment & dementia. *Neurochemistry international* **127**, 38–55 (2019).
- [48] Kim, S. E. *et al.* Sex-specific relationship of cardiometabolic syndrome with lower cortical thickness. *Neurology* **93**, e1045–e1057 (2019).
- [49] Elbejjani, M. *et al.* Depression, depressive symptoms, and rate of hippocampal atrophy in a longitudinal cohort of older men and women. *Psychological medicine* **45**, 1931 (2015).
- [50] Hua, X. *et al.* Sex and age differences in atrophic rates: an adni study with n= 1368 mri scans. *Neurobiology of aging* **31**, 1463–1480 (2010).
- [51] Irvine, K., Laws, K. R., Gale, T. M. & Kondel, T. K. Greater cognitive deterioration in women than men with Alzheimer's disease: a meta analysis. *Journal of clinical and experimental neuropsychology* **34**, 989–998 (2012).
- [52] of Neurology, A. A. Practice guideline update summary: Mild cognitive impairment. <https://www.aan.com/Guidelines/home/GuidelineDetail/881> (2018). (Accessed on 04/08/2021).
- [53] Belleville, S., Fouquet, C., Hudon, C., Zomahoun, H. T. V. & Croteau, J. Neuropsychological measures that predict progression from mild cognitive impairment to Alzheimer's type dementia in older adults: a systematic review and meta-analysis. *Neuropsychology review* **27**, 328–353 (2017).
- [54] Fowler, N. R. *et al.* Older primary care patients' attitudes and willingness to screen for dementia. *Journal of Aging Research* **2015** (2015).
- [55] Newcombe, E. A. *et al.* Inflammation: the link between comorbidities, genetics, and Alzheimer's disease. *Journal of neuroinflammation* **15**, 1–26 (2018).
- [56] Granger, C. W. J. & Joyeux, R. An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis* **1**, 15–29.

- 
- [57] General equivalence mappings (2019). URL [https://www.cms.gov/Medicare/Coding/ICD10/downloads/ICD-10\\_GEM\\_fact\\_sheet.pdf](https://www.cms.gov/Medicare/Coding/ICD10/downloads/ICD-10_GEM_fact_sheet.pdf).
- [58] Chattopadhyay, I. & Ray, A. Structural transformations of probabilistic finite state machines. *International Journal of Control* **81**, 820–835 (2008).
- [59] Chattopadhyay, I. & Lipson, H. Abductive learning of quantized stochastic processes with probabilistic finite automata. *Philos Trans A* **371**, 20110543 (2013).
- [60] Huang, Y. & Chattopadhyay, I. Data smashing 2.0: Sequence likelihood (sl) divergence for fast time series comparison. *arXiv preprint arXiv:1909.12243* (2019).
- [61] Cover, T. M. & Thomas, J. A. *Elements of Information Theory* (Wiley-Interscience, New York, NY, USA, 1991).
- [62] Kullback, S. & Leibler, R. A. On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86 (1951). URL <https://doi.org/10.1214/aoms/1177729694>.
- [63] Doob, J. *Stochastic Processes*. Wiley Publications in Statistics (John Wiley & Sons, 1953). URL <https://books.google.com/books?id=KvJQAAAAMAAJ>.
- [64] Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002). URL [http://dx.doi.org/10.1016/S0167-9473\(01\)00065-2](http://dx.doi.org/10.1016/S0167-9473(01)00065-2).
- [65] Hopcroft, J. E. *Introduction to automata theory, languages, and computation* (Pearson Education India, 2008).
- [66] Klenke, A. *Probability theory: a comprehensive course* (Springer Science & Business Media, 2013).
- [67] Doob, J. *Stochastic processes*. Wiley publications in statistics (Wiley, 1990). URL <https://books.google.com/books?id=7Bu8jgEACAAJ>.
- [68] Chattopadhyay, I. & Ray, A. Structural transformations of probabilistic finite state machines. *International Journal of Control* **81**, 820–835 (2008).
- [69] Chattopadhyay, I. & Lipson, H. Data smashing: uncovering lurking order in data. *Journal of The Royal Society Interface* **11**, 20140826 (2014).
- [70] Bondy, J. & Murty, U. Graph theory (2008). *Grad. Texts in Math* (2008).
- [71] Vidyasagar, M. *Hidden markov processes: Theory and applications to biology*, vol. 44 (Princeton University Press, 2014).
- [72] Kai, L. C. *Markov Chains: With Stationary Transition Probabilities* (Springer-Verlag, 1967).
- [73] Trahtman, A. N. The road coloring and Černý conjecture. In *Proc. of Prague stringology conference*, vol. 1, 12 (Citeseer, 2008).
- [74] Cover, T. M. & Thomas, J. A. *Elements of information theory* (John Wiley & Sons, 2012).
- [75] Matthews, A. G. d. G., Hensman, J., Turner, R. & Ghahramani, Z. On sparse variational methods and the kullback-leibler divergence between stochastic processes. *Journal of Machine Learning Research* **51**, 231–239 (2016).
- [76] Hardy, G. Divergent series, with a preface by je littlewood and a note by ls bosanquet, reprint of the revised (1963) edition. *Éditions Jacques Gabay, Sceaux* (1992).



## Key Resources Table

REAGENT	SOURCE	IDENTIFIER
Deposited data		
Truven Marketscan Database	IBM Watson® (with appropriate licensing)	<a href="https://www.ibm.com/watson-health/about/truven-health-analytics">https://www.ibm.com/watson-health/about/truven-health-analytics</a>
Small Patient Database	Excerpt from University of Chicago Medicine de-identified records between 2012-2021	<a href="https://github.com/zeroknowledgediscovery/EHRdata">https://github.com/zeroknowledgediscovery/EHRdata</a> <a href="https://doi.org/10.5281/zenodo.5348229">https://doi.org/10.5281/zenodo.5348229</a>
Software and algorithms		
PFSA inference algorithm implementation	Laboratory of Zero Knowledge Discovery ( <a href="http://zed.uchicago.edu">zed.uchicago.edu</a> )	<a href="https://pypi.org/project/zedsuite/">https://pypi.org/project/zedsuite/</a>
ZCoR modules	Laboratory of Zero Knowledge Discovery ( <a href="http://zed.uchicago.edu">zed.uchicago.edu</a> )	<a href="https://github.com/zeroknowledgediscovery/ZCOR-ADRD">https://github.com/zeroknowledgediscovery/ZCOR-ADRD</a> <a href="https://doi.org/10.5281/zenodo.5348219">https://doi.org/10.5281/zenodo.5348219</a>

# Supplementary Text: Rapid Universal Early Screening for Alzheimer's Disease and Related Dementia via Pattern Discovery in Diagnostic History

Dmytro Onishchenko<sup>1</sup>, Sam Searle<sup>7</sup>, Kenneth Rockwood<sup>7</sup>, James A. Mastrianni<sup>5,6</sup> and Ishanu Chattopadhyay<sup>1,2,3,4</sup>★

<sup>1</sup>Department of Medicine, University of Chicago, Chicago, IL USA

<sup>2</sup>Committee on Genetics, Genomics & Systems Biology, University of Chicago, Chicago, IL USA

<sup>3</sup>Committee on Quantitative Methods in Social, Behavioral, and Health Sciences, University of Chicago, Chicago, IL USA

<sup>4</sup>Center for Health Statistics, Department of Medicine, University of Chicago, Chicago, IL USA

<sup>5</sup>Department of Neurology, University of Chicago, Chicago, IL USA

<sup>6</sup>Committee on Neurobiology, University of Chicago, Chicago, IL USA

<sup>7</sup>Division of Geriatric Medicine, Department of Medicine, Department of Community Health and Epidemiology, School of Health Administration, Halifax, NS Canada

★To whom correspondence should be addressed: e-mail: [ishanu@u-chicago.edu](mailto:ishanu@u-chicago.edu).

## CONTENTS

### LIST OF TABLES

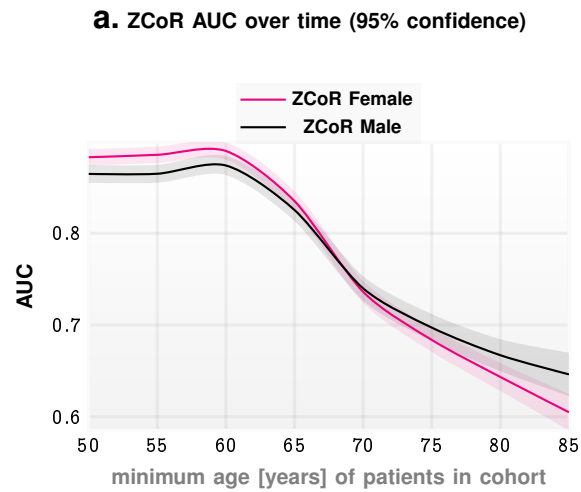
I	High risk cohort definition based on known comorbidities of AD and related dementia . . . . .	4
II	Disease Categories With Detailed Set of ICD Codes Used . . . . .	6

### LIST OF FIGURES

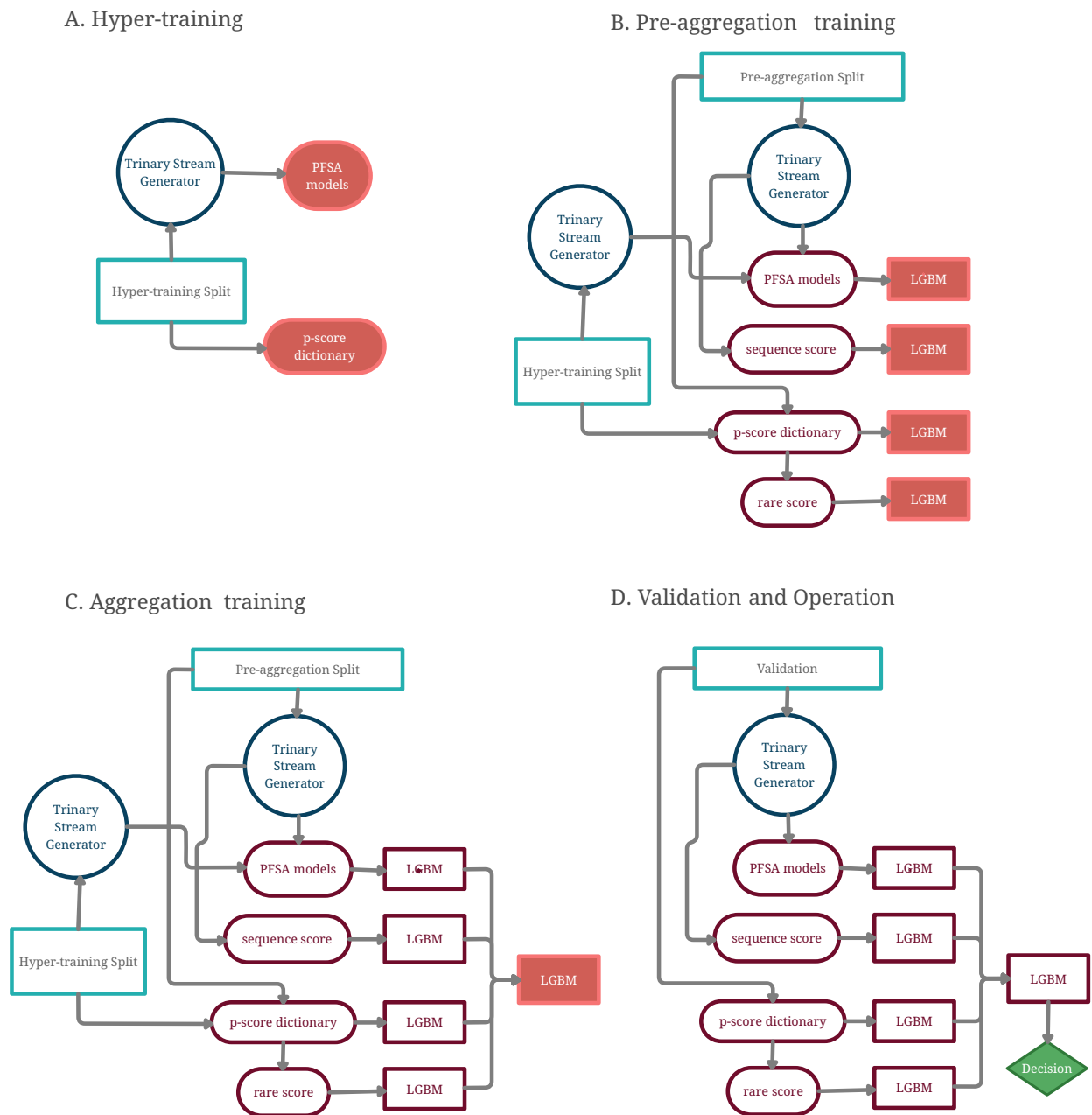
1	<b>Degradation of predictive performance with patient age.</b> With increasing patient age it become more difficult to distinguish age related cognitive decline from ADRD. This is reflected in the decreasing AUC with age, suggesting that comorbidity footprints associated with ADRD has confounding overlaps with conditions that arise more frequently as patients get older. . . . .	2
2	<b>Schematic of prediction pipeline.</b> Panels A, B and C show the sequential training steps, namely the hyper-training of the PFSA models and the p-score dictionary, the pre-aggregation training of the four LGBM models, and the aggregation training of the final LGBM model respectively. Panel D shows the configuration of the trained pipeline in operation. In the training steps, the filled box represents the component being trained in that step. . . . .	3

### LIST OF ALGORITHMS

1	<b>GenESeSS</b> . . . . .	18
2	Log-likelihood . . . . .	18



**SI-Fig. 1: Degradation of predictive performance with patient age.** With increasing patient age it become more difficult to distinguish age related cognitive decline from ADRD. This is reflected in the decreasing AUC with age, suggesting that comorbidity footprints associated with ADRD has confounding overlaps with conditions that arise more frequently as patients get older.



**SI-Fig. 2: Schematic of prediction pipeline.** Panels A, B and C show the sequential training steps, namely the hyper-training of the PFSA models and the p-score dictionary, the pre-aggregation training of the four LGBM models, and the aggregation training of the final LGBM model respectively. Panel D shows the configuration of the trained pipeline in operation. In the training steps, the filled box represents the component being trained in that step.

SI-Table I: High risk cohort definition based on known comorbidities of AD and related dementia

ICD code	description
250.0	DMII wo cmp nt st uncntr
250.02	DMII wo cmp uncntrld
252.0	Hyperparathyroidism NOS
252.02	Sec hyprprthyrd nonrenal
258.02	Mult endo neop type IIA
272.2	Mixed hyperlipidemia
278.0	Obesity NOS
278.01	Morbid obesity
278.02	Overweight
278.03	Obesity hypovent synd
296.2	Depress psychosis-unspec
296.21	Depress psychosis-mild
296.22	Depressive psychosis-mod
296.23	Depress psychosis-severe
296.24	Depr psychos-sev w psych
296.25	Depr psychos-part remiss
296.26	Depr psychos-full remiss
296.3	Recurr depr psychos-unsp
296.31	Recurr depr psychos-mild
296.32	Recurr depr psychos-mod
296.33	Recur depr psych-severe
296.34	Rec depr psych-psychotic
296.35	Recur depr psyc-part rem
296.36	Recur depr psyc-full rem
303.0	Ac alcohol intox-unspec
303.01	Ac alcohol intox-contin
303.02	Ac alcohol intox-episod
303.03	Ac alcohol intox-remiss
303.9	Alcoh dep NEC/NOS-unspec
303.91	Alcoh dep NEC/NOS-contin
303.92	Alcoh dep NEC/NOS-episod
303.93	Alcoh dep NEC/NOS-remiss
305.0	Alcohol abuse-unspec
305.01	Alcohol abuse-continuous
305.02	Alcohol abuse-episodic
305.03	Alcohol abuse-in remiss
401.0	Malignant hypertension
401.1	Benign hypertension
401.9	Hypertension NOS
402.0	Mal hyp ht dis w/o hf
402.01	Mal hypert hrt dis w hf
402.1	Benign hyp ht dis w/o hf
402.11	Benign hyp ht dis w hf
402.9	Hyp hrt dis NOS w/o hf
402.91	Hyp ht dis NOS w ht fail
403	Hypertensive chronic kidney disease
404	Hypertensive heart and chronic kidney disease
405.01	Mal renovasc hypertens
405.09	Mal second hyperten NEC
405.11	Benign renovasc hyperten
405.19	Benign second hypert NEC
405.91	Renovasc hypertension
405.99	Second hypertension NEC
427.31	Atrial fibrillation
440	Atherosclerosis
E11	Overweight
E66	Type 2 diabetes mellitus
E78	Disorders of lipoprotein metabolism lipidemias
F10.10	Alcohol abuse uncomplicated
F10.159	Alcohol abuse with alcohol-induced psychotic disorder unspecified
F10.20	Alcohol dependence uncomplicated
F10.21	Alcohol dependence in remission
F10.229	Alcohol dependence with intoxication unspecified
F10.231	Alcohol dependence with withdrawal delirium
F10.239	Alcohol dependence with withdrawal unspecified
F10.27	Alcohol dependence with alcohol-induced persisting dementia
F32.0	Major depressive disorder single episode mild
F32.1	Major depressive disorder single episode moderate
F32.2	Major depressive disorder single episode severe without psychotic features
F32.3	Major depressive disorder single episode severe with psychotic features
F32.4	Major depressive disorder single episode in partial remission

Continued on next page

SI-Table I – continued from previous page

ICD code	description
F32.5	Major depressive disorder single episode in full remission
F32.8	Premenstrual dysphoric disorder
F32.9	Major depressive disorder single episode unspecified
F33.0	Major depressive disorder recurrent mild
F33.1	Major depressive disorder recurrent moderate
F33.2	Major depressive disorder recurrent severe without psychotic features
F33.3	Major depressive disorder recurrent severe with psychotic symptoms
F33.41	Major depressive disorder recurrent in partial remission
F33.42	Major depressive disorder recurrent in full remission
F33.9	Major depressive disorder recurrent unspecified
G43.401	Hemiplegic migraine not intractable with status migrainosus
I10	Essential (primary) hypertension
I11.0	Hypertensive heart disease with heart failure
I11.9	Hypertensive heart disease without heart failure
I12.0	Hypertensive chronic kidney disease with stage 5 chronic kidney disease or end stage renal disease
I12.9	Hypertensive chronic kidney disease with stage 1 through stage 4 chronic kidney disease or unspecified chronic kidney disease
I13.0	Hypertensive heart and chronic kidney disease with heart failure and stage 1 through stage 4 chronic kidney disease or unspecified chronic kidney disease
I13.10	Hypertensive heart and chronic kidney disease without heart failure with stage 1 through stage 4 chronic kidney disease or unspecified chronic kidney disease
I13.11	Hypertensive heart and chronic kidney disease without heart failure with stage 5 chronic kidney disease or end stage renal disease
I13.2	Hypertensive heart and chronic kidney disease with heart failure and with stage 5 chronic kidney disease or end stage renal disease
I15.0	Renovascular hypertension
I15.8	Other secondary hypertension
I70.0	Atherosclerosis of aorta
I70.1	Atherosclerosis of renal artery
I70.209	Unspecified atherosclerosis of native arteries of extremities unspecified extremity
I70.219	Atherosclerosis of native arteries of extremities with intermittent claudication unspecified extremity
I70.229	Atherosclerosis of native arteries of extremities with rest pain unspecified extremity
I70.25	Atherosclerosis of native arteries of other extremities with ulceration
I70.269	Atherosclerosis of native arteries of extremities with gangrene unspecified extremity
I70.299	Other atherosclerosis of native arteries of extremities unspecified extremity
I70.399	Other atherosclerosis of unspecified type of bypass graft(s) of the extremities unspecified extremity
I70.499	Other atherosclerosis of autologous vein bypass graft(s) of the extremities unspecified extremity
I70.599	Other atherosclerosis of nonautologous biological bypass graft(s) of the extremities unspecified extremity
I70.8	Atherosclerosis of other arteries
I70.90	Unspecified atherosclerosis
I70.92	Chronic total occlusion of artery of the extremities
J12.0	Adenoviral pneumonia
J12.1	Respiratory syncytial virus pneumonia
J12.2	Parainfluenza virus pneumonia
J12.81	Pneumonia due to SARS-associated coronavirus
J12.89	Other viral pneumonia
J12.9	Viral pneumonia unspecified
J13	Pneumonia due to Streptococcus pneumoniae
J14	Pneumonia due to Hemophilus influenzae
J15.0	Pneumonia due to Klebsiella pneumoniae
J15.1	Pneumonia due to Pseudomonas
J15.20	Pneumonia due to staphylococcus unspecified
J15.211	Pneumonia due to Methicillin susceptible Staphylococcus aureus
J15.212	Pneumonia due to Methicillin resistant Staphylococcus aureus
J15.29	Pneumonia due to other staphylococcus
J15.3	Pneumonia due to streptococcus group B
J15.4	Pneumonia due to other streptococci
J15.5	Pneumonia due to Escherichia coli
J15.6	Pneumonia due to other Gram-negative bacteria
J15.7	Pneumonia due to Mycoplasma pneumoniae
J15.8	Pneumonia due to other specified bacteria
J15.9	Unspecified bacterial pneumonia
J16.0	Chlamydial pneumonia
J16.8	Pneumonia due to other specified infectious organisms
J17	Pneumonia in diseases classified elsewhere
K08.401	Partial loss of teeth unspecified cause class I
K08.402	Partial loss of teeth unspecified cause class II
S02.401A	Maxillary fracture unspecified side initial encounter for closed fracture
S02.401B	Maxillary fracture unspecified side initial encounter for open fracture
Y35.303A	Legal intervention involving unspecified blunt objects suspect injured initial encounter

SI-Table II: Disease Categories With Detailed Set of ICD Codes Used

	Description
Abnormal-Findings	<p><b>Constituent ICD9 Codes</b></p> <p>R89.5 R92.2 R71.0 794.7 R87.620 I20.8 R97.8 R82.5 794.01 R93.9 R87.612 I25.3 I21.11 R85.616 R94.30 R87.611 I25.42 R94.39 R92.8 R88.0 R93.1 R82.2 R85.81 R94.09 I25.10 R82.4 794.6 R87.811 794.31 R87.810 I25.811 R79.1 R80.2 794.4 794.09 I25.2 R87.616 R73.02 R74.8 R87.9 794.14 R83.9 R75 I21.29 794.00 794.10 794.19 R85.612 R93.5 R87.628 I21.4 794.13 R94.112 R78.89 R94.120 R79.81 794.39 R85.615 794.16 R94.31 R73.09 R94.8 R87.624 794.17 R87.623 R89.9 I25.83 R85.82 R93.2 R94.4 R82.99 R87.619 R94.110 R85.613 794.5 R76.8 R80.3 794.2 R86.9 R91.8 I25.9 R94.01 R71.8 R82.3 R97.0 I24.1 I25.41 R87.621 794.11 R87.613 R93.3 R94.113 R94.121 R94.131 R94.5 I24.0 R85.619 R93.4 R97.2 R76.12 R85.614 794.02 R94.111 794.15 R70.0 I21.19 I20.0 R82.0 794.8 I20.1 R79.82 R94.130 R87.820 R87.625 R85.611 R87.622 R78.0 I24.8 794.30 R94.7 R93.0 R89.7 R87.614 I25.89 R74.0 R87.615 R85.9 794.9 I25.810 R90.81 R82.1 R93.7 I21.3 R73.01 R94.2 R97.1 I25.82 R87.610 R94.118 R92.0 R78.81 R91.1 R85.610 R81 I21.09 I25.812 R89.8 R76.11 R93.8 R94.6 794.12</p>
Acute-Bronchitis	<p>466.19 466.11 J20.9 I24.8 I25.10 I25.89 I20.8 I25.9 I25.811 I24.1 I25.2 I25.41 I25.810 I25.3 I21.11 466.0 I21.29 I21.3 I24.0 I25.82 I21.4 I25.42 I21.19 I20.0 I20.1 I21.09 I25.812 I25.83</p>
Allergic	<p>477.2 493.81 T50.995A J67.2 495.6 T78.03x 372.14 J67 J67.0 M13.89 J30.1 995.63 995.65 558.3 T45.0X1A M13.859 716.27 D29.30 D29.1 L27.2 477.9 495.5 493.22 D69.2 T78.00x 287.33 995.60 J45.31 J45.51 D29.20 J67.7 T78.09xA D29.22 M13.80 J30.9 T78.08x 287.8 H10.45 B44.81 716.20 995.61 T78.05xA 493.92 693.1 493.90 T78.40x J45.20 493.82 J45.40 D69.42 495.7 J67.5 493.20 D69.49 J45.32 287.32 708.0 H65.119 995.64 D69.1 J45.21 D69.6 M13.819 716.23 495.4 995.67 287.1 T78.08xA T78.00xA 477.0 493.02 525.66 T78.02xA J67.1 D69.3 T78.04x T78.2xxA D29.4 716.25 T78.07xA 716.26 T78.07x M13.88 J67.3 495.9 J45.30 493.21 477 495.2 995.62 T78.40xA 995.27 287.2 495.8 495 287.5 995.0 493 T78.05x L50.0 493.11 J45.902 D29.0 J45.990 287.9 J45 D29.21 J30.0 963.0 495.1 D29.32 L25.9 J44.9 J44.0 477.1 M13.879 493.01 J45.41 T50.995 J45.998 692.9 M13.849 995.66 D69.8 995.69 T78.04xA J30 495.3 M13.869 287.30 J45.991 J44.1 995.3 287.4 J45.52 287.0 381.06 716.21 J45.901 J67.4 287.39 493.91 373.32 287.31 T78.06xA J30.89 287 K08.55 K52.2 D29.31 J45.50 495.0 J67.6 D69.9 D29.8 T78.02x 716.24 477.8 381.05 D29 493.12 T78.03xA J67.9 716.22 T78.2xx J30.5 999.4 493.00 M13.829 T78.01x T78.06x 493.10 518.6 716.28 J30.2 H01.119 995.68 M13.839 D69.0 T78.09x 381.04 D29.9 T78.01xA 716.29 J30.81 J45.22 J45.42 T45.0X1 J45.909 D69.41 J67.8</p>
Cardiovascular	<p>I35.0 I48.0 I25.728 444.8 P29.38 I94 I63.212 402.00 I70.669 440.30 I89.9 I60.9 I20.1 413.9 I24.1 I80.3 415.1 I77.811 785.9 I69.319 I69.339 I82.529 R04.1 429.6 G43.619 I82.503 I82.611 I70.512 I75.011 I69.834 I70.628 K64.9 I89.0 I21.09 428.42 447.5 442.8 I70.792 454.0 I70.318 I50.83 I70.744 405.0 426.2 455 I70.442 455.3 I82.B29 I12 415 433.8 I69.320 I27.81 444.21 I70.735 I82.602 I67.89 441.4 425.4 I35.9 I70.693 I69.234 I65.23 427.2 I70.244 I49.02 I82.91 P29.89 I70.719 I69.131 I36.8 I60.8 I60.11 442.82 I69.852 I75.029 438.22 I69.120 I70.641 I60.2 426.51 I70.302 417.9 I63.012 R04.2 R00.2 427.31 I25.718 I05.8 I70.791 I89 426.50 I63.349 I49.49 444.89 I63.213 I83.12 I77.0 I82.433 I70.608 P29 I47.1 428 I70.348 I82.C29 I82.532 I69.322 I63.311 I65.03 I82.291 I70.498 I97.791 I69.859 I70.644 I82.441 I63.413 I70.362 405.11 I62 I87.012 I80.292 411.1 433 I70.532 I97.638 I87.091 I69.998 I07.8 I11.9 I69.390 445.01 I69.223 I37.2 I87.319 I69.932 I70.208 I82.492 I82.891 I63.233 I70.319 I70.65 I70.341 435 441.9 I40.0 I63.539 K64.4 I95.89 I69.028 I82.5Y9 I50.32 I70.731 I70.768 458.0 G43.601 I21.01 410.71 429.3 I69.398 I87.399 441.0 I70.421 I73 I70.729 E86.0 I69.231 I28.8 I13.0 I70.568 I42.2 I63.431 411.89 I70.709 438.21 438.53 I70 I50.84 I42.0 I70.212 I77.74 I08.0 P29.2 455.0 410.1 416.1 I24.8 433.10 I70.393 413.1 I70.561 I87.092 I83.218 427.60 453.82 I70.534 I97.130 I97.821 I97.713 I13.2 I69.814 I50.812 I79 429.89 I69 I69.833 I23 410.20 I77.6 I69.165 I60.52 459.81 I47 410.10 I83.201 I82.419 I69.033 I15.9 404.0 I69.213 441.7 I82.412 I69.261 I82.432 417 I24.9 I69.315 438.19 I72.8 I36.2 I97.648 I70.539 I56 I69.392 I63.512 I69.820 I77.70 I63.00 I70.439 I70.643 433.80 276.50 459.3 I70.522 I69.121 433.21 426.4 346.61 I61 I75.021 I88.8 I55 785.1 I01.9 I25.729 I63.20 454.9 I63.521 I70.769 I69.334 I72.4 I50.31 438.50 434 I70.349 I48.2 435.1 I80.219 I10 I69.115 I70.544 414.00 I69.010 I66.12 445.8 I30.9 I27.1 I80.293 432 I97.190 I69.864 I86 I34.1 410.8 438.9 I63.133 I25.89 I63.032 I34.8 I77.4 I82.422 I87.391 411.0 I87.312 440.4 I63.232 I87.009 R00.8 I69.012 I70.468 Q82.5 I21.3 I87.303 I25.6 I69.842 290.41 455.8 I69.042 I82.C23 I87.099 I69.314 I74.10 I77.812 I69.393 I42.5 I69.141 I70.491 I82.210 785.50 I69.351 I82.513 785 I69.252 785.52 I77.819 415.11 I80.13 I71.5 I82.422 I42.6 P29.30 I69.031 I70.222 I31.4 I82.603 I82.729 437 I70.621 I70.509 I70.763 I95 I38 433.3 I69.133 K55.0 426.6 I80.212 I70.723 I70.229 404.10 I83.029 437.3 426.13 I65.22 I75.012 I80.202 I48.3 I27.83 I69.815 I75.022 I27.20 I70.711 453 I65.9 I80.233 410.01 I70.291 440.2 I63.29 404.11 I82.A21 R04.89 I08.3 785.0 438.10 I70.469 I91 I69.163 447.2 I50.1 414.4 I83.002 410.5 I70.243 I63.039 I70.249 I70.269 I80.211 I86.8 I70.499 I69.363 404.02 I22.0 I69.211 458.2 I83.213 I83.028 I74.2 428.0 I70.545 442 443.89 426.3 I69.362 I31 I70.398 I81 I69.810 I70.261 443.22 I70.1 I70.335 I92 I77.89 I69.00 I83.203 I72.3 I63.532 410.2 I71.02 I70.331 441.03 I69.262 410.6 I65.29 I54 453.71 I69.293 438.81 I69.110 I15 R65.21 428.31 453.89 I69.051 I67.7 I69.062 I69.111 I69.349 454.1 453.1 I63.39 I82.1 I66.03 I72 I46.8 426.12 I45.81 I73.1 I70.634 I69.918 I95.1 I87.393 410.02 410.0 I70.332 I82.549 443.29 I77.5 440 I31.8 I21 I82.B23 I11.0 433.01 455.7 403.00 I66.02 438.14 I06.0 G45.0 I97.51 I69.112 I70.369 I82.5Y2 I45.2 I74.01 415.19 I87.311 I63.22 I84 I20 I70.203 I83.009 I87 I71 433.1 I06.1 456.4 I69.822 I99 I33.0 I70.218 438.5 I60.10 I69.254 413 I51.2 I69.065 I59 410.62 I07.2 I80.8 I96 453.76 437.2 435.3 414.11 I83.019 I09.2 I83.93 426.81 I63.323 410.72 435.2 I63.441 I83.011 410.31 I82.B11 410.81 I27.23 438.41 I75.81 404.9 I13.11 I66.09 I95.0 I83.022 438.4 I82.431 I70.202 I70.342 I70.648 I23.3 I69.265 437.9 I82.812 I82.439 I63.59 I97.120 I69.249 I07.9 455.6 I82.501 I82.A23 I74.09 I63.329 I70.722 I82.423 I87.9 I82.C21 414.02 I42.9 I69.093 404 I24 455.9 I32 I70.798 I44.60 441.02 R04 I50.40 I87.302 I82.509 441 I77.3 I97.630 I69.993 I53 I83 I82.499 I69.034 I70.402 I60.01 458.1 438.82 I24.0 458 I69.144 R01.2 I70.231 I65.1 I70.429 I70.612 I50.33 I82.612 I63.422 I57 433.2 I69.221 I95.9 I78 I97.111 I04 I69.963 779.82 410.52 410.3 I70.298 438.20 I48 I50.82 I70.698 I70.533 I69.915 I69.052 425.9 I69.243 I80.229 427.32 438 410.41 I69.092 413.0 290.4 I27.22 I69.043 442.83 453.86 557 453.75 445.0 I23.2 438.1 458.21 I70.303 I70.541 I70.201 I69.154 411 429.1 I51.9 I60.7 I69.244 427.41 I72.6 I16.0 290.42 I70.793 I82.491 I69.022 443.8 I51.5 I26.09 I83.225 I37.9 I66.13 I70.329 I69.191 I70.638 444.2 I36.1 I69.91 R01.1 I08.1 I70.333 I70.462 I82.401 I82.A19 I63.519 K55.9 I82.4Y3 I82.5Z3 I33 410.7 I45.4 I69.333 I83.219 I69.198 I70.399 426.54 I25.710 442.81 557.9 I80.231 I82.443 I80.01 P29.12 447.3 I70.245 410.70 433.91 I69.812 I50.23 448.9 434.9 I26.99 427.5 I77.77 453.6 I69.118 I35 441.3 I82.C22 I08.8 414.07 I28.0 438.89 I97.42 I65.09 I37.0 I67.848 I70.449 I77.2 411.8 I70.603 G45.4 453.81 I44.5 426.0 I06.2 I74.5 427.9 427.42 I70.213 I61.2 444.9 I62.1 F01.51 I70.701 456.8 456 I87.013 I01.8 440.3 453.83 I69.242 I42.1 I83.811 I69.069 I69.313 I83.023 I45.9 I69.013 I69.813 453.77 I21.11 I25.82 I87.339 I70.749 I83.892 I23.5 I97.611 I66.3 I63.449 453.72 I88.0 405.9 I69.059 I40 I69.943 I70.748 I21.02 428.32 447.70 I78.8 403.9 I82.521 I69.321 405.09 453.40 I63.9 I69.854 453.42 I40.8 I69.162 426.1 I69.192 I83.222 I87.033 I25.83 I74.3 I82.523 458.9 K64.8 438.83 I42.7 I70.521 459 I83.221 I82.609 I82.703 I07.0 I97.110 I69.942 437.6 405.1 I60.6 I69.342 I70.728 405.99 I69.831 I69.30 I12.9 438.3 447.71 441.00 I69.828 434.91 I09.89 I23.4 I83.024 I47.0 I82.543 I50.810 I70.293 I82.211 405.19 I86.2 I69.954 I70.409 I83.812 I97.622 437.0 I25.799 403.01 I75.89 I50.20 I63.533 I78.1 428.33 I82.A22 I79.0 414 402 414.01 427.89 I69.964 I70.642 I63.542 I47.2 I87.309 I63.339 I77.72 I70.363 414.06 404.92 I71.2 441.1 I82.522 I20.8 402.11 I82.819 I65.8 I36.9 I99.9 785.3 429 I69.344 I97.811 I63.113 443.21 I69.020 I70.328 I72.9 I62.01 I70.613 447.72 I82.542 785.51 I69.132 I83.015 I82.493 I67.4 426.52 425.7 I80.00 I70.338 I63.30 I78.9 I74.19 I87.392 442.9 I82.449 I69.862 403.90 402.10 I25.758 402.1 I27.21 I83.202 I86.0 I70.401 I87.093 I49.8 I95.3 433.11 I02.9 425.2 I70.519 412 I70.739 I70.702 404.03 429.5 I88.9 410 I82.B21 410.22 I69.169 I70.368 429.82 I82.619 I82.592 I60 437.8 I80.10 414.0 433.9 I72.5 I70.591 404.12 I69.219 I73.9 I46 I63.50 434.1 I82.599 I74.9 I83.018 I82.429 I70.392 I66.11 I63.531 I70.435 I22.2 I65.02 I82.541 425.0 I69.298 I63.40 I69.263 I95.81 453.8 I82.409 R03.1 I25.759 I87.1 I69.353 434.01 557.1 I25.2 I70.598 I70.601 I79.1 R57.8 I63.131 I22 414.03 I46.2 I19 I63.439 403.91 I69.064 429.8 426.9 I63.09 I80.03 I45.89 I97 I82.531 429.79 I13 I69.032 I22.8 I26.90 434.10 415.12 I70.639 I69.232 I51.0 I60.51 I63.411 I97.131 I69.891 I70.8 I44.0 I39</p>
Cataract	<p>366.34 366.14 366.19 366.02 366.21 366.31 366.10 366.23 366.32 I24.8 I25.10 366.16 366.18 I25.89 I20.8 366.15 366.52 366.00 366.51 366.41 I25.9 366.13 I25.811 366.03 I24.1 I25.2 366.17 I25.41 I25.810 I25.3 I21.11 366.04 366.12 366.11 366.9 366.30 366.33 366.01 366.43 I21.29 366.09 366.53 I21.3 366.44 I24.0 I25.82 I21.4 I25.42 366.8 366.20 366.50 366.22 I21.19 I20.0 366.45 366.46 I20.1 I21.09 I25.812 I25.83 366.42</p>

Continued on next page



CNS	<p>346.32 335.10 G40.019 G47.54 G43.609 G44.029 G16 344 349.2 G47.20 G45.0 G21.4 347.00 G81.11 346.82 G95.19 G47.27 345.70 347.11 G25.89 341.21 322 344.30 346.93 G04.2 G40.909 G40.219 G43.619 G80.0 327.15 G43.011 333.1 349.82 327.51 G83.10 G12 G47.35 327.32 G03.8 327.41 337.00 G46.2 G96.12 G05.4 G43.101 G24.2 G47.29 G12.21 345.00 G93.1 346.72 G13.1 G95.9 331.7 G31.01 343.8 G40.209 346.40 334.9 G30.1 G44.309 G83.13 321.0 344.89 G47.23 G44.009 G11.9 324.9 G43.00 G81.90 346.20 G03.1 333.84 320.1 348.30 G30 G81.93 G95.0 G37.5 G47.12 G25.71 327.29 G40.201 346.41 G06.2 G37.9 767.0 G40.813 P91.63 G93.49 344.9 345.01 334 G47.52 322.9 G44.41 336.1 G44.82 335.11 348.39 G47.22 G37.2 G46.0 327.11 G04.89 G22 331 G41 G43.509 G83.30 327.30 G93.7 333.2 333.5 G47.419 344.41 P91.2 P91 328 349.1 G90.529 G25.79 344.5 G40.111 337.20 G83.89 G37.1 348.4 G23.2 G25.1 G44.81 G00.2 G40.804 G40.809 336 G44.83 G47.00 G82.50 344.60 337 343 G47.63 327.44 G44.40 336.9 G13.8 P91.1 333.82 G47.59 G31.1 G82.20 334.2 333.71 G35 G21.0 G24.1 G25.3 G23 G46.5 322.0 327.8 779.2 333.85 G25.0 G81.01 344.04 G43.C1 G44.53 G40.814 G00.3 G43.111 342.02 333.89 G23.1 G90.59 G80.2 F51.05 G91.0 G40.823 343.4 344.01 335.20 G40.411 G44.221 G47.14 G12.25 G04 343.3 G43.A0 333.79 327.22 346.01 G18 G03 G83.9 G40.409 G44.001 331.82 323.42 323.72 G82.53 337.3 G93.40 G99.0 G43.719 E75.4 G40.009 G47.30 G47.10 G40.802 336.0 327.01 G40.B01 P91.60 323.63 340 G43.B1 345.91 P52.8 G12.8 G43.419 G07 G93.9 G40.001 327.40 G44.209 337.1 G25.4 G99.2 333 333.72 323.81 327.23 323.02 344.2 G12.24 333.4 327.59 G12.23 G43.601 331.2 333.3 G83.4 G44.219 G40.101 B45.1 330.0 G44 G83.23 G93.2 G43.001 G43.C0 327.49 G21.8 G46.7 G30.9 G40.419 G82.52 G40.A01 345.61 347 G44.321 G11.8 G00.9 330.1 G43.B0 G44.011 323.1 G31.81 G43.401 320.89 G83.20 346.80 G47.9 G43.911 G47.62 348.9 G13.2 346.90 G47.31 G90.09 335 342.82 327.20 342.11 G11.2 327.19 346.02 341 G40.119 G47.09 322.1 G04.02 G11.9 348.5 G43.829 323.82 333.99 342.12 345.3 G40.011 G20 G45.9 G40.89 P91.0 346.42 346.60 335.23 G03.9 G40.309 336.3 333.90 349 G04.01 346.00 327.14 342 G05.3 G44.059 321.2 G31.89 G43.701 G37 G43.511 G40.822 G43.919 G44.51 G47.36 345.41 G31.2 G13.0 G91.1 344.31 G00.0 320.2 G25.81 344.61 G44.329 G47.429 G40.824 G80.8 346.63 G46 G30.8 G43.831 G80.3 348.2 333.83 G45.4 G13 320.0 345.10 344.40 341.1 G37.3 G31 G14 333.94 G24.3 G36.9 G45.2 349.39 G24.8 G47 G04.31 G05 G11.4 341.20 332.1 344.1 G47.39 G83.5 G46.1 G25.70 G21.19 G00.1 343.9 G25.61 G40.311 G21 331.11 G32.89 G47.51 332.0 335.22 G44.091 345.80 331.81 G36.8 G00.8 G46.8 G47.34 P91.88 G43.D1 323.71 330.9 G04.90 331.89 G06.1 G46.4 348.1 346.30 G43.019 G82.54 337.29 335.8 G04.1 337.22 G43.839 344.02 327.43 G12.1 G43.901 G47.69 G10 G47.421 G47.50 G44.84 G31.83 346.62 G47.01 345 G47.19 G44.031 G25 320.7 323.51 G33 G06.0 G32.81 346.50 G83.11 G27 337.09 G24.09 322.2 344.00 346.61 G43.801 327.37 G31.85 327.42 G36 346.31 G46.3 G24.9 G40.211 G24.01 327.26 G44.049 341.9 G44.039 327.33 346.83 334.1 335.29 344.09 323.2 346.13 327.12 G47.37 327.36 327.02 G43.909 G25.5 G40.803 346.12 G12.20 327.31 G37.8 345.2 327.10 G44.311 G43.409 344.42 343.0 G06 323.01 333.93 346.81 G40.509 325 G44.211 G43 G81.00 G93.0 341.8 335.0 G09 G96.9 320.81 G44.229 G04.32 G24.4 346.11 G25.69 333.6 G37.4 G04.81 G44.201 345.81 332 G12.22 346.51 346.33 331.5 342.90 G44.099 G40.401 G95.89 346.10 F51.13 G43.501 G01 334.4 324 327.13 G44.051 345.71 327.53 320.3 345.60 G12.0 G04.91 P91.4 341.0 G83.0 327.25 G47.25 346.52 G21.2 349.31 G25.9 G40.812 G24 337.21 331.4 G47.53 346.70 G91.2 345.11 330 327.09 344.03 G96.8 323.9 G44.019 G47.13 323.52 G44.301 324.0 G90.9 G42 346.73 320.9 G44.59 G25.82 G40.919 330.3 327.39 321.8 331.0 331.9 G30.0 G36.1 779.1 G81.03 334.3 G08 G47.21 331.3 G40 G11.3 G32 G04.39 342.10 G92 G36.0 323 G82.51 G45.3 G12.9 349.81 P91.819 321 G47.11 G40.A11 348.0 G93.81 G11 G40.801 G26 G80.9 342.00 G40.501 320 P91.3 G44.89 G90.519 G93.5 G81.91 G24.5 G45.8 331.19 323.61 G96.0 343.1</p>
Development	<p>520.4 742.59 764.96 743.66 750.19 M26.36 756.11 749.01 748.1 524.33 740 741.1 752.35 741.0 743.53 524.79 748.5 752.42 589.0 Q18.1 Q26.2 Q93.4 764.98 743.43 744.21 Q63.8 753.23 750.4 P92 N13.70 746.82 747.40 743.57 743.41 742.5 752 743.34 749.14 524.35 744.82 Q05.2 M26.9 Q55.63 Q10.1 752.43 748.3 756.16 M26.89 743.31 593.71 M26.221 M26.72 313.23 Q31.1 743.49 F94.0 750.27 744.41 Q35.7 Q23.0 Q15.8 Q61.3 749.1 Q16.9 759.0 743.12 329.41 G47.49 G61.8 P92.4 743.65 750.21 Q20.0 Q87.81 Q62.4 758.5 764.03 524.73 Q17.1 756.13 524.31 Q34.8 Q18.4 Q27.8 520.8 743.64 Q21.2 741.9 751.61 Q76.5 524.27 Q05.7 Q61.01 746.87 Q10.0 Q25.4 Q36.9 764.06 752.36 743.11 764.91 Q28.3 747 756.14 743.39 759.89 746.6 753.12 750 750.10 Q62.31 M26.39 750.6 Q51.4 Q56.3 Q24.8 747.82 745.7 752.65 750.16 749.02 Q28.8 P05.14 743.35 Q20.9 Q64.9 P05.08 749.23 742.9 747.10 756.12 Q16.4 743.44 K00.4 P05.15 722.5 747.20 745.19 753.4 764.07 Q51.3 N28.83 M26.56 Q21.1 744.8 M26.220 747.8 N13.722 Q23.2 752.40 Q00.1 Q87.1 Q06.4 764.92 P05.06 524.56 743.46 Q55.22 Q15.11 744.0 Q51.2 K00.2 M26.52 Q76.49 752.89 Q52.12 746.0 Q99.2 E30.1 P92.2 M26.54 745.12 Q02 753.10 745.8 Q33.4 Q24.2 764.12 Q24.6 754.0 315.8 744.89 307.6 Q55.23 750.1 759.4 M26.73 Q00.2 Q40.8 752.64 M26.211 593.7 Q64.10 758.32 Q33.6 524.3 Q26.5 Q18.2 Q55.64 746.5 P05.9 764.00 744 752.34 752.9 Q43.4 524.57 Q38.2 759.1 747.2 315.9 Q13.4 759.3 524.22 Q16.2 315.4 743.63 752.51 Q27.9 745.3 524.32 743.54 Q20.1 F93.9 Q12.0 746.8 524.8 P05.18 748.61 P92.9 Q42.9 Q52.0 743.56 742.1 Q20.8 Q89.3 524.21 Q51.0 749.20 753.3 Q18.7 315 Q51.811 747.3 Q03.8 748.60 744.09 Q61.9 Q93.81 745.69 741.92 759.2 Q05.8 746.2 747.81 750.29 741.02 Q76.0 744.04 743.36 764.24 P92.8 P05.10 749.11 Q12.4 524.55 Q54.4 746.4 520.1 743.59 764.10 Q37.8 M26.33 313.2 P05.13 524.7 P92.01 745.61 752.11 P05.16 M26.34 Q24.3 751.5 524.74 Q06.2 746.00 752.10 Q50.6 Q21.9 752.49 745.4 758.31 752.33 K00.0 E30.0 753.15 307.7 758.1 746.1 Q04.8 764.13 M26.30 752.41 Q10.6 744.23 759.9 764.17 764.90 Q14.0 Q13.0 764.99 593.70 Q43.0 753.21 P05.07 744.05 745.6 Q98.4 744.47 Q52.9 750.3 743.45 Q33.9 764.97 753.7 741.93 764 Q23.1 Q30.0 747.1 752.6 Q15.0 593.73 741.91 M26.79 742.51 524.81 Q51.818 764.21 Q33.0 743.1 Q22.0 Q10.7 Q51.5 752.69 752.1 Q22.1 748.6 Q12.1 750.8 F93.8 Q36.0 745.0 751.8 752.4 Q34.9 747.83 747.21 754.1 746.84 753.5 520.9 524.30 747.60 747.61 R62.51 K00.9 743.52 742.2 Q12.9 753.0 Q51.6 743.9 743.5 759.7 Q89.7 744.03 743.03 753.22 Q38.0 Q38.3 759.81 M26.23 745.9 M26.24 744.02 524.82 P92.5 Q01.9 751.60 Q26.8 Q17.0 Q41.9 524.20 Q97.1 764.20 Q52.3 Q13.5 Q38.4 752.8 752.3 M26.4 746.3 M26.55 747.22 520 743.62 741.00 M26.57 749.04 744.9 Q17.8 756.2 758.4 Q45.1 M26.32 F98.0 P92.6 P05.12 Q17.2 524.26 524.50 748 P92.09 743.2 P05.05 741 756.4 Q10.3 N27.0 747.9 309.21 Q05.0 745 P92.3 Q14.1 742.3 Q62.39 752.47 Q51.810 Q17.3 740.2 Q14.2 Q11.1 589.1 746.09 Q45.9 743.10 747.89 K00.5 751.0 Q15.9 Q64.39 750.11 Q68.0 750.13 Q62.10 749.24 751.9 313.8 743.42 Q77.1 750.26 Q22.2 744.83 524.5 593.0 259.1 747.4 744.01 524.29 744.5 524.53 K00.6 Q04.3 Q92.8 Q12.3 752.44 752.0 751.69 Q89.01 764.08 743.6 764.15 F88 315.5 748.69 589.9 Q11.2 Q55.8 752.7 764.16 747.42 743.33 744.81 Q25.2 741.01 Q62.12 Q05.5 748.4 744.24 Q13.3 Q95.0 P05.02 N27.1 313.9 Q38.5 748.4 742.4 744.43 Q50.4 Q54.9 520.5 P29.3 P05.01 524 Q99.9 749.12 758.9 758.33 764.11 753.8 Q07.9 745.5 748.0 747.63 589 758.6 P05.17 764.05 Q38.6 Q26.3 Q26.9 M26.70 524.34 747.0 Q20.5 M26.59 764.02 Q06.8 743.20 Q00.0 764.29 M26.213 Q33.1 752.5 Q76.419 Q91.3 K00.1 764.28 751.4 747.5 752.62 M26.25 524.70 748.8 Q43.3 743.58 Q35.9 749.10 Q21.0 752.2 N13.729 752.63 753.17 Q18.5 Q89.2 R62.50 M26.212 Q16.1 Q61.4 Q20.3 524.24 747.62 747.11 749.2 524.37 764.95 743.22 744.22 741.03 Q24.4 Q22.3 M26.82 744.84 Q16.0 Q24.5 753.1 259.0 748.2 752.46 747.64 764.25 764.01 Q39.5 524.72 756.17 743.21 756.15 750.23 Q51.820 Q38.1 Q18.8 Q60.2 P92.1 524.59 750.25 750.5 749.00 752.39 M26.50 752.45 744.3 743.32 F93.0 Q45.8 313.89 758.81 750.0 Q64.4 Q96.9 742 746.85 743.30 313 743.06 Q14.8 R62.7 743.3 Q05.4 751.6 746.89 758.39 749.25 R62.52 752.19 N13.721 752.52 744.1 Q76.2 764.04 Q16.3 745.2 P05.04 748.9 764.14 764.94 520.7 749.21 753.19 750.7 Q25.0 764.27 F98.1 752.31 Q20.4 P05.03 Q91.7 759.83 746.81 743.48 749.0 Q50.01 524.9 Q48.0 745.10 524.76 P05.11 524.89 753.2 750.2 Q27.31 R62.59 744.41 744.46 751.1 740.0 Q67.4 749 K00.8 F82 Q64.79 Q23.4 764.93 Q27.0 753.11 Q11.0 M26.74 Q18.0 Q55.62 M26.37 Q21.3 Q22.9 Q25.1 524.4 745.60 750.12 752.32 742.53 Q37.9 Q44.2 524.71 R62.0 Q51.828 N27.9 753.6 743.8 593.72 524.54 744.00 524.25 750.9 Q61.5 Q30.8 Q23.3 751.2 524.28 746.86 756.10 524.52 M26.35 746.83 743.61 520.2 764.19 Q28.9 744.42 Q18.9 746.01 749.22 744.49 M26.81 743.37 751.7 747.29 P05.2 520.3 524.36 750.22 764.22 Q64.0 Q76.1 742.0 747.6 744.2 Q14.3 746.7 Q53.9 749.13 743.55 Q62.11 Q87.0 743.51 Q16.5 Q27.2 Q05.1 Q50.1 752.81 752.81 758.2 520.6 M26.31 524.39 Q18.3 Q93.89 M26.20 Q38.7 755.55 524.75 764.26 M26.53 744.4 758.7 756 524.23 R62 K00.7 Q40.0 Q05.6 524.2 Q13.81 K00.3 745.1 756.19 752.61 Q31.0 758 520.0 741.90 746.9 Q93.88 Q89.9 753.29 Q13.1 Q18.6 F81.5 746.02 743.00 Q44.1 743.47 747.69 Q39.1 Q38.8 753 764.23 764.18 753.20 Q27.32 764.09 Q17.9 Q23.8 Q89.1 Q13.89 M26.71 750.15 753.9 743.69 744.29 P05.00 Q40.2</p>
Continued on next page	











Ischemic	<p>410.72 410.51 410.42 I20.8 410.20 414.12 410.00 414.00 I25.3 I21.1 410.40 I25.42 414.10 413.9 410.12 414.2 410.32 410.61 414.01 I25.10 411.0 414.19 410.22 410.60 410.02 410.92 414.02 I25.811 I25.2 I21.29 410.30 I21.4 410.31 410.21 413.0 I25.83 414.8 410.81 410.82 411.89 414.07 410.90 410.11 I25.84 410.50 I25.9 414.9 I24.1 I25.41 410.80 414.06 410.70 414.04 410.01 410.62 I24.0 410.71 I21.19 I20.0 I20.1 414.03 I23.0 412 410.42 I24.8 I25.89 414.11 I25.810 414.3 411.1 I21.3 I25.82 410.10 410.52 414.05 413.1 410.91 I21.09 I25.812 411.81</p>
Metabolic	<p>269 E83.30 270.5 P71.4 276.7 264.1 712.38 M1A.9xx0 274.82 E71.42 278.02 274.19 278.01 276.69 263.8 712.86 E72.8 268.1 275.40 274.03 712.28 M10.30 E66.2 783.1 E74.9 E71.41 270.8 712.19 E83.59 E87.70 M11.80 E50.7 E88.1 275.42 E67.0 260 M10.9 E46 274 E72.20 M11.88 276.3 712.84 E83.50 P19.0 264.3 E50.8 278.2 275.1 E50.1 M11.20 E83.81 R63.4 775.9 277.81 268 712.36 277.88 E83.40 E53.9 271.9 P71.3 E87.1 M11.879 270.7 E52 D84.1 273.9 266.2 264.4 265.1 274.8 712.1 M11.249 712.8 M11.849 277.82 E72.9 M11.269 712.16 E50.4 712.88 M10.40 M11.28 269.9 E88.9 264.9 268.0 E55.9 278.0 712.85 266.9 277.9 E75.21 712.3 712.89 712.27 712.97 277.8 P72.8 M11.859 278.3 264.0 278.03 G93.9 E53.0 E16.1 271.3 251.0 775.89 E70.0 P71.1 E51.11 P70.1 275.5 P70.4 277.7 272.3 712.15 E50.6 277.87 712.39 274.81 712.96 712.10 274.02 E87.8 712.90 E15 273.8 R63.0 271.2 E66.9 277.86 266.0 E53.8 E64.3 263.1 278 E87.0 272.5 E66.3 712.32 E67.1 262 712.92 M11.9 M11.29 712.31 272.1 P19.1 P71.8 M11.279 712.95 277.85 E65 E88.40 775.81 783.22 E61.4 783.2 263.2 269.0 261 E83.89 P71 M11.829 P74.0 E74.12 274.89 E70.21 E72.03 712.11 E50.2 M11.869 712.80 P19.9 E50.5 264.6 712.13 275.49 E16.2 E54 E67.8 712.83 276.8 278.4 775.7 712.37 E78.3 270.2 270.4 E53.1 E83.52 M83.9 269.2 269.1 783.21 712.33 M1A.00x E87.4 E87.5 P70.3 268.2 271.8 E87.3 M11.259 E76.01 E87.2 266.1 P71.2 E88.01 E55.0 251.2 266 273.4 274.00 272.2 275.01 275.8 712.98 E56.8 E56.1 E74.21 E44.1 272.7 712.91 263.9 E88.81 E78.89 E78.9 712.34 269.3 274.01 263.0 P70.9 712.22 277.2 E74.39 263 E67.3 276.6 E88.09 E50.3 E72.10 R63.5 E43 P19 265 275.9 264.5 E66.01 E83.00 E80.0 E88.89 712.18 E83.118 275.4 E63.8 264 274.10 277.5 275.3 712.94 712.12 E44.0 E40 M11.819 276.2 274.0 E51.8 270.6 E79.8 274.1 267 E83.9 712.23 276.9 272.4 P71.0 G93.89 712.2 712.87 E83.51 P70.2 271.1 276.4 E45 712.93 277.89 P84 P71.9 M11.89 E87.6 E41 712.17 265.0 E78.2 272.9 277.6 274.9 276.0 E56.9 E70.40 E83.110 M11.229 268.9 712.25 272 270 330.2 271 712.35 M11.839 264.2 275.09 712.30 276.1 R63.6 264.8 277.1 270.9 P70 712.21 278.8 272.6 712.9 275.03 M10.00 270.3 271.0 E88.3 E50.0 N20.0 775.8 P70.0 E63.9 E71.50 710.80 712.29 E74.4 274.11 E83.10 712.24 275.2 E78.1 P70.8 712.99 330.3 712.82 712.81 M11.219 M1A.9xx E50.9 251.1 E78.5 E78.6 272.8 P19.2 E71.318 264.7 E71.0 783.0 712.26 712.20 269.8 M1A.00x1 P72.9 265.2 M11.239 712.14 275.41 278.1</p>
Musculoskeletal	<p>553.9 M84.574D Q67.5 M84.442S M70.30 M25.761 M46.87 550.00 M60.231 M71.829 M85.622 M84.675A M84.753D M08.429 733.20 M84.472 M60.239 M84.346K 727.6 M12.332 735.1 M84.550K M62.241 M71.062 M66.219 M05.051 M24.275 552 M67.479 755.67 M89.322 M27.59 M25.842 M80.821S M84.753A M84.573P M54.02 718.80 M53.86 M85.58 Q65.00 733.43 717.41 M24.174 718.9 M85.012 M84.58XG M15.9 M10.452 M41.22 M20.21 M84.522K M16.10 M23.041 M66.211 M19 M13.822 M24.231 717.82 M90.871 M87.035 M48.57XD M96.679 M21.40 M32.12 M93.841 718.82 M46.96 M65.341 M11.10 M25.222 M67.279 M80.019S M71.99 M66.172 M84.462K M84.442K M96.639 M80.011S M08.231 K43.7 M08.849 M23.007 M12.212 M48.02 M84.619G M65.171 M67.229 M66.272 M60.819 719.90 M11.049 M84.563A M85.812 M89.312 719.70 M06.811 M12.449 M45.0 M79.4 M26.03 M24.075 M86.151 P13 M24.676 M99.77 M84.651 M10.329 M89.752 M23.91 M05.271 755.30 M80.052 M05.842 M08.829 M96.662 M00.232 M80.069A M85.311 M63 M94.279 M08.839 718.47 M77.30 M80.012G M05.372 M15.4 M62.511 719.04 M89.542 M84.561S M84.572 M94.9 736.05 551.01 M08.469 M94.0 M71.129 M80.011P M99.9 M07.649 M67.832 M05.519 M25.642 M84.533D M54.15 M05.069 M62.259 M84.662K M24.031 M20.001 M90.551 M60.031 M05.60 M61.146 M24.651 M41.112 M93.811 719.12 M89.619 M08.911 M14.879 M08.3 M05.69 M00.162 M05.722 756.55 M05.431 M18.32 M70.951 M80.859G 730.39 M92.12 M48.43X M65.852 M26.00 M23.052 M24.572 M40.14 M80.079D M86.542 M05.049 M65.261 M65.011 M84.663P M60.172 M05.811 711.822 M85.321 M99.21 M67.971 733.1 M84.364P M02.10 M60.074 M83.1 M35.9 M97.42XS M12.529 M05.879 M87.032 M46.38 754.5 M90.672 M50.23 719.50 M85.859 737.21 M12.9 M84.452S M84.443 M11.869 719.45 M93.879 M36.3 719.25 M89.49 M76.899 M80.929 M71.869 M89.042 M70.21 M87.334 M10.451 M26.23 M85.00 M06.831 727 M80.831 M90.852 M76.42 M89.28 M65.332 M86.661 M23.631 M50.823 M75.00 M05.479 M84.379 M10.321 M86.021 M61.129 M79.0 M84.569K M61.229 Q79.4 M08 M62.441 M66.811 M06.859 M54 M45.1 M84.571G M84.672D M23.005 M47.13 M84.612S M84.542 M67.272 K08.23 M12.379 M24.839 M10.019 M89.439 M48.51XS M89.30 M25.149 M22.3X2 M89.061 719.20 M84.756D 552.3 M60.831 717.0 M89.08 M32.11 M90.861 M42.09 M54.13 M12.519 M92.291 M93.851 719.83 M23.212 M71.80 M87.838 M06.059 M84.462A M99.43 M66.842 M23.322 722.30 M84.663K M01.X51 M07.60 M24.376 M80.069P M20.031 M72.2 767.5 M61.161 M02.361 M00.811 M26.213 M87.861 M84.673 M85.071 M84.475A M84.519P M99.18 M71.552 M24.575 M43.04 M12.579 717.8 M89.78 M12.219 M84.639 M84.563P M89.165 M60.261 M80.00X 730.37 M34.1 K40.10 P11.3 M26.82 M84.322G M84.564 M46.52 M51.05 718.89 M88.88 M05.121 M05.10 M84.674D M84.342D M89.571 M80.872A M84.841 M23.352 M71.051 M93.969 M99.01 M63.869 M48.061 M02.122 M21.379 Q65.30 719.92 M21.221 M25.361 M87.29 M84.359P M90.511 M72.0 M11.032 M85.639 M60.262 M80.00XP M84.350A 524.07 M94.261 718.41 M86.432 M54.81 M84.459 M10.422 M84.446K M84.572S M23.009 M6.4 M72.6 M87.263 729.39 M14.659 M87.274 M41.41 M84.750A M84.639A M84.672S 719.49 M91.10 M99.00 722.72 M14.621 M61.59 M67.942 M80.869G M89.032 M24.562 359.5 M60.019 M76.611 M24.371 M87.850 M92.219 M84.432G M13.841 M80.822 M84.474D M85.331 719.16 M25.051 M84.353S M94.229 718.73 M19.032 M84.38XK 717.83 M80.031 M84.632K M94.232 M90.532 737.41 M84.575 M10.349 M86.621 M21.859 M60.20 M61.512 M93.022 M84.649 M10.159 M71.811 M05.762 718.76 M23 M84.669D M48.8X5 719.91 M87.037 M52 M10.022 M84.511P M62.011 M26.31 M47.23 718.4 M31.0 M25.641 M11.219 M65.872 M24.112 M65.141 M21.331 M84.757 M20.092 M80.851D M27.0 M87.151 M99.08 M05.742 M70.971 M97 M84.534S M08.412 M43.8X8 736.02 M84.632A M87.011 M24.021 M61.461 M21.629 718.84 M40.56 M84.369G M87.378 M06.849 M89.157 732.8 M66.89 553.03 M92.299 M42.00 M71.572 M35.8 M02.152 755.65 726.70 M89.8X9 M12.822 M80.042K M84.433G M88.832 M67.411 M23.307 M24.7 M05.759 M67.864 M61.412 M54.08 M10.169 M84.434K M18.50 M24.276 M62.89 M84.833 Q65.4 M40.00 M02.022 M85.329 M60.259 733.10 754.81 M80.80XG M90.542 K08.20 M11.019 718.97 M71.121 756.53 M21.029 M62.131 M26.221 M25.129 M34.82 M86.261 M80.042S M86.462 M47.016 736.75 M46.81 G71.19 727.49 M61.071 M71.21 M02.18 M11.152 M80.829K M12.039 M13.131 M88.841 M85.48 M87.833 M84.361 M60.839 M84.350G M84.422G M19.079 M54.9 718.77 M23.003 M80.852S M60.073 M87.031 M71.479 M61.022 M05.321 719.19 M06.022 M84.352D M84.362 M62.419 M11.051 M84.369K M60.821 M90.551.29 M76.891 M66.352 M67.451 M86.061 M24.171 M86.079 M05.429 M97.42XD K44.0 M60.112 M90.552 M06.051 M05.552 M47.28 M08.48 M12.249 M84.439P M00.221 M85.332 M11.021 M85.029 M76.12 M26.52 M86.442 M84.334S M84.576P M13.112 M10.112 M86.252 M94.1 727.3 M85.342 M85.879 524.06 M85.821 M10.359 M84.475S M12.532 717.7 M27.62 M26.73 M43.01 M12.861 M60.869 M84.459A M17.5 M85.061 M48.55XD M88.839 M76.892 M19.022 M80.861A M84.321S M65.311 M80.852P M96.843 524.1 M80.019G M83.839 737.34 M84.632P M90.811 M18.11 M43.20 M24.673 717.81 M14.629 M61.032 M102.079 M60.162 M84.532 M84.633D M99.16 M06.09 M24.312 M35.4 M90.662 K45.8 M46.05 M08.012 M11.272 M61.19 M46.24 M84.671 M84.673D M48.8X9 M79.9 M84.755S M12.39 M84.58XA 736.73 M21.511 M00.271 M21.162 M66.329 M47.14 M84.443P 719.01 M05.741 M66.331 M84.659G M67.429 M20.22 M80.831P M84.431A M60.045 M25.339 M60.242 M20.62 724.01 M19.279 M48.56XG M65.172 M48.41XG K08.26 M71.821 M06.012 M66.322 M12.09 M66.369 M86.279 M47.813 M11.271 550.11 M12.462 M54.14 M05.122 M79.645 M71.039 M89.231 M87.372 M62.449 M11.111 M05.862 M92.01 Q72.10 M47.24 M99.42 M43.22 M20.10 M06.341 M33.29 736.72 M56 M71.819 M86.452 M65.28 M00.842 M50.80 M79 755.33 550.93 M80.039K M00.152 M25.271 M66.122 718.93 M50.120 M84.68XP M89.59 737.39 M25.772 M19.141 553.00 M65.151 M05.022 M80.811A M70.89 718.83 M10.011 M10.261 M90.50 M11.061 M89.158 M19.039 M86.132 M16.7 M84.350D M41.07 524.63 M46.1 M15.0 M12.131 M89.272 M93.832 M48.31 D48.1 M11.18 M05.771 M35.6 M12.141 M24.652 M87.029 M61.171 M33.11 524.61 M25.08</p>

Continued on next page



Oth-Joint-disord	<p>719.35 719.05 719.13 M25.40 M25.459 719.07 719.86 I20.8 719.33 M25.18 719.91 I25.3 I21.11 719.54 719.58 719.62 719.89 M25.073 719.29 719.38 719.27 719.25 719.92 M25.619 I25.42 M25.639 719.30 719.53 719.85 M25.48 719.90 M25.559 719.14 719.47 719.42 M25.439 719.34 719.86 I25.10 719.22 719.36 719.94 719.56 719.96 I25.811 719.81 M25.00 I25.2 M25.059 M25.08 719.99 I21.29 719.82 719.03 719.68 I21.4 719.11 719.09 M25.039 719.01 719.28 719.15 M25.019 M25.519 M25.60 719.60 M25.069 719.21 M25.029 719.31 719.66 M25.449 719.43 I25.83 719.08 M25.9 M25.879 M25.849 M25.649 719.93 M25.10 719.02 719.80 719.98 I25.9 719.04 719.32 I24.1 I25.41 719.18 719.55 M25.629 719.7 719.46 719.52 I24.0 M25.659 M25.669 719.19 M25.049 719.67 M25.579 M25.859 719.37 I21.19 I20.0 719.61 M25.80 719.24 I20.1 719.57 719.23 M25.429 M25.839 719.00 M25.469 719.84 719.97 719.59 719.39 719.45 719.65 I24.8 719.10 719.50 719.63 M25.473 M25.729 M25.70 719.26 M25.539 I25.89 719.16 719.44 719.41 719.49 719.83 M25.869 I25.810 719.87 719.59 M25.829 M25.50 719.17 719.20 I21.3 719.95 719.48 719.51 719.64 M25.673 I25.82 M25.119 719.12 719.69 M25.419 719.40 719.06 I21.09 I25.812 M25.529 M25.569</p>
Oth-Urinary	<p>N39.41 599.71 N39.45 N39.42 599.4 I24.8 I25.10 599.69 599.3 I25.89 I20.8 599.5 N39.44 I25.9 599.60 I25.811 N39.8 I24.1 I25.2 N39.46 I25.41 I25.810 I25.3 I21.11 599.70 N39.43 I21.29 I21.3 N39.3 599.9 I24.0 599.0 I25.82 I21.4 599.2 I25.42 N39.490 599.83 599.82 599.72 599.84 I21.19 I20.0 N39.498 599.1 I20.1 I21.09 I25.812 599.89 N39.0 599.81 I25.83</p>
Otic	<p>H66.006 H60.23 H93.249 H62.41 H91.02 388.71 H61.391 H80.20 H93.219 H93.212 H93.099 381.89 H95.811 H72.10 H73.91 H74.09 H66.21 H70.229 H90.A12 H95.123 384.00 H60.312 H61.302 H95.193 H65.199 H61.011 389.7 H70.009 380.39 H71.31 H69.91 H60.543 H94.01 H66.3X3 H80.91 H75.00 H65.191 H73.822 H68.019 385.8 H65.04 388.0 H68.112 H68.131 H92.20 H60.61 H95.54 H90.2 H74.392 H60.532 H70.211 H65.20 389.17 H69.90 H95.89 H90.3 388.70 H71.33 H74.41 H65.413 H83.12 H90.6 H67.3 H60.8X2 H72.93 H61.323 H80.23 H60.323 H80.92 H72.829 H90.A21 H72.01 H92.21 H93.13 H60.531 H61.103 H74.321 H72.11 H80.01 H70.092 380.5 H95.119 H72.813 381.9 H81.313 388.02 H81.43 H66.10 H70.811 H61.23 386.40 H83.90 H94.03 388.72 H94.80 H90.12 H74.40 386.32 H83.8X3 H73.811 H75.02 H95.121 H60.539 H73.093 385.13 H66.3X2 H90.A32 H61.003 H60.02 H73.012 H91 380.21 380.30 H95.03 H95.53 H83 386.31 H73.001 H95.111 H60.502 H88 H66.005 H65.01 H66.42 H81.12 H60.8X3 H60.42 388.2 389.02 H61.392 380.4 H61.031 H60.43 H91.8X9 H70.099 381.8 H66.009 380.03 H83.3X9 H93.293 H68.012 H62.8X2 H74.42 H72.12 H70.202 380.0 H74.8X3 H60.60 H61.113 H61.819 H93.243 386.48 H65.31 H73.003 380.00 H82.3 H95.88 389.03 H65.06 H79 H60.03 H81.41 H61.393 388.60 H92.13 388.00 H83.3X1 H60.10 389.0 H62.43 H83.19 H95.133 380.50 H61.019 H81.01 H93.091 H73 384.1 H66.23 H71.10 H74.391 H80.03 H72.02 H60.599 H66.91 380.51 380.52 381.62 H60.592 H71.32 H70.011 385.0 H73.099 H66.12 H71.93 H68.023 H93.232 H93.8X2 H77 388.11 H72.823 H60.00 H70.093 H65.91 H81.93 H67.2 H93.3X9 389.16 386.5 H60.559 389.08 H73.019 H61.013 H60.512 385.01 H95.113 H70.209 H81.49 H65.21 H74.323 H73.813 H70.012 389.14 H69 H62.42 H61.001 H90.A22 386.43 385.23 H61.102 H95.21 H68.129 H83.01 H61.129 H65.196 H72.90 H65.32 H83.2X2 384.8 H73.22 H95.139 H65.90 H74 384 H72.00 H91.20 H83.8X2 H71.23 H89 H93.3X3 H62.8X9 H73.829 H66.11 H71.00 389.11 H65.03 H74.399 H69.03 385.11 384.23 H60.511 H90.72 380.02 H60.391 H81.8X2 386.30 H91.09 H95 H66.003 H75.83 384.25 H83.02 H65.115 H73.10 H65.197 H60.541 H61.811 385.89 H60.331 H81.92 H70.219 385.24 380.9 H61.399 H68.022 384.21 H74.92 380.81 H60.40 H66.93 H92.23 H65.492 H69.02 H70.893 380.8 384.01 H73.821 389.05 H73.93 H91.01 H65.05 H74.393 387.1 H72.13 H81.21 H81.391 H73.013 H70.11 H60.399 H61.123 H69.80 H74.312 H61.93 H74.322 H83.11 H91.91 H75 383.32 H91.93 H74.311 H73.11 381.5 H93.241 H80.83 H90.5 H92.10 385.2 H61.039 380.89 H95.42 H68.133 H80.11 386.34 384.20 H81.11 384.9 H73.819 H93.A9 H80.90 H65.116 H65.193 H73.92 H65.113 H60.542 H70.813 385.21 H90.71 H81.23 386.58 385.12 380.31 381.6 H70.92 H66.22 H65.114 H80.93 H92.03 H65.419 H61.111 H93.223 H93.A2 H94.82 H60.521 H66.012 381.81 380 H93.12 H95.129 H93 H90.A11 H74.02 H81.22 385.35 H71.90 H93.239 H69.01 H95.02 386.56 387.8 389.06 H66.014 H70 H73.011 H60.333 384.82 H60.20 385.00 H60.392 385.09 H61.032 H74.329 H83.09 388.31 H74.22 H74.03 H93.019 H71.92 388.3 H70.12 H61.311 H93.221 H93.013 389.01 H61.023 H66.019 381.63 H60.529 H71.91 H95.813 H60.549 H75.82 H81.399 H61.119 H91.8X3 H65.22 H93.211 H70.212 H70.213 H60.92 H71.12 H61.92 H73.21 H61.303 H80.21 H83.91 H91.22 H73.091 H66.20 H71.11 H72.91 H90.0 H61.121 H70.222 H95.819 H61.301 H70.891 H73.891 H73.009 H65.07 H66.007 H61.029 H68.113 H71.02 H82.9 H61.813 H70.001 H81.393 H93.8X9 H61.009 H61.109 H94.02 H61.319 386.41 H93.012 H60 H60.8X1 H80.12 H81.312 385.33 H70.812 H74.12 H70.91 H95.132 H93.222 381.52 389.8 H80.10 H66.001 H65.119 H65.111 H93.299 H80.80 H65.195 H68.122 388.01 H95.22 H66.40 H74.13 H95.191 H81.8X9 H60.591 H61.892 H72.2X1 H74.313 H61.329 H65.30 H61.193 H81.13 380.32 387.2 H68.102 H60.523 H66 H74.43 H60.501 H83.13 H93.90 387 H91.3 H94.81 H95.52 H66.92 H80.02 H70.13 H73.893 H73.892 H65.00 H90.42 H60.62 H93.92 H60.41 H65.411 H91.11 H91.92 H71.30 389.9 H60.313 H62.8X3 H70.091 H83.2X3 389.00 H60.63 H74.20 389.12 385.82 H65.192 H93.092 H72 H68.109 H65.92 H73.90 H80.22 H69.00 H60.322 H66.004 H90 H68.132 H60.503 H71.01 H68.111 H91.13 H74.8X1 H64 388.6 H95.192 H72.821 H81.8X3 388.7 H61.012 H83.3X2 H74.01 H95.131 H66.3X1 H93.11 H83.8X1 386.50 H75.81 386.55 H61.112 H65.33 H60.332 H72.2X3 H81.02 H72.03 388.9 H60.11 380.01 H61.20 389.22 H62.8X1 H95.41 387.9 H73.812 388.8 H60.522 389.2 388.69 H91.8X2 H95.32 H95.812 H92.12 H93.291 H70.10 H70.201 H71.22 387.0 H71 H85 H72.2X9 H61.321 H65.194 H66.002 H61.91 H92.09 H61.313 H83.2X1 H66.43 H86 H83.8X9 H91.21 H70.002 389.13 H60.93 H67.1 H69.81 H81.42 H90.11 H71.21 H78 H80.13 H84 388.30 H61.812 H61.021 381.7 H74.8X9 H60.311 H74.90 H90.A31 H60.12 H60.8X9 H69.83 H65.493 H60.551 384.09 H61.893 385.03 H93.3X1 H60.513 H68.123 H72.822 H68.002 H63 H60.393 H91.8X1 H93.011 H61.891 H66.013 H70.899 H60.329 H61.309 H80.82 H62.40 H60.13 H66.016 H75.01 H91.23 H81.09 380.3 386.33 H80 H66.011 381.61 385.3 H93.093 H61.899 H81.91 H93.19 H81.319 H74.21 388 H92.01 H68.021 H83.92 H72.819 H92.02 H71.13 H61.192 H60.321 H68.009 H61 H74.319 H93.8X1 H68.119 H81.8X1 H68.001 H70.90 H83.03 385 H61.033 H65.491 386.54 H93.3X2 H70.003 H75.80 389.20 H93.A1 389 H68.013 H81.90 H60.01 H75.03 H72.811 H66.3X9 H70.203 H81 381.60 384.22 H95.199 H81.311 H92 389.04 H65.117 H68.101 381.50 H70.93 H73.823 H80.00 H60.519 H83.2X9 H61.122 H74.8X2 H60.509 386.52 385.9 H61.002 H82 385.02 388.32 H95.01 385.10 386.51 H61.191 H60.593 H73.13 H82.1 H95.122 H61.312 389.21 H62 H91.10 H73.23 384.81 H68.011 H95.00 H93.8X3 H70.892 H74.11 H80.81 H60.339 H65.02 386.3 H90.8 H65.412 H93.229 H93.231 H66.13 H73.002 H61.199 H67.9 H60.533 380.53 H65 H81.20 H92.11 H81.10 H65.112 H87 H61.21 389.15 H95.51 H93.93 H94.83 H68.003 H83.3X3 H60.21 385.32 H60.319 H61.90 H72.812 384.24 389.1 H74.23 H93.A3 H70.819 388.12 386.4 H60.22 385.19 H65.499 H69.82 H71.03 H74.93 H68.029 H70.013 H91.12 H70.019 H74.91 H70.223 H83.93 H81.392 386.53 H60.553 H61.101 H93.25 H60.91 H67 385.1 H68.103 H71.20 H65.23 388.10 H69.93 H74.19 H66.017 H73.20 H68 H90.41 H66.41 H91.03 388.1 H60.552 H95.31 H82.2 386.35 H61.322 H60.90 H81.03 H68.139 385.30 H65.91 389.18 H93.292 H95.112 H94 381 H93.233 386.42 H94.00 389.10 H72.92 H70.221 H72.2X2 H73.899 H93.242 H66.015 385.22 H92.22 384.0 H73.092 H93.91 H76 H69.92 H61.022 384.2 H91.90 381.51 H73.12 385.31 388.61 386.8 H93.213 G96.0 H66.90 H61.22 H68.121</p>
PNS	<p>G56.91 G73.3 G81.92 G62.82 355 359.21 P11.5 G81.11 352.5 G57.53 352.4 G82.22 G83.14 G57.00 354.3 G80.0 354.4 355.3 354.2 353.3 G83.10 G50.0 767.6 R26.9 G56.43 G79 G71.19 G60.9 G83.13 350.2 G81.02 G71.13 G81.90 G52.9 G57.32 G56.40 767.7 G81.93 G72.1 G62.89 G65.1 R26.0 356.2 G51.8 G57.92 P14.8 G57.21 G54.4 G72.2 357 G83.30 359.22 358.2 781.94 351.8 G71.11 G54.7 G72.9 G57.22 G83.89 359.81 G56.42 G60.2 G72.41 G82.50 G80 G56.02 359.71 G82.20 359.79 G83.32 G57.10 353.0 781.1 357.7 G57.80 G81.01 359.4 G57.43 354.8 G57.40 G56.00 G59 R26.1 G80.2 G81 G57.61 G56.80 355.2 353.4 G52.2 354.5 G81.14 358.9 G55 G72.89 G83.9 350.9 G57.83 G68 G58 358.8 G82.53 G57.82 G66 781.0 G69 355.0 351.0 R27.8 G56.83 R25.8 352.6 353.1 G70.89 G56.92 G83.84 G80.4 R29.5 352.1 G70.00 353.6 G61 G56.32 R41.4 R26.2 781.3 G83.4 R29.3 G51.0 G83.23 359.89 G82.52 G54.6 G56 G56.03 G83.20 G56.90 353.5 G60.1 G73 P14.3 G50.8 G60.0 350 358.01 G51.4 G57.02 G50 G61.82 G62 357.9 R43.0 359.3 G57.71 353 352.0 R25.0 G60.8 356.8 355.1 357.89 G54.8 G80.8 H90.2 G57.90 G71.9 G57.12 G80.3 357.1 781.5 G65 R29.1 R25.1 G61.81 356.4 357.0 359.29 781.2 G54.1 G70 P14.9 G53 G56.33 G65.0 356.3 G57.73 G56.23 G83.5 353.8 G57.93 781.99 R25.2 G57.13 352.9 G83.24 G54.5 G57.72 359.9 G61.89 355.71 G82.54 355.8 G56.11 G71.8 P14.0 G51.1 351.9 359.24 355.6 G56.93 355.4 G57.50 357.2 G70.01 G76 G73.1 G57.20 359.6 G83.11 351.1 781.8 G56.21 G82 G50.9 G83 R25 G57.63 G65.2 G57.62 G54.0 R25.3 G57.31 G83.82 357.3 355.79 G56.31 G80.10 352.2 G57.91 P11.4 G70.81 G78 R27.9 G62.9 G71 G56.20 358.1 G72.0 G74 P14 767.5 P14.1 G72.3 G83.31 357.81 359 E13.42 G57.70 R27 G57.33 G70.1 G80.1 359.0 356.9 G54 G81.10 P11.3 357.5 G71.12 356.1 781 G51.9 G52.0 350.1 G58.8 G77 355.5 G51 358.00 G81.94 358 G81.04 G56.22 G83.21 359.23 781.6 G83.34 G67 781.93 G57.41 R29.891 G57.11 G54.3 R29.810 R29.890 359.1 G57.42 352.3 355.9 G52.8 G57.01 353.2 767.4 G57.30 G62.1 P14.2 G54.9 G54.2 G58.7 G53.3 R26.89 G56.12 781.91 G52.7 G70.80 781.7 G70.9 G57.23 356.0 G83.33 R27.0 G52 G61.0 G50.1 G62.2 354.0 359.5 G51.2 R29.0 G60 354.1 356 G62.81 357.6 G83.22 G71.2 354 G61.1 G56.30 G82.21 G73.7 G81.13 G56.41 G56.01 G81.12 G51.3 G83.0 G83.83 R25.9 781.4 G57.51 R29.818 781.92 357.4 G71.3 G57.52 G62.0 R68.3 G57 G63 G83.81 G56.13 G52.1 G57.03 G56.81 G72.81 G81.03 352 G64 353.9 R26 G61.9 G71.0 G71.14 G72 G83.12 G82.51 G72.49 G60.3 354.9 R26.81 350.8 351 G57.60 G57.81 G80.9 G58.9 G75 357.82 G81.91 G56.82 G58.0 G70.2 G56.10</p>

Continued on next page





Respiratory	<p>J27 J63.0 J12.9 J37.0 P26.0 J39 R05 J69 J15.1 J67.0 516.0 P24.21 J95.859 770.3 502 J09.X9 J34.3 J52 770.81 J03.00 786.9 J95.01 277.09 J20.6 J21.8 J33.8 P28.0 J95.850 J84.848 770.0 J67.7 P28.2 R06.7 J34.89 P27.8 478.30 J01.21 J12.3 518.3 G47.35 495.7 J11.82 J95.71 J64 J82 P28.11 J01.20 J15.8 J45.21 495.4 J39.9 J95.89 P29 J26.9 J95.811 J70.1 J73 J95.862 J66.1 J79 J93.83 J20.9 512 J41.8 J84.112 J03 518 518.1 J94.9 471.8 J18.1 J34.81 786.7 478.33 J34.2 J93 J32.9 R06 J68.8 J45.30 519.9 327.29 J20.0 J63.5 J84.03 P91.63 770.15 R06.81 495 P19.0 J01.10 770.7 J95.863 J09.X1 J02.0 J34.9 J84.02 J15.212 J81.0 J94.0 478.3 J23 J10.81 J84.116 495.1 507.0 J05.10 P26.1 J11.00 J06.0 J96.91 J33 R06.6 J45.991 516.32 770.12 491 J32.0 768.4 J87 768.70 J01.00 R06.01 J94.1 512.1 786.51 786.6 J38.02 J95.830 J38.01 478.79 770.84 J68.3 J10.08 J84.115 J25 J01.41 M34.81 P27.0 J07 512.8 R06.00 J11.08 J46 R09.3 J49 J70.5 J95.4 J05.0 J95.02 J31.0 J95.821 P24.80 J91.8 768.71 J30.5 J38 J70.2 J63.4 J86 478.4 J10.00 J84.9 J96.20 J21 J30.2 J31 J01.01 J29 R07.9 J75 J15.7 327.22 J61 J93.11 J02.9 478.75 P28.5 491.2 J45.22 J96.11 J67.8 J85.0 J35.03 J96.21 J11.1 J35.01 505 491.1 P28.10 J15.3 J38.00 J67 492.8 R06.2 770.88 770.86 J15.20 J38.5 R07 J30.1 J95.62 J20.1 J22 512.0 G47.30 768.1 P23.0 J35.3 J95.88 P91.60 J95.860 J81 J43.8 J00 J68.0 J83 478.34 J18.9 J94.2 J70.8 J35.9 J33.0 R07.1 770 501.0 J43.2 478.74 J45.20 470 J45.40 786.06 327.23 P28.89 R06.9 J39.0 J98.9 J47.0 J98.3 P22.1 J39.8 R06.83 J65 P24.10 R07.0 J37 R07.89 517 503 786.52 J10 491.20 J95.72 J28 J09.X3 J70 P23.3 J01.80 J97 J12.89 J18.8 516 J67.3 495.9 J40 P24.01 E84.9 J10.01 500.0 J11.2 519 501 J93.0 J98.01 P22 495.8 J08 G47.31 518.8 327.20 J20.5 786.01 786.50 P24.00 J41 770.2 J84.843 J93.81 515.0 J95.03 J30.0 786.00 J92 491.21 515 J92.9 768 R06.89 J56 J01.11 770.5 P19.1 J70.3 516.3 J32 J41.1 J85.3 J02.8 J21.9 J10.89 J68.2 518.2 R06.3 P25.1 R06.5 J91 P25.2 G47.36 J45.901 J16.8 J84.83 P28.4 R07.82 J18.2 J34 J74 J30.89 478.7 768.0 J11.83 J95.84 J59 478.5 J15.5 J84.17 J35.02 J18.0 768.72 786 519.02 J95.04 J63.6 P28.9 P19.9 J95 J85.1 491.8 J66.2 517.8 J95.822 J01.40 J48 519.09 518.4 G47.39 J51 786.59 277.03 J84.111 J86.0 P23.9 J34.0 J18 J17 J43 786.2 J96 J96.22 J71 J01.31 J63.3 495.6 494.0 P24.81 J95.2 J42 J76 J95.09 J84.09 768.9 J35.8 J01.91 G47.34 J06.9 J62.0 J96.10 J20.4 J50 491.9 R06.82 J32.2 J03.80 516.36 770.11 P23.4 J01.30 J84 J84.89 J47.1 J98.4 J01.90 R06.03 491.22 786.8 R07.81 492.0 J14 J66 P28.81 J15.211 J09.X2 J93.9 J67.5 J70.4 J45.32 84.09 770.6 J89 J01.81 P28.19 518.5 J32.8 J33.1 J10.82 786.05 478.6 J32.4 J81.1 J98.8 327.26 J96.02 516.8 518.0 519.4 770.9 J38.3 R06.09 J03.91 G47.37 327.2 J34.1 J12.0 J21.1 J01 P22.0 J12.2 J39.3 P25.3 786.02 J95.851 J31.1 P25.0 J15 J43.0 P23.1 495.2 478.31 P24.30 770.18 P24.31 J57 J63 J95.00 J69.0 J38.4 J09 277.0 J04.30 J95.5 J45.902 504 P24.11 J45 J03.01 P23.6 R06.1 519.0 786.1 J37.1 J15.29 J98.6 J44.9 J84.10 P27.9 J44.0 J96.00 517.3 E84.0 J45.998 516.9 J86.9 471 J63.2 491.0 J68.1 786.07 J98.19 P25 J66.8 768.3 518.84 J95.3 J30 J72 J44.1 519.00 J16 J84.2 770.85 J66.0 J20.2 519.11 277.00 J47.9 P27.1 770.4 J06 770.87 P91.62 P22.9 786.03 J84.117 J84.113 516.30 770.16 P23.2 495.0 J67.6 J93.82 J98.51 516.37 770.89 768.6 510 J04.2 J44 J12 327.24 J43.9 J77 J84.82 J20.8 510.9 770.13 P28 J31.2 478.32 478.70 496 J38.6 J41.0 J24 J68 J11.89 J45.909 J03.81 J98.09 518.89 P91.61 770.10 J67.2 J15.0 J39.2 494.1 516.2 G47.33 277.02 J38.2 J60 E84.11 J09.89 J47 277.01 J95.831 J35 P23.5 J94.8 P24 J04.0 J69.8 517.2 R07.2 495.5 786.09 J20 J12.81 J45.31 J58 J45.51 J43.1 E84.19 P84 494 J55 J33.9 516.31 J30.9 P22.8 J98.59 J04.10 J38.1 J53 P25.8 P23.8 J62 516.33 J84.842 J54 J92.0 J32.1 769 J04.11 786.04 J68.9 J70.0 327.21 J90 P24.20 G47.32 770.83 P24.9 J12.1 J84.81 J67.1 J10.83 J32.3 J10.1 492 514 768.2 327.27 J95.61 J38.7 J16.0 J20.7 478.0 471.1 R06.02 J15.6 J95.1 J02 J95.812 J15.4 J68.4 J96.01 514.0 J03.90 J35.1 516.35 J45.990 786.3 J35.2 J62.8 J91.0 500 J93.12 770.82 J84.01 516.1 J45.41 J88 327.25 J84.114 P27 J20.3 471.9 J96.90 J11.81 518.82 495.3 J39.1 J98.11 J10.2 J80 519.8 J45.52 J96.12 518.83 J05.11 R06.4 770.17 J69.1 J67.4 J84.841 J99 J98.2 P26.8 J11 J63.1 J78 768.5 J85.2 J45.50 P26 P28.3 J15.9 J70.9 786.4 J36 J05 J95.861 J04.31 J67.9 J85 J04 478.71 471.0 518.81 E84.8 J98 J19 J21.0 J94 J43.9 770.14 516.34 J30.81 510.0 J45.42 J13 J26 768.73 R22.2 J96.92</p>
Rheumatism	<p>727.66 727.04 726.12 726.31 729.1 120.8 726.72 728.89 729.89 728.83 727.49 729.4 125.3 121.11 727.06 726.10 727.43 729.73 125.42 726.60 727.81 726.90 726.73 727.50 729.0 729.39 728.6 727.51 729.6 125.10 726.62 727.67 729.71 727.40 726.69 727.68 726.5 125.811 728.2 125.2 727.09 727.59 726.30 727.05 121.29 726.2 727.83 121.4 726.39 728.85 727.60 729.92 728.9 726.4 726.91 727.03 729.82 125.83 727.89 728.5 726.64 726.32 726.61 728.4 727.69 125.9 728.88 728.10 727.65 727.00 728.79 124.1 727.87 125.41 727.82 726.70 729.81 729.91 124.0 727.42 727.63 725 727.9 121.19 120.0 729.99 728.82 120.1 727.3 727.64 726.63 726.65 727.2 727.01 726.79 729.31 728.81 726.33 729.30 728.11 124.8 726.11 125.89 728.12 729.72 728.0 727.41 729.90 727.02 726.0 125.810 729.5 728.13 729.2 728.71 121.3 728.3 727.1 125.82 729.79 726.19 728.84 726.8 728.86 727.61 728.19 727.62 121.09 125.812 726.71</p>
Sleep-Disorders	<p>327.43 327.13 G47.37 G47.54 327.33 327.09 G47.01 120.8 327.40 G47.35 G47.8 327.02 327.42 125.3 121.11 G47.51 327.15 G47.33 G47.419 327.32 327.11 G47.10 125.42 327.20 327.27 327.51 G47.13 G47.27 327.41 G47.19 125.10 G47.9 G47.22 G47.39 125.811 327.00 327.36 327.30 G47.61 125.2 327.34 G47.31 G47.69 121.29 327.49 G47.14 121.4 G47.50 327.53 327.31 125.83 G47.29 G47.30 327.25 G47.36 327.35 G47.09 125.9 G47.32 G47.52 327.24 124.1 125.41 G47.63 G47.23 124.0 327.01 G47.11 G47.421 G47.20 327.10 G47.34 G47.00 327.29 121.19 120.0 120.1 327.21 327.44 G47.24 327.52 327.19 124.8 327.8 125.89 327.26 327.37 G47.12 G47.429 G47.53 327.22 125.810 G47.26 G47.411 G47.62 G47.21 327.39 121.3 327.12 125.82 327.23 327.59 G47.25 327.14 121.09 125.812 G47.59</p>
Symptoms-Abs-Pelvis	<p>R10.2 R10.11 R10.83 789.7 789.06 120.8 789.02 789.1 R10.10 125.3 121.11 R10.812 789.42 789.63 789.04 789.2 R10.815 125.42 787.3 R19.32 R13.13 R19.03 787.4 789.34 789.47 R10.13 789.60 789.09 125.10 789.67 125.811 789.69 789.65 789.33 125.2 789.01 121.29 789.35 121.4 R10.31 789.51 R10.12 125.83 789.49 789.05 125.9 R10.817 789.64 124.1 125.41 789.36 124.0 R10.33 789.9 121.19 120.0 120.1 R10.813 789.45 789.32 789.62 R10.9 789.43 789.40 124.8 125.89 789.46 789.41 R10.32 125.810 789.00 121.3 R10.84 789.31 125.82 789.37 789.30 789.03 789.66 789.39 R10.814 R10.816 789.07 121.09 125.812 789.59</p>
Symptoms-Digestive	<p>R14.1 R13.14 120.8 787.5 787.91 R19.4 R19.05 R19.01 R19.2 125.3 121.11 R11.11 R11.2 R19.07 R19.06 R19.09 R15.0 125.42 787.3 R19.32 R13.13 R19.03 787.4 R19.30 787.21 R11.13 R15.9 125.10 787.04 125.811 125.2 R11.0 787.20 121.29 R19.35 R11.10 121.4 R19.33 R19.04 787.29 787.03 125.83 R16.0 R19.31 R12 R13.19 R17 R19.02 R15.1 R18.0 787.6 125.9 R16.1 124.1 125.41 787.02 R15.2 124.0 R13.12 R11.14 787.99 R13.11 787.23 121.19 120.0 120.1 R19.36 R19.11 R19.8 124.8 R19.37 125.89 R13.10 R18.8 787.7 787.24 R19.34 787.01 125.810 787.22 121.3 R19.7 125.82 R19.5 R19.00 121.09 125.812 787.1</p>
Symptoms-General	<p>R53.82 R65.10 R56.01 R50.81 R68.3 R59.9 780.39 120.8 R53.1 R62.0 R55 R65.21 R62.51 780.64 125.3 121.11 780.1 780.63 780.65 R56.1 R68.0 R62.50 R64 780.59 R63.3 125.42 780.96 R68.83 780.8 R56.00 R68.84 R62.52 R53.81 125.10 R63.0 780.72 780.51 R68.82 780.91 R63.1 125.811 125.2 R68.12 121.29 R63.2 R57.0 121.4 R63.6 R63.4 R63.8 780.92 R57.9 125.83 780.62 780.52 R65.20 R65.11 780.55 R68.81 R50.82 780.32 780.94 125.9 780.95 780.71 124.1 R50.84 125.41 780.56 R68.11 780.09 124.0 R50.9 R63.5 121.19 120.0 120.1 780.99 780.97 780.61 R53.2 780.4 124.8 R58 R69 780.2 R51 125.89 R56.9 780.93 R52 780.03 R60.9 780.58 780.01 780.53 125.810 780.31 780.57 R62.7 121.3 125.82 R61 780.79 780.50 R57.8 R68.13 780.02 780.54 780.60 121.09 125.812 R50.83 R68.89</p>

Symptoms-Respiratory	<p>786.7 786.50 R09.2 786.3 R06.4 786.51 R06.7 R09.82 I24.8 I25.10 786.04 786.8 R06.9 786.2 I25.89 I20.8 R07.0 786.02 I25.9 786.06 786.05 786.1 I25.811 R06.3 R06.2 R09.02 I24.1 I25.2 786.4 I25.41 I25.810 I25.3 I21.11 R06.81 R06.02 R06.1 786.07 R06.01 R06.6 I21.29 I21.3 786.00 786.6 786.9 I24.0 R05 I25.82 I21.4 R06.89 I25.42 R06.82 R09.3 R06.00 786.09 786.59 R09.89 R09.01 I21.19 I20.0 R07.2 R07.1 I20.1 I21.09 I25.812 786.03 R07.89 786.52 786.01 I25.83 R07.9</p>
Symptoms-Skin	<p>R22.1 R23.4 R23.8 I24.8 I25.10 I25.89 I20.8 R23.2 782.2 I25.9 R22.2 I25.811 782.4 I24.1 I25.2 782.0 I25.41 I25.810 I25.3 I21.11 782.1 782.62 R21 R23.0 782.7 782.61 I21.29 I21.3 I24.0 R23.3 782.3 782.8 I25.82 I21.4 I25.42 782.5 I21.19 I20.0 I20.1 I21.09 I25.812 R23.1 R22.9 R20.3 782.9 I25.83</p>
Symptoms-Urinary	<p>R36.1 R36.9 788.35 788.61 788.36 I20.8 788.0 788.63 I25.3 I21.11 788.69 I25.42 R37 788.32 788.20 788.39 R32 R39.12 I25.10 R39.14 R31.9 788.65 788.38 R39.81 R35.8 R39.16 I25.811 R33.9 I25.2 I21.29 788.30 I21.4 788.37 R39.0 788.43 R39.13 R39.89 788.31 I25.83 788.99 788.7 R39.11 I25.9 I24.1 I25.41 R39.19 I24.0 788.91 I21.19 I20.0 788.8 R35.0 I20.1 788.21 788.33 788.42 I24.8 R35.1 R30.0 I25.89 R34 R33.8 788.62 I25.810 788.41 R39.15 I21.3 788.1 788.34 788.64 I25.82 788.5 788.29 I21.09 I25.812 R31.0 R31.1</p>
Thyroid	<p>E03.9 241.9 E04.1 E04.2 E05.90 I20.8 E05.21 I25.3 I21.11 245.2 243 241.0 E05.40 I25.42 E05.91 246.9 E07.89 E04.0 242.10 242.30 245.0 E07.0 I25.10 242.81 242.11 244.3 242.41 242.40 246.2 I25.811 240.0 E06.4 I25.2 I21.29 I21.4 242.21 241.1 E05.11 I25.83 E06.3 245.8 242.80 E05.20 E06.1 I25.9 E05.10 I24.1 I25.41 E07.81 E05.01 E01.8 245.4 I24.0 246.3 242.91 244.8 246.1 242.20 I21.19 I20.0 E01.2 242.01 I20.1 244.0 242.00 I24.8 245.1 E05.00 I25.89 244.9 246.0 E07.1 E06.9 I25.810 E00.9 244.1 E05.41 I21.3 244.2 242.90 240.9 245.9 E07.9 I25.82 E05.31 E05.30 245.3 E04.9 242.31 E06.0 246.8 E03.2 I21.09 I25.812 E06.5</p>

**Algorithm 1: GenESess**


---

**Data:** A sequence  $x$  over alphabet  $\Sigma$ ,  $0 < \varepsilon < 1$   
**Result:** State set  $Q$ , transition map  $\delta$ , and transition probability  $\tilde{\pi}$

*/\* Step One: Approximate  $\varepsilon$ -synchronizing sequence \*/*

- 1 Let  $L = \lceil \log_{|\Sigma|} 1/\varepsilon \rceil$ ;
- 2 Calculate the **derivative heap**  $D_\varepsilon^x$  equaling  $\{\hat{\phi}_y^x : y \text{ is a sub-sequence of } x \text{ with } |y| \leq L\}$ ;
- 3 Let  $\mathcal{C}$  be the convex hull of  $D_\varepsilon^x$ ;
- 4 Select  $x_0$  with  $\hat{\phi}_{x_0}^x$  being a vertex of  $\mathcal{C}$  and has the highest frequency in  $x$ ;

*/\* Step Two: Identify transition structure \*/*

- 5 Initialize  $Q = \{q_0\}$ ;
- 6 Associate to  $q_0$  the **sequence identifier**  $x_{q_0}^{\text{id}} = x_0$  and the probability vector  $d_{q_0} = \hat{\phi}_{x_0}^x$ ;
- 7 Let  $\tilde{Q}$  be the set of states that are just added and initialize it to be  $Q$ ;
- 8 **while**  $\tilde{Q} \neq \emptyset$  **do**
- 9     Let  $Q_{\text{new}} = \emptyset$  be the set of new states;
- 10    **for**  $(q, \sigma) \in \tilde{Q} \times \Sigma$  **do**
- 11     Let  $x = x_q^{\text{id}}$  and  $d = \hat{\phi}_{x\sigma}^x$ ;
- 12     **if**  $\|d - d_{q'}\|_\infty < \varepsilon$  **for some**  $q' \in Q$  **then**
- 13         Let  $\delta(q, \sigma) = q'$ ;
- 14     **else**
- 15         Let  $Q_{\text{new}} = Q_{\text{new}} \cup \{q_{\text{new}}\}$  and  $Q = Q \cup \{q_{\text{new}}\}$ ;
- 16         Associate to  $q_{\text{new}}$  the sequence identifier  $x_{q_{\text{new}}}^{\text{id}} = x\sigma$  and the probability vector  $d_{q_{\text{new}}} = d$ ;
- 17         Let  $\delta(q, \sigma) = q_{\text{new}}$ ;
- 18     Let  $\tilde{Q} = Q_{\text{new}}$ ;

19 Take a strongly connected subgraph of the labeled directed graph defined by  $Q$  and  $\delta$ , and denote the vertex set of the subgraph again by  $Q$ ;

*/\* Step Three: Identify transition probability \*/*

- 20 Initialize counter  $N[q, \sigma]$  for each pair  $(q, \sigma) \in Q \times \Sigma$ ;
- 21 Choose a random starting state  $q \in Q$ ;
- 22 **for**  $\sigma \in x$  **do**
- 23     Let  $N[q, \sigma] = N[q, \sigma] + 1$ ;
- 24     Let  $q = \delta(q, \sigma)$ ;
- 25 Let  $\tilde{\pi}(q) = \left[ \left[ (N[q, \sigma])_{\sigma \in \Sigma} \right] \right]$ ;
- 26 **return**  $Q, \delta, \tilde{\pi}$ ;

---

**Algorithm 2: Log-likelihood**


---

**Data:** A PFSA  $G = (\Sigma, Q, \delta, \tilde{\pi})$  and a sequence  $x$  over alphabet  $\Sigma$   
**Result:** Log-likelihood  $L(x, G)$  of  $G$  generating  $x$

- 1 Calculate the state transition matrix  $\Pi$  and observation  $\tilde{\Pi}$ ;
- 2 Calculate the stationary distribution over states  $\wp_G$  of  $G$  from  $\Pi$ ;
- 3 Calculate the stationary distribution of alphabet  $\phi_\lambda^T = \wp_G^T \tilde{\Pi}$ ;
- 4 Initialize  $\mathbf{p}$  by  $\wp_G$  and  $\mathbf{q}$  by  $\phi_\lambda$ ;
- 5 Let  $L = 0$ ;
- 6 **for**  $i$  from 1 to  $|x|$  **do**
- 7     Let  $\sigma$  be the  $i$ -th entry of  $x$ ;
- 8     Let  $L = L - \log \mathbf{q}|_\sigma$ ;
- 9     Let  $\mathbf{p}^T = \left[ \left[ \mathbf{p}^T \Gamma_\sigma \right] \right]$  where  $\Gamma_\sigma$  is defined in Eq. 8;
- 10    Let  $\mathbf{q}^T = \mathbf{p}^T \tilde{\Pi}$ ;
- 11 **return**  $L/|x|$ ;

---